Statistical and algorithmic developments for the analysis of Triple Negative Breast Cancers

Guillem Rigaill

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture
Agriculture
Alimentation
Food
Biologie
Biology
Environnement
Environment
Santé
Health

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

présentée et soutenue publiquement par

Guillem Rigaill

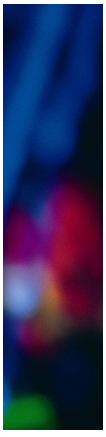
17 / 11 / 2010

Développements statistiques et algorithmiques pour l'analyse des cancers du sein de type triple négatif

Directeur de thèse : **Stéphane Robin**
Co-encadrement de la thèse : **Thierry Dubois et Emmanuel Barillot**

Jury

MD/Ph.D.	**Gordon MILLS,**	**Professeur, MD Anderson Cancer Center**	Rapporteur
Ph.D.	**Nancy ZHANG,**	**Professeur Assisant, Université de Stanford**	Rapporteur
Ph.D.	**Anestis ANOTONIADIS,**	**Professeur, Université Joseph Fourier**	Examinateur
Ph.D.	**Thomas SCHIEX,**	**DR, INRA**	Examinateur
Ph.D.	**Lodewyk WESSELS,**	**Chef d'équipe, The Netherlands Cancer Institute**	Examinateur
Ph.D.	**Stéphane ROBIN,**	**DR, INRA**	Directeur
Ph.D.	**Emmanuel BARILLOT,**	**DU, Institut Curie**	Co-directeur
Ph.D.	**Thierry DUBOIS,**	**IR, Institut Curie**	Co-directeur



AgroParisTech
UMR 518 AgroParisTech / INRA

16, rue Claude Bernard 75231 Paris CEDEX 05

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture
Agriculture
Alimentation
Food
Biologie
Biology
Environnement
Environment
Santé
Health

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

présentée et soutenue publiquement par

Guillem Rigaill

17 / 11 / 2010

Statistical and algorithmic developments for the analysis of Triple Negative Breast Cancers

Ph.D. Advisor : **Stéphane Robin**
Ph.D. Co-advisors : **Thierry Dubois et Emmanuel Barillot**

Jury

MD/Ph.D.	**Gordon MILLS,**	**Professor, MD Anderson Cancer Center**	Reviewer
Ph.D.	**Nancy ZHANG,**	**Assistant Professor, Stanford university**	Reviewer
Ph.D.	**Anestis ANOTONIADIS,**	**Professor, Université Joseph Fourier**	Examiner
Ph.D.	**Thomas SCHIEX,**	**DR, INRA**	Examiner
Ph.D.	**Lodewyk WESSELS,**	**Group Leader, The Netherlands Cancer Institute**	Examiner
Ph.D.	**Stéphane ROBIN,**	**DR, INRA**	Ph.D. Advisor
Ph.D.	**Emmanuel BARILLOT,**	**DU, Institut Curie**	Ph.D. Co-advisor
Ph.D.	**Thierry DUBOIS,**	**IR, Institut Curie**	Ph.D. Co-advisor

AgroParisTech
UMR 518 AgroParisTech / INRA

16, rue Claude Bernard 75231 Paris CEDEX 05

Développements statistiques et algorithmiques pour l'analyse des cancers du sein de type triple négatif

Résumé

Dans le monde, le cancer du sein est le cancer le plus fréquent de la femme. Plusieurs types de cancer du sein ont été mis en évidence. Les carcinomes infiltrants triple négatif (TNBC) sont l'un de ces types. Les TNBC sont parmi les plus agressifs cancers du sein et sont associés à un mauvais pronostique. Il n'y a pas encore de traitement dédié pour ces cancers. Cette thèse avait pour but d'identifier des gènes et des voies de signalisation dé-régulés dans les cancers de types TNBC en s'appuyant sur les profils transcriptomiques et génomiques de tumeurs TNBC bien caractérisées, obtenues par la technique des biopuces.

Mon travail comporte deux volets. D'abord, j'ai développé des méthodes pour l'analyse des données génomiques. J'ai proposé une méthode (ITALICS) pour la normalisation des données Affymetrix SNP 100K et 500K. J'ai travaillé sur la segmentation des profils génomiques. J'ai développé de nouveaux outils statistiques pour étudier la stabilité de la segmentation et j'ai obtenu des formules exactes pour des critères de sélection de modèle. Enfin, j'ai proposé un algorithme de programmation dynamique rapide qui retrouve la meilleure segmentation au sens de la norme euclidienne.

Dans un second temps, j'ai analysé les données omiques du projet. J'ai conçu le plan d'expérience. J'ai analysé les données transcriptomiques avec des méthodes déjà disponibles. J'ai comparé les classifications transcriptomique et immunohistochimique des TNBC. L'analyse des données transcriptomiques m'a permis d'identifier des gènes et des voies de signalisation dé-régulés dans les TNBC. Enfin, j'ai analysé les données génomiques avec les outils que j'ai développés.

Mots-clés Cancer du sein, Triple Negatif, biostatiques, profil transcriptomique, profil génomique

Laboratoire UMR 518 AgroParisTech / INRA, AgroParisTech 16, rue Claude Bernard 75231 Paris CEDEX 05 FRANCE

Statistical and algorithmic developments for the analysis of Triple Negative Breast Cancers

Abstract

Throughout the world and among the different types of cancer, breast cancer is one of the most prevalent ones. It can be subdivided in several types among which the triple negative invasive ductal breast carcinoma (TNBC). TNBC is one of the most aggressive types of breast cancer: it is associated to a poor prognosis and there is still no targeted therapy for this type of tumor. In this context, we aim to discover deregulated genes and signaling pathways in human TNBC using high-throughput omic data of well-characterized breast tumors to identify potential therapeutic targets.

My work can be divided in two main parts. First, I developed methods for the analysis of genomic data: I proposed a method (ITALICS) for the normalization of Affymetrix SNP 100K and 500K arrays, worked on the segmentation of DNA copy number profiles, proposed new algorithms and new statistical tools to assess the stability of segmentation and derive exact formulation of several model selection criteria and proposed an improved and faster dynamic programming algorithm that recovers the best segmentation exactly with respect to the quadratic loss.

Next, I worked on the analysis of the omic data. The first step of my analysis was to plan the experimental design of the omic experiments. I then analyzed the transcriptomic data using already developed and available tools. I sought to better characterize the distinctness of TNBC at the transcriptomic level and its overlap with immunohistochemistry data. I worked at the gene and pathway level to identify genes and pathways of interest. Finally, I analyzed the genomic data using the tools and methods that I have developed.

Keywords Breast cancer, Triple Negative, Gene expression profiling, DNA copy number profiling

Laboratory UMR 518 AgroParisTech / INRA, AgroParisTech 16, rue Claude Bernard 75231 Paris CEDEX 05 FRANCE

Remerciements

I would like to thanks Nancy Zhang and Gordon Mills for accepting to review my Ph.D. manuscript as well as Lodewyk Wessels, Anestis Antoniadis and Thomas Schiex for accepting to examine my Ph.D. defense.

Je remercie ensuite mon directeur de thèse, Stéphane Robin, et mes deux co-directeurs : Thierry Dubois et Emmanuel Barillot. Je les remercie pour leur confiance, leur franchise, leur aide, leurs conseils et pour m'avoir fait découvrir diverses facettes de la science : les statistiques, la biologie et la bioinformatique.

Je tiens également à remercier les trois membres de mon comité de thèse : Christophe Ambroise, Olivier Delattre et Alain Viari pour le temps qu'ils m'ont accordé et leurs conseils.

Je tiens aussi à remercier les habitués des réunions kinomes : Emilie, Gordon et Philippe de m'avoir écouté et conseillé, ainsi que les habitués des comités techniques Curie-Servier pour leurs conseils.

Je remercie Sergio Roman Roman de m'avoir accueilli dans le département de recherche translationnelle de l'institut Curie.

Je remercie Dominique, Jennifer, Marie, Carole, Sophie et Odile de m'avoir guidé dans les méandres administratifs.

Je remercie les trois laboratoires où j'ai travaillé durant ces trois années - l'équipe MIA 518, le laboratoire de signalisation et l'U900 - et surtout toutes les personnes qui y travaillent ou qui y ont travaillé ces dernières années. L'entourage technique et scientifique, ainsi que les repas chaleureux et les pauses conviviales ont eu (et garderont) une grande importance pour moi. Je remercie en particulier les personnes dont j'ai partagé le bureau, à St Louis : le bureau signalisation (qui dévoile derrière sa porte l'identité du père Noël) puis le bureau info, à l'U900 : le bureau biostat (où R et les M & M's sont rois) et tous les autres pour m'avoir accueilli quand il n'y avait plus de place et enfin à l'agro : le bureau tout au fond du couloir qui s'est ensuite déplacé au premier étage (bureau embelli par deux superbes écharpes de champions de France, l'une rouge, l'autre bleu). Je remercie toutes les autres personnes avec qui j'ai eu l'occasion de travailler durant ces trois ans. Ce fut un plaisir de travailler et d'apprendre avec eux.

Je remercie toutes les personnes qui ont relu et m'ont aidé à rédiger ce manuscrit : Stéphane, Thierry, Emmanuel, Anne, Gordon, Marie-Laure, Emilie, Alban, Patrick et Chloé, ainsi que tous ceux qui m'ont aidé à préparer ma soutenance, qui y ont assisté et/ou qui sont venus au pot de thèse qui a suivi.

Je remercie une certaine promo du master de bioinfo de Rouen, les jeunes statisticiens amateurs de jeux de société, quelques Bordelais venus du Nord, quelques Lyonnais anciennement Niçois et quelques Spinoliens.

Un grand merci à tous ceux que je ne mentionne pas ici, faute de place. Ils me le pardonneront j'espère ...

Enfin, je remercie ma mère, mon père, mon frère, mon oncle et une certaine lectrice attentive et patiente pour tout le reste, jamais mes mots ne sauront exprimer ma gratitude. Je vous dédie cette thèse.

Résumé substantiel

Le cancer du sein est l'un des cancers les plus répandus qui soit dans le monde occidental, 1 femme sur 8 se voit un jour diagnostiquée d'un cancer du sein. Cette maladie est aussi très hétérogène. Elle peut être divisée entre : les tumeurs qui expriment les récepteurs aux œstrogènes (ER+ : estrogen receptor positive) et celles qui n'expriment pas ces récepteurs (ER-, voir figure 1). Les tumeurs ER+ comprennent les tumeurs Luminal A et Luminal B. Les tumeurs ER- comprennent les tumeurs ER- / HER2+ qui surexpriment le récepteur du facteur de croissance épidermique humain 2 (récepteur *HER2*), ainsi que les tumeurs triple négatives (TNBC) qui sont également négatives pour les récepteurs à la progestérone (PR) et ne surexpriment pas le gène de *HER2*. Les tumeurs TNBC ont un taux important de pertes et de gains chromosomiques mais présentent moins d'amplifications d'ADN que les autres sous-types de cancer du sein. Les tumeurs TNBC souffrent d'un très mauvais pronostic. Des thérapies ciblées, innovantes et prometteuses sont actuellement en train d'être étudiées. Un exemple est l'inhibition de la polymérase poly ADP-ribose (PARP). Malheureusement, il n'y a pas encore de thérapies ciblées pour les TNBC qui soient utilisées en routine comme il en existe pour les tumeurs surexprimant HER2 (des anticorps monoclonaux anti-HER2) et pour les tumeurs Luminal (thérapie endocrine). A travers le monde, de nombreuses équipes scientifiques composées à la fois de cliniciens et de biologistes cherchent à mieux comprendre les aspects cliniques et la biologie des TNBC dans le but d'appliquer leurs résultats en clinique et de proposer des thérapies innovantes et sur mesure.

Mon projet de thèse faisait partie d'une collaboration entre l'Institut Curie et le groupe pharmaceutique Servier. Le but de cette collaboration est de découvrir des gènes et des voies de signalisation dérégulés dans les TNBC humains, afin d'identifier de nouvelles cibles thérapeutiques (résumé sur la figure 2).

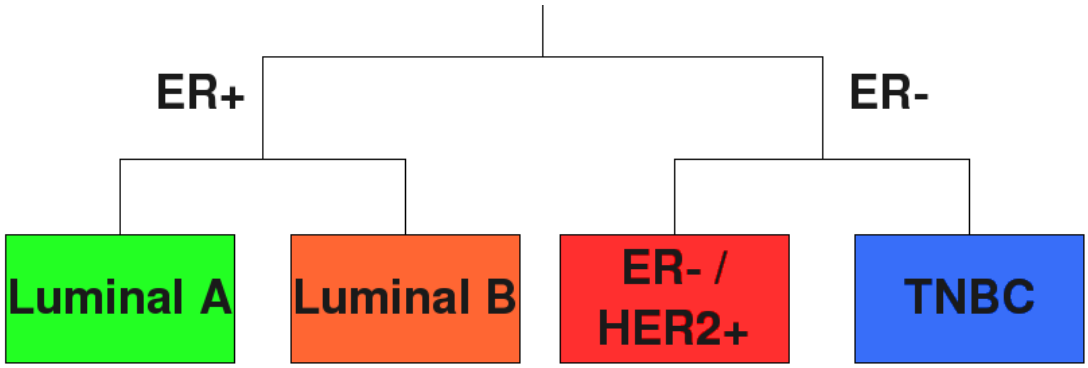


Fig. 1 – Classification des cancers du sein entre les tumeurs qui expriment les récepteurs aux œstrogènes (ER+) et celles qui ne les expriment pas ces récepteurs (ER-). Les tumeurs ER+ contiennent les tumeurs Luminal A et Luminal B. Les tumeurs ER- comprennent les tumeurs ER- / HER2+ et TNBC.

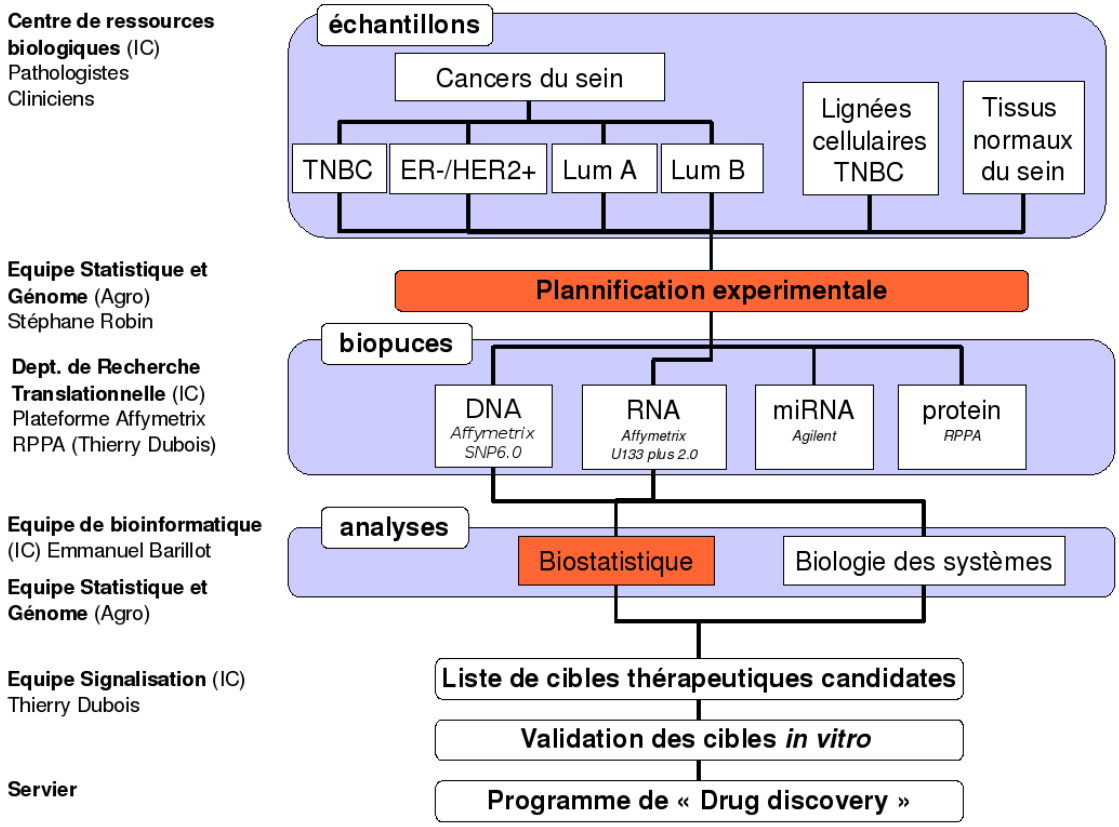


Fig. 2 – Des tumeurs de type Luminal A, Luminal B, ER- / HER2+ et TNBC ont été choisies et caractérisées par une pathologiste (Anne Vincent Salomon, MD/Ph.D.) de l'Institut Curie (IC) par immunohistochimie (IHC). Des lignées cellulaires de TNBC et des tissus mammaires sains ont également été collectés. Des profils de biopuces transcriptomiques, génomiques, microARN et protéomiques ont été générés par le département de transfert de l'Institut Curie. Cette importante quantité de données a été analysée par l'équipe de bioinformatique de l'Institut Curie et l'équipe de biostatistiques de l'AgroParisTech (Agro) à l'aide d'outils de biostatistiques et de biologie des systèmes. Ces analyses ont généré des listes de cibles potentielles. Certaines de ces cibles ont déjà été caractérisées et leurs fonctions validées dans des lignées cellulaires par une équipe de biologistes du département de transfert de l'Institut Curie. **Dans ce projet, j'ai été responsable de l'analyse biostatistique des données omiques** représentées en rouge.

Mon travail peut être divisé en deux parties. J'ai développé des méthodes statistiques et des algorithmes pour analyser les données génomiques et j'ai en parallèle analysé les données transcriptomiques avec des méthodes biostatistiques et bioinformatiques déjà disponibles.

Analyse des données génomiques J'ai travaillé sur la normalisation de profils de nombre de copies d'ADN et j'ai proposé une méthode (appelée ITALICS) pour la normalisation des puces Affymetrix SNP 100K et 500K (Rigaill et al. (2008)). Nous avons montré qu'ITALICS était plus performant que les autres méthodes existantes à l'époque de notre étude en termes de rapport signal sur bruit et qu'il permet une meilleure classification entre les nouvelles tumeurs primaires et les récurrences de tumeurs dans un jeu de données de cancers du sein (Bollet et al. (2008), en collaboration avec Marc Bollet, MD/Ph.D.).

J'ai également travaillé sur la segmentation de profils de nombre de copies d'ADN. Avec Emilie Lebarbier (Ph.D.) et Stéphane Robin (Ph.D.), j'ai proposé de nouveaux algorithmes et de nouveaux outils statistiques pour déterminer la stabilité des segmentations et la formulation exacte de plusieurs critères de sélection de modèles (Rigaill et al. (2010), ce travaill a été séléctionné pour les actes de la conférence COMPSTAT 2010). J'ai aussi conçu et implémenté un algorithme de programmation dynamique plus rapide que les précédents qui permet de trouver la meilleure segmentation au sens de la norme Euclidienne (Rigaill (2010), ce travail a été soumis dans un journal d'algorithmique et de mathématiques appliquées).

J'ai appliqué ces différents outils statistiques et algorithmiques aux données génomiques de cancers du sein Curie-Servier. Cela m'a permis en particulier d'identifier la perte de PTEN dans plus de 50% des tumeurs TNBC en collaboration avec Bérengère Marty (Marty et al. (2008)). Plusieurs autres régions du génome ont ainsi été identifiées. J'ai ensuite extrait (de la base de données Ensembl) les gènes de ces régions pour fournir une liste de gènes candidats.

Analyse des données transcriptomiques Ensuite, j'ai travaillé sur l'analyse des données Curie-Servier transcriptomiques. La première étape de cette analyse a été de concevoir le plan d'expérience. Une étape presque évidente et pourtant souvent négligée. Sans surprise, la réalisation de ce plan d'expérience a permis de rendre l'expérience plus robuste et d'améliorer notre capacité à détecter les « véritables » différences entre les tumeurs TNBC et les autres types de tumeurs.

Après cela, j'ai analysé les données transcriptomiques à l'aide d'outils biostatistiques et bioinformatiques déjà disponibles. Le but était de mieux caractériser les particularités transcriptomiques des TNBC et leur recoupement avec les données immunohistochimiques (IHC) (Rigaill et al. (2011), en préparation en collaboration avec Anne Vincent Salomon, MD/Ph.D). J'ai réalisé l'analyse statistique des profils transcriptomiques tumoraux à l'échelle du gène et à celle de la voie de signalisation, afin d'identifier des gènes et des voies de signalisation candidats. Le rôle de certains de ces candidats dans le développement tumoral a pu être confirmé in vitro ; en particulier dans le cas des tumeurs TNBC le rôle des formines liées aux diaphanes (DRF, Lizárraga et al. (2009)) en collaboration avec l'équipe de Philippe Chavrier (Ph.D.) et l'implication des voies du stress oxidatif dans les tumeurs du sein ER- / HER2+ (Toullec et al. (2010)) en collaboration avec l'équipe de Fatima Mechta Grigoriou (Ph.D.).

L'analyse des données omiques à haut-débit est au croisement de la biologie, des sciences cliniques, de la biotechnologie, des statistiques et de l'informatique. Afin d'appréhender ces différents aspects, cette thèse a été réalisée entre 3 laboratoires : un laboratoire de biologie travaillant sur les TNBC et dirigé par Thierry Dubois (Ph.D.), un laboratoire de bioinformatique travaillant sur les données liées au cancer et dirigé par Emmanuel Barillot (Ph.D.) et un laboratoire de statistique travaillant sur les données biologiques à haute densité et dirigé par Stéphane Robin (Ph.D.). De plus, à travers cette thèse, j'ai collaboré avec plusieurs autres groupes de biologistes, de cliniciens et de scientifiques du groupe pharmaceutique Servier.

Idéalement, les biostatistiques cherchent à répondre à des questions d'intérêt biologique et/ou clinique, à l'aide de techniques statistiques robustes et d'algorithmes efficaces. En réalité, il est extrêmement difficile, voire impossible, d'accomplir tout cela en même temps, sans doute à cause de la complexité intrinsèque de la biologie. Ainsi, il est important d'accepter certaines simplifications biologiques nécessaires, certaines limitations statistiques et certains imperfections algorithmiques. Prendre en compte toutes ces incertitudes permet de mieux analyser et mieux comprendre les résultats bioinformatiques et biostatistiques. Une bonne façon d'atteindre un équilibre entre tous ces éléments est le concept de modèle biostatistique. Un modèle biostatistique est conçu pour répondre à une question biologique spécifique. On peut le percevoir comme une collection de règles mathématiques qui sont

idéalement justifiées et compréhensibles biologiquement, prennent en compte les aléas pour permettre leur étude statistique, et enfin justifient les algorithmes (et si nécessaire les méthodes heuristiques) utilisées. Tout au long de ma thèse, j'ai essayé d'utiliser ce concept de modèle aussi souvent que possible afin de permettre une meilleure intégration des différents aspects (biologique, statistique et informatique) et finalement répondre aux questions biologiques ou cliniques de départ.

Conclusion

Cette thèse fait partie d'un projet plus large qui tente d'identifier de nouvelles cibles thérapeutiques pour les cancers du sein. Je me suis concentré sur l'analyse des données génomiques et transcriptomiques. J'ai conçu les plans d'expérience, j'ai ensuite analysé les données transcriptomiques avec des méthodes déjà disponibles et enfin j'ai proposé de nouveaux outils biostatistiques et des algorithmes efficace pour l'analyse des données génomiques. Mes analyses ont conduit à des listes de gènes et de voies de signalisation candidats. Le rôle de certains de ces candidats dans le développement tumoral a pu être confirmé in vitro.

Au-delà du cadre de cette thèse, en tant que partie intégrante d'un projet plus large, l'analyse omique va se poursuivre. En particulier, les données micro ARN et protéomiques n'ont pas encore été analysées (voir la figure 2). De plus, il sera important d'intégrer les informations apportées par ces différentes sources (ADN, ARN, microARN et protéines) pour chaque échantillon afin de mieux comprendre la pathologie moléculaire des TNBC, dans l'espoir d'identifier ainsi de nouvelles cibles thérapeutiques.

Contents

Part I

Introduction

Chapter 1

Overview

1.1 Introduction

Breast cancer is one of the most prevalent types of cancers and 1 woman out of 8 is diagnosed with breast cancer at some time in her life in the western world. Breast cancer is a very heterogeneous disease and can be divided into estrogen receptor (ER) positive and ER negative tumors (see Figure 1.1). ER positive tumors comprise Luminal A and Luminal B tumors. ER negative tumors include ER- / HER2+ tumors, which overexpress the Human Epidermal growth factor Receptor 2 (*HER2*) gene, and Triple negative (TNBC) tumors, that are ER negative, progesterone receptor (PR) negative and do not overexpress *HER2*. TNBC have a very high rate of chromosomal loss and gain, harbor less DNA amplifications than other breast cancer subtypes and have a very poor prognosis. Innovative and promising targeted therapies are currently explored for TNBC, such as poly ADP-ribose polymerase (PARP) inhibition, but there is still no targeted therapy for TNBC in routine clinical practice as there is for both HER2+ tumors (HER2 monoclonal antibodies) and for Luminal tumors (endocrine therapy). Many scientific teams throughout the world, both of clinicians and biologists, are working to understand better the clinical aspects and the biology of TNBC and would like to apply their research to clinical prospects and to suggest innovative and tailored, therapy.

The project I have been involved in is a collaboration between the Institut Curie (IC) and the

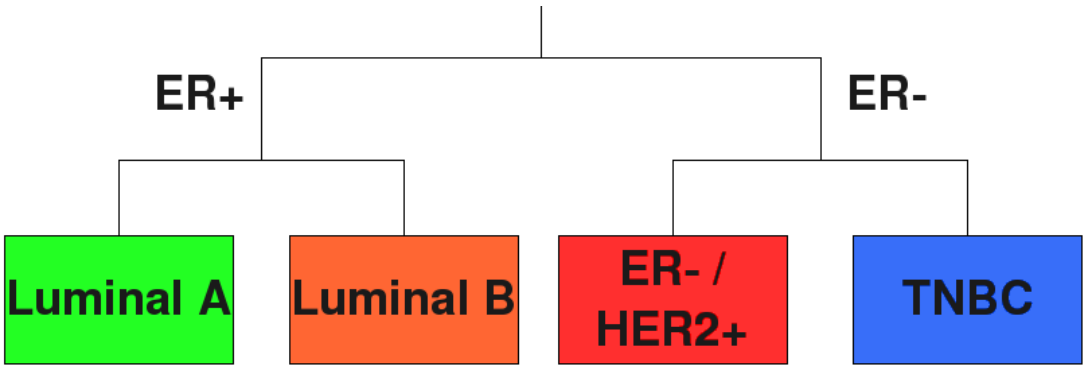


Figure 1.1: Classification of breast cancers between estrogen receptor (ER) positive and ER negative tumors. ER positive tumors comprise Luminal A and Luminal B tumors. ER negative tumors include ER- / HER2+ tumors and TNBC tumors

Servier pharmaceutical group. The goal of the project was to discover deregulated genes and signaling pathways in human TNBC to identify new therapeutic targets. This project is summarized in Figure 1.2.

My work can be subdivided in two main parts. First, I developed methods for the analysis of genomic data. I worked on the normalization of DNA copy number profiles and I proposed a method (ITALICS) for the normalization of Affymetrix SNP 100K and 250K arrays. I also worked on the segmentation of DNA copy number profiles. With Emilie Lebarbier (Ph.D.) and Stéphane Robin (Ph.D.), I proposed new algorithms and new statistical tools to assess the stability of segmentation and derive exact formulation of several model selection criteria. I developed and implemented an improved and faster dynamic programming algorithm that recovers the best segmentation with respect to the Euclidean norm. I applied these tools to the analysis of the Curie-Servier dataset.

Next, I worked on the analysis of the Curie-Servier transcriptomic data. The first step of my analysis was to plan the experimental design of the transcriptomic, genomic and miRNA experiments. It is a very standard problem yet it is often overlooked and unsurprisingly it resulted in an improved power to detect true differences between TNBC and other sample types and a more robust design. I

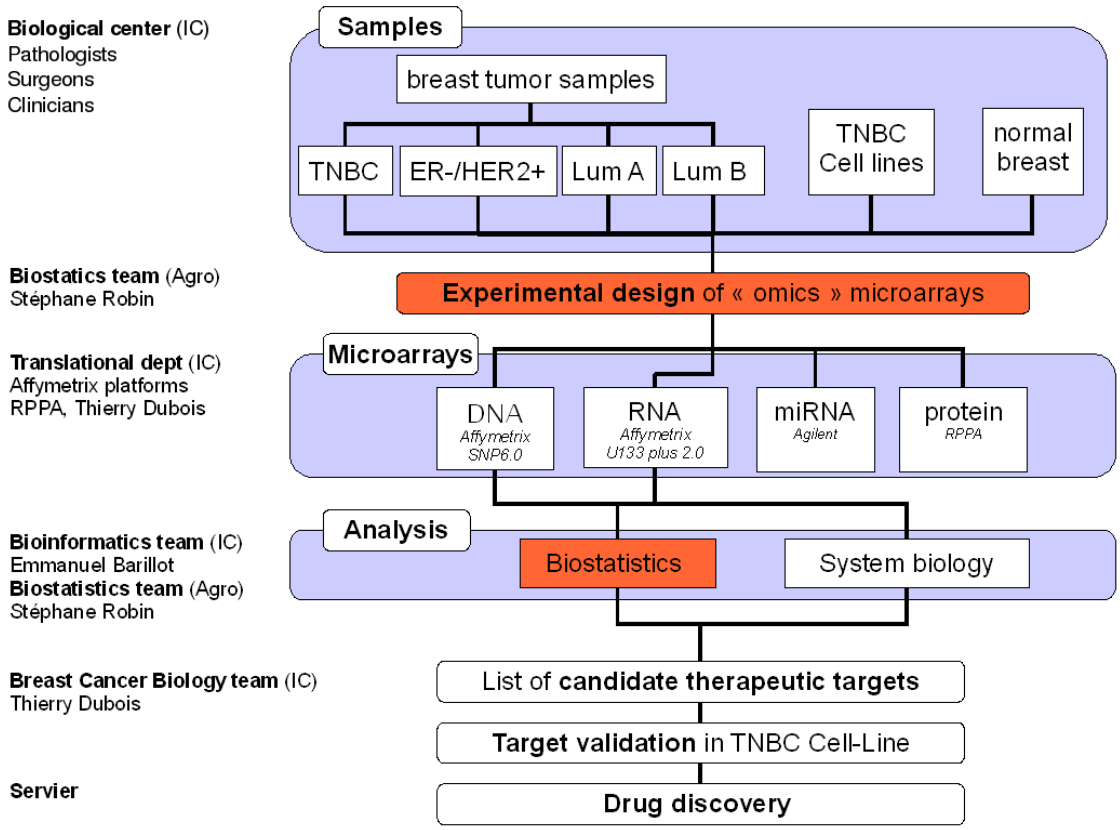


Figure 1.2: Both breast tumors of Luminal A, Luminal B, ER- / HER2+ and TNBC subtypes were selected and characterized by a pathologist from the IC using immunohistochemistry (IHC, Anne Vincent Salomon, MD/Ph.D.). TNBC cell lines and normal breast tissues were also collected. Transcriptomic, genomic, miRNA and proteomic microarray profiles were generated at the translational department of the IC. This huge amount of data has been analyzed in the IC bioinformatics team and the biostatistics team of Agroparistech (Agro) using biostatistical and system biology tools. This analysis has generated lists of potential targets. Some of these potential targets have already been further characterized and functionally validated in cell lines by a team of biologists at the translational department of the IC. **In this project, I have been in charge of the biostatistical analysis of the omic data** represented in red.

then analyzed the transcriptomic data using already available biostatistical and bioinformatical tools. I sought to characterize better the distinctness of TNBC at the transcriptomic level and its overlap with the immunohistochemistry (IHC) data. I worked at the gene and pathway level to identify genes or pathways that are deregulated in TNBC and propose lists of genes and lists of pathways of interest.

The analysis of high-throughput omics data is at the crossroad between biology, clinic, biotechnology, statistics and computer science. To achieve a balance between these different aspects, this thesis has been realized through a close collaboration between three main laboratories: a laboratory of biologists working on TNBC headed by Thierry Dubois (Ph.D.), a laboratory of bioinformaticians working on cancer data headed by Emmanuel Barillot (Ph.D.) and a laboratory of statisticians working on high-throughput biological data headed by Stéphane Robin (Ph.D.). Moreover throughout the thesis I have collaborated with several other groups of both biologists and clinicians and with scientists from the pharmaceutical group Servier.

Ideally biostatistics aim at answering biologically and/or clinically relevant questions, using well grounded statistical techniques and efficient computational schemes or algorithms. Probably due to the intrinsic complexity of biology it is extremely difficult, perhaps impossible, to achieve all these goals simultaneously, thus it is important to acknowledge the necessary biological simplifications, statistical limitations and algorithmic mishaps and account for all these uncertainties to better analyze and understand bioinformatical and biostatistical results. A good way to achieve this balance is the concept of a biostatistical model. A biostatistical model is made to answer a specific biological question. It can be viewed as a collection of mathematical rules that ideally should be biologically understandable and justified, account for randomness and enable its statistical study, and finally justify algorithmic computations and if necessary heuristics methods. Throughout the thesis I have tried to use this concept of a model as often as possible to enable a better understanding between biology, statistics and computer science and in the end answer the initial biological or clinical questions.

1.2 Methods for the analysis of DNA copy number profiles

TNBC have a very high rate of chromosomal gain and loss. These genomic alterations can be measured using various technologies such as CGH and SNP arrays and next-generation sequencing. The correct detection of these numerous alterations is important as we hope to identify tumor suppressor genes in frequently lost regions and oncogenes in frequently gained regions. The biostatistical analysis and biological interpretation of this kind of data is difficult for several reasons.

As for all microarray technologies, measurements are influenced by various non-relevant factors (for example the probes GC-content) and there is a need for efficient normalization methods. In collaboration with Philippe Hupé (Ph.D.), I worked on the normalization of Affymetrix SNP arrays and proposed a new method: ITALICS (Rigaill et al. (2008, 2007)). We have shown, at the time of the study, that ITALICS outperforms existing methods in terms of signal to noise ratio and enable a better classification of true recurrence and primary on a breast cancer data set (Bollet et al. (2008)). Moreover, for TNBC due to the many genomic rearrangements, recovering the ploidy of tumors is an important and difficult issue that we took into account in collaboration with Tatiana Popova (Ph.D., Popova et al. (2009)) from the group of Marc-Henri Stern (MD/Ph.D., IC).

Both CGH and SNP profiles are modeled as a succession of regions sharing the same copy number or LOH status. These regions are delimited by change-points or breakpoints corresponding to chromosome rearrangements. These profiles are usually analyzed using multiple change-points and segmentation methods. Most segmentation methods return a single segmentation, characterized by a set of breakpoints. Their qualities are rarely questioned. However, for an n-point profile there are 2^{n-1} possible segmentations, thus picking one segmentation out of so many is obviously a difficult task. To make a valid biological interpretation we would like to be sure that the best segmentation is by far the best fit to the data. If it is not the case we would like to check that the second best, third best and more generally other good segmentations do not have a completely different set of change-points. I have been working on this problem with Emilie Lebarbier (Ph.D.) and Stéphane Robin (Ph.D.) and proposed new algorithms and statistical tools (Rigaill et al. (2010c,d)) to assess and take into account the uncertainty of change-point estimation. From these algorithms and statistical tools we derive

exact formulation of model selection criteria (to select the number of breakpoints) that used to be asymptotically approximated.

The Affymetrix SNP 6.0 technology scans around 2.10^6 million positions along the human genome and thus around 2.10^5 probes per chromosome. For these very dense and large profiles even recovering the most likely segmentation is a very difficult task and the fastest algorithm had a runtime quadratic in the size of the data and it took several days to analyze one SNP 6.0 profile. Thus most methods rely on heuristics to reduce the computation time. However, this is done at the price of some errors as heuristics do not recover the best segmentation but rather a good candidate segmentation. This is clearly a problem for biological interpretation as we cannot guarantee that there is not a better way to segment the data. I proposed a new algorithm that recovers the best segmentation in an almost linear runtime and it takes a few minutes only to analyze an SNP 6.0 profile (Rigaill (2010a,b)).

All these statistical and algorithmic developments were applied to the Curie-Servier breast cancer genomic data and allowed us to identify in particular the loss of PTEN (Marty et al. (2008)) in more than 50% of TNBC tumors. Several other regions of the genome were identified and resulted in a list of candidate genes.

1.3 Biostatistical analysis of the transcriptomic Curie-Servier dataset

Experimental design

From my experience and the experience of others, the most critical step by far in any data analysis is the experiment itself and a bad experiment will always lead to bad analyses and poor results. The goal of experimental design is to ensure that the way the experiment is conducted will actually enable us to answer the main biological question. Omic experiments, as others, must therefore be carefully planned to take into account various, identified and unsuspected, non-relevant factors. From a statistical perspective, it is possible without any data but with some biostatistical model in mind to compare two experimental designs and assess their respective power to detect some biological pattern

of interest.

The biostatistical model I designed for the project is relatively simple, yet it helped us to clarify commonly made assumptions. First we expect that samples of the same histological type will have similar mRNA measurements. Unfortunately, on the day of the experiment, the temperature during the experiment and many other non-relevant factors influence these measurements. However, hopefully, these non-relevant effects are relatively independent of the biological signal and they can be corrected with a good experimental design. The main biological question of the transcriptomic experiments was what are the mRNA differences between TNBC and normal samples but we were also interested in detecting differences between TNBC and other tumor types.

Keeping all this in mind I constructed experimental designs to answer these questions. The objective was to maximize our ability to detect true differences between TNBC and other histological types. The main issue is that the set of possible designs is large even for computers. For the transcriptomic, genomic and miRNA experiments, I explored either exhaustively when it was possible or stochastically the set of all possible designs to choose one with a powerful ability to detect differences between TNBC and other sample types. Moreover, I randomized samples of the same type to account for unsuspected non-relevant factors.

Transcriptomic data analysis

One of the main problems when analyzing TNBC is that they are a very distinct subgroup of tumors and harbor many genetic, genomic and transcriptomic alterations. Therefore, it is easy to find genes or/and pathways that are differentially expressed in TNBC. For example, from our transcriptomic data I found that almost half of the analyzed genes were differentially expressed in TNBC compared to normal samples. However, the goal is to find driver alterations or key events in the TNBC tumorigenesis and it seems biologically reasonable to think that most differences between TNBC and other sample types are passenger alterations and will not necessarily lead to potential therapeutic targets.

Thus, in collaboration with Anne Vincent Salomon (MD/Ph.D., pathologist), we sought to characterize better the distinctness of TNBC at the transcriptomic level and understand the transcriptomic

classification of our tumors using unsupervised classification methods. In particular, I compared the transcriptomic-based classification with the IHC-based classification and found that they were in relatively good concordance (less than 15% discordance). Moreover, I assessed the influence of the set of genes used for the transcriptomic classification. Interestingly, the transcriptomic classification and in particular the TNBC cluster seems relatively independent of the set of genes used for the classification (Rigaill et al. (2010b), in preparation). This independence also suggests that identifying a small robust set of genes characteristic of the TNBC is intrinsically difficult.

Once we had acknowledged the specificity of TNBC, we decided to refine our search for genes of interest using biological or clinical information. First, I focused on specific sets of drugable genes such as kinases, and in a collaboration with the team of Philippe Chavrier (Ph.D.), on the Diaphanous-Related Formins (Lizárraga et al. (2009)). This mechanically increased our statistical power to detect differences in these sets of genes and thus our chances of detecting interesting transcriptomic modifications. I also used pathway/genesets analysis using the globaltest software and the KEGG and GO databases. However, at the pathway level the distinctness of TNBC is even more of a critical issue because in a list containing a lot of differentially expressed genes it is extremely easy to find a geneset with many deregulated genes. Thus, I focused on some highly significant pathways of biological interest to characterize and understand better their expression pattern. These in-depth analyses of smaller sets of genes also mechanically increased our statistical power and allowed us to identify some key transcriptomic patterns and/or regulation events, more specifically in the Wnt pathway (Rigaill et al. (2010a), in preparation) and in oxidative stress pathways (Toullec et al. (2010)) in collaboration with the team of Fatima Mechta Grigoriou (Ph.D.).

These analyses generated different lists of genes. These genes of interest were validated on other publicly available transcriptomic datasets and some have been experimentally validated using TNBC cell lines (using for example clonogenic, survival and apoptosis assays).

1.4 Conclusion

This thesis is part of a larger project aiming at identifying new therapeutic targets for TNBC. I focused on the analysis of transcriptomic and genomic data. I designed the experiments, used already available methods for the transcriptomic data and proposed new biostatistical methods and algorithms for the genomic data. These analyses resulted in lists of genes or pathways that are deregulated in TNBC, some of which have been further validated in vitro.

As part of a larger project the omic analysis will continue after this thesis. In particular the miRNA and proteomic data have not yet been analyzed and obviously all these different sources of information (DNA, RNA, miRNA, proteins) generated for each sample will have to be integrated to understand better the molecular pathology of TNBC and hopefully this will lead to the identification of new therapeutic targets.

Chapter 2

A small introduction to Triple Negative Breast Cancers

This section is a short introduction to the biology of breast cancer and more specifically of TNBC (Triple Negative Breast Cancers).

2.1 Breast cancers

Epidemiology and risk factors of breast cancers

Throughout the world, breast cancer is the most prevalent type of cancer among women and there are approximately 1.1 million new cases of breast cancer every year. Breast cancer is also the leading cause of cancer deaths in women with 410 000 deaths every year (Vincent-Salomon (2008)). The incidence of breast cancers has regularly increased. In the USA and Europe, this is partly due to the setting up of mammography screenings. Breast cancer survival rates are around 73% in occidental countries and 57% in under-developed countries. The prevalence and incidence of breast cancers greatly differ from one country to another one. Developed countries tend to have higher incidences (Parkin (2004)). Yet, socio-economical factors are not the only ones that should be taken into account. For example, Japan has a very low incidence of breast cancer. Apart from socio-economical factors, biological factors

such as genetic background and environmental factors are important to explain this variation between countries.

Indeed, some genetic factors have been known to be related to breast cancers: the Collaborative Group on Hormonal Factors in Breast Cancer (2001) reported that the relative risk of breast cancer significantly increases with the number of first degree relatives that have been affected. Several genetic factors have been clearly identified. They are discriminated between the following groups (Mavaddat et al., 2010):

- High penetrance mutations. In this case a single allele conferring a high risk is responsible for the disease. Typical examples are mutations occurring in *BRCA1* and *BRCA2* genes.
- Moderate penetrance mutations. They regroup uncommon variants associated to moderate risk increase. *CHEK2* and *ATM* are two examples.
- Low penetrance variants. They encompass variants associated to a very small risk. Yet, it is likely that most of the unexplained fraction of familial risk could be explained by these low penetrance variants assuming they are sufficiently numerous.

To conclude on genetic factors, only a minority of familial risk factors are explained by known genetic variants. It is hoped that large sequencing projects and high-density SNP arrays will enable us to discover new variants.

Various environmental factors have been correlated to the development of breast cancer and they might account for 75% of all cases of breast cancer (Ellsworth et al., 2004). Instability in genes that maintain genomic integrity, as well as exogenous chemicals and environmental pollutants are involved. For example, long estrogen exposure (e.g. caused by early puberty, late menopause or hormonal replacement therapy) increases the risk of breast cancer through increased cell proliferation and/or DNA destabilization through depurination (Yager and Davidson, 2006).

Heterogeneity of breast cancers

An important feature of breast cancers is their heterogeneity. This heterogeneity can be seen at many levels. First, breast cancers can be segregated in subgroups according to their histological grade. The

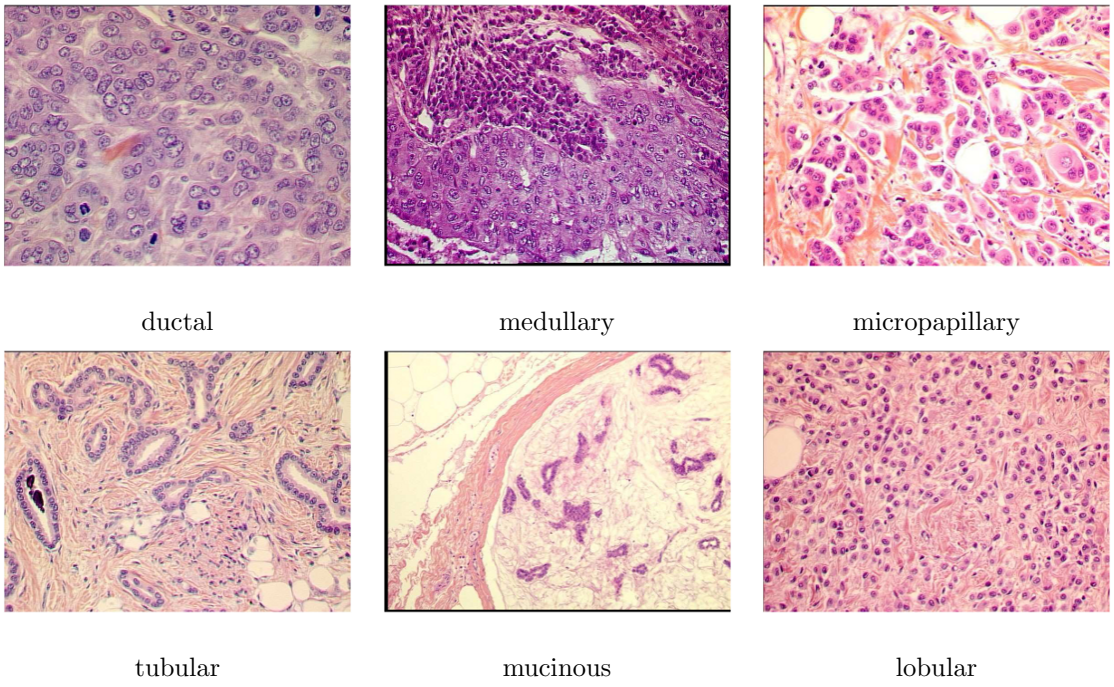


Figure 2.1: Morphological heterogeneity of breast tumors. Six histological sections of six invasive breast cancers stained with Hematoxilin Eosin Saffron (HES) representing six of the histological types defined by the WHO (Tavassoli et al., 2003). Images provided by Anne Vincent-Salomon (MD/Ph.D., Institut Curie).

histological grade takes into account both tumor differentiation and proliferation and it is a validated prognostic factor to determine breast cancer therapy. For example, it is used in the Nottingham prognostic index (Galea et al., 1992; Blamey et al., 2007). Breast cancers can also be classified according to their histological type. These types correspond to specific morphological and cytological patterns (see Figure 2.1). The most common of these types is the Invasive Ductal Carcinomas of No Special Type (IDC-NST). These IDC-NST represent approximately 75% of all breast cancers and correspond to cancers that do not exhibit any characteristic of the special histological types (Weigelt et al., 2010b). Overall, the WHO (World Health Organization) defined at least 17 different histological types of breast cancer (Tavassoli et al., 2003).

The heterogeneity of breast cancers can be further decomposed using immunohistochemical features. For example, breast cancers are subdivided according to the estrogen receptor (ER) status

and/or the Human Epidermal growth factor Receptor 2 (HER2) status.

IDC-NST breast tumors can also be classified according to their transcription profile (Perou et al., 2000; Sørlie et al., 2001a; Sørlie, 2003; Chin et al., 2006). This classification is called the molecular classification (see Figure 2.2). Hierarchical clustering of breast cancer transcription profiles segregates ER+ from ER- tumors. The ER+ class is further subdivided in Luminal A and B. Luminal A tumors have high levels of expression of ER-activated genes and low proliferation signature. Luminal B cancers usually have a higher histological grade and proliferation rates, and a worse prognosis. Some Luminal B tumors overexpress the *HER2* gene and thus are ER+ / HER2+. The ER- group is subdivided in Normal-like, ER- / HER2+ and Basal-like (ER- and absence of HER2 overexpression). The Normal-like group might be an artifact due to normal tissue contamination (Parker et al. (2009); Peppercorn et al. (2008)). ER- / HER2+ tumors over-express the *HER2* gene. Basal-like tumors usually have a high histological grade, high mitotic index, central necrosis and pushing borders. The molecular analyses of breast cancers has revealed and brought to the forefront various types of breast cancers and provide new insights on the biology of breast cancers. Nevertheless, as we will see (in section 10.2), the stability, reproducibility and clinical use of this classification have been questioned. In the last years, other ER- groups have been identified: the apocrine (Farmer et al., 2005; Doane et al., 2006), interferon (Hu et al., 2006) and claudin-low groups (Herschkowitz et al., 2007; Hennessy et al., 2009).

To conclude, all these various classifications (molecular, histological...) suggest that breast cancer is, in fact, a collection of different diseases affecting the same organ. This heterogeneity of breast cancer raises one obvious question: what is at the origin of these different types? This is a debated question. One hypothesis is that molecular groups of breast cancer correspond to different cell types originally becoming cancerous (Polyak, 2007; Vargo-Gogola and Rosen, 2007). More specifically, the cells from which the tumor originates could either be breast stem cell (Stingl, 2009) or their progenies, which would be the cause of the heterogeneity. Breast stem cells have the ability to renew themselves through mitotic division and differentiate into any specialized breast cell type.

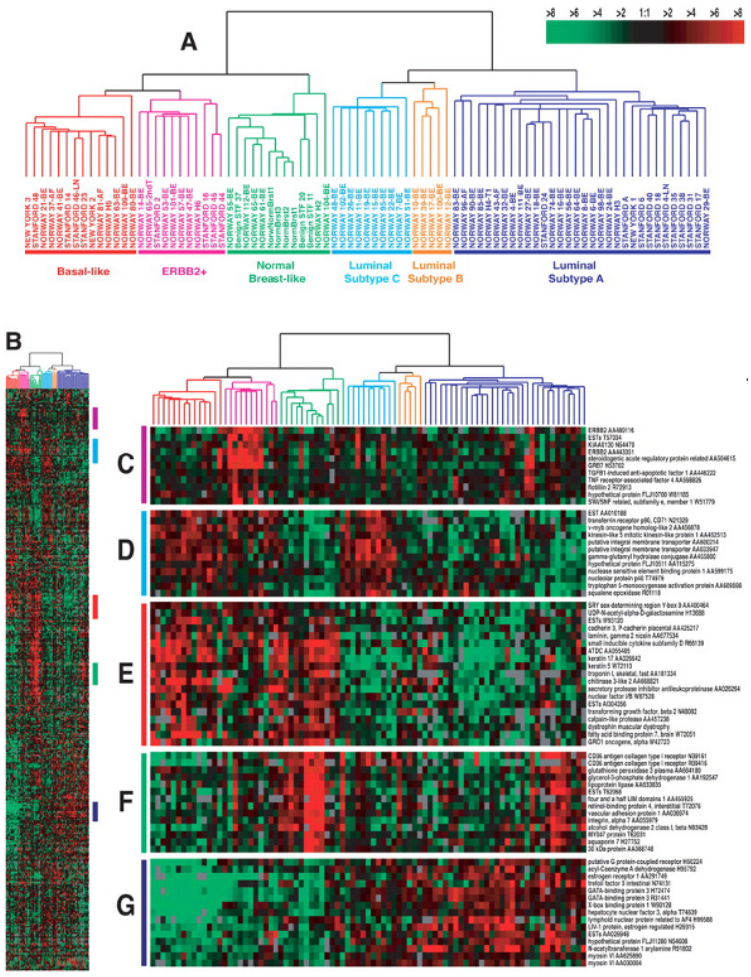


Figure 2.2: Molecular classification of breast cancer from mRNA expression profiles - Gene expression patterns of 85 experimental samples representing 78 carcinomas, 3 benign tumors, and 4 normal tissues, analyzed by hierarchical clustering using the 476 cDNA intrinsic clone set. (A) The tumor specimens were divided into 6 subtypes based on differences in gene expression. The cluster dendrogram showing the 6 tumor subtypes are colored as: luminal subtype A, dark blue; luminal subtype B, yellow; luminal subtype C, light blue; Normal-like, green; Basal-like, red; ERBB2+, pink. (B) The full cluster diagram scaled down. The colored bars on the right represent the inserts presented in C-G. (C) ERBB2 amplicon cluster. (D) Novel unknown cluster. (E) Basal epithelial cell-enriched cluster. (F) Normal-like cluster. (G) Luminal epithelial gene cluster containing ER. (images and legend from Sørlie et al. (2001b)).

2.2 Triple Negative and Basal-like breast cancers

Triple Negative Breast Cancers (TNBC) are immunohistochemically characterized by the absence of ER and progesterone receptors (PR) and the lack of HER2 overexpression. Due to its aggressiveness, poor prognosis and lack of targeted therapy, these particular tumors are the focus of many research studies. Although the match is not perfect, there is a good correspondence between TNBC and basal-like tumors. Basal-like tumors were identified based on the hierarchical clustering of IDC-NST gene expression profiles while TNBC can be either IDC-NST or one of the special histological types. Overall, the exact definition of Basal-like tumors in comparison to TNBC and the use of the term "basal" is still subject to debate (Gusterson et al., 2005; Gusterson, 2009; Moinfar, 2008). Indeed, no consensus has been reached to identify this group using immunohistochemistry (Rakha et al., 2008; Reis-Filho and Tutt, 2008). In the Curie-Servier dataset, Basal-like tumors were identified as ER-, PR-, lack of HER2 overexpression IDC-NST tumors that express either cytokeratin 5/6 and/or cytokeratin 14 and/or Epidermial Growth Factor Receptor (EGFR). In the following, I will use both "TNBC" and "Basal-like" names, even though they are not strictly equivalent, to describe IDC-NST breast tumors that have a basal or TNBC related pattern.

Overall TNBC have high histological grades with a high mitotic index and they frequently harbor central tumor necrosis. These tumors are characterized by an impaired DNA repair process and harbor complex genomic rearrangements and more gains and losses than the luminal subtypes (Chin et al., 2006; Vincent-Salomon et al., 2007). It has also been shown that 85 % of the tumors of patients with *BRCA1* mutations have a TNBC immunophenotype (Foulkes et al., 2003). Moreover, TNBC are associated to high levels of various proliferation genes such as Ki-67, and very frequent p-53 mutation (Manié et al., 2009).

From a clinical point of view, TNBC are relatively chemo-sensitive. Indeed, these tumors show more pathological complete response to neoadjuvant chemotherapy than other types of tumor (Rouzier et al., 2005). Showing pathological complete response means that the tumor is no longer detectable. The poor overall survival rate of patients with TNBC is explained by the fact that among those patients that do not show a complete response, there is a very high number of relapses (Podo et al.

(2010) and references therein). Alternative approaches to chemotherapy are currently explored such as targeting the EGFR, the topoisomerase 2A (TOP2A), c-MYC and vascular endothelial growth factor (VEGF) receptor (Podo et al. (2010) and reference therein). One of the most promising treatments at the moment is poly ADP-ribose polymerase (PARP) inhibition. In BRCA1-defective cells, inhibition of PARP leads to the accumulation of DNA double-strand breaks that are not correctly repaired due to the lack of functional BRCA1. This leads to tumor cell death (McCabe et al., 2006). In normal cells PARP inhibition has a limited effect due to active BRCA1. The general principal behind PARP inhibition is synthetic lethality (Tucker and Fields, 2003). Synthetic lethality occurs when two otherwise non-lethal changes result in cell death when present together. PARP inhibition showed promising results in *BRCA1*-mutation carriers (Fong et al., 2009). As mentioned earlier, many *BRCA1*-mutation carriers present TNBC and it has been hypothesized that at least a fraction of TNBC are BRCA1 deficient (Turner et al., 2006) due to the expression of ID4 (a negative regulator of *BRCA1*) or the epigenetic silencing of *BRCA1* (Veeck et al., 2010; Evers et al., 2010). To conclude there is still no targeted therapy for TNBC available in routine clinical practice. Hopefully a better understanding of the biology of these tumors and their links to *BRCA1* mutations will lead to the development of new treatments for these cancers.

2.3 Breast tumors of the Curie-Servier cohort

For the Curie-Servier project breast tumors of Luminal A, Luminal B, ER- / HER2+ and TNBC subtypes were selected and characterized by a pathologist (Anne Vincent Salomon, M.D./Ph.D.) of the IC using immunohistochemistry (IHC, Anne Vincent Salomon and Marion Richardson, M.Sc.). These tumors were obtained from patients treated at the IC (Biological Resource Center) and contain between 50% and 90% tumor cells. Many features of these tumors were collected such as the size of the tumor and the overall survival of the patients. Additionally normal tissues from mammoplastic surgery were collected by Anne Vincent Salomon and Fabien Reyal (M.D./Ph.D.). Finally, cell-lines characterized as TNBC in Neve et al. (2006) were obtained: 184B5, MDA-MB-436, HCC1143, HCC1187, BT20, HCC1937, MCF-12A, HCC38, Hs 578T, MCF-10A, MDA-MB-468, BT-549, HCC70, MDA-MB-157,

	TNBC	ER-/ HER2+	Luminal A	Luminal B	TNBC cell-lines	Normal Tissue
ER/PR status	-	-	+	+	-	
overexpresion of HER2	-	+	-	+ / -	-	
CK14 or CK5/6 or EGFR	+	-	-	-		
Grade	III	III	I	III		
Number	46	33	35	40	16	19

Figure 2.3: Summary of the samples of the Curie-Servier dataset and their Histological and Immunohistochemical characterization.

MDA-MB-231. All the information on the samples are summarized on Figure 2.3.

This means that the different subtypes are known before any of our analyses. This information can be used to confirm the groups we find, and also earlier for experimental design, in particular the information can be taken into account to determine batches and make sure batch effects are not responsible for the differences we observe between subtypes.

Part II

Genomic Analysis

Chapter 3

Chromosome aberrations

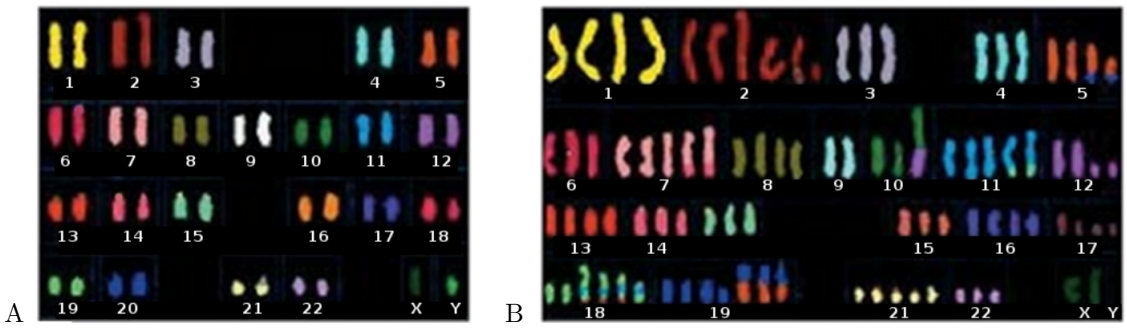


Figure 3.1: Karyotype of two cells. (A) The karyotype of a normal cell, which contains two copies of each autosome. (B) The karyotype of a colon cancer cell, which typically shows many aberrations in chromosome number and structure. (From Alberts et al. (2002))

In a normal human cell, chromosomes go by pairs, excluding sex-determining chromosomes (see Figure 3.1 A). Thus, most regions of the genome are present in two copies, one coming from the mother and the other one coming from the father. This balanced state is called euploidy. Deviation from this normal state is called aneuploidy and is often observed in cancer cells (see Figure 3.1 B). This aneuploidy is often the consequence of the genomic instability of tumor cells. More precisely, due to an accumulation of defects in DNA repair pathways, in cell cycle check-points and in mitotic

segregation pathways, tumor cells often fail to properly carry out the duplication and segregation of chromosomes and accumulate chromosome aberrations (Aguilera and Gomez-Gonzalez, 2008).

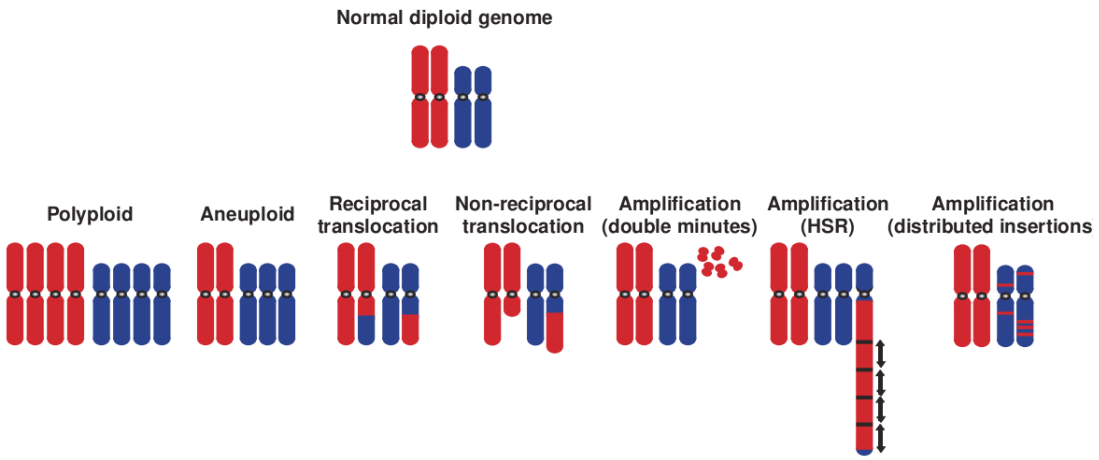


Figure 3.2: Schematic view of typical chromosome aberrations that are found in cancer cells. Based on Albertson et al. (2003)

Typical tumoral alterations are described below and schematically represented in Figure 3.2:

Polyploidy A number of chromosomes have p copies with p greater than 2.

Euploidy At least one chromosome has an abnormal number of copies.

Translocation A chromosome translocation is a rearrangement of parts between chromosomes, whether or not they are from the same original pair. Translocations can be reciprocal, i.e. there is an exchange between the chromosomes and no regions are lost or gained. But translocations can be non-reciprocal, i.e. a part is gained and/or lost during the rearrangement. Translocations might create fusion genes or truncated genes.

Amplification A small contiguous portion of the genome is present in a high number of copies (from 4 to over 50 copies). These copies can be isolated fragments without centromeres and are called double minutes. Otherwise, they can be incorporated into chromosomes, either in nearly

contiguous homogeneously staining regions (HSR) or interspersed in the genome.

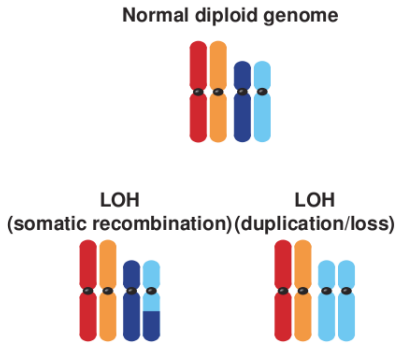


Figure 3.3: Schematic view of LOH aberrations that are found in cancer cells. Based on Albertson et al. (2003)

Moreover, chromosomal aberrations do not necessarily produce abnormal karyotypes. Indeed, it has been observed in cancer cells that both chromosomes of a given pair come from the same parent. One of the chromosomes has been lost and the other one has been duplicated. This is called Loss of Heterozygosity (LOH) without DNA copy number change (see Figure 3.3). When the LOH rearrangement concerns a whole chromosome, it is called isodisomy and when it impacts only a portion of the chromosome, it is called partial isodisomy or somatic recombination.

Some of these alterations are responsible for the development of cancer. For example, in ER- / HER2+ breast tumors, an amplicon around the *ERBB2* gene (also called *HER2* gene) on chromosome 17 leads to the over-expression of the ERBB2 protein and this over-expression in turn induces the activation of the PI3K / AKT pathway, a pathway which is known to affect tumor development. This particular case is well known and the ERBB2 protein is now targeted by therapy. More generally, it is thought that lost regions harbor tumor suppressor genes while gained regions harbor oncogenes. Thus, the study of chromosome aberrations in tumor cells, and more specifically the identification of

frequent aberrations, is a way to identify new oncogenes or tumor suppressor genes.

3.1 Some technologies to study genomic rearrangements

There are many different technologies to study these events such as Comparative Genomic Hybridization (CGH) arrays and Single Nucleotides Polymorphism (SNP) arrays. Historically, the genome-wide study of DNA copy number changes was performed using the CGH technique, which was developed in the early 1990s. In this technique, total genomic DNA is isolated from tumor and normal control cells, labeled with different fluorochromes and hybridized to normal metaphase chromosomes (Kallioniemi et al., 1992). This technique is therefore called chromosomal CGH. Differences in the tumor fluorescence with respect to the control fluorescence along the metaphase chromosomes are then quantified to reflect changes in the DNA copy number of the tumor genome. Subsequently, array CGH, where arrays of genomic sequences replaced the metaphase chromosomes as hybridization reporters, was established (Solinas-Toldo et al., 1997; Pinkel et al., 1998) and solved many of the technical difficulties and problems caused by working with cytogenetic chromosome preparations. The main advantage of array CGH is its ability to perform copy number analyses with a much higher resolution compared to chromosomal CGH (resolution smaller than a megabase compare to several megabases for chromosomal CGH). Array CGH has already been widely used in oncology for many purposes such as global analysis of copy number aberrations, identification of putative target genes, tumor classification or assessment of clinical significance of copy number changes (Kallioniemi, 2008). Pinkel and Albertson (2005) give details in their review about the technology and its application in oncology. Here, we will present only the general outline of the protocol (see Figure 3.4):

1. Total genomic DNA is isolated from a tumor sample (i.e. the test DNA) and from a normal sample (i.e. the reference DNA). Genomic DNA is then digested with a restriction enzyme and the obtained DNA fragments are labeled. The tumoral DNA is usually labeled with a red fluorochrome and the normal DNA with a green fluorochrome.

2. Both the tumoral and normal DNA are hybridized on the same chip. For each spot, there is

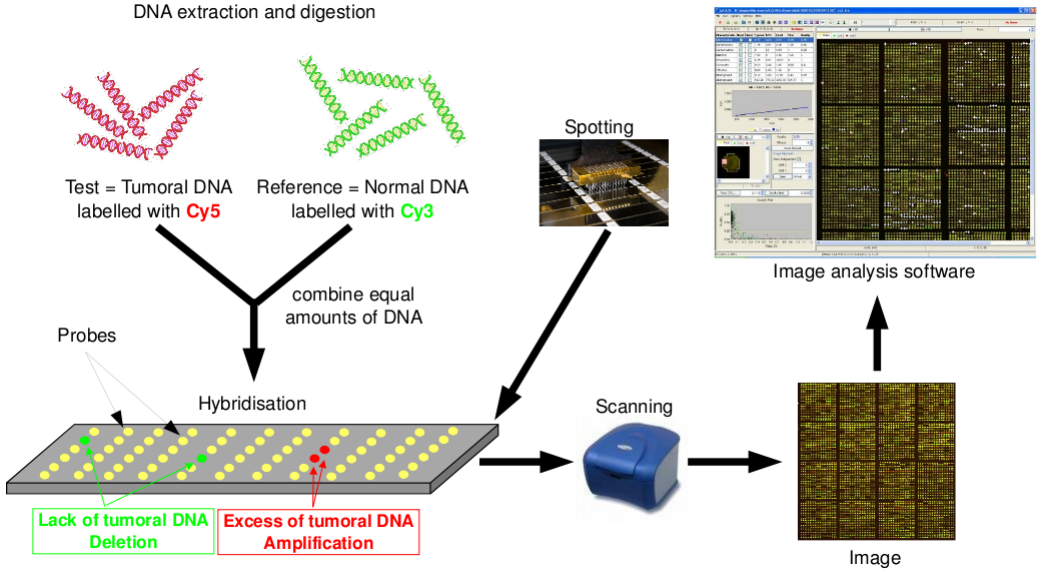


Figure 3.4: Schematic view of the CGH protocol (From Philippe Hupé).

a competitive hybridization between the tumoral DNA target sequences and the normal DNA target sequences.

3. After hybridization, the chip is scanned and the signal intensity is quantified for both the red and green wavelengths. Image files are created in which each pixel is given a red and a green intensity.

4. An image analysis software reconstructs the signal intensity for each spot.

In the case of SNP arrays, the protocol is quite similar except that there is no normal DNA reference.

3.2 DNA copy number profiles of SNP and CGH arrays

In CGH arrays, the DNA copy number is obtained by comparing the test sample with a normal reference sample. This is often done with the ratio of the measured intensity of the test sample and reference. For example, a ratio of 1 means that the usual 2 copies of DNA are present in the test

sample (see Figure 3.5). While doing this, one assumes that the DNA copy number of the reference is 2. However, it is not necessarily the case as copy number polymorphisms are common in the healthy population.

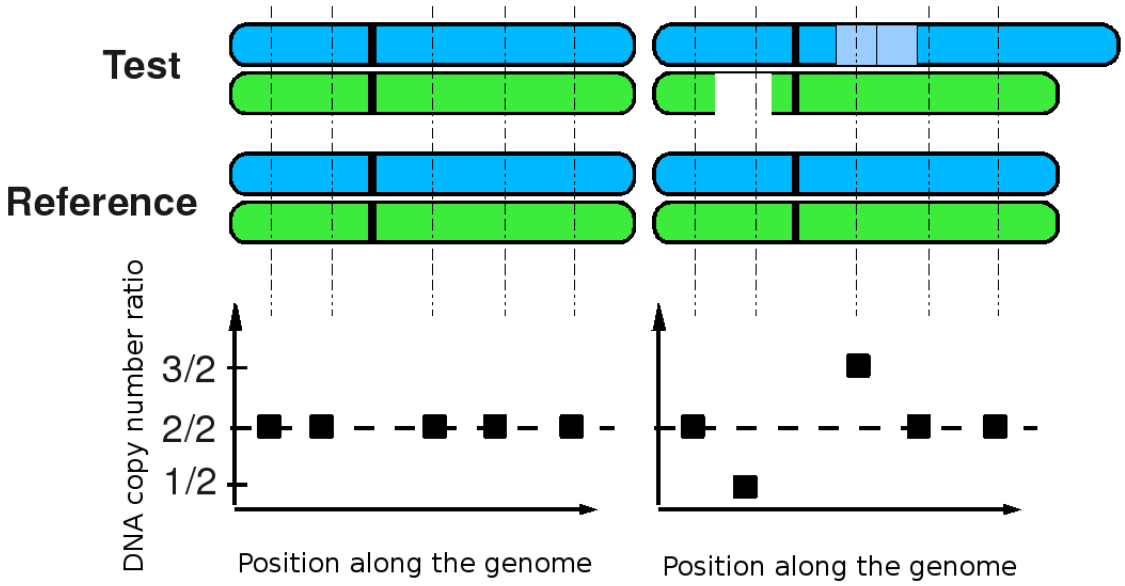


Figure 3.5: Illustration of chromosomes and the corresponding CGH profiles. Using CGH arrays, a test sample is compared to a reference. (Left) If the two samples have the same DNA copy number, the profile should be a set of points on the 2/2 ratio line. (Right) If there are differences between the samples, points in gained regions appear above the 2/2 ratio and points in lost regions appear below the 2/2 ratio.

The main difference between CGH and SNP arrays is that SNP arrays usually do not use a reference (see Figure 3.6). Moreover, the DNA copy number is not measured directly but rather computed as the sum of the intensities of both alleles. In this way, one observes differences between regions of the genome but it is not necessarily easy to determine the intensity that corresponds to 2 DNA copies. This is all the more true for TNBC as they harbor many rearrangements and one cannot assume that the mean or median intensity of all probes corresponds to the intensity of 2 DNA copies.

Studying SNP arrays also gives information about LOH. This is valuable information to recover

the reference intensity of 2 DNA copies in TNBC (Popova et al., 2009).

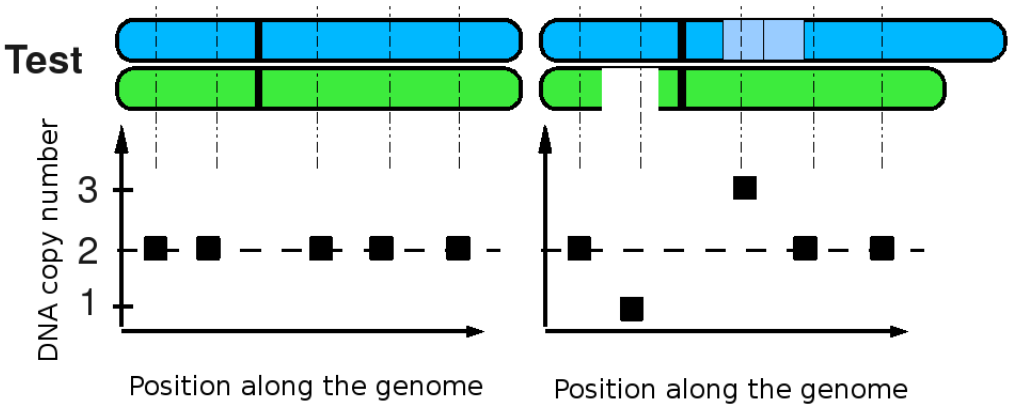


Figure 3.6: Illustration of chromosomes and the corresponding SNP profiles. Using SNP arrays, a test sample is analyzed without any references. The level corresponding to 2 DNA copies is usually chosen as the mean or median intensity of the whole profile. (Left) If there is no DNA copy number changes, the profile should be a set of points on the same line. (Right) If there is some changes, points in gained regions appear above others and points in lost regions appear below others.

For both CGH and SNP arrays, we expect a limited number of possible values for the measured intensity. If we could measure the copy number almost continuously along the genome, we would expect a constant signal, except for a few abrupt changes corresponding to gains and losses. However, there are measurement errors and noise is observed around the signal, which complicates the analyses.

3.3 An overview of CGH data analysis

Many methods have been developed specifically to analyze CGH arrays (see the review by van de Wiel et al. (2010)). They can be divided into two categories: pre-processing and downstream methods.

Pre-processing methods are a critical step because their results affect any following analyses and their biological interpretation. Pre-processing usually consists in the following steps:

Control Assess the quality of the experiment via a number of checkpoints.

Normalization Remove artifacts that hamper our ability to extract the biological signal.

Segmentation Divide the genome into regions sharing the same DNA copy number.

Calling Recover the DNA copy number (0, 1, 2, 3...) or at least try to make the difference between normal, gained and lost regions.

Once these different steps have been performed, many different types of downstream analyses can be performed depending on the biological or clinical questions. Many specific methodologies have been proposed to identify:

- recurrently aberrant regions (across tumors);
- new subgroups of cancer (unsupervised classification);
- markers associated to prognosis, diagnosis or other clinical variables of interest (supervised classification or regression).

In the following chapters, I will highlight some of my contributions to the normalization (Rigaill et al., 2008) and segmentation (Rigaill et al., 2010c; Rigaill, 2010b) of these DNA copy number profiles.

Chapter 4

Normalization of DNA copy number profiles

In microarray experiments, as in many experimental protocols, measurements are influenced by non-relevant factors that hamper our ability to extract the signal of interest. The intensity of a probe is affected by three elements:

- the biological signal, which is the level of the fragment of interest (either mRNA or DNA);
- some systematic biases such as the probe GC content or spatial artifacts;
- some random factors that take into account the inevitable variability between repeated measurements.

In this context, normalization aims to remove the systematic biases while preserving the biological signal, namely the biological signal. The random factors cannot specifically be taken into account because they are random.

In this section, I will first give a short overview of current issues regarding microarray normalization. Then I will pinpoint some of the normalization specificities for tumor DNA copy number array. Finally, I will present the ITALICS method (Rigaill et al. (2008), the article is provided in subsection 4.4) that I developed to normalize Affymetrix 50K and 250K SNP arrays.

4.1 Short overview of microarray normalization

Microarray normalization is often referred to as a preprocessing step. Indeed, it is the first step of many microarray analysis pipelines. Therefore it influences the results of all further analyses and more importantly the biological interpretation of these results. Thus, it is a critical issue and an extensively studied problem. Many normalization methods have been proposed, especially for one-color gene expression microarrays, see Binder et al. (2010) for a review. There are two main reasons which make it difficult to find an efficient and understandable normalization method:

- knowledge about the underlying hybridization mechanisms is incomplete;
- tools to assess the quality of a given normalization procedure are unsatisfactory.

These two issues are further detailed below.

First, any microarray normalization procedure relies on a model that describes the relationship between the probe intensity, the level of the mRNA or DNA fragment of interest and some non-relevant phenomena. Some of the phenomena are relatively well understood, such as probe duplex formation in solutions which depends on the probe sequence. This can be described using a nearest neighbor thermodynamic model (SantaLucia, 1998) and seems to work quite well on microarrays (Binder et al. (2009) and reference therein). Moreover, quite recently, it has been shown that surface hybridization could be modeled using an adsorption model such as the Langmuir adsorption equation (Binder et al., 2008). Many other phenomena are less known and they control the specificity and sensitivity of a given probe such as steric hindrance, RNA or DNA secondary structure formation and probe-probe interactions (Zhang et al., 2003). Therefore, many normalization models are based on (sometimes questionable) statistical considerations rather than physical or thermodynamical considerations. For example, for mRNA expression arrays, it is generally assumed that the majority of genes are not differentially expressed and that the proportions of down-regulated and up-regulated genes are similar. Similarly, in the ITALICS method (Rigaill et al., 2008), the "mean intensity" of a given quartet (PM probe of the A and B allele) across a reference dataset was used to correct the measured intensity (see subsection 2.1 paragraph "Non-relevant sources of variation" on page 2 and subsection 2.2 paragraph

"Non-relevant effect estimation" on page 3 of the ITALICS paper). We have empirically shown that this correction dramatically increases the quality of the data (see subsection 3.2 on page 4 of the ITALICS paper). This "mean intensity" is a very good surrogate of the sensitivity and specificity of the probe, yet we poorly understand what non-relevant effects it takes into account. Overall, normalizing microarrays is based on correcting their signal using two complementary types of information: biological and physical knowledge of the mechanisms and empirically validated normalization tricks.

The second issue, is the way to assess the quality of a given normalization procedure. Indeed, one would like to know which normalization method is the best or which method should be used in a given context. For two-color microarrays this is relatively easy and one can assess the performance of a given normalization method using an Anova (Kerr et al., 2000; Cui et al., 2003). That is not the case for one-color microarrays. To assess the performances of a one-color microarray normalization method one can use a benchmark dataset. In the case of expression profiling arrays, there are several, e.g. spike-in studies (Irizarry et al., 2003b) and dilution series (Bolstad et al., 2003). Based on these datasets, it is possible to compare the precision and accuracy of different methods or in other words their ability to reduce the variance without introducing any biases. These benchmark datasets are certainly not perfect to assess the quality of normalization methods and several other strategies and statistical criteria have been proposed (Galfalvy et al., 2003; Harr and Schlötterer, 2006; Jiang et al., 2008; Ploner et al., 2005). Note that these other strategies are certainly not perfect either. Like the benchmark datasets, they should be considered with caution. For example Ploner et al. (2005) proposed an interesting criteria based on the overall correlation of random sets of genes and argued that on average the correlation should be 0. However, this criterion cannot be used alone because it is quite clear that to achieve this goal it is enough to normalize all intensities to 1 and thus one removes simultaneously all non-relevant and relevant effects. An alternative approach is the use of quantitative real-time PCR as a gold standard technique. Unfortunately, it can be used for only a few measurements. Overall, assessing the quality of a normalization procedure is a complex question and subject to many controversies.

4.2 Specificities of tumor DNA copy number profile normalization

For gene expression profiling normalization, it is usually assumed that the majority of genes are not differentially expressed and that the proportions of down-regulated and up-regulated genes are similar. This hypothesis is questionable, but normalization methods relying on this assumption were shown to be quite efficient (Do and Choi, 2006). However, for DNA copy number profiling of tumors, this is clearly not the case. Indeed, some tumor samples, especially TNBC, are genomically unstable and harbor many genomic rearrangements. For these tumors, there is no reason to think that the number of gains equals the number of losses. Moreover, it has been empirically shown that not taking into account DNA copy number alterations in CGH arrays of tumor samples causes problems for conventional normalization methods (Staaf et al., 2007). More specifically, it leads to over-fitting and a decreased signal to noise ratio. We confirmed this result for Affymetrix SNP array 50K and 250K (data not shown).

Another specificity of tumor DNA copy number profile normalization is the possibility to assess (without knowing the true DNA copy number) the signal to noise ratio of a given normalization procedure (Neuvial et al., 2006). The idea is that, after using a given normalization procedure, it is possible to identify gained, lost and normal regions of the genome. This is the "calling" step. It is then possible to compute:

- the "signal" as the difference between the mean gain intensity and the mean normal intensity;
- the "noise" as the residual error of the signal.

When comparing two different normalization methods, it is important to compute their "signal" and "noise" with the same definition of gained and normal regions. Indeed, a method detecting more gained regions would not be favored. This is because some of these extra gained regions would necessarily correspond to small differences between normality and gain, resulting in a smaller signal to noise ratio. Therefore, only consensus gained and normal regions should be used (see Figure 4.1). Overall this is certainly not an unbiased estimation of the signal to noise ratio as it heavily relies on the calling step.

However, for a given calling procedure, it seems a good way to assess the relative advantages of various normalization procedures.

In conclusion, when normalizing tumor DNA copy number profiles, it is important to take into account both the non-relevant factors and the DNA copy number alterations. Moreover, without knowing the true signal it is possible to evaluate and compare the signal to noise ratio of two different normalization methods. Keeping all this in mind, we worked on the normalization of Affymetrix Genechip 50K and 250K SNP arrays.

4.3 Normalization of Affymetrix Genechip 50K and 250K SNP arrays

In this section, I will give an overview of the ITALICS normalization method that I proposed to normalize Affymetrix Genechip 50K and 250K SNP arrays (Rigaill et al. (2008), the article is provided in the following section: 4.4). Besides normalization, ITALICS performs the analysis of the DNA copy number profiles using the GLAD methodology (Hupé et al. (2004)). GLAD performs both the segmentation and calling step. The ITALICS method is available as an R package in Bioconductor.

As in any microarray, Affymetrix Genechip 50K and 250K SNP arrays are influenced by non-relevant factors such as the probe GC content, spatial artifacts and others (see Figure 4.2 for an overview of the experimental protocol). To take into account both the non-relevant factors and the DNA copy number as suggested by Staaf et al. (2007), ITALICS iteratively and alternatively segments the DNA copy number profile and estimates the influence of the non-relevant factors (see Figure 4.3). Having a rough first estimation of the DNA copy number profile, it is possible to correct for non-relevant factors using a multiple linear regression. Continuing from this corrected profile, we then re-iterate the segmentation and correction steps to improve their qualities (see subsection 2.2 and Table 1 on page 2-3 of the paper for a more detailed description). We have empirically shown that two iterations are required to achieve a good signal to noise ratio (see subsection 3.1 and Figure 2 on page 4 of the ITALICS paper).

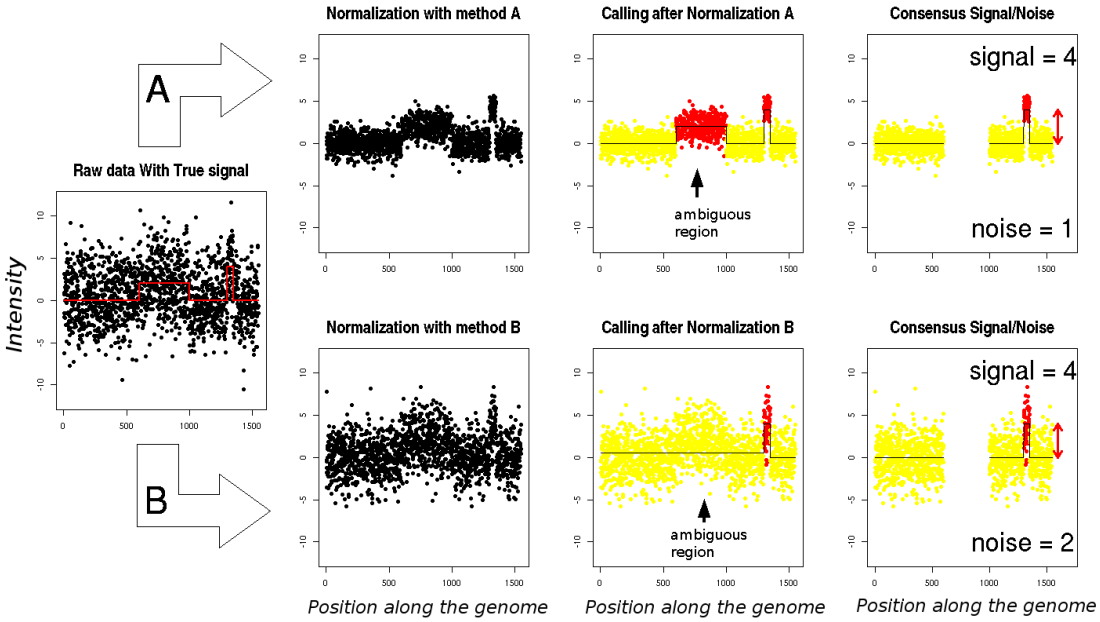


Figure 4.1: Comparing the signal to noise ratio of two normalization methods. For all graphs, the x-axis is the position along the genome and the y-axis is the normalized intensity except for the leftmost graph where it is the raw intensity. The raw data (leftmost graph) is normalized using either the A method (top) or B method (bottom). Then the calling step of normal and gained regions is done on the two normalized profiles with the same method. Gained regions are in red and normal regions are in yellow. Ambiguous regions (gained in A and normal in B or normal in A and gained in B) are then discarded to obtain a consensus calling and compute the signal to noise ratio. In this example, method A (signal to noise ratio of $4/1 = 4$) is found to be better than B (signal to noise ratio of $4/2 = 2$).

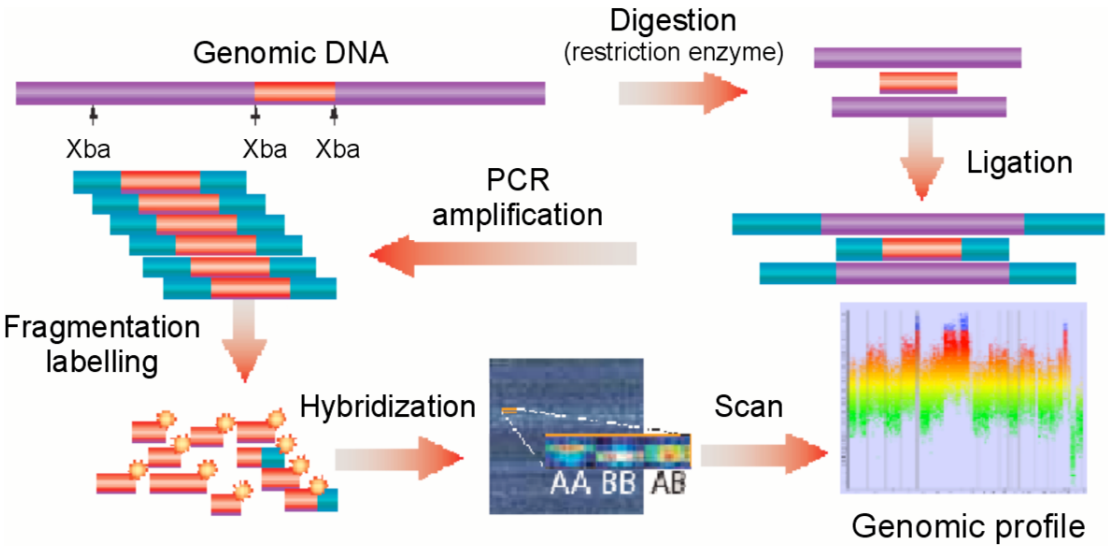


Figure 4.2: Affymetrix SNP 50K and 250K experimental protocol

We assessed ITALICS performances by various means. First, we showed that ITALICS outperforms other available methods in terms of signal to noise ratio. The ratio was measured using the *dyn* criteria proposed by Neuvial et al. (2006) (see subsection 3.2 and Figure 3 on page 4-5 of the ITALICS paper). Second, in collaboration with Anna Almeida (Ph.D.), we used quantitative real-time PCR to assess the performance of ITALICS compared to other normalization methods. In particular we showed for a few examples that ITALICS was able to lead to a better assessment of breakpoint positions compared (see paragraph "Quantative PCR validation" and Figure 4 in subsection 3.5 on page 5-6 of the paper). Third, in collaboration with Marc Bollet (MD/Ph.D.) and Nicolas Servant (M.Sc.) we showed that using ITALICS allows to identify breakpoints at the exact same position between primary and true recurrence (see paragraph "Patients with breast cancer relapses" and Figure 5 in subsection 3.5 on page 5-6 of the paper). It was not the case with other available methods. With ITALICS the breakpoint positions can be used to classify new tumors between new primary and true recurrence (Bollet et al. (2008), the article is provided in Appendix). We also applied the ITALICS methodology to a pilot

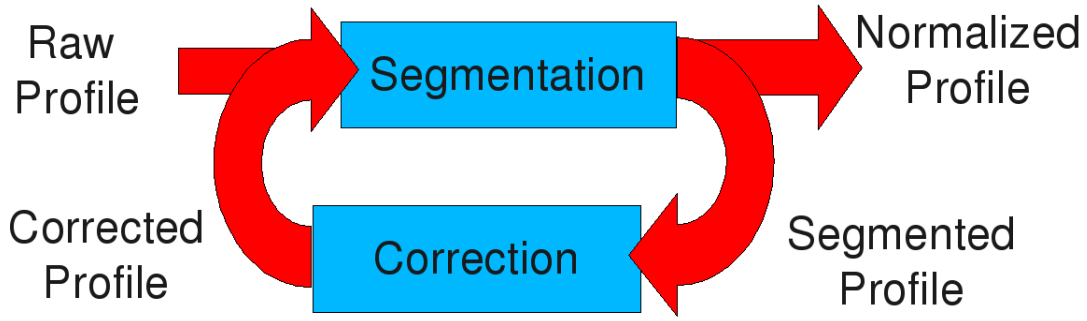


Figure 4.3: Outline of the ITALICS normalization method. The raw profile is iteratively segmented and then corrected based on this segmentation. The signal to noise ratio is improved with every iteration of these two steps. After two or three iterations, a normalized profile is obtained.

study of the Curie-Servier project that included 14 TNBC and 11 ER- / HER2+ tumors. We showed that the *PTEN* gene was lost in more than 50% of all TNBC samples (Marty et al. (2008), the article is provided in subsection 11.1.3). Interestingly, it was not possible to recover such a high percentage with other available normalization procedures.

At the time of the study, we used the GLAD methodology (Hupé et al., 2004) to perform the segmentation of DNA copy number profiles. However, in principle any efficient segmentation procedure should work. To conclude, ITALICS is hopefully a sound methodology, it has given good empirical results and we have shown that ITALICS outperformed other available methods.

4.4 Paper: ITALICS

In this section is the Bioinformatics paper describing ITALICS.

BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 6 2008, pages 768–774
doi:10.1093/bioinformatics/btn048

Genome analysis

ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays

Guillem Rigaill[1,2,5,†], Philippe Hupé[1,2,3,5,*,†], Anna Almeida[4], Philippe La Rosa[1,2,5], Jean-Philippe Meyniel[4], Charles Decraene[3,4] and Emmanuel Barillot[1,2,5]

[1]Institut Curie, Service de Bioinformatique, [2]INSERM, U900, [3]CNRS UMR144, [4]Institut Curie, Translational Research Department, 26 rue d'Ulm, Paris F-75248 and [5]Ecole des Mines de Paris, ParisTech, Fontainebleau, F-77300 France

Received on August 21, 2007; revised and accepted on January 29, 2008

Advance Access publication February 5, 2008

Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Affymetrix SNP arrays can be used to determine the DNA copy number measurement of 11 000–500 000 SNPs along the genome. Their high density facilitates the precise localization of genomic alterations and makes them a powerful tool for studies of cancers and copy number polymorphism. Like other microarray technologies it is influenced by non-relevant sources of variation, requiring correction. Moreover, the amplitude of variation induced by non-relevant effects is similar or greater than the biologically relevant effect (i.e. true copy number), making it difficult to estimate non-relevant effects accurately without including the biologically relevant effect.

Results: We addressed this problem by developing ITALICS, a normalization method that estimates both biological and non-relevant effects in an alternate, iterative manner, accurately eliminating irrelevant effects. We compared our normalization method with other existing and available methods, and found that ITALICS outperformed these methods for several in-house datasets and one public dataset. These results were validated biologically by quantitative PCR.

Availability: The R package ITALICS (ITerative and Alternative normaLIzation and Copy number calling for affymetrix Snp arrays) has been submitted to Bioconductor.

Contact: italics@curie.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The development of high-throughput technologies, and of microarrays in particular, has made it possible to analyze DNA copy number throughout the entire genome, with ever-increasing resolution. Various techniques for detecting DNA copy number alterations are available (for a review, see Ylstra *et al.*, 2006). Affymetrix SNP arrays, such as the Affymetrix GeneChip Human Mapping 100K Set (Kennedy *et al.*, 2003), seem to be one of the most widely used tools. These chips can be used for simultaneous genotyping and copy number determination for single nucleotide polymorphism (SNP), at high resolution. This technology has various uses, including studies of copy number variations in populations and the identification of genomic alterations in developmental genetics or cancer (for a review, see Pinkel and Albertson, 2005). In cancer studies, Affymetrix SNP arrays provide new insight into the mechanisms of tumor progression; they can be used to pinpoint new candidate genes for tumor-suppressor genes (Liu *et al.*, 2007) and oncogenes (thought to be present in loss and gain regions, respectively), and to classify tumors, improving diagnosis for new patients and the evaluation of prognosis.

Like all microarrays, Affymetrix SNP arrays are affected by systematic non-relevant sources of experimental variation. For accurate extraction of the biologically relevant effect (i.e. the true DNA copy number of each SNP in the genome, corresponding to the biological signal), the raw data must be corrected, taking these different effects into account. We present here a normalization algorithm for this purpose, which can be used for the simultaneous correction of different sources of experimental variation and biological signal estimation when trying to infer DNA copy number.

Several methods have already been developed for correcting non-relevant sources of variation. These methods include CNAG (Nannya *et al.*, 2005), GIM (Komura *et al.*, 2006) and CARAT (Huang *et al.*, 2006). However, none of these methods take into account that the range of variation due to the non-relevant effects is similar or higher than the biologically relevant effect. Therefore, the impacts of the biologically relevant effect and non-relevant effects may easily be confused. Correct estimation of the non-relevant effects also depends on the correct estimation of copy number. We therefore propose an alternative, iterative method for estimating the biologically relevant effect and non-relevant effects, to improve biological signal estimation. We will begin by briefly presenting Affymetrix SNP arrays. We will then describe our algorithm (*ITerative and Alternative normaLIzation and Copy number calling for affymetrix Snp arrays*: ITALICS) for data normalization in detail. We then discuss the results obtained with this algorithm, comparing them with those obtained with other algorithms. Finally, we discuss the advantages of ITALICS and possible improvements to this method.

*To whom correspondence should be addressed.

†The authors wish to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

2 MATERIALS AND METHODS

2.1 Affymetrix SNP arrays

Technology: Affymetrix SNP arrays can be used to detect DNA copy number alterations at a resolution of 6–210 kb, using around 11 000–500 000 human SNPs. The Affymetrix GeneChip Human Mapping 100K and 500K Sets are comprised of two arrays. Each array is based on specific restriction enzymes: *Xba*I and *Hind*III for the 100K set and *Sty*I and *Nsp*I for the 500K set. The Affymetrix 50K *Xba*I and *Hind*III arrays contain no common SNPs and their combination provides the DNA copy numbers of more than 115 000 SNPs.

Each allele of each SNP is represented by n_i perfect match (PM) probes and n_i mismatch (MM) probes. Reverse or forward probes may be used and these probes may be centered on the SNP position or offset by −4 to +4 base pairs. Thus, all the PM probes of an SNP allele have different DNA sequences. Probes are grouped into probe quartets of four probes: one PM and one MM probe for each of alleles A and B. All four probes have the same orientation and offset.

The Affymetrix SNP arrays assay is carried out as follows. Genomic DNA is digested with a restriction endonuclease. Adaptors are ligated to all fragments. These fragments are amplified by PCR and then fragmented, labeled with biotin and hybridized with the chip. The chip is then washed and scanned to generate the cell intensity file (.CEL) which is used as input to the proposed algorithm.

Hereafter, the raw signal $Y_{i.}$ of a given SNP i is given by:

$$Y_{i.} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \text{ with } Y_{ij} = Y_{ij}^A + Y_{ij}^B$$

where Y_{ij}^A and Y_{ij}^B are the log-intensity of the PM probe A and B of the j-th probe quartet for the SNP i, and Y_{ij} is the sum of PM log-intensities for the j-th quartet. $Y_{i.}$ is the mean PM log-intensity of the n_i quartets for the SNP i. MM probes are not taken into account in our algorithm. The two PM probes defining the entity Y_{ij} are referred subsequently as $Quartet_{PM}$, the subscript i is referred to as SNP i, and the subscript j as one of the n_i quartets.

Non-relevant sources of variation: ITALICS deals with known systematic sources of variation, such as the GC-content of the $Quartets_{PM}$ (QGC_{ij}), the length of the PCR-amplified fragment (FL_i) and the GC-content of the fragment amplified by PCR (FGC_i) (Nannya *et al.*, 2005; Komura *et al.*, 2006). It also takes into account the $Quartet_{PM}$ effect (Q_{ij}), resulting from the systematically low intensity of some $Quartets_{PM}$ and the systematically high intensity of others.

We also found that some Affymetrix SNP arrays suffer from spatial artifacts, as reported by Neuvial *et al.* (2006) for CGH array data. A spatial artifact is illustrated in Figure 1A: neighboring $Quartets_{PM}$ on the chip present abnormal intensities. The corresponding SNPs which appear as outliers in the genomic profile, as shown in Figure 1C, D and E, and should be removed. We have addressed this issue using a filtering criterion, making it possible to discard bad probes, as described subsequently.

2.2 The ITALICS algorithm

Overview: In Affymetrix SNP arrays, non-relevant sources of variation ($NonRel_{ij}$) have comparable or greater influence on the raw signal variability than the biological signal ($CopyNb_i$) (see Section 3.2 to compare the type III sum of squares of the different effects in a multiple linear model). We therefore propose an iterative, alternative normalization method, making it possible to estimate the biological signal and non-relevant effects and, therefore, to eliminate most of the non-relevant effects while preserving most of the biological information. During each iteration, ITALICS:

(1) Estimates the biological signal $CopyNb_i$ using the GLAD algorithm (Hupé *et al.*, 2004),

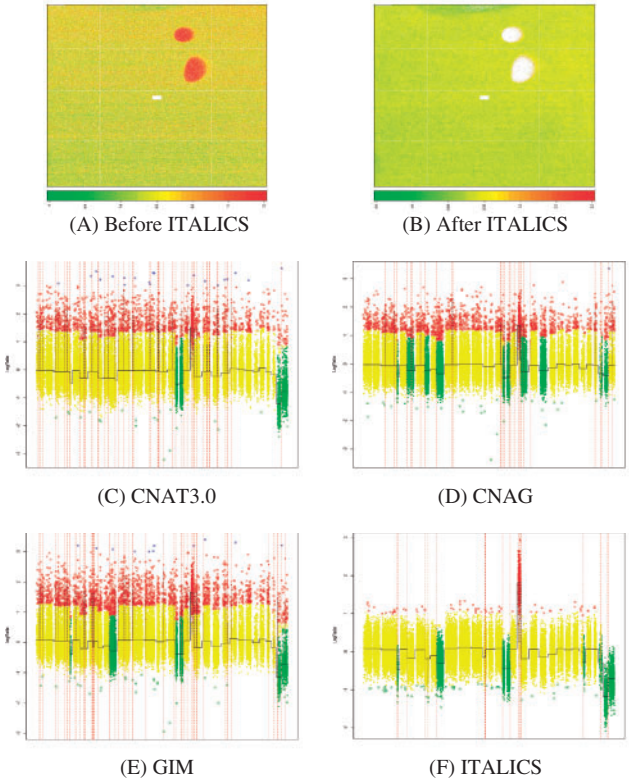


Fig. 1. Impact of spatial artifacts on genomic profiles. Image of an *Xba*I 100K Set chip (HF0844_Xba, Kotliarov *et al.* (2006)) before (**A**) and after normalization with ITALICS (**B**) (flagged $Quartets_{PM}$ in white). The Y_{ij} value of each $Quartet_{PM}$ is represented, using a gradient from green to red. (**C**), (**D**), (**E**) and (**F**) are the genomic profiles normalized with CNAT 3.0, CNAG, GIM and ITALICS. Vertical dashed red lines represent the breakpoints detected with GLAD and the assigned statuses are indicated by a color code: green for loss, yellow for normal and red for gain. Two stains of abnormally high $Quartets_{PM}$ values (in red) are visible in (A) and their corresponding SNP values correspond to outliers (colored in red) in the genomic profiles (C), (D) and (E), for which 1661, 1818 and 2331 outliers respectively, were detected. ITALICS flagged most of these $Quartets_{PM}$ (B) but evaluated the signals for their SNPs using the $Quartets_{PM}$ from the rest of the chip, resulting in the removal of only 13 of the 57 500 SNPs. ITALICS eventually identified only 88 outliers (F).

(2) Assuming the biological signal to be known, it estimates the non-relevant effects $NonRel_{ij}$ on raw data, by multiple linear regression.

After the last iteration, the $Quartets_{PM}$ for which multiple linear regression predicts the signal poorly are flagged. They correspond to $Quartets_{PM}$ with abnormal values and are excluded from the final step, in which ITALICS uses GLAD to estimate the biological effect $CopyNb_i$ on the remaining normalized $Quartets_{PM}$. The algorithm is presented in more detail below.

Biological signal estimation (CopyNb_step): ITALICS applies the GLAD algorithm to $Y_{i.}$ values to estimate the biological signal. The GLAD algorithm segments the genomic profile, defining regions of homogeneous DNA copy number. For each of these regions, it provides a smoothing value and a status (gain, normal or loss). The smoothing

value is the median of the $Y_{i.}$ values within the region concerned, and corresponds to the inferred copy number $CopyNb_i$.

Non-relevant effect estimation (NonRel_step): After estimating the biological effect $CopyNb_i$, ITALICS infers the non-relevant effects by multiple linear regression. The model used is as follows:

$$Y_{ij} = \mu + \alpha CopyNb_i + f(NonRel_{ij}) + \varepsilon_{ij}$$
$$f(NonRel_{ij}) = P_1(FL_i) + P_2(FGC_i) + P_3(QGC_{ij}) + \beta Q_{ij}$$

with:

$$i = 1, \ldots, N \text{ (the number of SNPs)}$$
$$j = 1, \ldots, n_i \text{ (the number of } Quartets_{PM} \text{ per SNP)}$$
$$P_k(x) = \sum_{l=1}^{l=3} \gamma_{kl} x^l, k = 1, \ldots, 3$$
$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

The multiple linear regression can also be expressed in classical matrix notation:

$$Y = X\theta + \varepsilon$$

with:

$$\theta = (\mu, \alpha, \gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{31}, \gamma_{32}, \gamma_{33}, \beta)$$

The parameter θ is estimated using the ordinary least-squares method. The degrees of the polynomial functions P_k were chosen using the BIC criterion (Schwarz, 1978) on a training data set of 128 reference diploid chips (Matsuzaki *et al.*, 2004).

The $Quartet_{PM}$ effect is dealt with by calculating Q_{ij} as the mean of each $Quartet_{PM}$ on the 64 female chips of the same Affymetrix reference data set (Matsuzaki *et al.*, 2004).

Once the non-relevant effects have been estimated, the Y_{ij} values are corrected as follows:

$$Y_{ij}^{\text{cor}} = Y_{ij} - \hat{f}(NonRel_{ij}),$$

where $\hat{f}(NonRel_{ij})$ corresponds to the estimate of non-relevant effects based on multiple linear regression. The corrected $Y_{i.}^{\text{cor}} = (\sum_{j=1}^{j=n_i} Y_{ij}^{\text{cor}}/n_i)$ is used in the next step of the GLAD procedure, to re-estimate the biological effect. This algorithm is repeated until the number of iterations reaches the predetermined fixed number of iterations *itermax*.

ITALICS uses GLAD and therefore we investigate if the normalization was influenced by the choice of GLAD parameters. In Supplementary information, we give guidelines for choosing parameters and expose the result of sensitivity analysis that shows a large robustness of ITALICS to parameter settings.

Elimination of poorly predicted $Quartets_{PM}$: After the last iteration, $Quartets_{PM}$ Y_{ij} poorly predicted by multiple linear regression are flagged out. This is achieved by calculating the 95% prediction interval. All Y_{ij} outside this interval are flagged. SNPs with less than three non-flagged $Quartets_{PM}$ in a total of n_i are then discarded. If more than three Y_{ij} are not flagged, $Y_{i.}^{\text{cor}}$ is recalculated as:

$$Y_{i.}^{\text{cor}} = \frac{\sum_{j \notin F_i} Y_{ij}^{\text{cor}}}{n_i - NbF_i},$$

with F_i the set of flagged $Quartets_{PM}$ for the SNP_i and NbF_i the number of flagged $Quartets_{PM}$ for the SNP_i.

Data scaling: The data are scaled to allow between-chip comparison. After the first GLAD step, the biological signal is subtracted and the standard deviation s of $(Y_{i.} - CopyNb_i)$ is calculated for each chip using all SNPs i of the chip. The data are then scaled as follows:

$$Y_{ij}^{\text{scaled}} = \frac{Y_{ij}}{s}$$

Table 1. ITALICS algorithm overview

```
iter: = 0
while iter < itermax do
  CopyNb_step()
  if iter = 0 then
    Data_Scaling()
  end if
  NonRel_step()
  iter: = iter + 1
end while
elimination_of_poorly_predicted_quartet_PM( )
CopyNb_step()
```

The ITALICS procedure is summarized in Table 1.

2.3 Comparison with other methods

Other methods: Several other methods have already been developed. Most use linear regression to estimate and correct for non-relevant effects. They differ in the effects taken into account and in their pre- and post-processing steps.

CNAG: Copy Number Analysis for GeneChip (Nannya *et al.*, 2005). CNAG corrects the raw signal intensity of a sample, by introducing the notion of *averaged best fit*, corresponding to a pseudochip constructed from the five samples most similar to the reference samples. CNAG subtracts this averaged best fit from the raw signal and then corrects for the length of the PCR-amplified fragment and GC-content effects by linear regression. This method is available within CNAG 2.0 and is also used in CNAT 4.0 (Copy Number Analysis Tool, see below).

CNAT 3.0: Chromosome Copy Number Analysis Tool 3.0. Affymetrix developed this method for the extraction of DNA copy number. No specific step for the correction of non-relevant effects is included. This method uses samples with varying chromosome X copy number for intensity calibration and transforms SNP intensity into copy number values.

CNAT 4.0: Chromosome Copy Number Analysis Tool 4.0. This tool uses CNAG to normalize the data and then smoothes the data with a user-defined window. This step artificially reduces the variance of the data and visibly improves the quality of the profile.

CARAT: Copy Number Analysis with Regression And Tree (Huang *et al.*, 2006). CARAT uses a reference data set to select probes showing a high-allelic response and to remove those with no such response. For each new sample, it first standardizes the probe signal, based on mismatch probe information. It then corrects for probe GC-content and PCR fragment length effects, by linear regression. Finally, each SNP intensity is regressed against the average intensity of the reference samples with the same genotype.

GIM: Genomic Imbalance Map (Komura *et al.*, 2006). GIM roughly estimates the biological effect and subtracts it from the raw signal, using a simpler version of ChARM (Myers *et al.*, 2004). It removes defective probes with a high local GC-content and then re-estimates the biological effect without using the defective probes and subtracts this effect from the raw signal. It takes into account probe GC-content, the length of the PCR-amplified fragment and its GC-content, and mean SNP intensity for the reference dataset, by linear regression. GIM is implemented in Matlab and is freely available.

We compared ITALICS with CNAG, CNAT 3.0 and GIM. We did not compare ITALICS with CARAT, because no software was

available for CARAT at the time of the study, or with CNAT 4.0, which presents no improvement over CNAG. For the CNAG, CNAT 3.0 and GIM genomic profiles, copy number and the status of the genomic regions were inferred with the GLAD algorithm, using the same parameters as for the ITALICS algorithm.

Quality criteria: As described by Neuvial *et al*. (2006), we used several quality criteria to compare the various normalization algorithms.

As defined by Neuvial *et al*. (2006), the *dyn* criterion estimates the dynamics of the DNA copy number signal. Its value is:

$$dyn(a) = \frac{(median(Y_{i.}^{\mathrm{cor},a})_{i\in G} - median(Y_{i.}^{\mathrm{cor},a})_{i\in N})}{smt}$$

with G and N the regions considered to correspond to *Gain* and *Normal* and $Y_{i.}^{\mathrm{cor},a}$ the corrected signal of SNP i using the normalization method a. $smt = median(|Y_{i.}^{\mathrm{cor},a} - Y_{(i-1).}^{\mathrm{cor},a}|)$ for ordered $Y_{i.}^{\mathrm{cor},a}$ throughout the genome. *smt* quantifies the smoothness of the signal over the genome, and *dyn* assesses the dynamics of the signal, as defined by the signal-to-noise ratio (SNR). If no gain region have been identified, the *dyn* criteria is computed over loss regions. A high *dyn* should be obtained with good normalization methods.

The criterion *out* is the number of outliers detected by GLAD. GLAD defines regions of homogeneous DNA copy number and outliers are SNPs with values different from those of other SNPs in the same region. These abnormal values may be accounted for by point mutations in the genome. However, a large number of such changes is unlikely, so the total number of outliers should be relatively low and the *out* parameter close to zero.

The criterion *flag* is the number of flagged SNPs. We introduced this criterion for the comparison of methods that remove SNPs, such as GIM and ITALICS. These methods may artificially improve the quality of the signal (as measured by *dyn* and *out*), by removing SNPs with abnormal behavior. The number of flagged SNPs should, therefore, not be too high. When faced with a choice between two methods with equal SNR, the method with the lowest *flag* should be preferred.

Comparison of two normalization methods: These three criteria can be used to determine which of the two normalization methods gives the best results for a given array. In this pairwise comparison context, *dyn* must be calculated with the same definition of gain, normal and loss regions for both normalized arrays. We therefore define consensus gain, normal and loss regions associated with an array processed with two different normalization methods, as the intersection of the two corresponding gain, normal and loss regions obtained with the two different normalization methods [see also Neuvial *et al*. (2006) for details].

For the comparison of two different methods, a and b, in terms of a certain criterion, we calculate relative performances as follows:

$$RP^{dyn}(a,b) = (dyn(a) - dyn(b))/dyn(a)$$
$$RP^{out}(a,b) = -(out(a) - out(b))/out(a)$$
$$RP^{flag}(a,b) = -(flag(a) - flag(b))/flag(a)$$

RP measures the percentage improvement observed with method a, with respect to method b. The minus signs for the *out* and *flag* criteria ensure that a positive $RP^{cri}(a,b)$ always means that method a is better than method b for criterion *cri*.

2.4 Datasets

We carried out our study on two public datasets: a dataset for 128 reference diploid chips (Matsuzaki *et al*., 2004) and a glioma dataset corresponding to 356 chips (Kotliarov *et al*., 2006). We also used datasets produced by the Affymetrix platform of the Institut Curie obtained with 22 uveal melanoma samples, 40 ovarian cancer samples and 26 breast cancer samples.

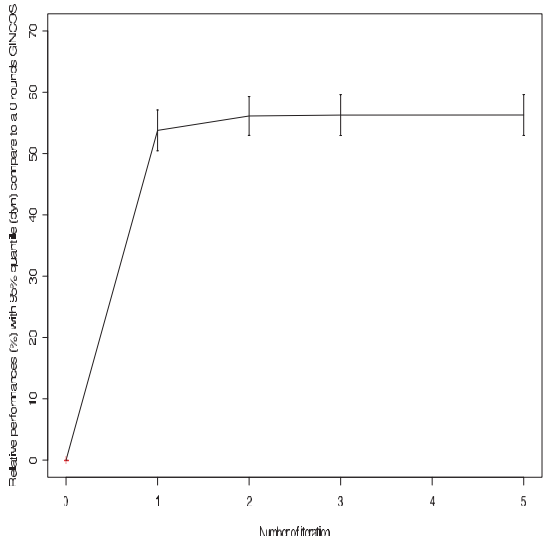


Fig. 2. Improvement in SNR with the number of ITALICS iterations The improvement in SNR obtained with each iteration was assessed by calculating the percentage improvement RP^{dyn} for 1, 2, 3, and 5 iterations with respect to no iterations. The results are summarized in this graph, showing RP^{dyn} as a function of the number of iterations. The SNR improved with the first two iterations, with no major improvement observed for subsequent iterations.

3 RESULTS

3.1 Choosing the number of iterations

We assessed the extent to which each iteration within the ITALICS algorithm improved the SNR, by calculating the *dyn* criteria for different values of *itermax* (0, 1, 2, 3 and 5) for each chip of the 356-glioma chips dataset. The percentage improvement RP^{dyn} for different values of *itermax* (1, 2, 3 and 5) with respect to no iteration was then calculated (Fig. 2). One iteration gave 53.8% improvement, two gave 56.1% improvement and three and five gave 56.3% improvement. As the third and subsequent iterations gave only a very slight improvement, we set *itermax* to two in the ITALICS algorithm.

3.2 Importance of each effect on the signal

For each chip of the glioma dataset, we calculated the type III sum of squares for each effect in our multiple linear regression model. A low type III sum of squares indicates that the difference between the full model and the model excluding the studied effect is very small. The $Quartets_{PM}$ effect gave the highest type III sum of squares, with a mean of 550×10^3 versus 10.4×10^3, 16×10^3 and 14×10^3 for $Quartets_{PM}$ GC-content, fragment length and fragment GC-content. The biological effect was the second most important effect, with a mean of 24×10^3.

3.3 ITALICS outperformed the other methods

We calculated *dyn* and *out* with ITALICS, GIM, CNAT 3.0 and CNAG, using three different cancer datasets: two in-house datasets corresponding to 22 choroidal melanoma chips and

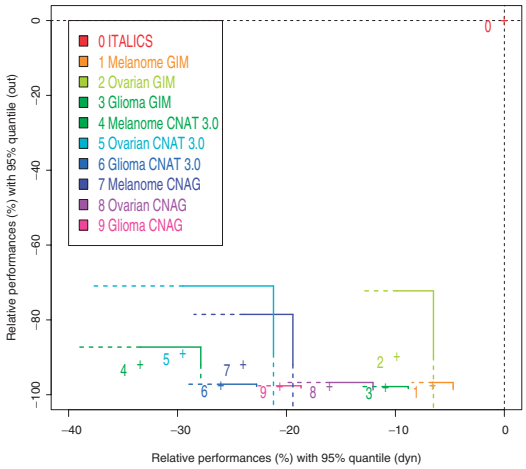


Fig. 3. Comparison of ITALICS with other normalization methods. We compared ITALICS with CNAT 3.0, CNAG and GIM for two quality criteria—*dyn* and *out*—using three different cancer datasets: two in-house data sets corresponding to 22 choroidal melanoma chips and 40 ovarian cancer chips and one public dataset corresponding to 356 glioma chips (Kotliarov *et al.*, 2006). Each color corresponds to the comparison of ITALICS with a different method or data set. ITALICS is taken as the reference [red point 0 at (0, 0)]. For each method, the cross indicates the mean relative performance on the data set concerned, for the *dyn* and *out* criteria, and the lines give the corresponding 95% quantile for relative performance. ITALICS significantly outperforms all methods for both quality criteria, *dyn* and *out*.

40 ovarian cancer chips and one public data set of 356 glioma chips. All methods were used with their default parameters.

We calculated the percentage improvement (*RP*) for CNAT 3.0, CNAG and GIM, in terms of *dyn* and *out*, with respect to ITALICS (Fig. 3). For the three competitors $RP^{cri}(competitor,ITALICS)$ is calculated and we performed *t*-tests to assess the significance of the improvement. We found that ITALICS outperformed CNAT 3.0, CNAG and GIM, in terms of *dyn* and *out*, with *t*-test *P*-values below 10^{-5} for all three data sets. For GIM, RP^{dyn} ranged from −10.9% to −6.5%, for CNAG, it ranged from −23.9% to −16.0% and for CNAT 3.0 it ranged from −33.4% to −26.0%. RP^{out} ranged from −98.1% to −89.0% for all three methods. Chip data normalized with ITALICS therefore had a significantly better SNR than those normalized with CNAT, CNAG and GIM, with fewer outliers.

Both ITALICS and GIM flag certain SNPs for elimination. The improvement in SNR obtained with these methods may therefore be partially due to the mechanical effect of this removal. We compared the number of SNPs flagged between GIM and ITALICS and found that ITALICS flagged significantly fewer SNPs than GIM, with a mean of 300 SNPs per chip for ITALICS versus 3000 for GIM. The $RP^{flag}(GIM,ITALICS)$ is −90%.

3.4 Spatial artifact correction

Some Affymetrix SNP arrays suffer from spatial artifacts. The step flagging poorly predicted $Quartets_{PM}$ removes most $Quartets_{PM}$ with abnormal intensity detected by visual inspection, as shown in Figures 1A and B. To our knowledge, ITALICS is the only method capable of doing this. Moreover, the removal of these abnormal $Quartets_{PM}$ increases the quality of the signal, by removing many outliers from the genomic profile: 1661, 1818 and 2331 outliers were detected for CNAT 3.0, CNAG and GIM (Figure 1C, D and E). With ITALICS, there were only 88 outliers (Figure 1F), but only 13 of the 56 000 SNPs were removed because they had less than three non-flagged $Quartets_{PM}$.

3.5 Biological validation

Quantitative PCR validation: We used QPCR (see Supplementary Material for more detail) to validate our method with a different technology. As a test case, we used a set of paired breast cancer samples (primary tumor and relapse, Bollet *et al.* 2008) and tried to identify a breakpoint in chromosome 20. We compared the results obtained with QPCR with those obtained with ITALICS, CNAG, GIM and CNAT, for the *Xba*I and *Hind*III arrays. We also carried out QPCR on two breast cancer tumors, each with a normal chromosome 20 (white and striped bars in Fig. 4) to assess noise for QPCR and to validate the significance of copy number change. As shown in Figure 4, ITALICS was more accurate than CNAG, GIM and CNAT 3.0 for comparisons of copy numbers, based on the estimates obtained with PCR. ITALICS, CNAG, GIM and CNAT 3.0 detected changes in copy number in this region of chromosome 20. However, ITALICS breakpoints were closer to QPCR breakpoints than CNAT breakpoints (see Fig. 4A, C and D) and CNAG and GIM breakpoints (see Figure 4A). In Figure 4A, QPCR and ITALICS breakpoints are found at identical positions (between P14 and P15). In Figure 4C and D, CNAG, GIM and ITALICS detect a copy number change between P12 and P13, close to that detected by QPCR between P14 and P15, whereas CNAT detects this breakpoint further away, between P06 and P07 in Figure 4C and between P08 and P09 in Figure 4D. In Figure 4B, QPCR, CNAT, GIM, CNAG and ITALICS found the same breakpoint.

Patients with breast cancer relapses: The problem tackled was determining whether the second cancer was a true recurrence of the first cancer or a new primary tumor, based on the two Affymetrix SNP array profiles (Bollet *et al.*, 2008). We tried to identify common breakpoints between the cancer chips for the two tumors. The breakpoints detected with CNAT 3.0 or ITALICS normalization are represented in Figure 5A and B for chromosome 6 and 9, respectively, for one patient. GIM and CNAG results are similar to ITALICS for chromosome 6 and similar to CNAT for chromosome 9 (data not shown). ITALICS identified breakpoints at identical locations for both cancers and this is true for the two chromosomes presented in Figure 5A and B. It is important to notice that this was not possible with CNAT 3.0, CNAG and GIM. The precise match between the breakpoints mapped in the two cancers with ITALICS suggests that the second cancer is a true recurrence, whereas the opposite conclusion would have been drawn with CNAT 3.0. As CNAG and GIM detect less precise matches, they lead to the same conclusion as ITALICS, but the evidences for this conclusion are weaker. Expert assessment based on clinical data also indicated that this was a true recurrence, and

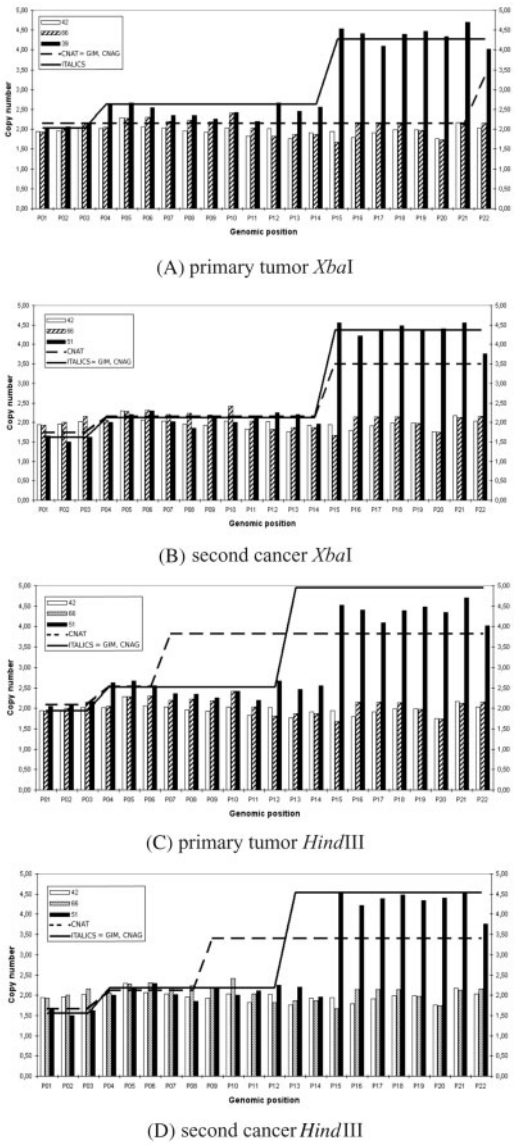


Fig. 4. Affymetrix SNP arrays and QPCR DNA copy number profiles for a patient with breast cancer relapse. CNAT 3.0 (dashed line) and ITALICS (solid line) DNA copy number determination along chromosome 20, from position 17453432 (P01) to position 49386812 (P22), for the primary tumor (**A**, **C**) and the relapse (**B**, **D**) using the *Hind*III (C, D) and *Xba*I (A, B) Affymetrix SNP arrays. CNAG and GIM results are identical to CNAT for (A) and identical to ITALICS for (B, C and D). We performed QPCR on two breast cancer tumors with a normal chromosome 20, to estimate the noise associated with QPCR and to validate the significance of copy number change. The bar charts generated show the QPCR estimation of DNA copy number in two breast cancer tissues with a normal chromosome 20 (white and striped bars, A, B, C and D), the primary breast tumor (black bars, A and C) and the corresponding relapse (black bars, B and D). In (A), both ITALICS and QPCR detect a copy number change between P14 and P15, whereas GIM, CNAG and CNAT detects a change between P21 and P22. In (C) and (D), ITALICS detects a copy number change between P12 and P13, close to that detected by QPCR between P14 and P15, whereas CNAT detects a breakpoint further away, between P06 and P07 in (**C**), and between P08 and P09 in (D). In (B) QPCR, CNAT and ITALICS found the same breakpoints.

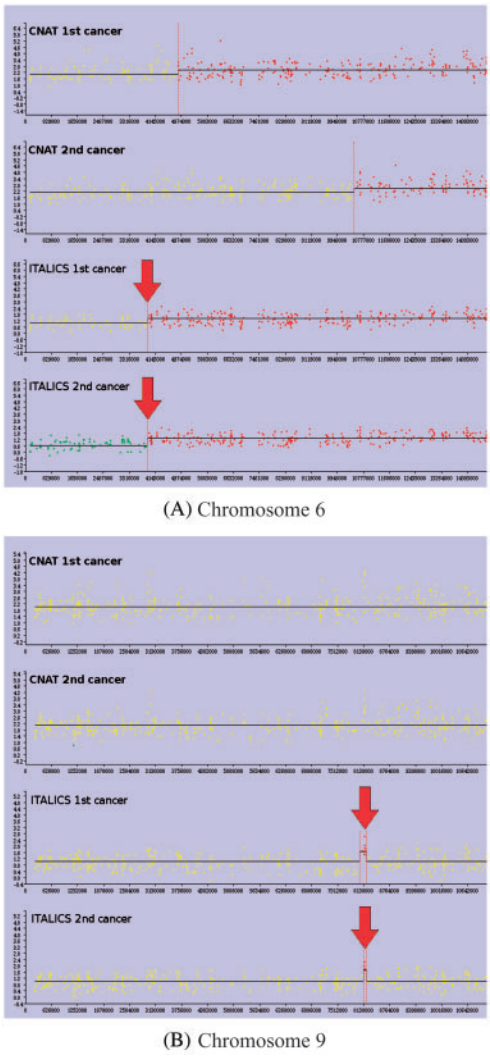


Fig. 5. Detection of breakpoints common to first and second cancers, using ITALICS. We present part of the chromosome 6 (**A**) and 9 (**B**) profiles obtained with VAMP (La Rosa *et al.*, 2006) for a patient with two breast tumors. For both (**A**) and (**B**), the first two profiles are CNAT 3.0 profiles of the first and second cancers and the last two profiles are ITALICS profiles of the first and second cancers. GIM and CNAG results are similar to ITALICS for chromosome 6 and similar to CNAT for chromosome 9 (data not shown). CNAT 3.0 identified no breakpoints (red dashed lines) common to the two cancers, whereas ITALICS did (red arrows), strongly suggesting that the second cancer was a true recurrence. Moreover, the results obtained with ITALICS are supported by an expert classification based on clinical data.

was therefore consistent with the results obtained with ITALICS. Similar conclusions were drawn for the rest of the data set (13 first and second cancer pairs). Thus, ITALICS improves the classification of true recurrences and new primary tumors.

4 DISCUSSION AND PERSPECTIVES

We present here a new method for normalizing Affymetrix SNP arrays: ITALICS. This method is highly efficient and

outperforms other normalization methods, such as CNAT 3.0, CNAG and GIM, in terms of SNR, giving a more accurate localization of breakpoints validated by QPCR. This improvement may be due to various features of the ITALICS algorithm. This algorithm estimates alternatively and iteratively both non-relevant and biologically relevant effects. The correct estimation of relevant effects depends on correct estimation of the biological signal and vice versa, as the relevant effects induce similar or higher ranges of variation than the biologically relevant effect. By estimating both the non-relevant and biologically relevant effects in an iterative manner, we avoid overestimation of the non-relevant effects and a loss of biological signal. The first estimation on raw data is necessarily rough, but improves the subsequent estimation of non-relevant effects. Each new estimation of the biological or non-relevant effects leads to a better estimation of the other effects. In practice we iterate our algorithm twice, as additional iterations were found to lead to no significant improvement in the SNR. This algorithm also includes a flagging step, making it possible to remove aberrant SNPs. Indeed, some PM intensity values are subject to spatial artifacts. The PM intensity of their $Quartets_{PM}$ is therefore abnormal, poorly predicted by the regression model and flagged. The discarding of poorly predicted $Quartets_{PM}$ does not necessarily lead to the discarding of the corresponding SNP, provided that enough $Quartets_{PM}$ remain elsewhere on the chip. As a result, very few SNPs are removed from the final genomic profile. This filtering step detects spatial artifacts only indirectly, but nevertheless gives good results in practice. Methods for the precise detection of spatial artifacts and the removal of all probes within spatial artifacts have already been developed (Neuvial *et al*., 2006). However, their direct application to SNP chips is impossible due to the very high density of these chips (more than 2 million probes per chip). Computing $Quartets_{PM}$ effect on an in-house reference dataset would certainly improve the quality of the normalization. Nevertheless, the $Quartets_{PM}$ effect is the most important effect and ignoring it would decrease the efficiency of the normalization.

We normalized *Xba*I and *Hind*III chips separately. The same major changes were detected with both chips. However, it is difficult to merge *Xba*I and *Hind*III data due to the difference in signal amplitude for consecutive alterations between the two chips. The merging of the *Xba*I and *Hind*III genomic profiles would result in a higher resolution profile, but also in a lower SNR. The ITALICS algorithm could be improved by taking into account the enzyme effect (*Xba*I and *Hind*III) to overcome this problem.

Technically, the ITALICS algorithm could be applied to higher density chips, such as the Affymetrix GeneChip Human Mapping 500K Set and even the Genome Wide SNP array 5.0 and 6.0, which do not have MM probes, as ITALICS is based solely on PM probes. Of course, we would have to check whether the non-relevant effects in our model are also observed with these higher density chips. We would also need to obtain a reference dataset for calculating the quartet effect.

5 CONCLUSION

We developed ITALICS, a new normalization algorithm for Affymetrix SNP arrays. This method was designed for the normalization and analysis of DNA copy number and significantly outperformed other methods, such as CNAT 3.0, CNAT 4.0, CNAG and GIM, in terms of SNR and can also be used to correct for experimental artifacts due to spatial effects. This method was validated by QPCR and accurately detected the breakpoints in genomic profiles. It could therefore be used to improve the characterization of samples in genomic studies.

ACKNOWLEDGEMENTS

This work was supported by the Institut Curie and the Centre National de la Recherche Scientifique. We thank Sophie Piperno-Neumann and Simon Saule, Jean-Paul Thiery and Marc Bollet, who were kind enough to provide us with access to their uveal melanoma, ovarian cancer and breast cancer datasets, respectively. We thank Marc Bollet, Nicolas Servant and Pierre Neuvial for fruitful discussions. We thank Audrey Rapinat and David Gentien for performing the Affymetrix Genechip experiments.

Conflict of Interest: none declared.

REFERENCES

Bollet,M. *et al*. (2008) High resolution mapping of breakpoints to define true recurrences among ipsilateral breast tumor recurrences. *J. Natl Cancer Inst*., **100**, 48–58.

Huang,J. *et al*. (2006) CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics*., **7**, 83.

Hupé,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Kennedy,G.C. *et al*. (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol*., **21**, 1233–1237.

Komura,D. *et al*. (2006) Noise reduction from genotyping microarrays using probe level information. *In Silico Biol*., **6**, 79–92.

Kotliarov,Y. *et al*. (2006) High resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. *Cancer Res*., **66**, 9428–9436.

La Rosa,P. *et al*. (2006) VAMP: visualization and analysis of array- CGH, transcriptome and other molecular profiles. *Bioinformatics*, **22**, 2066–2073.

Liu,W. *et al*. (2007) Deletion of a small consensus region at 6q15, including the MAP3K7 gene, is significantly associated with high-grade prostate cancers. *Clin. Cancer Res*., **13**, 5028–5033.

Matsuzaki,H. *et al*. (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*., **1**, 109–111.

Myers,C.L. *et al*. (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**, 3533–3543.

Nannya,Y. *et al*. (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*, **65**, 6071–6079.

Neuvial,P. *et al*. (2006) Spatial normalization of array-CGH data. *BMC Bioinformatics*., **7**, 264.

Pinkel,D. and Albertson,D. G. (2005) Comparative genomic hybridization. *Annu Rev. Genomics Hum. Genet*., **6**, 331–354.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat*., **6**, 461–464.

Ylstra,B. *et al*. (2006) BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucl. Acids Res*., **34**, 445–450.

Chapter 5

Segmentation of DNA copy number profiles

From a biological point of view, the segmentation of DNA copy number profiles into segments of equal copy number is an obvious task. Indeed, using such profiles greatly reduces the dimensionality of the problem. However, both from a statistical and a computational point of view, we will see that, perhaps counter-intuitively, segmentation is not a simple problem.

A simple way to grasp the complexity of this problem is to count the number of possible segmentations of a small DNA copy number profile of $n = 1000$ points. Each segmentation is completely defined by its set of breakpoints and one can put a breakpoint in between any two points. Overall there are $n - 1 = 999$ possible positions for any breakpoint and thus there are $2^{n-1} = 2^{999}$ possible segmentations of the data. Thus, the goal of any segmentation method is to choose one segmentation out of over 2^{999} possibilities for a profile of $n = 1000$ points. From a computational point of view, the difficult aspect is to find an efficient way of exploring this exponentially large set of possibilities. Meanwhile, from a statistical point of view, the difficulty lies in the selection of one solution when the number of possibilities to choose from is exponentially bigger than the amount of data available. All this means that this procedure is complex and can also be rather hazardous.

Nonetheless, many segmentation methodologies have been proposed, among which:

- Jong et al. (2003) used a Genetic Algorithm (GA) that tried to maximize a likelihood with a penalty term containing the number of breakpoints;

- Hupé et al. (2004) used a Gaussian model-based approach;

- Fridlyand et al. (2004) used Hidden Markov Model (HMM) in which the underlying DNA copy numbers are the hidden states with certain transition probabilities;

- Olshen et al. (2004) used the Circular Binary Segmentation (CBS), where the maximum of a likelihood ratio statistic is used recursively to detect narrower segments of aberration;

- Wang et al. (2005) used hierarchical tree-style clustering (Clustering Along Chromosomes);

- Picard et al. (2005) used a dynamic programming algorithm with a penalized likelihood in order to choose the most appropriate number of breakpoints;

- Engler et al. (2006); Broet and Richardson (2006); Guha et al. (2008) used latent variable approaches with Gaussian mixture models;

- Ben-Yaacov and Eldar (2008) used wavelet decomposition and thresholding;

- Lai et al. (2008) used Bayesian segmentation models.

The relative performances of some of these methods have been assessed empirically (Willenbrock and Fridlyand, 2005; Lai et al., 2005).

In any segmentation method for DNA copy number profiles, there are fundamentally three different problems that are often considered as a whole:

- the modeling problem is: defining a biologically relevant model or a collection of relevant (biostatistical) models;

- the statistical problem is: developing valid statistical criteria to select one model out of this collection and estimate the parameters of the model;

- the computational problem is: defining an efficient strategy to compute these statistical criteria and to explore the set of models to recover a good one with respect to these criteria (if possible, the best one).

If a particular methodology fails, the question is obviously: which of these three elements is to blame? As we have seen, the segmentation problem is a very difficult problem from a computational perspective, especially for very large profiles. Thus, many segmentation methods leave no choice but to rely on a non-optimal computational scheme or heuristic which might sometimes explain some erratic behaviors. In this respect, CGHseg (Picard et al., 2005) is the only method to have an optimal computational strategy, i.e. it explores the full set of possible segmentations to recover the best model according to its definition.

In the following section, I will first explain why a segmentation model is well justified from a biological point of view. I will then briefly describe the CGHseg methodology (Picard et al. (2005), modeling, statistical and algorithmic strategies) that has been shown to be one of the best methods available to analyze CGH arrays (Lai et al., 2005) but was limited to the analysis of relatively small profiles (less than 10000 points). I will then describe two of my contributions to improve this methodology (Rigaill et al. (2010c); Rigaill (2010b)), these two papers are provided in subsection 5.4 and 5.6). The first one tackles the problem of assessing the quality of a given segmentation or a given breakpoint. The second extends the optimal computational scheme of CGHseg for the analysis of very large DNA copy number profiles such as Affymetrix SNP 6.0 (with more than 10^5 points).

5.1 A piecewise constant model for the analysis of DNA copy number profiles

Among other things, CGH and SNP arrays both enable us to measure the DNA copy number of tumoral cells along the genome. In theory, only a few values are possible DNA copy numbers (0, 1, 2, 3, 4, ...) and the signal is piecewise constant. In practice, due to some stochasticity in the measurements, it is not the case. Thus, the signal is often modeled as a piecewise constant signal affected by some noise.

This piecewise constant model seems rather simple but is well adapted. Indeed, if we consider a clonal population and put aside technical biases, we expect the DNA copy number to be piecewise constant.

Some data shows smoother transitions that might be due to cell heterogeneity (Huang et al., 2007). Indeed, in the case of tumor samples procured through a surgical procedure, there are often several types of cells: normal cells, peritumoral cells and tumoral cells with possibly several subclones (Kronenwett et al., 2004). Note that, in that case, we still expect the true signal to be piecewise constant because a finite sum of piecewise constant functions is piecewise constant. Nonetheless, if there are several cell populations with breakpoints at almost the same position (but not exactly the same), we can get the impression of a smooth transition because of the low probe density of the microarray and/or the variability of the data (see Figure 5.1). However, it seems unlikely for this phenomenon to be frequent.

Wave patterns have been identified in both SNP and CGH arrays. These waves are likely due to technical artifacts and appear to correlate with GC content. Various methodologies have been specifically proposed to normalize these waves (Marioni et al., 2007; Diskin et al., 2008; van de Wiel et al., 2009). To conclude, it seems reasonable to assume that the underlying DNA copy number profile is a succession of segments or regions sharing the same DNA copy number and this justifies the use of a segmentation model.

5.2 The CGHseg methodology

The model The CGHseg methodology (Picard et al., 2005) relies on piecewise constant models. It considers all possible segmentations of the data up to a user-defined maximum of K segments. CGHseg makes several common statistical hypotheses, namely that the residual errors are independent from each other, that they follow normal distributions centered on 0 and that the variance of these distributions is at least constant by segment. These hypotheses are rather technical, as they simplify both the statistical inference and the computational analysis. From a modeling point of view, none of these hypotheses are well justified. In particular, the empirical distributions of residual errors do not seem to be normal (van de Wiel et al., 2010), though one can hope they are not too far from normality.

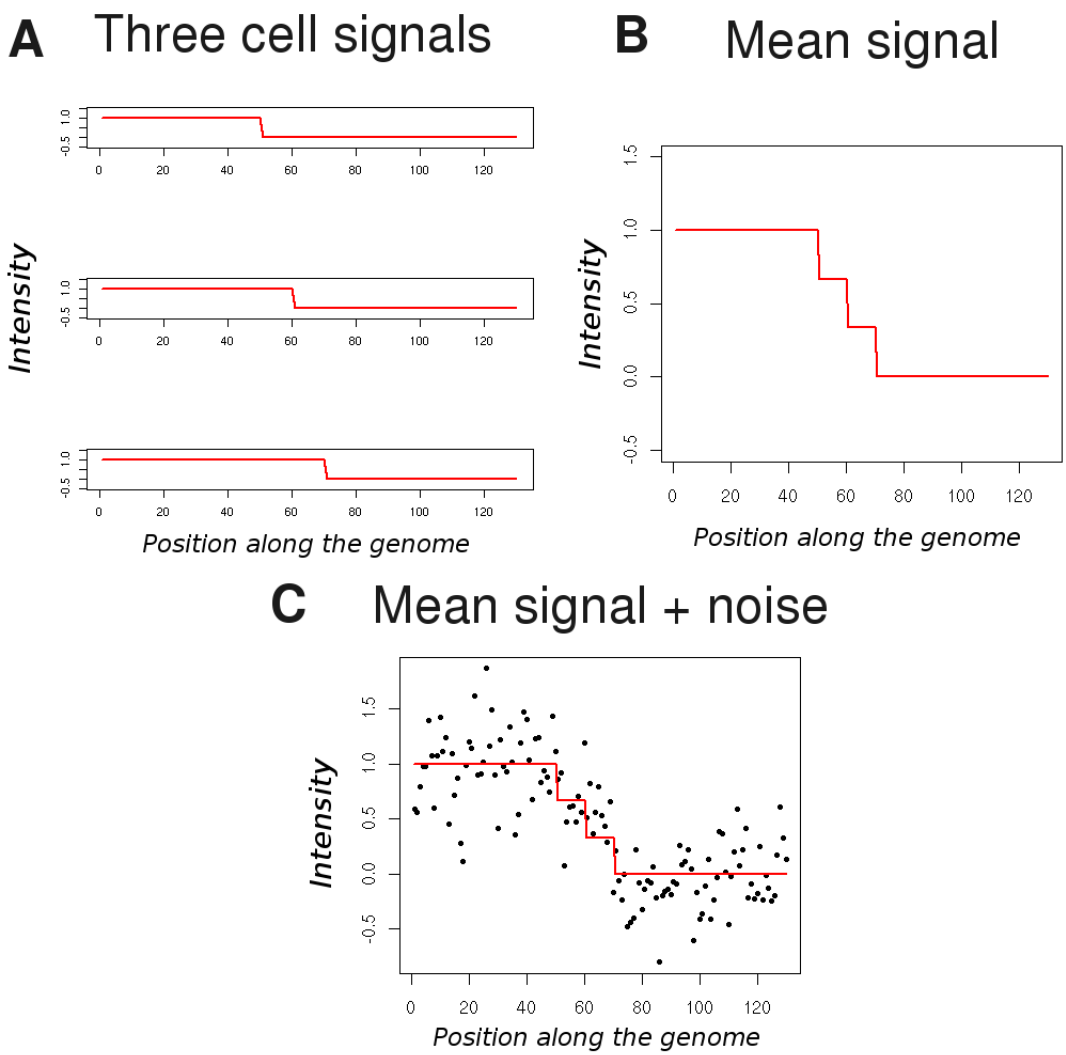


Figure 5.1: Impact of cell-to-cell variability if there are several cell populations with close breakpoints. In all graphs, the x-axis is the position along the chromosome and the y-axis is the intensity. (A) In this example, we consider three different cell populations with three very close breakpoints, each represented in a graph. (B) The average of these three signals is still piecewise constant. (C) If we consider some noise around the mean signal (black dots), we get the impression of a relatively smooth signal. However, it seems unlikely for this phenomenon to be frequent.

To conclude, these hypotheses are commonly made in many applications, including DNA copy number profile analysis, and give good results empirically. Two major issues arise from this model:

- a statistical one, the estimation of the number of segments;
- a computational one, the localization of the segments on the genome.

The statistical criteria Having defined these possible segmentations, it is possible to derive measures of adequacy between a segmentation in k segments and the data. Here, using the maximum likelihood method, one unsurprisingly obtains the classical Mean Square Error (MSE) as a measure of adequacy. The MSE is the sum of all squared distances between observed and predicted values. However, segmentations with more segments tend to better fit the data and based only on the MSE, one would over-fit the data and always retrieve a maximum number of breakpoints. This is a typical issue in statistics and one of the solutions is to penalize the MSE increasingly with the number of breakpoints. For CGHseg, two penalization schemes adapted to segmentation issues were proposed (Lebarbier, 2005; Lavielle, 2005).

The computational strategy Having defined these criteria, the goal becomes quite clear. The objective is to recover the best segmentation, i.e. the one with the smallest mean squared error, for each number of segments up to K. Knowing these "best" segmentations, it would then be possible to select the best one according to the penalty term. Recovering the best segmentations with respect to the MSE for each number of segments up to K is possible using a dynamic programming algorithm (Bellman, 1961). This algorithm is well adapted for profiles up to 10000 points but cannot be applied to larger profiles with a modern computer.

The computational and statistical issues are tightly connected. For a given statistical criteria there is a need for an efficient algorithm. Reversely, given an algorithm to explore the segmentation space it becomes possible to access new quantities and then study their statistical properties.

To conclude on CGHseg, it relies on a simple piecewise constant model, uses model selection criteria to select the number of breakpoints and uses an optimal computational scheme to recover the best

segmentation. Interestingly, it has been shown to be one of the best performing methods for the analysis of CGH data (Lai et al., 2005). I think there are two reasons for this:

- it relies on an optimal computational scheme;
- there is only one, easy-to-calibrate parameter: the maximum number of breakpoint K.

5.3 Assessing the quality of a given segmentation

This section highlights a contribution I made with Emilie Lebarbier and Stéphane Robin (Rigaill et al. (2010c)) to assess the quality of a given segmentation. From here on I will refer to the said paper as the "exploration paper". Most segmentation methods give as an output a particular segmentation. This segmentation has been selected using some defined statistical criteria and identified using a particular computational scheme out of 2^{n-1} possible segmentations (with n the number of points in the DNA copy number profiles). From a statistical point of view, two obvious questions arise:

- If the computational scheme is a heuristic, how confident are we that we have indeed found a good segmentation?
- Can we assess the confidence we have in a given breakpoint or segment?

For a valid biological interpretation, we would like to be sure that the best segmentation is by far the best fit to the data. If not, we would like to check that the second best, third best and more generally other good segmentations do not have a completely different set of change-points. Thus, assessing the quality of a given segmentation is an important issue (Lai et al., 2008).

These questions have already been studied from a statistical point of view, e.g. Yao (1984); Fearnhead (2005); Guédon (2008); Lai et al. (2008). The main idea behind the existing methodologies is the use of an algorithm (most of the time a forward-backward-like algorithm) to fully explore the segmentation space (computational issue). From this exploration, it is possible to derive quantities of statistical interest such as the probability of a breakpoint (statistical issue).

From a model selection point of view, the segmentation problem is a difficult one. To choose the number of breakpoints one usually considers the set of all segmentations with k segments ($\mathcal{M}_k$)

for any k in between 1 and a user specified maximum. Within those sets, a segmentation is further defined by the exact position of its breakpoints. Breakpoints are discrete parameters and many of the often-used model selection criteria fail to accommodate this discrete status. In particular, the Bayesian Information Criteria (BIC) is not theoretically justified due to this discrete status. Recently a modified BIC was proposed (Zhang and Siegmund (2007)). This modified BIC was derived from an asymptotic approximation of the Bayes Factor. We recently proposed an exact formulation of this criterion. The key idea behind this result is to consider one given segmentation in k segments instead of all possible segmentations in k (see section 3 page 7 of the exploration paper).

From a computational point of view, we showed that exploring the segmentation space can be viewed as a simple vector-matrix product (see Theorem 2.1 page 3 of the exploration paper). This exploration enables us to compute exactly the following quantities: the posterior probability of a breakpoint, the posterior probability of a segment, the posterior probability of the number of breakpoints and the entropy of the segmentation space (see Proposition 2.5 page 6 of the paper). The entropy is a key quantity to assess the quality of the segmentation. Intuitively, a small entropy means that the best segmentation is by far the best. Thus, if the entropy is small it is possible to interpret the best segmentation with confidence because it stands out. Using this entropy information to assess the quality of the chosen segmentation is interesting, but it would actually be more interesting to use this criterion earlier in the process, to select the number of breakpoints. We followed this line of thought more rigorously and adapted the Integrated Completed Likelihood (ICL) criteria first developed in the mixture model context (Biernacki et al., 2000). As described in subsection 3.2 page 7 of the exploration paper, the ICL for segmentation takes into account three elements to select the number of breakpoints:

- how well it fits to the data;
- a penalty for large numbers of segments;
- a penalty for large entropies.

An important remaining issue is the application of this assessment methodology to very large DNA copy number profiles. No optimal computational scheme is available. But several approximate

procedures have been proposed (Fearnhead, 2005; Lai et al., 2008). Another strategy would be to restrict the exploration to some good candidates or to the m best candidates as proposed by Guédon (2008).

5.4 Paper: Exploration of the segmentation space

This paper is available on arxiv: 1004.4347 and it is under consideration for publication in a journal in the field of Computational Statistics. A shorter version of the paper has been accepted in COMPSTAT 2010 and published in the proceedings of the conference (Rigaill et al., 2010d).

Exact posterior distributions over the segmentation space and model selection for multiple change-point detection problems

G. Rigaill[1,2,3], E. Lebarbier[1,2], S. Robin[2,1]

([1]) AgroParisTech, UMR 518, F-75005, Paris, France
([2]) INRA, UMR 518, F-75005, Paris, France
([3]) Institut Curie, Département de Transfert, F-75005 Paris, France

Abstract

In segmentation problems, inference on change-point position and model selection are two difficult issues due to the discrete nature of change-points. In a Bayesian context, we derive exact, non-asymptotic, explicit and tractable formulae for the posterior distribution of variables such as the number of change-points or their positions. We also derive a new selection criterion that accounts for the reliability of the results. All these results are based on an efficient strategy to explore the whole segmentation space, which is very large. We illustrate our methodology on both simulated data and a comparative genomic hybridisation profile.

Keywords: BIC, change-point detection, ICL, model selection, posterior distribution of change-points

Short title: Posterior distribution over the segmentation space

1 Introduction

Segmentation and change-point detection problems arise in many scientific domains such as econometrics, climatology, agronomy or molecular biology. The general problem can be written as follows. It is assumed that the observed data $\{y_t\}_{t=1,\dots,n}$ is a realization of an independent random process $Y = \{Y_t\}_{t=1,\dots,n}$. This process is drawn from a probability distribution G, which depends on a set of parameters denoted by θ. These parameters are assumed to be affected by $K-1$ abrupt changes, called change-points, at some unknown positions $\tau_2, \dots, \tau_K$(with the convention $\tau_1 = 1$ and $\tau_{K+1} = n+1$). Thus, the change-points delimit a partition m of $\{1, \dots, n\}$, called here a segmentation, into K segments $r^{(k)}$ such that $r^{(k)} = [\![\tau_k, \tau_{k+1}[\![= \{\tau_k, \tau_k + 1, \dots, \tau_{k+1} - 1\}$ and

$$m = \{r^{(k)}\}_{k=1,\dots,K}$$

The segmentation model has the following general form for a given m:

$$Y_t \sim G(\theta_r) \qquad \text{if } t \in r \quad \text{and} \quad r \in m$$

where θ_r stands for the parameters of segment r. In this study, all the change-points are detected simultaneously, a strategy called off-line detection (as opposed to on-line detection). With this strategy, the question of finding the best segmentation in a given number of segments has already been largely studied (see for example Lavielle (2005), Braun and Müller (2000), Bai and Perron (2003)). But two important issues remain: assessing the quality of the proposed segmentation and selecting the number of segments (also called dimension). In both cases, the main problem is the discrete nature of the change-points, which prevents the use of routine statistical inference.

On the one hand, the quality of a given segmentation can be assessed by studying the uncertainty of the change-point positions. From a non-asymptotic and non-parametric point of view, the standard likelihood-based inference is very intricate, since the required regularity conditions for the change-point parameters are not satisfied (Feder (1975)). Different methods to obtain change-point confidence intervals have been proposed. Most of them are based on the limit distribution of the change-point estimators (Feder (1975), Bai and Perron (2003)) or the asymptotic use of a likelihood-ratio statistic (Muggeo (2003)). Others proposed confidence intervals are based on bootstrap techniques (Husková and Kirch (2008) and references therein). A practical comparison of these methods can be found in Toms and Lesperance (2003).

On the other hand, choosing the number of segments is also a critical issue. This is usually done by minimising a penalised contrast function and the problem is to find a good penalty. General penalized criteria have been developed, such as AIC (Akaike (1973)) and BIC (Schwarz (1978)). In the segmentation framework, these criteria are not adapted since an exponential model collection is considered (Birgé and Massart (2007), Baraud *et al.* (2009)) and these criteria tend to overestimate the number of segments (see for example Lavielle (2005)). Recently, some penalised criteria have been proposed specially for the segmentation framework. Some depend on constants to be calibrated (Lavielle (2005) and Lebarbier (2005)), but others do not (Zhang and Siegmund (2007)). More precisely, Zhang and Siegmund (2007) discussed the fact that the classical BIC was not theoretically justified in the segmentation context. Indeed, the BIC criterion is derived from an asymptotic approximation of the posterior model probabilities and requires the likelihood function to be three times differentiable with respect to the parameters of the model (Kass and Raftery (1995), Lebarbier and Mary-Huard (2006)). As the change-points are discrete parameters, the previous condition is not satisfied. A modified BIC criterion has thus been developed by Zhang and Siegmund (2007) by considering a continuous-time version of the problem.

The purpose of our work is to provide exact, non-asymptotic, explicit and tractable formulae for both the posterior probability of a segmentation and that of a change-point occurring at a given position. More specifically, we consider the segmentation problem in a Bayesian framework so that the posterior probability of a segmentation is well defined. To tackle the discrete nature of change-points, we work at the segment level, where statistical inference is straightforward. From these segments, the issue is to get back to the segmentation or dimension level. Provided that the segments are independent, it will be necessary to calculate quantities such as:

$$\sum_{m \in \mathcal{M}^\star} P(Y|m)P(m) = \sum_{m \in \mathcal{M}^\star} P(m) \prod_{r \in m} P(Y^r|r) \tag{1}$$

where Y^r stands for all observations in segment r and $\mathcal{M}^\star$ is usually a very large set of segmentations. We propose a close-form (in terms of matrix products) and tractable formulation of such quantities. Some similar quantities were computed by Guédon (2008) in a non-Bayesian context, using a forward-backward-like algorithm. However, this author computes all these quantities for fixed values of the segment parameters, which are the maximum likelihood estimators. From our formula, we derive key quantities to assess the quality of a segmentation and select the number of segments.

On the one hand, we obtain the exact formulae for both the posterior probability of a segmentation and that of a change-point occurring at a given position. This enables the construction of credibility intervals for change-points. Moreover, we retrieve the exact posterior probability of a segment within a given dimension, the exact entropy of the posterior distribution of the segmentations within a given dimension and the exact posterior mean of the signal.

On the other hand, we derive a so-called 'exact' BIC criterion for choosing the number of segments K, taking $\mathcal{M}^\star = \mathcal{M}_K$ which is the set of all possible segmentations with K segments. In the same way, we derive the ICL criterion of Biernacki *et al.* (2000) in the segmentation framework. This last criterion takes into account the reliability of the results.

In Section 2, we give some exact formulae to explore the segmentation space and assess the quality of a segmentation. In Section 3, we focus on the model selection problem: we derive an exact BIC criterion and propose a new ICL criterion. In the last section, we illustrate our results first on Poisson simulated data and second on comparative genomic hybridization (CGH) data in a Gaussian framework.

2 Exploring the segmentation space

A naive computation of (1) is impossible when $\mathcal{M}^\star$ is large, which is usually the case. For example, if $\mathcal{M}^\star = \mathcal{M}_K$, there are $\binom{n-1}{K-1}$ segmentations of n data into K segments. In this section we propose a tractable and close-form formula of (1). The following assumption enables us to derive an exact matrix product formulation of (1) enabling its straightforward computation in $O(Kn^2)$ time.

Factorability assumption: A model satisfies the factorability assumption if

$$(\mathbf{H}) : P(Y, m) = C \prod_{r \in m} a_r P(Y^r | r) \tag{2}$$

where $P(Y^r|r) = \int P(Y^r|\theta_r)P(\theta_r)\mathrm{d}\theta_r$. In the following, for the sake of clarity, we will simply denote $P(Y^r)$. This is true when all segment parameters are different but this is false, for example, for the normal homoscedastic model $G(\theta_r) = \mathcal{N}(\mu_r, 1/\tau)$ with unknown precision τ.

We denote by $\mathcal{M}_K(\llbracket i, j \llbracket)$ the set of all possible segmentations of $\llbracket i, j \llbracket$ into K segments. The simplified notation $\mathcal{M}_K$ refers to $\mathcal{M}_K(\llbracket 1, n+1 \llbracket)$.

Theorem 2.1 *Consider a function F such that, for all $k \in \llbracket 1, K \rrbracket$ and for all segmentation $m \in \mathcal{M}_k(\llbracket 1, j \llbracket)$ (for $1 \leq j \leq n+1$), there exists a function f such that: $F(m) = \prod_{r \in m} f(r)$. Let $\mathbf{A}$ be a square matrix with $n+1$ columns such that*

$$\begin{array}{rcll} \mathbf{A}_{ij} & = & f(\llbracket i, j \llbracket) & \text{if } 1 \leq i < j \leq n+1 \\ & = & 0 & \text{otherwise.} \end{array}$$

Then,

$$\sum_{m \in \mathcal{M}_k(\llbracket 1, j \llbracket)} F(m) = (\mathbf{A}^k)_{1,j}$$

and the $K \times (n+1)$ elements of

$$\left\{ \sum_{m \in \mathcal{M}_k(\llbracket 1, j \llbracket)} F(m) \right\}_{k \in \llbracket 1, K \rrbracket \cap j \in \llbracket 1, n+1 \rrbracket}$$

can all be computed in $O(Kn^2)$

The proof is given in Appendix A.1. It is based on a linear algebra lemma. The lower triangular part of matrix $\mathbf{A}$ is set to 0 to fit the segmentation context. Note that, similarly, we have $\sum_{m \in \mathcal{M}_k(\llbracket i, j \llbracket)} F(m) = (\mathbf{A}^k)_{i,j}$ for all $1 \leq i \leq j \leq n+1$. Theorem 2.1 will be used many times in the following sections, using a specific function $f(r)$ for each quantity of interest.

2.1 Joint distribution of the data and the segmentation or the dimension

$P(Y, m)$ and $P(Y, K)$ are key ingredients to calculate various quantities of interest, such as (1). To calculate $P(Y, m)$ and

$$P(Y, K) = \sum_{m \in \mathcal{M}_K} P(Y, m), \tag{3}$$

we first need to define priors for the segmentation m. We consider here two typical priors.

Uniform conditional on the dimension: For any prior on the dimension $P(K)$, we define a uniform prior distribution for m given its dimension K:

$$P(m|K(m)) = \binom{n-1}{K(m)-1}^{-1} \qquad \Rightarrow \qquad P(m) = P(K(m)) \Big/ \binom{n-1}{K(m)-1} \tag{4}$$

that is $a_r = 1$ in (2), denoting $K(m)$ the number of segments (i.. the dimension) of the segmentation m.

Homogeneous segment lengths: Segmentation with balanced segment lengths are sometimes desirable. They are favoured by the following prior:

$$P(m) = C \prod_{r \in m} n_r^{-1}, \qquad \text{where } C \text{ ensures that } \sum_{m \in \mathcal{M}} P(m) = 1. \tag{5}$$

that is $a_r = n_r^{-1}$ in (2), where n_r denotes the length of segment r and $\mathcal{M}$ the set of all considered segmentations. In this case, the prior distribution of m is directly defined and the prior distribution of the dimension $P(K)$ is not explicit. Determining the constant C requires summing over all possible segmentations. This sum can be handled using the properties given below.

Proposition 2.2 *Under assumption* (**H**)*, for prior distributions* (4) *and* (5)*, $P(Y, K)$ can be computed in $O(Kn^2)$ as $P(Y, K) = C(\mathbf{A}^k)_{1,n+1}$ with $\mathbf{A}_{i,j} = 0$ for $j \leq i$ and, for $j > i$, for prior distribution* (4)*:*

$$\mathbf{A}_{i,j} = P(Y^{[\![i,j]\!]}) \quad \textit{and} \quad C^{-1} = \binom{n-1}{K-1};$$

and for prior distribution (5)*:*

$$\mathbf{A}_{i,j} = n_{[\![i,j]\!]}^{-1} P(Y^{[\![i,j]\!]}) \quad \textit{and} \quad C^{-1} = \sum_{m \in \mathcal{M}_K} \prod_{r \in m} n_r^{-1}.$$

Proof. For prior distribution (4), we use Theorem 2.1 with $f(r) = P(Y^r)$, implying $\mathbf{A}_{i,j} = f([\![i, j]\!]) = P(Y^{[\![i,j]\!]})$.
For prior distribution (5), we first retrieve C using Theorem 2.1 with $f(r) = n_r$. The result follows, using Theorem 2.1 again, taking $f(r) = n_r^{-1} P(Y^r)$.■

The preceding results require the calculation of $P(Y^r)$. Hence, $n(n-1)/2$ integrals need to be evaluated, corresponding to each possible segment. For general priors, they can be evaluated numerically or via any stochastic algorithm. A close form can be obtained if conjugate priors are used.

Poisson and Gaussian models. We recall classical results for two models that will be used later. First is the segmentation problem of a piecewise constant Poisson rate model:

$$\begin{aligned} \{\mu_r\} \text{ i.i.d.}, \quad \mu_r &\sim \mathcal{G}\text{am}(\alpha_r, \beta_r); \\ \{Y_t\} \text{ independent}, \quad Y_t &\sim \mathcal{P}(\mu_r) \quad \text{if } t \in r. \end{aligned} \tag{6}$$

Second is the segmentation of a Gaussian signal where both the mean and the variance are affected by the change-points:

$$\begin{aligned} \{\tau_r\} \text{ i.i.d.}, \quad \tau_r &\sim \mathcal{G}\text{am}(\nu_0/2, 2/s_0); \\ \{\mu_r\} \text{ independent}, \quad \mu_r|\tau_r &\sim \mathcal{N}(\mu_0, (n_0\tau_r)^{-1}); \\ \{Y_t\} \text{ independent}, \quad Y_t &\sim \mathcal{N}(\mu_r, 1/\tau_r) \qquad \text{if } t \in r. \end{aligned} \tag{7}$$

For the Poisson model, we get

$$P(Y^r) = \frac{\Gamma(\alpha + \sum_{t \in r} Y_t^r)\ \beta_r^{\alpha_r}}{(\beta_r + n_r)^{\alpha_r + \sum_{t \in r} Y_t^r}\ \Gamma(\alpha_r) \prod_{t \in r} (Y_t^r!)}.$$

For the Gaussian heteroscedastic model, we get

$$P(Y^r) = \frac{{n_0}^{1/2}\ (s_0/2)^{\nu_0/2}\ \Gamma((\nu_0 + n_r)/2)}{(2\pi)^{n_r/2}\ \Gamma(\nu_0/2)\ \sqrt{n_r + n_0}}\ \theta^{(\nu_0 + n_r)/2} \tag{8}$$

where $\theta = 2(n_r S_r^2 + s_0 + \frac{n_r n_0 (\bar{y}_r - \mu_0)^2}{n_r + n_0})^{-1}$, $S_r^2 = \sum_{t \in r} (Y_t - \bar{y}_r)^2 / n_r$ and $\bar{y}_r$ is the empirical mean of the signal within segment r.

2.2 Posterior distribution of the change-points and segments

We now give explicit formulae for the posterior distribution of change-points and segments. We first define the corresponding segmentation subsets:

$\mathcal{B}_{K,k}(t)$ is the subset of segmentations from $\mathcal{M}_K$ such that the k-th segment starts at position t, i.e. that the $(k-1)$-th change-point is at t:

$$\mathcal{B}_{K,k}(t) = \{m \in \mathcal{M}_K : \tau_k = t\};$$

$\mathcal{B}_K(t)$ is the subset of segmentations having a change-point at position t:

$$\mathcal{B}_K(t) = \bigcup_k \mathcal{B}_{K,k}(t);$$

$\mathcal{S}_{K,k}([\![t_1, t_2[\![)$ is the subset of segmentations having segment $r = [\![t_1, t_2[\![$ as their k-th segment:

$$\mathcal{S}_{K,k}([\![t_1, t_2[\![) = \{m \in \mathcal{M}_K([\![1, n+1[\![) : \tau_k = t_1, \tau_{k+1} = t_2\};$$

$\mathcal{S}_K([\![t_1, t_2[\![)$ is the subset of segmentations including segment $[\![t_1, t_2[\![$:

$$\mathcal{S}_K([\![t_1, t_2[\![) = \bigcup_k \mathcal{S}_{K,k}([\![t_1, t_2[\![).$$

We denote the conditional probability given the data Y and the dimension K of each of these subsets by the corresponding capital letters with same indices, e.g.

$$B_{K,k}(t) = \Pr\{m \in \mathcal{B}_{K,k}(t) | Y, K\}.$$

$B_K(t)$, $S_{K,k}(t)$ and $S_K(t)$ are defined similarly. The following proposition gives explicit formulae for these probabilities.

Proposition 2.3 *For all $[\![t_1, t_2[\![$ such that $t_1 < t_2$, we define, for $K \geq 1$,*

$$F_{t_1,t_2}(K) = \sum_{m \in \mathcal{M}_K([\![t_1,t_2[\![)} P(Y^{[\![t_1,t_2[\![}|m) P(m|K),$$

*and we set $F_{t_1,t_2}(K) = 0$ if $t_1 \geq t_2$. Under assumption (**H**), the probabilities $B_{K,k}(t)$, $B_K(t)$, $S_{K,k}(t)$ and $S_K(t)$ are*

$$\begin{aligned} B_{K,k}(t) &= \frac{F_{1,t}(k-1) F_{t,n+1}(K-k+1)}{P(Y|K)}, \\ S_{K,k}(t_1,t_2) &= \frac{F_{1,t_1}(k-1) F_{t_1,t_2}(1) F_{t_2,n+1}(K-k)}{P(Y|K)} \end{aligned}$$

$B_K(t) = \sum_{k=1}^K B_{K,k}(t)$ and $S_K(t_1,t_2) = \sum_k S_{K,k}(t_1,t_2)$.

The proof is given in Appendix A.2. It is mainly based on set decompositions, such as

$$\mathcal{B}_{K,k}(t) = \mathcal{M}_{k-1}([\![1,t[\![) \times \mathcal{M}_{K-k+1}([\![t,n+1[\![) \tag{9}$$

and all sums over $\mathcal{M}_{k-1}([\![1,t[\![)$ and $\mathcal{M}_{K-k+1}([\![t,n+1[\![)$ can be obtained with Theorem 2.1.

$\{B_{K,k}(t)\}_t$ provides the exact posterior distribution of the starting point of the k-th segment, given dimension K. From that, we get the exact credibility of interval $[\![t_1, t_2]\!]$ for change-point τ_k:

$$C_{K,k}([\![t_1,t_2]\!]) = \Pr\{\tau_k \in [\![t_1,t_2]\!] | Y, K\} = \sum_{t=t_1}^{t_2} B_{K,k}(t).$$

2.3 Retrieving the mean signal

In many applications, the mean value μ_t of the signal at a given position can also provide some insight about the phenomenon under study. This mean signal can be retrieved via model averaging over the segmentation space. The posterior mean of the signal is

$$\bar{s}_K(t) = \sum_{m \in \mathcal{M}_K} P(m|Y,K)\widehat{s}_m(t), \tag{10}$$

where $\widehat{s}_m(t) = \mathbb{E}[\mu_t|m,Y]$.

Proposition 2.4 *The posterior mean of the signal given the dimension is*

$$\bar{s}_K(t) = \sum_{r \ni t} S_K(r)\widehat{\mu}_r,$$

where $\widehat{\mu}_r = \mathbb{E}[\mu_r|Y^r]$. Under assumption (**H**)*, it can be computed with a quadratic complexity.*

Proof. If a segment r belongs to a segmentation m and if position t lies in segment r then $\widehat{s}_m(t) = \widehat{\mu}_r$. The rest of the formula is straightforward. Assumption (**H**) ensures that the $S_K(r)$ can be computed in $O(Kn^2)$. ■

2.4 Posterior entropy

Segmentation problems are often reduced to choosing $\widehat{m}_K$, the best segmentation (i.e. the one with maximal posterior probability) with dimension K. The other segmentations with dimension K are rarely considered. The entropy of the distribution $P(m|Y,K)$

$$\mathcal{H}(K) = -\sum_{m \in \mathcal{M}_K} P(m|Y,K)\log P(m|Y,K)$$

measures how the posterior distribution is concentrated around the best segmentation. Intuitively, a small entropy $\mathcal{H}(K)$ means that the best segmentation is a much better fit to the data than any other segmentation. We use this information in Section 3 for model selection.

Proposition 2.5 *Under assumption* (**H**)*, the posterior entropy $\mathcal{H}(K)$ is*

$$\mathcal{H}(K) = -\sum_r S_K(r)\log f(r) + \log A_K$$

where $f(r) = a_r P(Y^r)$ and $A_K = \sum_{m \in \mathcal{M}_K} \prod_{r \in m} f(r)$, which can be computed using Proposition 2.2.

Proof. Since all distributions can be factorized, we have

$$\begin{aligned}
\mathcal{H}(K) &= -\sum_{m \in \mathcal{M}_K}\sum_{r \in m} P(m|Y,K)\log f(r) + \sum_{m \in \mathcal{M}_K} P(m|Y,K)\log A_K \\
&= -\sum_r \log f(r) \sum_{m \in \mathcal{M}_K, m \ni r} P(m|Y,K) + \log A_K \sum_{m \in \mathcal{M}_K} P(m|Y,K)
\end{aligned}$$

and the result follows. ■

3 Model selection

In a Bayesian framework, the BIC criterion aims to choose the model which maximises $P(M|Y)$, where M is the model. To calculate the BIC criterion, one needs to know $P(Y|M) = \int P(Y|\theta_M, M)P(\theta_M|M)\mathrm{d}\theta_M$, where θ_M is the set of parameters of the model M. Similar quantities are involved in the Bayes factor for model comparison (Kass and Raftery (1995)).

In our case, the word 'model' is too broad and we have to distinguish between the selection of the dimension K and the selection of the segmentation m. When considering the choice of K, a direct application of the Laplace approximation is not theoretically justified to calculate the previous integral because the required differentiability condition is not satisfied for change-points (Zhang and Siegmund (2007)). However, we can bypass the problem by working at the segment level and then going back at the dimension level using Proposition 2.2. Thus, the derivation of BIC criteria only requires the calculation of $P(Y^r) = \int P(Y^r|\theta_r)P(\theta_r)\mathrm{d}\theta_r$, which can be obtained in a close form for the Poisson model and the heteroscedastic Gaussian model as shown in Section 2.1. Moreover, we derive an adaptation of the ICL criterion, first proposed for mixture models, to the segmentation context (Biernacki *et al.* (2000)).

3.1 Exact BIC criterion for dimension and segmentation selection

Choice of the dimension. In segmentation problems, the selection of the 'best' number of segments K can be addressed per se, or as a first step toward the selection of the 'best' segmentation. The Bayesian framework suggests to choose

$$\widehat{K} = \arg\min_K \mathrm{BIC}(K), \quad \text{where} \quad \mathrm{BIC}(K) = -\log P(Y, K). \tag{11}$$

$\mathrm{BIC}(K)$ can be computed in a quadratic time, using Proposition 2.2.

Choice of the segmentation. The best segmentation can be chosen in two ways.

Two-step strategy: The 'best' segmentation m can be chosen, conditionally to the pre-selected dimension $\widehat{K}$ as

$$\widehat{m}(\widehat{K}) = \arg\min_{m \in \mathcal{M}_{\widehat{K}}} \mathrm{BIC}(m|\widehat{K}), \quad \text{where} \quad \mathrm{BIC}(m|\widehat{K}) = -\log P(Y, m|\widehat{K}). \tag{12}$$

One-step strategy: The 'best' segmentation m can also be directly chosen among a larger collection $\mathcal{M} = \bigcup_{k=1}^{K} \mathcal{M}_k$ as

$$\widehat{m} = \arg\min_{m \in \mathcal{M}} \mathrm{BIC}(m), \quad \text{where} \quad \mathrm{BIC}(m) = -\log P(Y, m). \tag{13}$$

Both $\mathrm{BIC}(m|K)$ and $\mathrm{BIC}(m)$ can be computed efficiently thanks to Proposition 2.2.

3.2 ICL criterion for dimension selection

In the framework of incomplete data models (e.g. mixture models), Biernacki *et al.* (2000) suggest to use the criterion $\mathrm{ICL}(M)$, which is an estimate of $\mathbb{E}[\log P(Y, Z, M)|Y]$ where Z stands for the unobserved variables. Based on the equation

$$\mathbb{E}[\log P(Y, Z|M)|Y] = \log P(Y|M) + \mathbb{E}[\log P(Z|Y, M)|Y],$$

they argue that the entropy $H(Z|Y, M) = -\mathbb{E}[\log P(Z|Y, M)|Y]$ is an intrinsic penalty term. The ICL criterion will tend to select models that provide a reliable prediction of Z, i.e. with a small entropy. This may be desirable, for example in the classification context.

In the segmentation context, the segmentation m can be considered as an unobserved variable. The dimension K can then be chosen according to the ICL as

$$\widehat{K} = \arg\min_K \mathrm{ICL}(K) \quad \text{where} \quad \mathrm{ICL}(K) = -\log P(Y, K) + H(m|Y, K).$$

Biernacki *et al.* (2000) We expect ICL to favour the dimension K where the best segmentation $\widehat{m}(K)$ clearly outperforms the other segmentations in K segments, so that $\widehat{m}(K)$ is more reliable.

3.3 Comparison with other penalized criteria

Many model selection criteria have the following form:

$$\log P(Y|\widehat{\theta}, m) - \text{pen}(m)$$

and use a two-step strategy. Interestingly, since the penalty generally depends only on the dimension (Lebarbier (2005), Lavielle (2005)), the best segmentation $\widehat{m}(K)$ does not actually depend on the penalty.

The calculation of the exact BIC does not provide any explicit penalty enabling a direct comparison with such criteria. For such comparison, we derive two approximations of $\log P(Y^r) = \log \int P(Y^r|\theta_r)P(\theta_r)\mathrm{d}\theta_r$ in the heteroscedastic Gaussian case. The first one is based on a Laplace approximation:

$$\log P(Y^r) \approx \log P(Y^r|\widehat{\theta}_r) - \frac{D}{2}\log n_r$$

where D stands for the number of parameters involved in each segment (here, $D = 2$). This approximation is valid only for large segments, i.e. where $P(Y^r|\theta_r)$ satisfies regularity conditions. For the second approximation, we let the hyperparameters n_0, ν_0 and S_0 go to 0 in (8) and we obtain

$$\log P(Y^r) \approx -\frac{n_r}{2}\log S_r^2 - \frac{D}{2}\log n_r \approx \log P(Y^r|\widehat{\theta}_r) - \frac{D}{2}\log n_r.$$

We emphasize that these approximations are both questionable since the asymptotic framework of the Laplace approximation is not correct for small segments and because the priors are improper for null hyperparameters. Our purpose is only to show that they both provide the same penalty form:

$$\log P(m|Y) \approx \log P(m) + \log P(Y|\widehat{\theta}, m) - \frac{D}{2}\sum_{r\in m}\log n_r.$$

Using uniform prior (4), we get

$$\text{pen}(m) = \log P(K(m)) - \log\binom{n-1}{K(m)-1} - \frac{D}{2}\sum_{r\in m}\log n_r.$$

A similar form is obtained in the Poisson case. The complexity term, $\log\binom{n-1}{K-1}$, is similar to the one of Lebarbier (2005). The regularity term, $\sum_{r\in m}\log n_r$, favours segments with equal lengths and is similar to the one of Zhang and Siegmund (2007). Using the alternative prior (5) reinforces the regularity term. Due to this term, the best segmentation $\widehat{m}(K)$ within $\mathcal{M}_K$ does depend on the penalty.

4 Applications

In this section, we first present a simulation study to assess the ability of the exact BIC and ICL criteria to select the dimension and the ability of model averaging to retrieve the mean signal. We then analyse a real CGH profile and use our formulae to assess the quality of the segmentation.

4.1 Simulations

Simulation design. We performed the simulation study in the Poisson model (6) so that only one parameter had to be chosen. We simulated a sequence of 150 observations affected by six change-points at the following positions: 21, 29, 68, 82, 115, 135. Odd segments had a mean of 1, while even segments had a mean of $1+\lambda$, where lambda varies from 0 to 10. The higher λ is, the easier it should be to recover the true number of change-points. The hyperparameters α and β were set to be equal and we considered three values for them: 0.01, 0.1 and 1. For each configuration, we simulated 300 sequences.

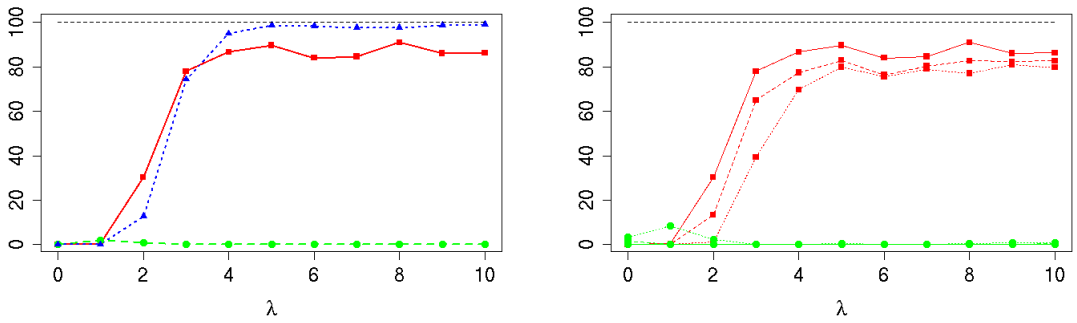


Figure 1: Percentage of true dimension recoveries as a function of λ. Left panel: for the three criteria. BIC($\widehat{m}_K$) : ■−, BIC(K) : • − − and ICL(K) : ▲ $\cdots$. Right panel: for the BIC criteria; •: uniform prior over all segmentations, ■: uniform prior over all segmentations of a dimension, − : $\alpha = \beta = 1$, −− : $\alpha = \beta = 0.1$, $\cdots$: $\alpha = \beta = 0.01$.

4.2 Recovering the number of change-points

4.2.1 The ICL criterion performed better than the BIC criterion

Model selection. The BIC criterion for dimension selection, BIC(K), almost never returned the true dimension, even for high values of λ (Figure 1, where α and β were set to 1). On the other hand, both the BIC criterion for model selection, BIC(m), and the ICL criterion, ICL(K), tend to recover the true dimension more often when λ became larger. ICL(K) even increased to a maximum of 99% true recoveries compared to a maximum of 91% for the BIC(m) criterion for model selection.

Influence of the priors. The ability of BIC(m) to retrieve the true dimension was greatly affected by the prior distribution of the segmentation (Figure 1). To illustrate this effect, we considered a prior that gave equal probability to all segmentations, whatever their dimension: $P(m) = \text{cst}$. This led to a 90% decrease in the ability to return the true dimension compared to a conditional uniform prior given the dimension (4) (with $P(K(m)) = \text{cst}$ whatever m). The impact of the two hyperparameters α and β seemed relatively limited in comparison: less than 10% difference in the ability to return the true dimension (Figure 1).

Estimation of the mean signal. We then compared the ability of the maximum likelihood estimators (MLE) and that of the posterior mean signal to recover the true signal in terms of the Kullback-Leibler distance. For each simulation, we computed the following:

$$d(\widehat{\mu}, \mu) = \sum_t KL[\mathcal{P}(\widehat{\mu}_t); \mathcal{P}(\mu_t)]$$

for both the MLE estimate $\widehat{\mu} = \widehat{\mu}_{\text{MLE}}$ and the posterior mean $\widehat{\mu} = \bar{s}_K(t)$ (see equation (10)).

When K was lower than the true dimension (7 segments), the two estimates were almost equivalent (Figure 2). However, for larger dimensions, the distance of the MLE to the true signal increased whereas the distance of the posterior mean did not (Figure 2). The posterior mean seemed less prone to over-fitting. Moreover, for a very small signal-to-noise ratio ($\lambda = 1$), the distance between the posterior mean of the signal and the true signal still decreased when K was higher than the true dimension. Therefore, when the signal was of poor quality and led to a poor assessment of the true dimension, the posterior mean of the signal led to better results. Moreover, the standard deviation of d for the posterior mean is almost always smaller than the one of the MLE (not shown).

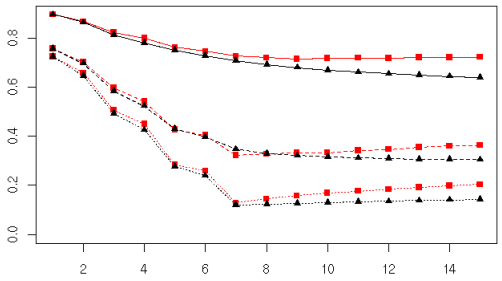


Figure 2: Kullblack-Leibler-based distance d to the true signal as a function of the dimension. ■: $d(\widehat{\mu}_{\mathrm{MLE}}, \mu)$, ▲: $d(\overline{\mu}, \mu)$ for three value of λ 1: $-$, 2: $--$ and 6: $\cdots$. The true number of segments was 7.

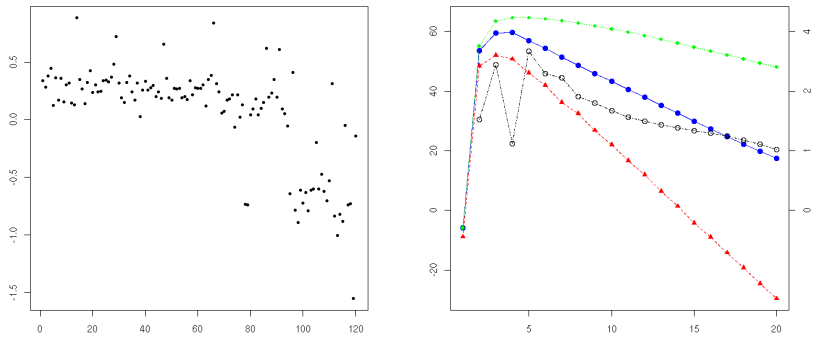


Figure 3: Left panel: Chromosome 10 profile of cell line BT474. The DNA copy number logratio is represented as a function of its position along the chromosome. Right panel: (Left axis) BIC(m): ▲, BIC(K): • and ICL(K): ■ as a function of the dimension. (Right axis) $\mathcal{H}(K) - \mathcal{H}(K-1)$: ∘ as a function of the dimension.

4.3 Analysis of a CGH profile

In the following subsection, we used a comparative genomic hybridation (CGH) profile to illustrate our methodology. CGH enables the study of DNA copy number gains and losses along the genome (Pinkel *et al.* (1998)). We used the Gaussian segmentation model defined in (7) that is often used for this type of data (Picard *et al.* (2005)). The profile shown in Figure 3 represents the copy number logratio of cell line BT474 to a normal reference sample, along chromosome 10.

Model selection. Since the true dimension was unknown, the first issue was to choose one. The ICL(K) criterion selected 4 segments whereas BIC(m) selected a segmentation with 3 segments (Figure 3). The additional penalty term involved in ICL does not necessarily penalise larger dimensions. In our example, ICL selected a segmentation with a larger dimension because it was more reliable. The choice of ICL was motivated by the relatively small gain of entropy between dimensions 3 and 4. This choice was also supported by the posterior distributions of the change-points and that of the segments shown below. The best segmentations for 3 and 4 segments are shown on Figure 4 (i).

Posterior probability of the change-point positions. The distribution of the successive change-points for dimensions 3 and 4 are shown on Figure 4 (*ii*). For dimension 3, the exact intervals with credibility 95% were $[\![64, 78]\!]$ and $[\![92, 97]\!]$ for τ_2 and τ_3, respectively. For dimension 4, the intervals were $[\![66, 78]\!]$, $[\![78, 97]\!]$ and $[\![91, 112]\!]$ for τ_2, τ_3 and τ_4, respectively.

The existence of a change-point at a given position t is assessed by posterior probability $B_K(t)$. Note that, contrarily to $B_{K,k}(t)$, $B_K(t)$ is not a probability distribution over the positions, because its sum is the number of change-points: $K-1$. In our example, the posterior probabilities $B_4(t)$ presented sharper peaks than $B_3(t)$ (see Figure 4 (*iii*)), which was consistent with the choice of the ICL criterion that favours reliable segmentations.

Posterior probability of a segment. Similar conclusions were drawn from the posterior probability of the segments. In Figure 4 (*iv*) each point corresponds to a segment. A reliable dimension should display K sharp peaks. The position of the first two segments are very uncertain for $K = 3$, due to the uncertainty of τ_2. Their position were much more certain with $K = 4$. In particular, the smallest segment from $K = 4$ at positions $[\![78, 79]\!]$ had a relatively high probability of 0.34.

Posterior mean of the signal. Similarly, the posterior mean for 3 segments was different from the one for 4 segments (Figure 5); the former failed to capture the small deletion at $[\![78, 79]\!]$. As soon as K exceeded 4, the posterior mean of the signal was very stable, see the example for $K = 5$ segments in Figure 5.

All presented results show that, the segmentation in 4 segments selected by the ICL(K) is more reliable than the segmentation in 3 segments selected by the BIC(m).

Acknowledgements. We thank Marie-Pierre Etienne (AgroParisTech, UMR 518, Paris) for her helpful advice for the writing of this paper. We also thank Thierry Dubois (Institut Curie, dpt de Transfert) and Emmanuel Barillot (Institut Curie, MinesParisTech, INSERM, unité U900) for their support.

References

[Akaike (1973)] Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, (B. Petrov and F. Csaki, ed.), 267–281. Akademiai Kiado, Budapest.

[Bai and Perron (2003)] Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econ.* **18** 1–22.

[Baraud *et al.* (2009)] Baraud, Y., Giraud, C. and Huet, S. (2009). Gaussian model selection with unknown variance. *AS.* **37 (2)** 630–672.

[Biernacki *et al.* (2000)] Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel.* **22 (7)** 719–725.

[Birgé and Massart (2007)] Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. *Probability Th. and Related Fields.* **138** 33–73.

[Braun and Müller (2000)] Braun, R.-K., J.-V. Braun and Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika.* **87** 301–314.

[Feder (1975)] Feder, P. I. (1975). The loglikelihood ratio in segmented regression. *"AS".* **3 (1)** 84–97.

[Guédon (2008)] Guédon, Y. (2008), Explorating the segmentation space for the assessment of multiple change-points models. Technical report, Preprint INRIA n°6619.

[Husková and Kirch (2008)] Husková, M. and Kirch, C. (2008). Bootstrapping confidence intervals for the change-point of time series. *Journal of Time Series Analysis.* **29 (6)** 947–972.

[Kass and Raftery (1995)] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

[Lavielle (2005)] Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing.* **85 (8)** 1501–1510.

[Lebarbier (2005)] Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing.* **85** 717–736.

[Lebarbier and Mary-Huard (2006)] Lebarbier, E. and Mary-Huard, T. (2006). Une introduction au critère BIC : fondements théoriques et interprétation. *J. Soc. Française Statis.* **147 (1)** 39–57.

[Muggeo (2003)] Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Stat. Med.* **22 (19)** 3055–3071.

[Picard *et al.* (2005)] Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, J.-J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics.* **6 (27)** 1. `www.biomedcentral.com/1471-2105/6/27`.

[Pinkel *et al.* (1998)] Pinkel, D., Segraves, R., Sudar, D., S.Clark, Poole, I., D.Kowbel, C.Collins, Kuo, W., C.Chen, Zhai, Y., Dairkee, S., Ljung, B. and Gray, J. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics.* **(20)** 207–211.

[Schwarz (1978)] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6 (2)** 461–4.

[Toms and Lesperance (2003)] Toms, J. D. and Lesperance, M. L. (2003). Piecewise regression: A tool for identifying ecological thresholds. *Ecology.* **84 (8)** 2034–2041.

[Zhang and Siegmund (2007)] Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics.* **63 (1)** 22–32.

A Lemma and Proofs

A.1 Proof of Theorem 2.1

The proof of the theorem relies on the following lemma.

Lemma A.1 *Let* $\mathbf{A}$ *be a square matrix with* n *columns. For all* $k \in \mathbb{N}$*, we define the function* $f_{\mathbf{A},k}$ *as:*

$$\forall (i,j) \in [\![1,n]\!]^2 \qquad f_{\mathbf{A},k}(i,j) = \sum_{(t_2 \cdots t_k) \ \in \ [\![1,n]\!]^{k-1}}^{t_1=i, \ t_{k+1}=j} \prod_{i=1}^{k} \mathbf{A}_{t_i,t_{i+1}}$$

The n *elements of* $\{f_{\mathbf{A},k}(i,j)\}_{\{i \ \in \ [\![1,n]\!]\}}$ *for* $1 \leq k \leq K$ *can all be computed in* $O(Kn^2)$ *as*

$$f_{\mathbf{A},k}(i,j) = (\mathbf{A}^k)_{i,j}$$

.

Proof of the Lemma. $f_{\mathbf{A},k}(i,j) = \mathbf{A}^1_{i,j}$ holds for $k=1$. Suppose that $f_{\mathbf{A}}(k,i,j) = \mathbf{A}^k_{i,j}$ holds for a given $k \in \mathbb{N}$. For $k+1$, we have:

$$f_{\mathbf{A},k+1}(i,j) = \sum_{(t_2 \cdots t_{k+1}) \in [\![1,n]\!]^k}^{t_1=i, t_{k+2}=j} \prod_{i=1}^{k+1} \mathbf{A}_{t_i,t_{i+1}} = \sum_{t=1}^{n} \sum_{(t_2 \cdots t_k) \in [\![1,n]\!]^{k-1}}^{t_1=i, t_{k+1}=t} \prod_{i=1}^{k} \mathbf{A}_{t_i,t_{i+1}}. \ \mathbf{A}_{t,j} = \sum_{t=1}^{n} f_{\mathbf{A},k}(i,t). \ \mathbf{A}_{t,j}$$

Using our induction hypothesis and by definition of the matrix product, we obtain:

$$f_{\mathbf{A},k+1}(i,j) \quad = \quad \sum_{t=1}^{n} \mathbf{A}^k_{i,t} \mathbf{A}_{t,j} = \mathbf{A}^{k+1}_{i,j}$$

Thus, the $K \times n$ elements of the form

$$\{f_{\mathbf{A},k}(t_1, t_{k+1})\}_{\{k \ \in \ [\![1,K]\!] \ \cap \ t_{k+1} \ \in \ [\![1,n]\!]\}}$$

can be computed in $O(Kn^2)$ as the t_1-th line of matrices $\mathbf{A}$, $\mathbf{A}^2 \cdots$, $\mathbf{A}^K$ respectively. ■

Proof of the Theorem. For any $(t_1, ..., t_{k+1})$ in $[\![1, n+1]\!]^{k+1}$ such that we do not have $t_1 < t_2 \cdots < t_{k+1}$, $\prod_{i=1}^{k} \mathbf{A}_{t_i,t_{i+1}} = 0$. Therefore, for all $k \in [\![1,K]\!]$ and for all j in $[\![1,n]\!]$:

$$\sum_{m \in \mathcal{M}_k([\![1,j]\!])} F(m) = \sum_{t_1 < t_2 \cdots < t_{k+1}}^{t_1=1, t_{k+1}=j} \prod_{i=1}^{k} \mathbf{A}_{t_i,t_{i+1}} = \sum_{(t_2, \ldots t_k) \ \in \ [\![1,n+1]\!]^{k-1}}^{t_1=1, t_{k+1}=j} \prod_{i=1}^{k} \mathbf{A}_{t_i,t_{i+1}}$$

Using Lemma A.1 on matrix $\mathbf{A}$ and integer K, we see that the $K \times (n+1)$ terms of the form

$$\left\{ \sum_{m \in \mathcal{M}_k([\![1,j]\!])} F(m) \right\}_{k \ \in \ [\![1,K]\!] \ \cap \ j \in [\![1,n+1]\!]}$$

can be computed as $\sum_{m \in \mathcal{M}_k([\![1,j]\!])} F(m) = (\mathbf{A}^k)_{1,j}$ and that therefore they can all be computed in $O(Kn^2)$ as the first line of the successive powers of matrix $\mathbf{A}$.

A.2 Proof of Proposition 2.3

Proof. We first consider the posterior distribution of the change-points. With Equation (9), we obtain

$$B_{K,k}(t) = \frac{\sum_{m \in \mathcal{B}_{K,k}(t)} P(Y|m) P(m|K)}{P(Y|K)} = \frac{F_{1,t}(k-1) F_{t,n+1}(K-k+1)}{P(Y|K)}.$$

Using Theorem 2.1, we see that all the F functions can be computed in $O(Kn^2)$. $O(K^2 n)$ products and divisions remain to be done to compute all $B_{K,k}(t)$, so the overall complexity is in $O(Kn^2)$. The probability $B_K(t)$ follows straightforwardly.

We now consider the posterior distribution of the segments. We first quote that if $t_1 = 1$, then $S_{K,1}(1, t_2) = B_{K,2}(t_2)$. Similarly, when $t_2 = n+1$, we have $S_{K,K}(t_1, t_2) = B_{K,K}(t_1)$. So we only have to consider the case where $1 < t_1 \leq t_2 < n+1$. Since $\mathcal{S}_{K,k}([\![t_1, t_2]\!])$ can be decomposed as

$$\mathcal{S}_{K,k}([\![t_1, t_2]\!]) = \mathcal{M}_{k-1}([\![1, t_1]\!]) \times \{[\![t_1, t_2]\!]\} \times \mathcal{M}_{K-k}([\![t_2, n+1]\!]),$$

we have

$$S_{K,k}(t_1, t_2) = \frac{\sum_{m \in \mathcal{S}_{K,k}([\![t_1,t_2]\!])} P(Y|m) P(m|k)}{P(Y|K)} = \frac{F_{1,t_1}(k-1) F_{t_1,t_2}(1) F_{t_2,n+1}(K-k)}{P(Y|K)}.$$

Again using Theorem 2.1, we see that all the F functions can be computed in $O(Kn^2)$. We then need to compute $O(n^2)$ products and divisions to get the $S_{K,k}(t_1, t_2)$, thus the overall complexity is in $O(Kn^2)$. The last probability comes from the definition of $S_K(t_1, t_2)$. $O(Kn^2)$ additions remain to be done the overall complexity is therefore in $O(Kn^2)$. ■

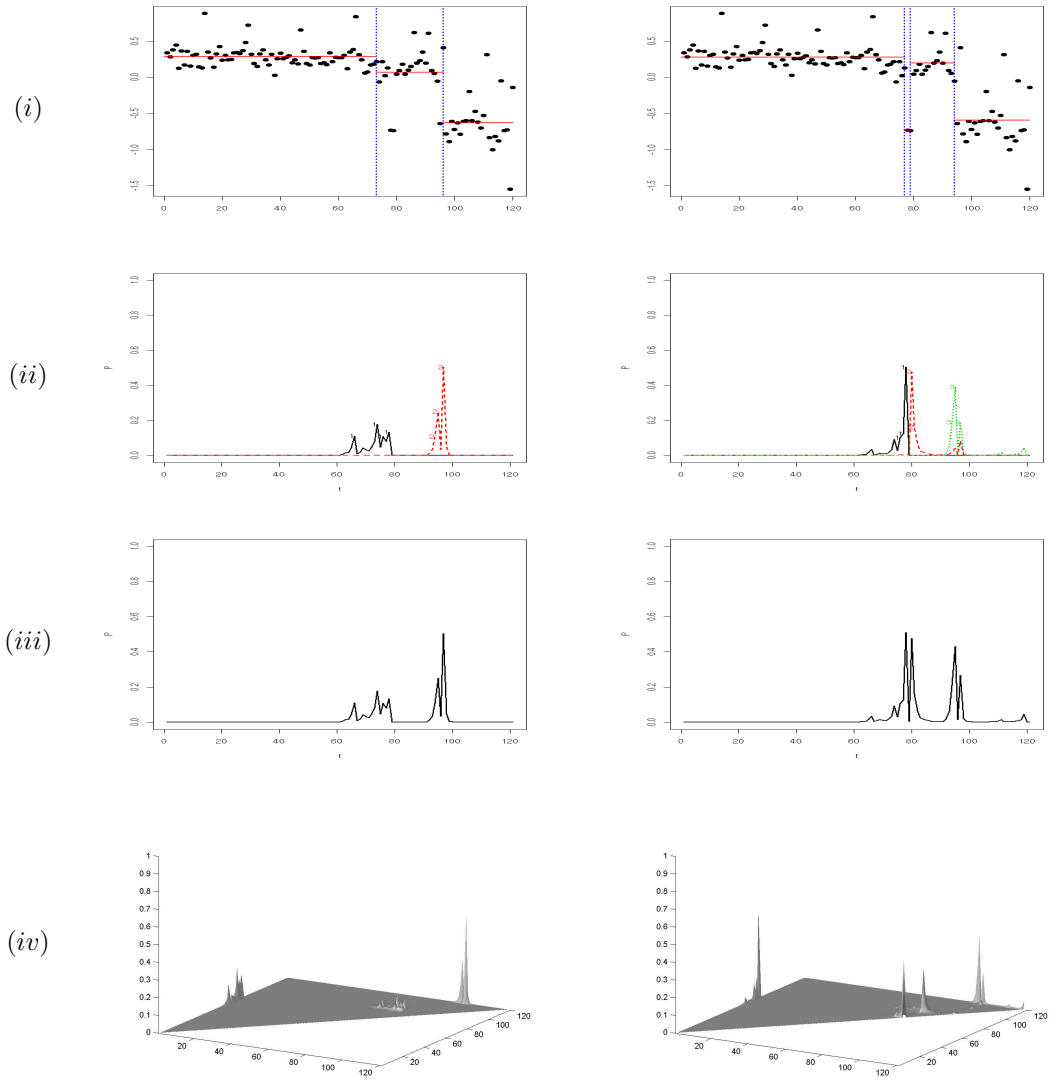


Figure 4: (i): Best segmentation of the profile in 3 (left) and 4 (right) segments. $\bullet$ represent the logratio as a function of the position along the chromosome. $-$: averaged signal of the segment. $\cdots$: change-point positions. (ii): Posterior probability that the k-th change-point is at position t knowing that there is either 3 (left) or 4 (right) segments. Probability of the first change-point: $-$, probability of the second change-point: $--$ and probability of the third change-point: $\cdots$. (iii): Posterior probability that there is a change-point at position t knowing that there is 3 (right) or 4 (left) segments. (iv) : 3D plot of the probability of all segments. Left panel: $K = 3$ segments; right panel: $K = 4$ segments. x-axis: t_1, y-axis: t_2, z-axis: $S(\llbracket t_1, t_2 \rrbracket)$.

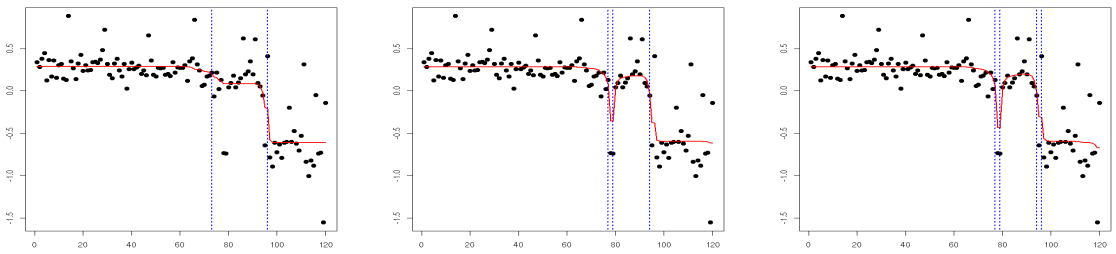


Figure 5: Posterior mean of the signal; Left: $K = 3$ segments; Center: $K = 4$ segments; Right: $K = 5$ segments. $\bullet$: logratio as a function of the position along the chromosome. $-$: posterior mean of the signal. $\cdots$: change-point positions of the best segmentation.

5.5 Optimal computational scheme for large DNA copy number profiles

This section highlights the pruned Dynamic Programming Algorithm (DPA) I developed to enable the analysis of very large DNA copy number profiles (at least up to 10^6 points) with respect to the MSE criterion on a regular computer (Rigaill (2010b), the paper is provided in the next section 5.6). I will refer to this paper as the "pruned DPA" paper). The computational strategy has become one of the major problems regarding CGH and SNP profile data analysis, due to the increasing size of this data (Venkatraman and Olshen, 2007; Ben-Yaacov and Eldar, 2008; Lai et al., 2008; Tibshirani and Wang, 2008). Execution time of the strategy is now one of the foremost issues when developing new methods. Indeed, several methods developed for CGH arrays cannot be used for large profiles because their execution time is too long (several days). For example, it is the case of CGHseg. Indeed the original DPA used in CGHseg to recover the best segmentation with respect to the MSE is too complex.

There are two reasons for this complexity: a space complexity and a time complexity problem. For large (more than 10000 points) DNA copy number profiles, it requires too much memory (RAM) to run on a regular computer. This problem has been solved quite recently (Guédon, 2008) by adapting a forward-backward dynamic programming algorithm proposed by (Auger and Lawrence, 1989). I implemented this trick in the case of the CGHseg methodology enabling the analysis of larger profiles, up to 10^5 points (unpublished results).

However, this trick does not solve the time issue. It does enable the analysis of profiles up to 10^5 points but it takes several hours to do so with a modern computer. Still, when compared to the time needed to collect the tumor sample and perform the experiments (usually several months), it does not seem that long. The relationship between the time and the size of the profile is quadratic, i.e. the time is proportional to the square of the size. Overall, the time required to analyze a single Affymetrix SNP 6.0 on one computer should be several days.

Thus this time problem is still an issue and recovering the best segmentation with respect to the MSE is a known problem not only for DNA copy number profile analysis but for other fields as well

(see Harchaoui and Levy-Leduc (2008); Gey and Lebarbier (2008) and references therein). Several strategies have been proposed to avoid this problem, e.g.:

- changing the optimization problem slightly (Harchaoui and Levy-Leduc, 2008);
- using a heuristic (Gey and Lebarbier, 2008).

It is important to note that, these methods do not recover the optimal solution with respect to the MSE. Using them allows faster run-times, but at the cost of some errors. With some luck, one would hope to recover a solution that is not too far from the best solution, or even the best solution . In relatively simple examples, it can be the case, but it cannot be guaranteed. When it comes to biological interpretation, this is big issue, as you cannot be sure that the identified breakpoints are the best, and can have no idea how close or how far you are from them.

I will now give an example to show that segmentation is not a simple computational problem, though it intuitively seems to be so. One of the usual ideas to recover a good segmentation of the data is to use a recursive scheme. The simplest recursive scheme is the following. First you cut the signal optimally into two pieces. Then you iterate the process on each of the two pieces. This seems like a sound idea, but it is not optimal. Indeed, to cut a cake in three equal pieces, you should not start by cutting the cake into two equal pieces. But, we all know, that cutting a cake is a problematic issue! As illustrated by this seemingly simple example, as well as by Figure 5.2, this procedure is bound to fail. Indeed, choosing where to cut before knowing how many pieces will be needed is hazardous. While these are specific counterexamples, there are some more robust recursive segmentation methodologies. Still, many of these need to rely on a post-processing step to discard irrelevantly identified breakpoints (Olshen et al., 2004; Harchaoui and Levy-Leduc, 2008; Gey and Lebarbier, 2008) and it seems unlikely for these errors not to have influenced the following decisions. Overall, and if possible, one should always prefer an optimal computational scheme (an algorithm *strictu senso*) over a heuristic.

In the piecewise constant segmentation problem there are two types of parameters the positions of the breakpoints and the mean level in between each of these breakpoints. The original DPA first optimize the level within each segment. Then knowing these optimal values, it searches for the optimal breakpoint positions. The key idea of the pruned DPA is to work the other way around. First, the

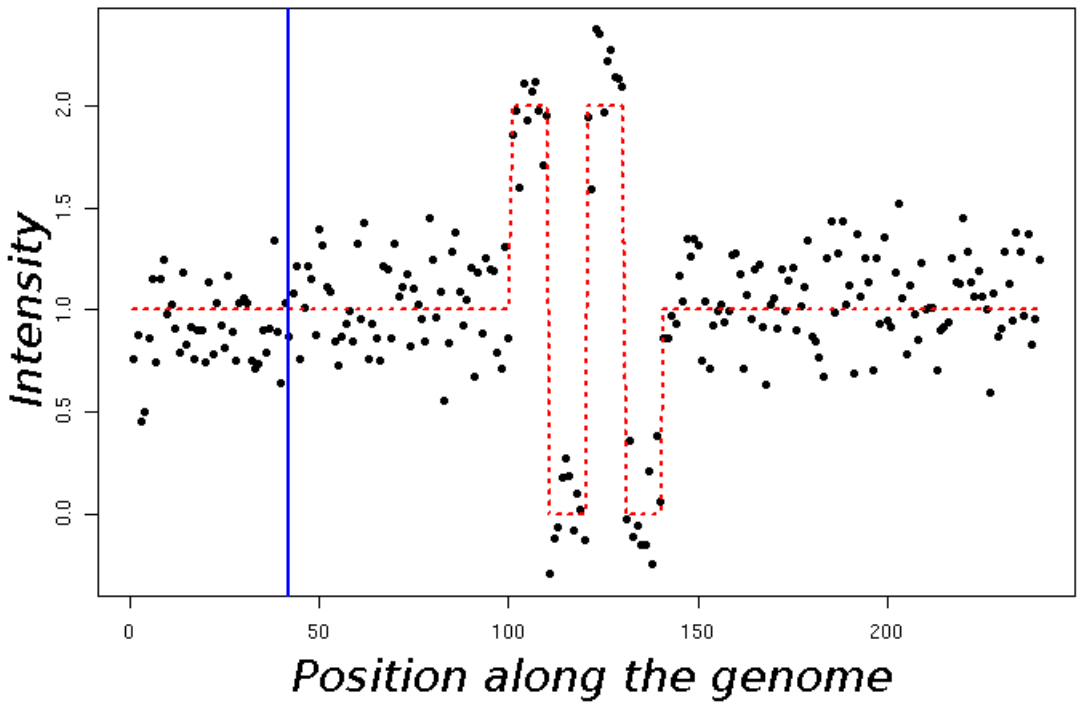


Figure 5.2: Cutting a signal into two segments based on a recursive methodology. The x-axis is the position along the genome and the y-axis is the intensity. The measured signal is represented by black dots. The true signal is represented as a red dotted line. The position of the best cut in two pieces of the signal is represented by the vertical blue line. Here, the signal is small and the noise is relatively low so it should be quite easy to recover the true position of breakpoints manually. The best segmentation in 7 segments with respect to the MSE is indeed the true signal.

pruned DPA optimizes the position of the breakpoints for any possible values of the signal. Then knowing these solutions it optimizes the mean level within each segment. For a more complete intuition of the algorithm see section 2 on page 2 of the pruned DPA paper. For a full description see section 3 on page 3 of the paper. The overall run-time to scan an Affymetrix SNP 6.0 array chromosome by chromosome with this algorithm is about 3 minutes with a regular computer (see subsection 5.2 at page 8 of the paper). To my knowledge, the only other optimal schemes available to optimize the MSE are:

- the original DP algorithm (Bellman, 1961), which cannot be run due to the RAM space issue;

- the linear space DP algorithm (Guédon, 2008), which would take several days to scan the previous array.

The pruned DPA is faster than the original one and obtains exactly the same results. Since the original DPA has been validated for small CGH profiles (less than 10000 points), one can assume that the results obtained by the pruned version are also valid for such values as well as larger ones. Also, for larger values, the results are based on more points and therefore, might actually be even more reliable. To conclude, using the pruned DPA I have developed, it is now possible to recover the best segmentation of the data with respect to the MSE in a reasonable amount of time. We have seen that this MSE criteria is relatively well adapted for the analysis of SNP or CGH arrays (Lai et al., 2005), though sometimes SNP and CGH arrays are plagued by outliers. In such cases, it would be interesting to use criteria more robust to outliers such as the L1 loss. It can be hoped that adapting the pruned DP algorithm to the L1 loss is possible with only a few more computational tricks. Also, note that a very efficient algorithm was proposed to recover the best segmentation with respect to the L_∞ loss (Fournier and Vigneron, 2008). Finally, the pruned DP algorithm has been implemented in an R package and will soon be made available. It has already been used for the analysis of the Curie-Servier dataset and was made available for several collaborators of the Institut Curie.

5.6 Paper: Pruned dynamic programming for segmentation

This paper is available on arxiv: 1004.0887.

Pruned dynamic programming for optimal multiple change-point detection

Guillem Rigaill

August 18, 2010

Abstract

Change-point detection problems are a common occurrence, and therefore have been extensively studied. Yet, finding the best change-points w.r.t. the quadratic loss of a one dimensional piecewise-constant signal remains a major computational bottleneck, especially for large datasets. Indeed, the complexity of the Dynamic Programming Algorithm (DPA) that is generally used is far too important to analyze large datasets. Faster methods exits, but they do not recover the optimal solution. We propose a pruned DPA that recovers the best change-points w.r.t. the quadratic loss, and prove that its complexity is at worst equivalent to the original DPA. Moreover, we show empirically, both with simulated and real data of up to a million points, that the pruned DPA can actually be used to analyse large datasets. The algorithm is proposed for an entire class of loss functions, including the quadratic and Poisson functions. Moreover, seeing that it processes one point after the other, it could be adapted to on-line data.

1 Introduction

Segmentation and change-point detection problems are a common occurrence in various fields of research, such as econometrics [1], audio [2] and molecular biology [3]. The task consists in splitting the signal in K homogeneous and contiguous segments of variable length. These segments are delimited by $K-1$ change-points and can be identified on-line (sequentially) or off-line (retrospectively) [4]. It is well known that for the off-line detection problem, a Dynamic Programming Algorithm (DPA) recovers the optimal solution w.r.t. the quadratic loss of an n-point signal in $\Theta(Kn^2)$ time ([5], [6], [7]) and $\Theta(Kn)$ space ([8]). In practice, the quadratic time complexity in n restricts the use of such an algorithm to small or intermediate values of n.

In order to handle large datasets and get around the problem of complexity, several approaches were proposed, relying on efficient heuristics (see [9]) or on small modifications of the optimization problem (see [10], [11]). However these methods do not recover the optimal solution w.r.t. the quadratic loss. Using them allows faster run-times, but at the cost of some errors. Many other fast methods exist, for example wavelets denoising (see for example [12]), or particle filters (see [13], [14]) but they were not developed to optimize the same criteria.

Our contribution In this paper, we will present a pruned DPA that efficiently recovers the optimal $K-1$ change-points w.r.t. the quadratic loss of a one-dimensional piecewise-constant signal. We will

prove that its complexity is at worst equivalent to the original DPA both in time ($O(Kn^2)$) and space ($O(Kn)$, [8]). Moreover, we will show empirically that the pruned DPA is in many cases faster than the original DPA, especially for large datasets. For example, the pruned DPA processes large SNP (Single Nucleotide Polymorphism) profiles of a million points in a few minutes while the original DPA would take several days. The pruned DPA algorithm works for a whole class of convex loss functions, including the quadratic and Poisson functions. The algorithm processes one observation after the other, and therefore could be adapted to on-line data.

Outline In Section 2, we will describe our framework, give a brief overview of the original DPA ([5], [6], [7]) and outline the key principle of the algorithm we worked on. In Section 3, we will describe this pruned DPA, in Section 4 we will demonstrate its worst case complexity and in Section 5 we will empirically compare the runtime of our pruned DPA and the original DPA on both simulated and real data. In the last section we will give a detailed proof of the algorithm's complexity.

2 From the original to the pruned DPA

In this section we first outline the original DPA. We then propose a linear time dynamic programming heuristic as a first step towards the exact pruned DPA.

2.1 Framework and original DPA

We consider a sequence of n observations $\{Y_i\}_{i\in[\![1,n]\!]}$ in a set A (e.g. $\mathbb{N}$, $\mathbb{R}$, ...). We call $\mathcal{M}_{K,t}$ the set of all possible segmentations in K segments up to point t. The number of possible segmentations, $card(\mathcal{M}_{K,t})$, is $\binom{t-1}{K-1}$. We denote $r_k(m) = [\![\tau_k(m), \tau_{k+1}(m)[\![= \{\tau_k(m), \dots, \tau_{k+1}(m) - 1\}$ the k-th segment of segmentation m delimited by change-points $\tau_k(m)$ and $\tau_{k+1}(m)$. With the convention that $\tau_1(m) = 1$ and $\tau_{K+1}(m) = t+1$, any segmentation m of $\mathcal{M}_{K,t}$ is defined as $\{r_1(m), \dots, r_K(m)\}$. Both the original and proposed pruned DPA recover:

$$argmin_{\{m \in \mathcal{M}_{K,n}\}}\{ \sum_{r \in m} \quad min_{\{\mu \in \mathbb{R}\}}\{\sum_{i \in r} \gamma(Y_i, \mu)\}\}$$

where $\gamma : A \times \mathbb{R} \to \mathbb{R}$ is a convex function of μ. Throughout the paper we will take the quadratic loss, $\gamma(Y_i, \mu) = (Y_i - \mu)^2$, as a leading example.

For any segment r, we define its cost as $g_r(\mu) = \sum_{i \in r} \gamma(Y_i, \mu)$ and its optimal cost as $c_r = min_\mu\{g_r(\mu)\}$. We define $C_{k,t} = min_{\{m \in \mathcal{M}_{k,t}\}}\{ \sum_{r \in m} c_r \}$. As $C_{k,t}$ is segment-additive, it is easy to prove for $k \geq 2$ the following update equation:

$$\forall\ t \geq k \qquad C_{k,t} = \min_{k-1 \leq j < t}\{ C_{k-1,j} + c_{[\![j+1,t+1[\![}\} \tag{1}$$

Using update equation (1), the original DPA performs t comparisons at each step and therefore retrieves $C_{K,n}$ in exactly $\Theta(Kn^2)$ runtime.

2.2 A simple linear-time dynamic programming heuristic

If we know the value of the last segment, the optimization problem becomes much simpler. Indeed, the task consist in minimizing for a given μ, $H_{k,t}(\mu)$:

$$H_{k,t}(\mu) = min_{\{m \in \mathcal{M}_{k,t}\}} \{ \sum_{r=r_1(m)}^{r_{k-1}(m)} c_r + g_{r_k(m)}(\mu) \},$$

Note that $g_{r_k(m)}(\mu)$ is point-additive and, similarly to $C_{K,t}$, $H_{k,t}(\mu)$ is segment-additive. This way we retrieve in a straightforward way the update equation:

$$\forall\ t \geq k \qquad H_{k,t+1}(\mu) = \min(\ H_{k,t}(\mu),\ C_{k-1,t}\ ) + \gamma(Y_{t+1}, \mu) \tag{2}$$

Using update equation (2), only one comparison is done at each step and $H_{k,n}(\mu)$ is retrieved in linear-time. Unfortunately, in most cases we do not know the value of μ. A basic heuristic method is to test a large but finite set of possible values for μ denoted here $\{\mu_p\}_{p \in [\![1,P]\!]}$ and use equation (2) repeatedly to recover all $H_{K,n}(\mu_p)$ in $\Theta(KPn)$ runtime and retrieve an upper bound of $C_{K,n}$ as $min_{p \in [\![1,P]\!]} \{H_{K,n}(\mu_p)\}$.

However, this method is heuristic based and therefore does not necessarily retrieve the optimal solution. If we want to get a solution that is closer to the optimal one, we simply increase the number of tested values μ_p. For each of these tested μ_p the proposed heuristic method stores an optimal last change-point. Intuitively, two very close tested values of μ probably share the same optimal last change-point. In other words, the heuristic probably stores several times the same information and it seems that only critical values of μ, corresponding to a change in the last optimal change-point, are needed. The pruned DPA is built on this idea and updates at each step the set of critical values.

3 Pruned DP algorithm for segmentation

In this section, we describe the pruned DPA. Briefly, for each total number of segments, k, from 2 to K, the algorithm works on a list of candidate last change-points, $Candidates_k$. For each of these change-points, t', the algorithm stores:

- its cost, $Cost_{k,t'}$, as a function of μ which is the mean of the last segment;
- its set of μ winning intervals, $Set_{k,t'}$, for which the change-point t' is optimal.

The bounds of these winning intervals are the critical values of μ described in subsection 2.2. With every new data point, the pruned DPA efficiently updates and prunes the list of candidates and their associated costs and winning intervals.

3.1 The algorithm

More precisely, we define $h_{k,t,t'}(\mu)$ as the cost of the best candidate segmentation in k segments with a last change-point at position t' and a last segment mean value of μ:

$$\forall\ t' < t \qquad h_{k,t,t'}(\mu) = C_{k-1,t'} + \sum_{i=t'+1}^{t} \gamma(Y_i, \mu),$$

We have: $H_{k,t}(\mu) = min_{\{t' \in [\![k-1,t-1]\!]\}}\{ h_{k,t,t'}(\mu) \}$. We define $S_{k,t,t'}$ as the set of μ values such that a last change-point at t' is optimal:

$$S_{k,t,t'} = \{\mu \mid h_{k,t,t'}(\mu) = H_{k,t}(\mu)\}.$$

We denote $I_{k,t,t'}$ as the set of μ values such that a last change at t' is better than a change at t:

$$I_{k,t,t'} = \{ \mu \mid h_{k,t,t'}(\mu) \ \leq \ C_{k-1,t} \}.$$

As γ is convex, $h_{k,t,t'}$ is convex and thus $I_{k,t,t'}$ is an interval.

Here we review the key properties of $h_{k,t,t'}$ and $S_{k,t,t'}$ that allow, in the course of the pruned DPA, to simply update the cost functions ($Cost_{k,t'}$) and the winning intervals ($Set_{k,t'}$) and to efficiently prune the candidate last change-points.

Proposition 3.1

Cost $\quad \forall\ t\ >\ t' \qquad h_{k,t+1,t'}(\mu) = h_{k,t,t'}(\mu) + \gamma(Y_{t+1}, \mu)$

Interval $\quad \forall\ t > t' \geq k, \quad S_{k,t+1,t'} \ = \ S_{k,t,t'} \ \cap I_{k,t,t'}$
$\quad \forall\ t' \geq k, \qquad S_{k,t',t'}, \ = \ \complement_{\mathbb{R}}(\cup_{t \in [\![k-1,t'-1]\!]} I_{k,t,t'})$

Pruning $\quad S_{k,t,t'} = \emptyset \qquad \Rightarrow \qquad \forall\ t^* \geq t \quad S_{k,t^*,t'} = \emptyset$

The Cost property is obvious. The proofs of the other properties rely on the Cost property and are left to the reader. Interestingly, using the interval property we see that, as all $I_{k,t,t'}$ are intervals, all $S_{k,t,t'}$ are finite union of intervals. In other words the pruning rule says that, if at observation t candidate t' is beaten for every possible μ then whatever the observations after t the best segmentation does not change at t'.

In the DPA, $Cost_{k,t'}$ and $Set_{k,t'}$ store respectively the successive $h_{k,t,t'}$ and $S_{k,t,t'}$ using the Cost and Interval properties (see Proposition 3.1). As soon as $Set_{k,t'}$ is empty t' is discarded from the list of candidates using the Pruning property (see Proposition 3.1). Importantly, in the case of the quadratic loss, $\gamma(Y_i, \mu) = (Y_i - \mu)^2$, cost functions, $Cost_{k,t'}$, are stored as a second degree polynomial function of μ and sets of winning intervals, $Set_{k,t'}$, are all initialized as $[min_i(Y_i), max_i(Y_i)]$.

For each possible number of segments, k, from 2 to K the DPA proceeds schematically as follows. First the list of candidates is initialized as $\{k-1\}$ because the first possible last change-point is $k-1$. Then for every new data point t:

1. a new candidate change-point t is initialized ;
2. all previous candidate cost functions are updated ;
3. these cost functions are compared to the new one to update the winning intervals ;
4. all candidates with an empty set are discarded ;
5. the best candidate at point t is retrieved.

The pruned DPA is described in more details in Algorithm 1 for a given $k \geq 2$. For $k = 1$, all $C_{1,t}$ are computed in $\Theta(n)$ as in the original DPA. Importantly, the algorithm gradually includes the observed data points and therefore is suitable for the detection of change-points in on-line data.

Algorithm 1 Pruned DPA

$Candidates_k := \{k-1\}$
$Cost_{k,k-1} := C_{k-1,k-1}$; $Set_{k,k-1} := D$
for t **from** k **to** $n-1$ **do**
  $Cost_{k,t} := C_{k-1,t}$; $Set_{k,t} := D$
  for $l \in Candidates_k$ **do**
    $Cost_{k,l} := Cost_{k,l} + \gamma(Y_t, .)$
    $I = \{\mu \mid Cost_{k,l}(\mu) \leq Cost_{k,t}\}$
    $Set_{k,l} := Set_{k,l} \cap I$
    if $Set_{k,l} = \emptyset$ **then**
      $Candidates_k := Candidates_k \setminus \{l\}$
    end if
    $Set_{k,t} := Set_{k,t} \setminus I$
  end for
  if $Set_{k,t} \neq \emptyset$ **then**
    $Candidates_k := Candidates_k \cup \{t\}$
  end if
  $C_{k,t} = min_{\{l \in Candidates_k\}}\{min_\mu (Cost_{k,l})\}$
end for
for $l \in Candidates_k$ **do**
  $Cost_{k,l} := Cost_{k,l} + \gamma(Y_t, .)$
end for
$C_{k,n} = min_{\{l \in Candidates_k\}}\{min_\mu (Cost_{k,l})\}$

3.2 An example

In this subsection, we illustrate the proposed pruned DPA with a four-point signal for $k = 2$ segments. These four points are respectively $0, 0.5, 0.4, -0.5$ (see Figure 1-A).

Step 1 The algorithm initializes the candidate $t' = 1$. It is represented in Figure 1-B and stored as:

Candidate	Cost function	Set of Intervals
$t' = 1$	$Cost_{2,1} = \ C_{1,1} = 0$	$Set_{2,1} = \ [-0.5, \ 0.5\ ]$

Step 2 Then, the pruned DPA initializes candidate $t' = 2$. It adds $(Y_2 - \mu)^2 = (0.5 - \mu)^2$ to the cost function of $t' = 1$. Next, it compares candidate $t' = 1$ and $t' = 2$ and updates their winning intervals. These two candidates are represented in Figure 1-C and stored as:

Candidate	Cost function	Set of Intervals
$t' = 1$	$Cost_{2,1} = \ 0.25 - \mu + \mu^2$	$Set_{2,1} = \ [\ 0.146\ ,\ 0.5\ ]$
$t' = 2$	$Cost_{2,2} = \ C_{1,2} = 0.125$	$Set_{2,2} = \ [-0.5, 0.146\ ]$

Step 3 Then, the pruned DPA initializes a new candidate $t' = 3$. It adds $(Y_3 - \mu)^2 = (0.4 - \mu)^2$ to the cost functions of both candidate $t' = 1$ and $t' = 2$. Next, it compares these two candidates to the new candidate $t' = 3$ and updates their winning intervals. Figure 1-D shows that candidate $t' = 2$ is beaten for all possible μ and thus is discarded by the algorithm. In the end, the algorithm stores:

Candidate	Cost function	Set of Intervals
$t' = 1$	$Cost_{2,1} = \ 0.41 - 1.8\mu + 2\mu^2$	$Set_{2,1} = \ [\ 0.190\ ,\ 0.5\ ]$
$t' = 3$	$Cost_{2,3} = \ C_{1,3} = 0.14$	$Set_{2,3} = \ [-0.5, 0.190\ ]$

Step 4 Finally, the algorithm adds $(Y_4 - \mu)^2 = (-0.5 - \mu)^2$ to all candidates and retrieves :

Candidate	Cost function	Set of Intervals
$t' = 1$	$Cost_{2,1} = \ 0.66 - 0.8\mu + 3\mu^2$	$Set_{2,1} = \ [\ 0.190\ ,\ 0.5\ ]$
$t' = 3$	$Cost_{2,3} = \ 0.39 + \mu + \mu^2$	$Set_{2,3} = \ [-0.5, 0.190\ ]$

Figure 1-E shows that the best segmentation in 2 segments for the four-point signal is obtained for $\mu = -0.5$ and a last change-point $t' = 3$.

4 Worst case complexity

In this section, we prove that, if the loss function γ is convex, the complexity of the pruned DPA is at worst in $O(Kn^2)$ time and in $O(Kn)$ space. We obtain this complexity by bounding the total number of intervals stored by the algorithm (see section 7).

More precisely, for a given number of segments k and a given point i of the algorithm at most $i-k+2$ change-point candidates need to be updated. For each candidate change-point t the algorithm will first

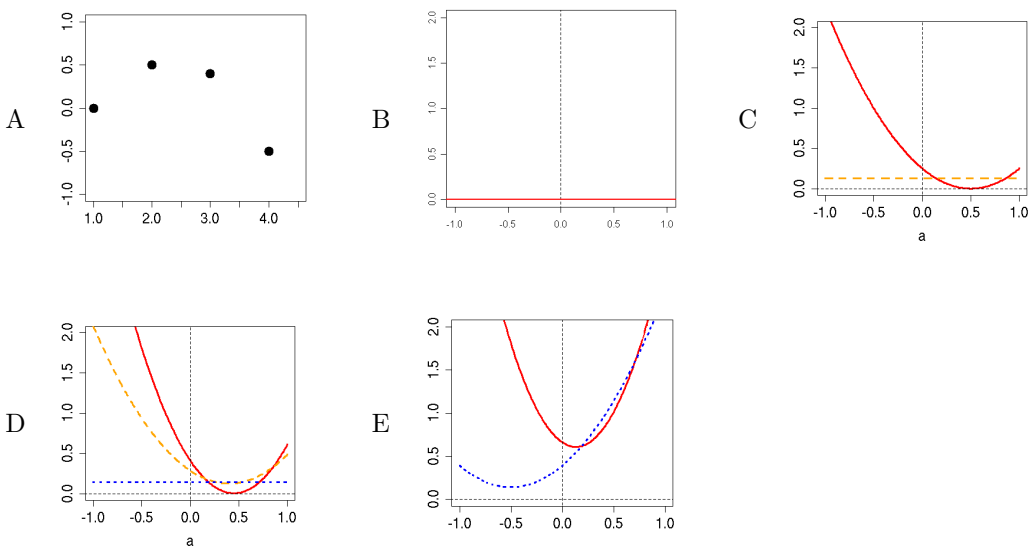


Figure 1: A: Four point signal, with Y_i as a function of i. B, C, D, E: Cost as functions of μ and the last change-point. Successive candidate functions stored by the pruned DPA are represented: $Cost_{2,1}$ in straight line, $Cost_{2,2}$ in dashed line and $Cost_{2,3}$ in dotted line. The four graphs correspond to the different steps described in subsection 3.2, B: Step 1, C: Step 2, bottom D: Step 3, E: Step 4.

update the cost function $Cost_{k,t}$, compute its minimum value and the roots of $Cost_{k,t}(\mu) = C_{k-1,i}$. All these steps are of complexity $O(1)$, at least for the quadratic loss. Then, the algorithm updates the set of winning intervals $Set_{k,t}$. One would think the number of intervals to update would increase too fast. But, in fact, it doesn't. Indeed, theorem 7.3, demonstrated in the section 7, shows that if γ is convex and if there are i candidates then the total number of intervals is at most $2i-1$ and thus at each step there are at most $O(i)$ intervals to update. Thus at point i the time complexity is of $O(i)$ and we retrieve an overall $K\sum_{i=1}^{n} O(i) = O(Kn^2)$ time complexity. Similarly, at point i the DPA stores $O(i)$ cost functions, $O(i)$ intervals and, as in the original DPA, the best segmentation in k up to point t ($C_{k,t}$). Thus we retrieve an overall $\Theta(Kn)$ space complexity.

In the case of the quadratic loss, the worst bound $O(Kn^2)$ is reached with the sequence $Y_i = i$. Indeed, in this case, all new candidate change-points are kept so that at step i there are at least $i-k+2$ candidates and intervals to update and we can retrieve a lower bound on the time complexity of $\Omega(Kn^2)$. However, most interestingly, for a constant signal with no noise, it is easy to see that the complexity is in exactly $\Theta(Kn)$ as the algorithm will keep only one candidate change-point at each step. Therefore we can expect that for more or less piecewise-constant signal the complexity will be closer to $O(Kn)$ than to $O(Kn^2)$.

5 Empirical complexity

For this section, we assessed, in the case of the quadratic loss, the efficiency of the pruned DPA to analyze both simulated and real data. The algorithm was implemented in C++ and was run on a

3.16GHz Intel(R) Xeon X5460.

5.1 Simulated data

Using a constant, sinusoid or rectangular signal we have simulated a series of sequences. For the sinusoid and rectangular waves we consider various amplitudes and frequencies. We considered a Gaussian noise of variance 1, a uniform noise of variance 1, a chi-square noise of variance 1 and a Cauchy noise. We compared the pruned DPA to the original one for rather small sample sizes, $n \leq 2^{14}$. We checked that the two algorithms retrieved the same result and compared their runtimes. We evaluated the runtime of the pruned DPA to process signals of a million points. Finally, we computed at each step of the pruned DPA the maximum number of intervals stored. For all these tests we set $K = 50$.

Figure 2-A shows that the pruned DPA was clearly faster than the original DPA for a constant signals with a normal noise as soon as $n = 4000$. Similar results were obtained for other simulated signals. For million-point signals, the pruned DPA had a runtime around 80-250 seconds depending on the nature of the signal and the type of noise. Figures 3-A and 3-B show, for constant and sine wave signals, the limited number of intervals stored by the pruned DPA and thus demonstrates the efficiency of pruning candidate change-points. For all simulated signals, the maximum number of intervals stored by the pruned DPA was 87 instead of a theoretical maximum of $2 \times 10^6 - 1$ (see section 4).

5.2 Real data

We use the publicly available $GSE17359$ project from GEO (http://www.ncbi.nlm.nih.gov/geo/). This data set is made of 18 SNP (Single Nucleotide Polymorphism) array experiments. SNP arrays enable the study of DNA copy number gains and losses along the genome. For this kind of data a multiple change-point detection procedure based on the quadratic loss is often used [3]. For each SNP array experiment, there are two signals of almost a million points each (SNP and CNV: Copy Number Variant). These two signals correspond to 24 chromosomes that were analyzed separately and together with $K = 50$.

As shown on Figure 2-B, the $18 \times 2 \times 24$ profiles of up to 8.10^4 points could be analyzed in less than 20 seconds by the DPA. The maximum runtimes to analyze the million-point CNV profiles and SNP profiles were respectively 113 and 109. The maximum number of intervals stored by the algorithm never exceeded 50 and is summarized in Figure 3-C for the CNV profiles.

Given the nature of these results on both synthetic and real data, we believe the pruned DPA we presented performs an efficient pruning of the candidate change-points and can actually be used to recover the best change-points w.r.t. the quadratic loss of large datasets.

6 Conclusion

Using the quadratic loss for off-line multiple change-point estimation is something various fields of research have in common, from econometrics [1] to molecular biology [3]. It has been extensively studied from a theoretical point of view, in both an asymptotic [15] and a non-asymptotic [16] setting. However retrieving the best set of K-1 change points w.r.t. quadratic loss remained a major computation bottleneck due to the very large number of possible segmentations: $\binom{n-1}{K-1}$. Until now, the best algorithm in use was a DPA with a quadratic complexity of $O(Kn^2)$, therefore hard to use for large

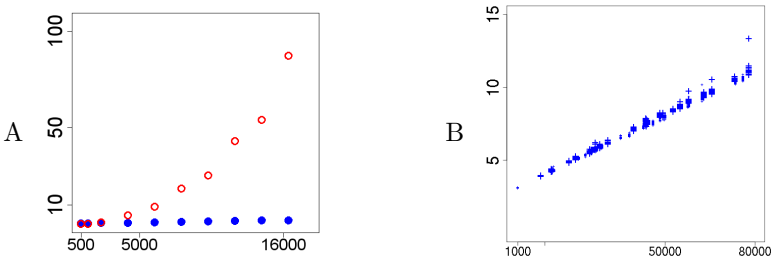


Figure 2: A: Mean runtime in seconds of the original (∘) and pruned DPA (•) for simulated constant signal with a normal noise for sequences up to 2^{14} points, B: Runtime in seconds of the pruned DPA for the $18 \times 2 \times 24$ profiles of the $GSE17359$ dataset.

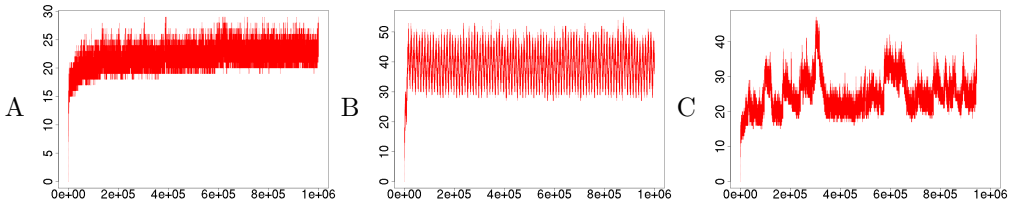


Figure 3: Maximum number of intervals stored by the pruned DPA at each point of the signal for $k = 2$. A: For 100 simulated constant signals of 10^6 points with a normal noise of variance 1. B: for 100 simulated sine wave signals of 10^6 points with a normal noise of variance 1. C: for 18 CNV profiles of almost 10^6 points.

datasets. Faster methods were proposed ([9], [10] and [11]) to cope with large datasets but they do not recover the best solution. In this paper we have presented a pruned DPA that recovers the best change-points w.r.t. the quadratic loss. We have proved that it is at least as efficient as the original DPA. Moreover, we have shown empirically, both with simulated and real datasets, this algorithm's ability to process large datasets of up to a million points in only a few seconds or a few minutes. These results lead us to think that the pruned DPA we presented overcomes the quadratic complexity bottleneck, and provides an efficient way to recover the best change-points w.r.t. the quadratic loss for large piecewise-constant signals.

7 Bound on the number of intervals

In this section, we study a special class of function that we denote $\mathcal{B}$ and demonstrate theorem 7.3. A direct application of this theorem shows that if there are t candidate change-points, then there are at most $2t-1$ intervals. We used this theorem in section 4 to prove the worst case complexity of the pruned DPA.

7.1 $\mathcal{B}$ functions

Definition 7.1.1 *Let $\mathcal{B}_n$ denote the set of all functions $B : \mathbb{R} \to \mathbb{R}$ such that*

$$\forall\ \mu\ \in\ \mathbb{R},\ B(\mu) = min_{\{t\ \in\ [\![1,n]\!]\}}\ \{u_{B,t} + \sum_{j=t+1}^{n+1} f_{B,j}(\mu)\}$$

where all $u_{B,t}$ are real numbers and all $f_{B,j}$ are convex functions of μ. Note that $\mathcal{B}_n \subset \mathcal{B}_{n+1}$. Let $\mathcal{B} = \bigcup \mathcal{B}_n$.

Definition 7.1.2 *For any B in $\mathcal{B}_n$ and A a subset of $[\![1,n]\!]$ we define the function B_A as*

$$B_A(\mu) = min_{\{t\ \in\ A\}}\ \{u_{B,t} + \sum_{j=t+1}^{n+1} f_{B,j}(\mu)\}$$

Proposition 7.1 $B_A \in \mathcal{B}_{card(A)}$

It is easily shown for $A = [\![1,n]\!] \setminus \{i\}$ with i in $[\![1,n]\!]$ and thus by induction it is true for any A.

Definition 7.1.3 *The rank of a function $B \in \mathcal{B}$ is $\mathcal{R}(B) = min\{n\ \in \mathbb{N}^*\ | B \in \mathcal{B}_n\}$*

7.2 Decomposition in intervals and order of $\mathcal{B}$

Definition 7.2.1 *Let $\mathcal{I}$ be a partition of $\mathbb{R}$ in a finite set of intervals $\mathcal{I} = \{I_j\}_{\{j \in [\![1,k]\!]\}}$. $\mathcal{I}$ is a k-decomposition of a function $B\ \in\ \mathcal{B}$ if*

$$\forall\ I_j,\ \exists\ i,\ \forall x\ \in\ I_j \qquad B_i(x) = B(x)$$

The set of all $\mathcal{B}$ functions with a k-decomposition is denoted $\mathcal{B}^k$. Similarly the set of all $\mathcal{B}_n$ functions with a k-decomposition is denoted $\mathcal{B}_n^k$.

Proposition 7.2 *If B is $\mathcal{B}$ then there exists a k-decomposition of B.*

Definition 7.2.2 *The order $O(B)$ of a $\mathcal{B}$ function is $min\{k \in \mathbb{N}^* | B\ \in\ \mathcal{B}^k\ \}$*

Theorem 7.3 *For all $B \in \mathcal{B}$, we have $O(B) \leq 2 \times \mathcal{R}(B) - 1$.*

Proof We demonstrate this theorem by induction. It is true if $O(B) = 1$. Assume it is true for any B with $O(B) = n$. Let $B \in \mathcal{B}$ with $O(B) = n + 1$. We have:

$$\begin{aligned} \forall \mu \in \mathbb{R}, \qquad B(\mu) &= min\{B_{[\![1,n]\!]}(\mu),\ B_{\{n+1\}}(\mu)\} \\ B(\mu) &= min\{C(\mu),\ u_{B,n+1}\} + f_{B,n+2}(\mu), \end{aligned}$$

where $C \in \mathcal{B}_n$:

$$C(\mu) = \{u_{B,t} + \sum_{j=t+1}^{n+1} f_{B,j}(\mu)\}$$

Let $\mathcal{I} = \bigcup_{j \in [\![1,k]\!]} I_j$ be the smallest set of intervals such that:

$$\forall\ I_j\ \in\ \mathcal{I},\ \forall\ \mu\ \in\ I_j \qquad u_{B,n+1} >\ C(\mu)$$

Let A_k be the subset of $[\![1, n]\!]$ defined as $\{i \quad | \exists\ x\ \in\ I_k \quad C_{\{i\}}(x) < u_{B,n+1}\}$. As $\mathcal{R}(B) = n + 1$ and as for all i in $[\![1, n]\!]$ $C_{\{i\}}$ is convex, there exists a unique j such that $i \in A_j$ and therefore $\sum_{j=1}^{k} card(A_j) = n$.

In each interval I_k, we have $C(\mu) = C_{A_k}(\mu)$. By induction, $O(C_{A_k}) \leq 2 \times card(A_k) - 1$.

Overall, for any B with $O(B) = n + 1$, we have:

$$O(B) \leq \quad \sum_{j=1}^{k} O(C_{A_k}) + (k+1) \leq \quad 2\sum_{j=1}^{k} card(A_k) + 1 \leq 2n + 1 \leq 2\mathcal{R}(B) + 1 \quad \blacksquare$$

References

[1] Jushan Bai and Pierre Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, January 1998. ArticleType: primary_article / Full publication date: Jan., 1998 / Copyright Â© 1998 The Econometric Society.

[2] O. Gillet, S. Essid, and G. Richard. On the correlation of automatic audio and visual segmentation of music videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2007.

[3] Franck Picard, Stephane Robin, Marc Lavielle, Christian Vaisse, and Jean-Jacques Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6(1):27, 2005.

[4] Mich Ele Basseville, MichÃ¨le Basseville, and Igor V Nikiforov. Detection of abrupt changes: Theory and application. 1993.

[5] Richard Bellman. On the approximation of curves by line segments using dynamic programming. *Commun. ACM*, 4(6):284, 1961.

[6] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, August 2005.

[7] Jushan Bai and Pierre Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.

[8] Yann Guédon. Exploring the segmentation space for the assessment of multiple change-point models. http://hal.inria.fr/inria-00311634/fr/, 2008.

[9] Servane Gey and Emile Lebarbier. Using CART to detect multiple change points in the mean for large sample. http://hal.archives-ouvertes.fr/hal-00327146_v1/, February 2008.

[10] Robert Tibshirani and Pei Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostat*, page kxm013, May 2007.

[11] Zaid Harchaoui and Céline LEVY-LEDUC. Catching change-points with lasso. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 617–624. MIT Press, Cambridge, MA, 2008.

[12] Erez Ben-Yaacov and Yonina C. Eldar. A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*, 24(16):i139–145, August 2008.

[13] Nicolas Chopin. Dynamic detection of change points in long time series. *Annals of the Institute of Statistical Mathematics*, 59(2):349–366, 2007.

[14] Paul Fearnhead and Zhen Liu. On-line inference for multiple change points problems. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY B*, 69:589—605, 2007.

[15] Marc Lavielle and Eric Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, 21(1):33–59, 2000.

[16] P. Massart. A non asymptotic theory for model selection. In *European Mathematical Society, 2005.*, pages 309–323, 2005.

Chapter 6

Analysis of the Curie-Servier Genomic dataset

In this chapter, I give a review of genomic alterations in breast tumors. Then I present the first results of my analysis of the genomic dataset gathered by the Curie-Servier alliance.

6.1 Genomic alterations in breast cancers and in TNBC

The genomic alterations of breast cancer cells as well as the clinical implications of recurrently alterated regions have been studied for a long time (Gray et al., 1994). Several recurrently alterated regions were identified using cytogenetics, Fluorescence In Situ Hybridisation (FISH) and chromosome-based CGH. Two well-known features are the gain of 8q24 (including the *MYC* gene) and the gain of 17q12 (including the *ERBB2* gene). It was observed that some of these alterations often occur together (see Bärlund et al. (1997); Courjal et al. (1997)). Courjal et al. (1997) analyzed 15 different positions along the genome in 1875 breast tumors. This very high number of samples undoubtedly gave them remarkable statistical power to detect recurrent alterations at those positions and to analyze correlations between those 15 positions or between those positions and other clinical parameters.

The development of CGH arrays allowed the analysis of individual breast tumor genomic rear-

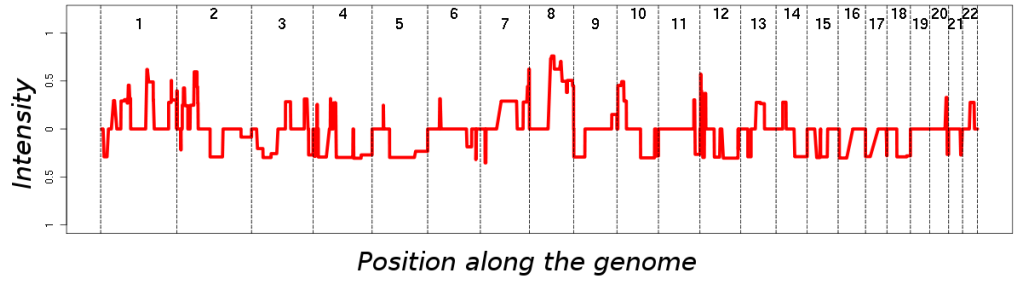
rangements with a much higher resolution (Pinkel et al., 1998). CGH arrays and SNP arrays contain from 10 thousand up to 1 million probes. Many groups have studied the complexity of breast cancer DNA copy number and LOH rearrangements using this technology (see for example Loo et al. (2004); Wang et al. (2004); Naylor et al. (2005); Bergamaschi et al. (2006); Chin et al. (2006); Stange et al. (2006); Hicks et al. (2006); Fridlyand et al. (2006) and more recently Chin et al. (2007); Haverty et al. (2008); Han et al. (2008); Loo et al. (2008); Argos et al. (2008); Andre et al. (2009); Jönsson et al. (2010); Staaf et al. (2010)). These analyses have successively revealed ever more refined lists of genomic rearrangements. They also revealed the existence of different types of genomic rearrangement patterns known as "sawtooth", "firestorm" and "simple".

Sawtooth Sawtooth profiles are characterized by a very complex pattern with many small rearrangements of low amplitude. This pattern is very common in TNBC. This type of profile recurrently gains 10p and losses 3p, 4p and 4q, 5q, 14q and 15q. This pattern is illustrated in the genomic Curie-Servier dataset on Figure 6.1 A.

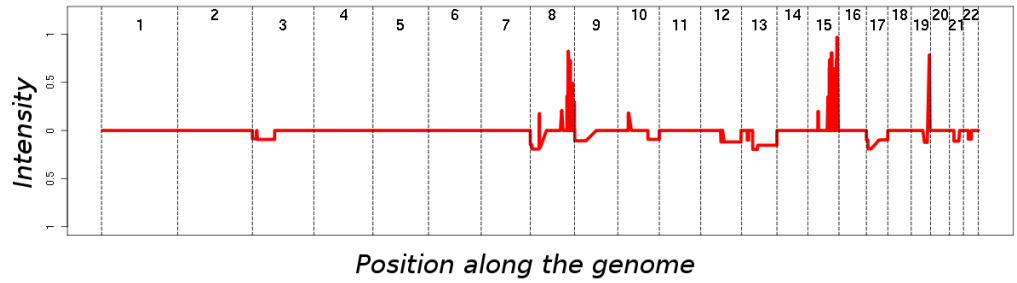
Firestorm Firestorm patterns harbor focal DNA copy number amplifications of high amplitudes. It is typical of both Luminal B and ER- / HER2+ tumors. Among the recurrent focal amplification sites are: 8q24 (that includes *MYC*), 17q12 (*ERBB2*), 8p12 (*FGFR1*), 11q13 (*CCND1*). This pattern is illustrated in the genomic Curie-Servier dataset on Figure 6.1 B.

Simple Simple profiles harbor very few rearrangements except for a recurrent gain of 1q and 16p and loss of 16q. This type of profile is often found in Luminal A tumors and a typical example is shown on Figure 6.1 C.

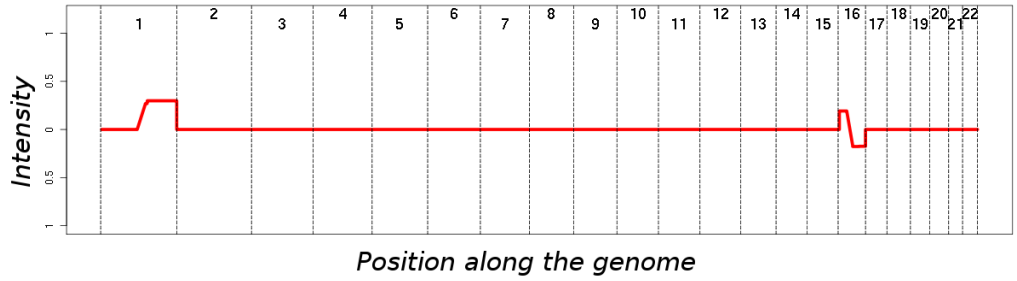
To conclude this quick review of genomic rearrangements in breast tumors (for more information, see the review by Kwei et al. (2010)), I will highlight some of the insight brought by high-throughput sequencing and more specifically by the paired-end sequencing strategy (Volik et al., 2003). Briefly, in this technique, pieces of the tumoral genome are retrieved and the ends of these pieces are sequenced. Using paired-end sequencing, it is possible to measure DNA copy number and assess rearrangements. The DNA copy number is measured by counting the number of reads in each region of the genome.



A



B



C

Figure 6.1: Breast genomic patterns. Three examples using the Curie-Servier genomic dataset. (A) A sawtooth profile of a TNBC sample with many low-amplitude alterations. (B) A firestorm profile of a Luminal B tumor with sharp high-amplitude alterations. (C) A simple profile of a Luminal A tumor with very few alterations except the gain of 1q, 16p and loss of 16q. The x-axis is the position along the genome with the 22 autosomes separated by dashed vertical lines. The y-axis is the intensity. The segmented profiles are represented as red piecewise-constant lines.

Rearrangements are detected when two sequenced ends match conflicting regions of the reference genome (for example two different chromosomes). This technique has revealed yet another level of complexity (Campbell et al., 2008; Stephens et al., 2009). It showed that rearrangements are much more frequent than what was previously thought. Moreover, various patterns of rearrangements have been identified, more or less matching the genomic patterns identified with CGH. In the case of TNBC, it also revealed the existence of relatively small rearrangements ranging from a few kilobases to a few megabases that could not be detected using CGH arrays.

While many TNBC samples present a sawtooth genomic profile, the mechanisms underlying these numerous rearrangements are not yet understood. The number of these rearrangements nonetheless hints at a problem in terms of DNA repair mechanisms. As previously mentioned, sporadic TNBC are similar to BRCA1-associated tumors in terms of their histology (Foulkes et al., 2003). Interestingly, BRCA1-associated tumors also have sawtooth genomic patterns (Fridlyand et al., 2006) and similarly to TNBC they lack markers of an inactive X chromosome (Richardson et al., 2006). All this lead to the hypothesis that BRCA1 is defective in sporadic TNBC (Turner et al., 2006). A defective BRCA1 induces a defect in homologous recombination (HR) that favors error-prone repairs of DNA and in the end leads to chromosome rearrangements. This defect in BRCA1 would thus explain the very complex DNA profile patterns of both sporadic TNBC and BRCA1-associated tumors. Yet, other phenomena can explain the similarity between TNBC and BRCA1-related tumors. For example, the very frequent mutation of TP53 (Manié et al., 2009) in both types of tumor or the loss of PTEN (Marty et al., 2008; Saal et al., 2008). Indeed, both TP53 and PTEN have been shown to modulate defects in HR (see Ralhan et al. (2007); Mendes-Pereira et al. (2009); Kwei et al. (2010)).

6.2 Analysis of the genomic Curie-Servier dataset

In this subsection, I present the first results of my analysis of the genomic Curie-Servier dataset. The experimental design was carefully planned. The goal of experimental design is to ensure that the way the experiment is conducted will actually enable us to answer our question of interest. Here, the design was simple to do and thus I will not describe it in detail (see Chapter 8 for a short introduction to

	TNBC	ER-/ HER2+	Luminal A	Luminal B	TNBC cell-lines	Normal Tissue
ER/PR status	-	-	+	+	-	
overexpresion of HER2	-	+	-	+ / -	-	
CK14 or CK5/6 or EGFR	+	-	-	-		
Grade	III	III	I	III		
Number	43	35	40	44	14	17

Figure 6.2: Summary of the samples in the genomic Curie-Servier dataset and their histological and immunohistochemical characterization.

experimental design and the design of the transcriptomic experiment). Briefly, a total of 193 samples were processed in 5 batches and the different sample types were balanced across these five batches. The characteristics of the samples processed in this experiment are summarized in Figure 6.2.

Affymetrix SNP 6.0 arrays contain almost 1 million Copy Number Variant (CNV) probes and 1 million Single Nucleotide Polymorphism (SNP) probes. CNV probes directly measure the DNA copy number. SNP probes make is possible to measure the DNA copy number as the sum of allele A and allele B. It is also possible to measure the allelic difference of Loss of Heterozygosity (LOH). In the following, I will focus on the analysis of DNA copy number profiles. Overall, for each sample, we have twice one million measurements of the DNA copy number. These measurements are scattered across the 22 autosomes and the 2 sexual chromosomes. The measurements of the 2 sexual chromosomes are quite different from the autosomes and they were removed from this analysis.

To study these 193 samples, I integrated the pruned Dynamic Programming Algorithm (DPA) in the CGHseg methodology. I analyzed CNV and SNP profiles separately. The DNA copy number

profiles for each of the 22 autosomes were processed by the pruned DPA. For each chromosome, the pruned DPA recovered the best segmentations (with respect to the Mean Square Error, MSE) with 0, 1, 2, ... 99, 100 breakpoints. We assume that 100 breakpoints is a gross over-estimation of the likely number of breakpoints on a chromosome so this method should enable us to be quite confident that no breakpoints are being left out. Once we have the 100 best segmentations for each of the 22 autosomes, we can recover with a dedicated algorithm the best segmentation across all chromosomes in 22, 23, ..., 2200 segments. At the genome level, there are at least 22 segments, corresponding to the 22 autosomes. Then, we need to select the number of breakpoints for the whole genome of a given sample. This number is selected as in the CGHseg methodology and in this case the maximum number of breakpoints per sample (K_{max}) is 1000. The influence of this parameter has been assessed (see the next paragraph). The largest chromosome profiles have 200 000 points and the total runtime for one sample is approximatively 3 minutes. In the end, we retrieved 193 segmented profiles. A simple look at those segmented profiles is enough to distinguish between sawtooth, firestorm and simple profiles (as shown on Figure 6.1).

The pruned DPA is fast and recovers the best segmentation with respect to the MSE. The only uncertainty lies in the selection of the number of breakpoints. The only two parameters of the method are the maximum number of breakpoints per chromosome and the maximum number of breakpoints per sample. They have a simple biological interpretation and are quite easy to calibrate. It is possible to assess the influence of the maximum number of breakpoints (K_{max}) for a very limited additional computational cost (a few seconds). For all possible K_{max} between 50 and 2200, I launched the CGH breakpoint selection procedure and stored the selected number of breakpoints. Overall the selection is stable and the selected number of breakpoints is constant for large ranges of K_{max} values (data not shown).

The next step of the analysis is the calling step, which consists in identifying gained, lost and normal regions of the genome. But here we have the segmented DNA copy number profiles of 17 normal breast tissues. Based on their profiles, it is possible to devise a data-driven threshold for gains and losses. For example, we can choose a gain threshold so that the probability of a normal genome

to be above is smaller than $p = 5\%$ (by taking the 95% percentile). In the end, we would like to retrieve the percentage of gains in each tumor subtype for each position of the genome. To obtain conservative estimates, I considered a probability of $p = 2.5\%/43$ for both gains and losses. 43 is the number of TNBC samples, which form the largest group of the Curie-Servier genomic dataset. The probability of $p = 2.5\%/43$ ensures that for any subtype and at any considered position of the genome the probability of attributing more than one gain or loss by mistake is smaller than 5%. I used this threshold across all samples. It is important to realise that across the whole genome we are bound to make errors. I did not control the probability of attributing more than one gain or loss by mistake across the whole genome because it would result in a too stringent threshold.

The results are summarized per subtype in Figure 6.3. We identify a number of elements that were previously known:

- for TNBC, the frequent gain of 10p and loss of 3p, 4p and 4q, 5q, 14q and 15q;
- for both ER- / HER2+ and Luminal B tumors, the frequent gain of the *ERBB2* gene (17q12);
- for Luminal A (simple profiles), the recurrent gain of 1p and 16p and the loss of 16q.

Peaks in these graphs also identify some interesting positions along the genome. For example, the amplification of the *HER2* gene is detected in all ER- / HER2+ tumor samples and in 80% of Luminal B tumor samples. This is not surprising since 75% of our Luminal B tumor samples were selected as ER+ and HER2+. This peak is relatively sharp and delimits a small region of the genome. The genomic view of this region in the Ensembl database (http://www.ensembl.org/index.html) is depicted in Figure 6.4 A. Similarly, there is a very sharp peak pinpointing a 86% recurrent amplification around *MYC* in TNBC (see a genomic view on Figure 6.4 B) and there is a wider peak on chromosome 10 corresponding to a 51% recurrent loss around *PTEN* in TNBC (see a genomic view on Figure 6.4 C). The importance of such recurrent alterations can be confirmed by looking at individual profiles. Continuing on our *MYC* and *PTEN* examples, it is possible to identify alterations precisely on *MYC* and alterations centered on *PTEN* for some TNBC samples (see Figure 6.5)

More generally, it is possible to identify all such peaks of recurrent amplifications or deletions. For example, I considered all such peaks with a percentage of alteration larger than 50%. There are not

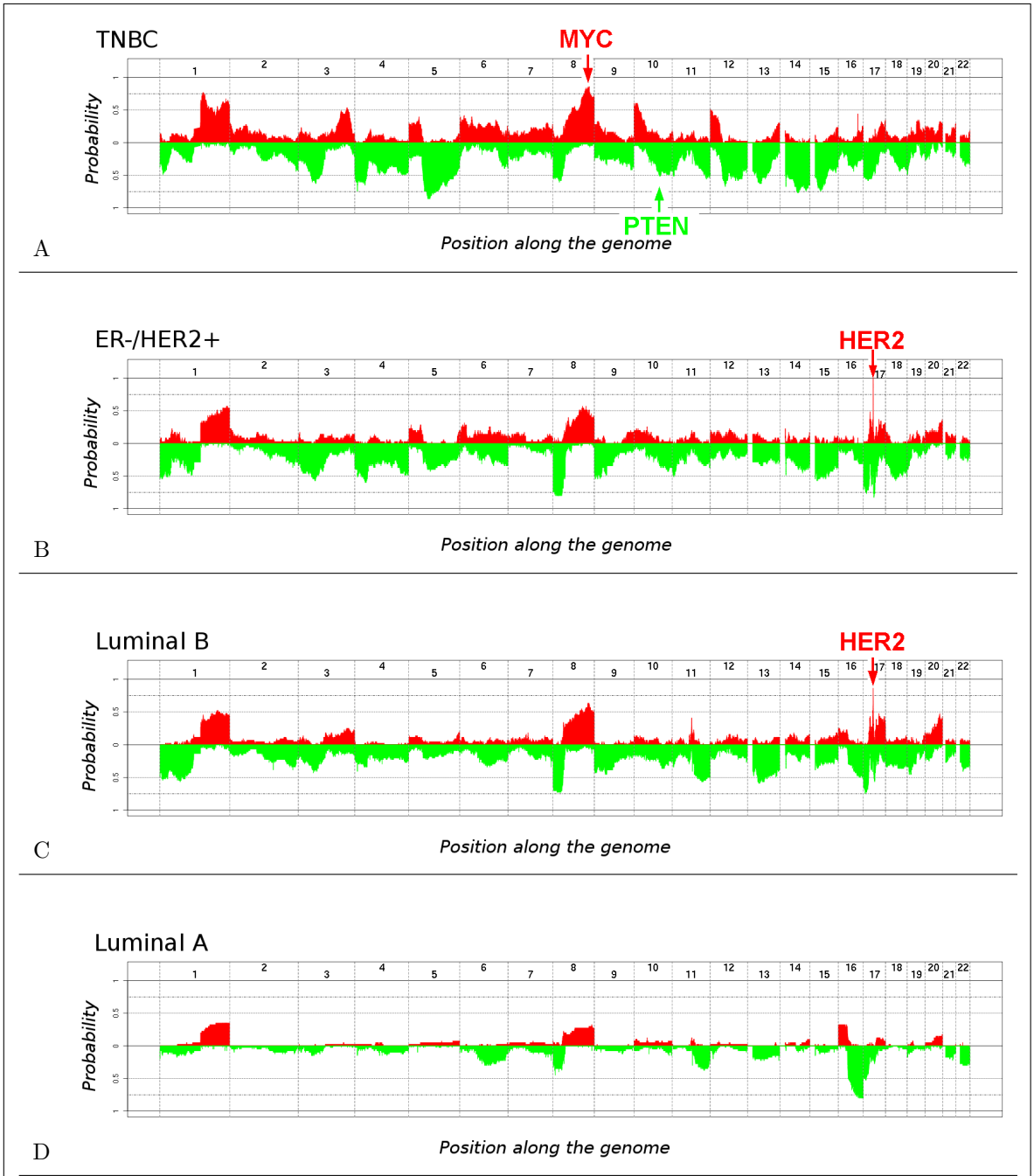


Figure 6.3: Proportions of detected gains (red) and losses (green) across the 22 autosomes per subtype of breast tumors. (A) TNBC. (B) ER- / HER2+ tumors. (C) Luminal B tumors. (D) Luminal A tumors. The x-axis is the position along the genome with the 22 autosomes separated by dashed vertical lines. The y-axis is the proportion of gains in red and losses in green.

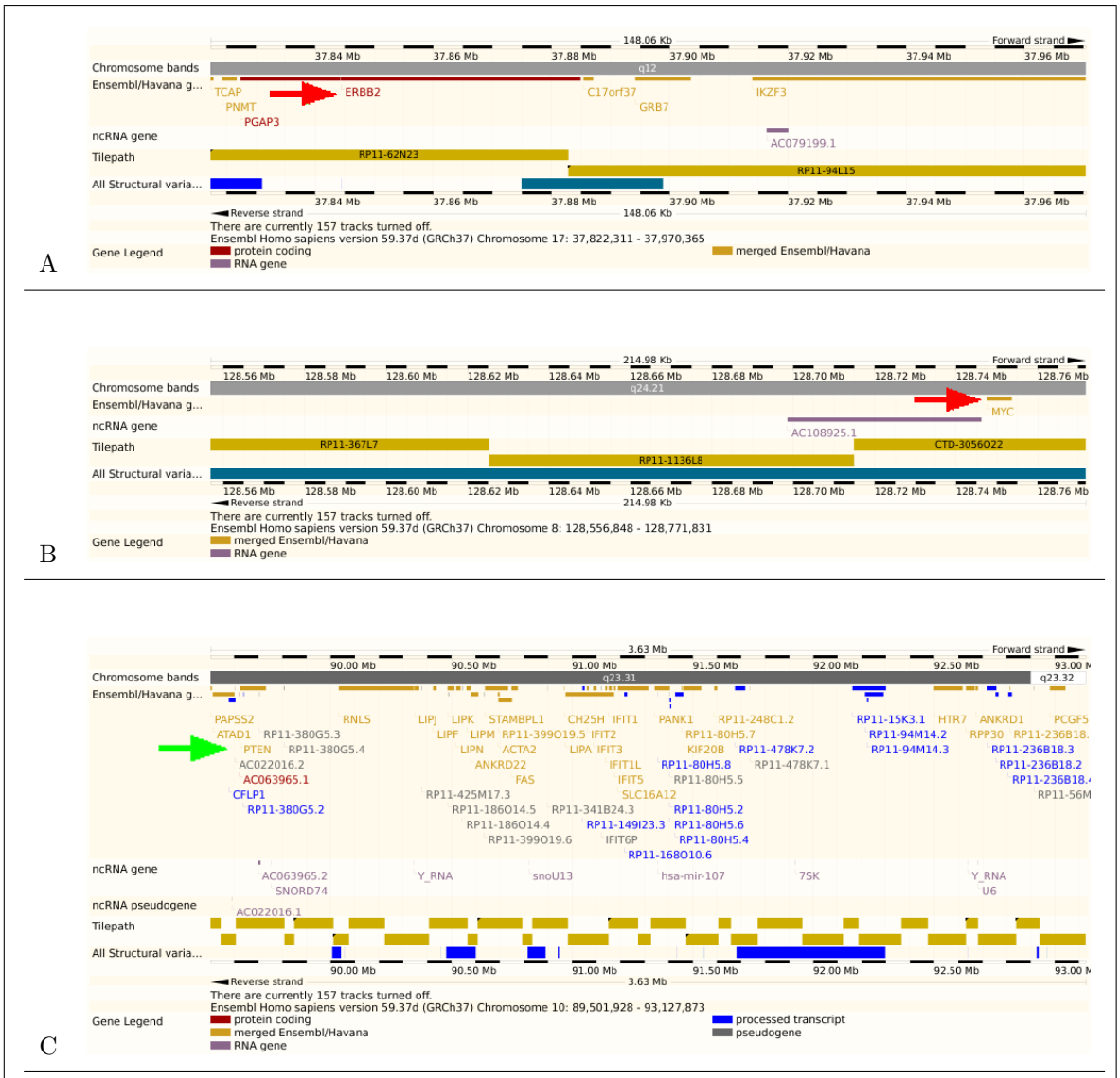


Figure 6.4: Genomic view of three frequently aberrant regions. These views were retrieved using the Ensembl database (http://www.ensembl.org/index.html). (A) Region amplified in 100% of ER- / HER2+ tumors and that includes the *ERBB2* gene. (B) Region amplified in 86% of TNBC and that includes the *MYC* gene. (C) Region lost in 51% of TNBC and that includes the *PTEN* gene. On these views, one can see the chromosome bands and the position of known genes.

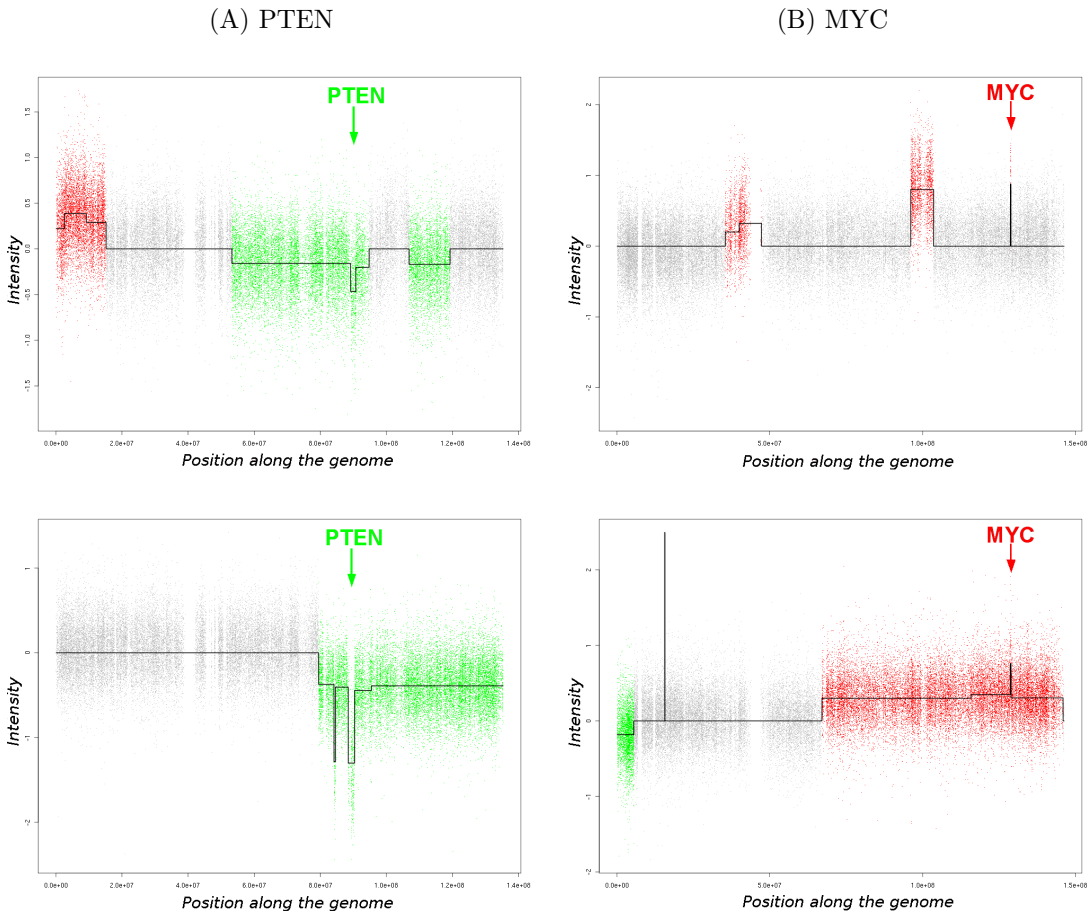


Figure 6.5: Example of focal amplifications of *PTEN* and *MYC* in TNBC. (A) Chromosome 10 profiles of two TNBC samples. (B) Chromosome 8 profiles of two TNBC samples. The x-axis is the position along the genome. The measured DNA copy numbers are represented by dots. Gained regions are in red. Lost regions are in green. Normal regions are in gray. The segmented profiles are represented as black piecewise-constant lines.

Number of SNPs	Chr	Proba	Position start	Position end
344	1	0,77	153540320	155318308
15	1	0,77	155841517	155935244
254	10	0,6	104427	1030987
1246	10	0,6	5800811	8188595
532	10	0,6	10754532	11840993
298	12	0,51	850288	1955552
848	3	0,53	169415294	172153170
35	8	0,86	128556848	128771831
453	8	0,86	124616796	125745994

A

Number of SNPs	Chr	Proba	Position start	Position end
1362	10	0,51	89501928	93127873
768	11	0,56	120775110	122750044
101	12	0,67	54926329	55160710
192	13	0,63	42843054	43322006
176	13	0,63	44648656	44998330
76	13	0,63	45933243	46225449
42	13	0,63	46440647	46520873
34	13	0,63	46949335	47115608
358	13	0,63	49079503	50506066
581	14	0,77	62802710	64654431
370	14	0,77	65784860	67483521
76	15	0,74	40251741	40529130
612	17	0,58	15501133	18917513
1053	3	0,63	60742456	63026375
2268	5	0,86	66997601	75616799
112	8	0,6	22884231	23249947

B

Figure 6.6: Peaks of losses and gains in TNBC with more than 50% alterations. (A) Gained regions. (B) Lost regions.

many of them and they can easily be recovered by hand. A short list of these peaks and their positions along the genome is shown on Figure 6.6.

To conclude, this analysis is undoubtedly simple and straightforward, but it relies on few and simple hypotheses and there are very few parameters: only the maximum number of breakpoints considered for the segmentation space and the threshold for detecting gains and losses. This analysis does not take into account the LOH information, even though this information should be very useful in the case of TNBC. T. Popova (Ph.D.) recently proposed a methodology named GAP (Genomic Alteration Print) for the joint analysis of DNA copy number profiles and LOH profiles of complex genomes like

those of TNBC (Popova et al. (2009), the paper is provided in Appendix A.2). I worked with her on this methodology. Briefly, based on the segmented DNA copy number and LOH profiles (provided by the pruned DPA), GAP recovers the copy number vs. LOH pattern of the tumor. Using this pattern allows us to distinguish between tetraploid and diploid tumors and to attribute to each segment a precise DNA copy number and an allelic count. In other words, for each segment, GAP predicts an integer value for allele A and for allele B (SNP-A and SNP-B). I must admit that this method is at a heuristic stage and there is no clear model yet, in a biostatistical sense. Indeed, from a biostatistical point of view, it is difficult to summarize and ponder the different hypotheses, to precisely understand the meaning of the different parameters involved in this methodology and to assess the predictive power of GAP even in very simple cases such as simulations. Nevertheless, GAP empirically shows very good results. Using it, I hope to identify more precisely regions of the genome involved in the development of TNBC and this could lead to the identification of new targets.

Part III

Transcriptomic Analysis

Chapter 7

Introduction

Ten years ago, with the first microarray-based classification of breast tumors (Perou et al., 2000), it was hoped that transcriptionnal analysis of breast cancer and the possibility of looking at many genes at the same time would rapidly change our understanding of the disease and make conventional diagnostic techniques, such as histopathology, obsolete (Aparicio et al., 2000). Today, the scope of microarray gene expression profiling has been somewhat reduced. However, it has clearly contributed to our general understanding of breast cancer pathology (see Weigelt et al. (2010a) for a review). Briefly, much profiling of breast tumors has been done with more than 400 gene expression datasets of breast cancer studies publicly available on the ArrayExpress database (http://www.ebi.ac.uk/microarray-as/ae/).

These different studies can be subdivided into three main categories:

Class discovery aims at unraveling the heterogeneity of breast tumors at the molecular level. It tries to decompose tumors into homogeneous subgroups that are hopefully biologically and/or clinically relevant.

Class comparison is a supervised approach that aims at deciphering the key molecular differences between two or more previously identified subgroups of tumors.

Class prediction is also a supervised approach that looks for a "gene signature" or "classifier" able to correctly predict the class membership of a new sample based on previously identified subgroups

of tumors.

All these studies have undeniably brought new insights into the biology of breast tumors, but the validity of these analyses has been questioned (Ein-Dor et al., 2005; Michiels et al., 2005; Ioannidis et al., 2009). In particular, it seems that it is difficult to reproduce the results of these analyses. This highlights the importance of assessing the quality of experiments and explaining the various bioinformatical and biostatistical steps.

From chapter 8 to 11, I will describe some of the biological and/or clinical questions that I have addressed using our Curie-Servier dataset of gene expression profiling on breast cancer. The analyses I made were a multi-step process, which can be viewed as a pipeline. The validity and importance of each step is rarely questioned. The choice of a given methodology is often a subtle decision and there rarely is a definite answer. In addition to time and money constraints, there are at least three different decision-making criteria that should not be confused: biological and clinical ones, statistical ones and computational ones. I will thus try to bring forward these different aspects and explain as much as possible the methodologies I used and the necessary concessions I made.

This part also illustrates my day-to-day life as a biostatistician working with biologists and medical doctors of the Institut Curie. It was an important part of my work during my PhD candidacy (two to three days a week). The analysis of high-throughput biological data is a long process that requires some knowledge of biology, computer science and statistics. To understand all these aspects as much as possible, I worked in three laboratories: a biology lab, a bioinformatics lab and a statistics lab.

It was a very enriching experience.

This work is presented in the following order. I will first give an overview of the pre-processing step of the analysis, namely the experimental design, and then the normalization and the exploratory analysis of the data. Second, I will describe some of my work in collaboration with Anne Vincent Salomon (MD/Ph.D., Institut Curie) to compare the immunohistochemistry-based classification of breast tumors and the gene expression profiling-based classification. Third, I will describe the gene-by-gene and pathway-by-pathway comparisons of the different breast tumor subtypes. These comparisons have led to collaborations with the group of Philippe Chavrier (Ph.D., Institut Curie) and the group

of Fatima Mechta Grigoriou (Ph.D., Institut Curie).

Chapter 8

Experimental Design

In this chapter, I will briefly describe how I constructed the experimental design of the transcriptomic experiment of the Curie-Servier project. First, I will give a very short introduction to experimental design. For a more thorough "non-mathematical" introduction to experimental design, see Cox (1992). Its mathematical counterpart is Cochran and Cox (1992).

8.1 A small introduction to experimental design

The goal of experimental design is to ensure that the way the experiment is conducted will actually enable us to answer our question of interest. The main idea behind experimental design is that it is possible to assess the relative power of two different designs (with a model but without any data) to answer a well-defined question. This might come as a surprise, but in fact it is quite natural: for example, in architecture, one does not need to build two houses to know which one looks better, one judges from the blueprints.

Importantly, with a carefully planned design you can improve your ability to detect some patterns of interest or the robustness to experimental problems. Reversely, if the design is not well planned it can lead to useless experiments, from which no information can be obtained. From a statistical perspective, there are two successive problems. First, can we estimate the parameters of interest from

the data? Next, if we can, how precisely can we estimate them? To further illustrate these two points, I will use two simple examples.

Estimability, pitfalls of experimental design: an example

In this subsection, I will describe a bad experimental design. Suppose I want to compare the growth of two plant species A and B. I plant a thousand seeds of A in field number 1 and a thousand seeds of B in field number 2. Ten days later, I measure the size of the plants in field 1. Because it is quite a long process, I barely manage to finish measuring all these plants in one day. On the next day, I go to field 2 and measure the size of the plants there. In the end, I compare the measurements of species A and species B. Since I have many measurements for each species, I get a very good estimate of the size of species A after 10 days in field 1 and of species B after 11 days in field 2. From a mathematical point of view, I could compute the difference between those two estimates. However, it does not make much sense to compare the size of two different plants that have grown in two different fields for a different period of time especially if you want to compare their growth.

I would agree that it is improbable to choose such a poor experimental design to compare species A and B. However, this example is very simple. In more complex cases, such pitfalls are not so easy to detect.

Improving the precision: an example

Imagine that you have one object of mass μ_A. To get an estimate of this mass, you weigh the object using scales. You measure a mass m_1. Due to some randomness in the measurement and the precision of the scales, m_1 is not exactly equal to μ_A and we have $\mu_A = m_1 + \varepsilon_1$. ε_1 is the error. It is a random variable. If there is no bias, the variance σ^2 of ε_1 is related to the precision of the scales. The smaller it is, the more precise are the scales.

Suppose you have three objects A, B and C of mass μ_A, μ_B and μ_C respectively. A simple way to estimate these three masses is to weigh each object one at a time (as illustrated in Figure 8.1 (top)). In the end, you get an estimate for each of the three masses. Each estimate has an error of variance

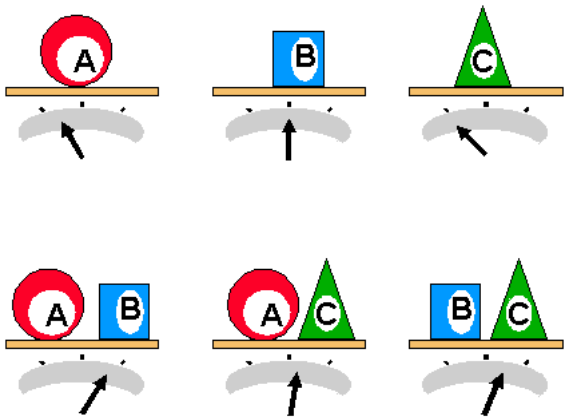


Figure 8.1: Weighing three objects A, B and C. (Top) First method: A, B and C are weighed one after the other. (Bottom) Second method: The three objects are weighed in pairs.

σ^2. It is an obvious experimental design.

However, if one assumes that masses are additive, i.e. that the mass of two objects is the sum of their mass and that the errors are independent, one can obtain better precision for the mass estimates. A simple way to do so is to weigh the objects in pairs as illustrated in Figure 8.1 (bottom). Retrieving an estimate of the three masses is less obvious. However, the simple calculations described below show that the variance is reduced to $\frac{3}{4}\sigma^2$. Thus, with exactly the same number of measurements, we get a smaller variance. However, this second experimental design is less robust. Indeed if any of the three measurements fails, it is impossible to get an estimate for any of the three masses. The first design is more robust because in such a situation, one still obtains 2 estimates out of the 3.

Quick proof If we assume that masses are additive and using the design described in Figure 8.1 (bottom), we get the following system of equations:

$$\begin{aligned} \mu_A + \mu_B &= m_1 + \varepsilon_1, \quad (L1) \\ \mu_A + \mu_C &= m_2 + \varepsilon_2, \quad (L2) \\ \mu_B + \mu_C &= m_3 + \varepsilon_3, \quad (L3) \end{aligned}$$

where each line corresponds to one measurement, m_1, m_2, m_3 are the three measurements and ε_1, ε_2, ε_3 are the random errors associated to these three measurements. Combining the three equations as follows $(L1) + (L2) - (L3)$ we get:

$$2\mu_A \quad = \quad m_1 + m_2 - m_3 + \varepsilon_1 + \varepsilon_2 - \varepsilon_3$$

Thus, we estimate μ_A as $\frac{1}{2}(m_1+m_2-m_3)$. If the errors are independent, the variance of this estimation is:

$$V[\frac{1}{2}(\varepsilon_1 + \varepsilon_2 - \varepsilon_3)] = \frac{1}{4}V(\varepsilon_1 + \varepsilon_2 + \varepsilon_3) = \frac{1}{4}[V(\varepsilon_1) + V(\varepsilon_2) + V(\varepsilon_3)] = \frac{3}{4}\sigma^2.$$

To conclude, although it is time consuming to design an experiment correctly, the design has a dramatic impact on the rest of the analysis and it should be done carefully.

8.2 Design of the transcriptomic experiment

The main question of the Curie-Servier project was to identify therapeutic targets for TNBC. To do so, one usually looks for genes that are overexpressed in TNBC compared to normal tissue. More generally, we were interested in comparing any two subtypes of samples. Thus, we looked for a design that enabled an efficient comparison between TNBC and normal tissue but also to a lesser degree between any two subtypes.

Choosing the number of replicates

Prior to my arrival in the project, two pilot studies had already been carried out. The first pilot study was carried out on 13 TNBC and 11 ER- / HER2+ tumors. The second study was carried out on 24 TNBC and 16 ER- / HER2+ tumors. It was decided to make a last transcriptomic experiment including Luminal tumors, TNBC cell-lines and normal samples. We decided to include several TNBC and ER- / HER2+ replicates of the first two studies in this last experiment to easily compare and aggregate the three studies (estimability problem see subsection 8.1). Due to various constraints, only 24 TNBC and ER- / HER2+ samples could be reprocessed in the last experiment. At that time, we had very little information on the number of Luminal, ER- / HER2+, TNBC and normal samples

that we would be able to obtain. We expected around 40 Luminal A, 40 Luminal B, 15 normal and 15 TNBC cell-lines. To choose how many of the 24 replicates should be TNBC, I proposed to examine different possible designs.

The statistical model

In order to aggregate the three studies, I considered a simple mixed linear model. In statistics, using a mixed linear model is a common way to take into account the existence of replicates. While building such a model is a tedious process for those not used to it, it is nevertheless necessary to have a well-defined model to compare different possible experimental designs. Moreover, without a well-defined model, the hypotheses we need usually remain ill-defined or even undefined and it is thus very difficult to test them or grasp their importance.

The linear model takes into account 3 things:

Type The measured intensity depends on the sample types (t). Each sample type (TNBC, Luminal A, ...) has a different mean level. In the model the effect of being of type t is associated to a fixed parameter α_t. While this is certainly not true for every gene, it is a practical hypothesis and it is precisely the existence of such a sample type effect that we want to test.

Study The measured intensity is also explained by the study (s: first study, second study and last study). Each study is associated to a different mean level. In the model, the effect of being in study s is associated to a fixed parameter β_s. This might be a false assumption. However, we want to be absolutely certain that differences observed between sample types are not a mere artifact of differences between studies.

Biological sample To take into account the existence of some variability between samples of the same type t, each biological sample b of this type is associated to a random variable A_{tb}. To a given tb corresponds a unique biological sample namely the b-th sample of type t. A_{tb} takes into account the biological variability.

Residual error To take into account the existence of some variability between replicates of the same sample tb, we add a residual error ε_{tbs}. The index is tbs because a given replicate of sample tb is

perfectly characterized by its study. Indeed, for a given sample tb there is at most one replicate per study. ε_{tbs} takes into account the technical variability.

All these effects are combined additively. This means that the influence of a sample being both Luminal A and from pilot study 1 is the same as the sum of the effect of a sample being Luminal A and the effect of a sample belonging to pilot study 1. One might have considered a non-additive model with interactions, meaning for example that some special combinations of study and type have an erratic behavior. However, including interactions in the model would have resulted in a significant increase in the number of parameters and in the complexity of the model. From previous analyses of the two pilot studies, the additive model seemed a reasonable assumption.

In a statistical framework, the model can be written in the format that follows:

$$Y_{tsb} = \mu + \alpha_t + \beta_s + A_{tb} + \epsilon_{tbs}, \quad V(\epsilon_{tbs}) = \sigma^2, \quad V(A_{tb}) = \gamma^2$$

The parameter with a capital letter (A_{tb}) is random while the others (μ, α_t, β_s) are fixed parameters.

Testing the difference between any two types of sample

Now that we have a model, we need to characterize our question in terms of our model. We want to aggregate the three experiments. We want to efficiently compare the TNBC or ER- / HER2+ types from the first two studies to Luminal tumors, normal and TNBC cell-lines from the final study. For example, we want to assess the differences between the effect of being TNBC and normal or between the effect of being TNBC and Luminal A. From a statistical perspective, we want to estimate all these differences with a maximum precision, i.e. a minimum variance. It is possible to compute this variance (up to a factor of proportionality) given the ratio between the biological and the technical variability.

From a more technical point of view, given an experimental design, the biostatistical model can be described in matrix formulation:

$$\mathbf{y} = \mathbf{X}\theta + \mathbf{E}, \quad V((E)) = \sigma^2\mathbf{I} + \gamma^2\mathbf{Z}\mathbf{Z}',$$

where $\mathbf{y}$ is the vector of all measurements, the matrix $\mathbf{X}$ represents the experimental design for fixed parameters, the vector θ represents the fixed parameters of the model, $\mathbf{E}$ is the vector of all residual

errors and $\mathbf{Z}$ takes into account the presence of replicates. The set of all differences that we want to estimate (for example TNBC vs. normal) can be described using a matrix usually called $\mathbf{C}$ for contrast matrix.

Given all these matrices, it is possible to assess the covariance matrix of all the comparisons as a matrix $\mathbf{V}(\mathbf{X})$ for the design $\mathbf{X}$: $\mathbf{V}(\mathbf{X}) = \mathbf{C}'(\mathbf{X}'(\mathbf{I} + \frac{\gamma^2}{\sigma^2}ZZ')\mathbf{X})^{-1}\mathbf{C}$. Suppose now that we have two experimental designs described by the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. For these two designs, we can compute $\mathbf{V}(\mathbf{X_1})$ and $\mathbf{V}(\mathbf{X_2})$ and compare the efficiency of the designs by comparing the two matrices. For example, if the element in the first column and first line of $\mathbf{V}(\mathbf{X_1})$ is bigger than the one in $\mathbf{V}(\mathbf{X_2})$, it means that for the first comparison the second design has a smaller variance.

To illustrate these matters, let us return to the simple problem of the three weights previously described in section 8.1. The vector $\mathbf{y}$ is the column vector with values (m_1, m_2, m_3) and the vector θ is the column vector (μ_A, μ_B, μ_C). For the simple design described in Figure 8.1 (top), the matrix $\mathbf{X_1}$ is the identity matrix:

$$\mathbf{X_1} = \mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

For the more complex design described in Figure 8.1 (bottom), we obtain:

$$\mathbf{X_2} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

To estimate the three weights (μ_A, μ_B, μ_C), the contrast matrix $\mathbf{C}$ is the identity. In this example, there is no need to include the matrix $\mathbf{Z}$ because there are no replicates. Thus, for the simple design we recover:

$$\mathbf{V}(\mathbf{X_1}) = \mathbf{C}'(\mathbf{X_1}'(\mathbf{I})\mathbf{X_1})^{-1}\mathbf{C} = \mathbf{I}'(\mathbf{I}'\mathbf{I})^{-1}\mathbf{I} = \mathbf{I}.$$

For the complex design, we retrieve after a few matrix products:

$$\mathbf{V}(\mathbf{X_2}) = \mathbf{C}'(\mathbf{X_2}'\mathbf{X_2})^{-1}\mathbf{C} = \frac{1}{4}\begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}.$$

We recover, in the diagonal of this matrix, that the three weights are estimated with a variance of $\frac{3}{4}\sigma^2$.

Exploring the set of possible designs

Having this model and a clear definition of which quantities we want to test, I explore the set of possible designs. In this case, the set of all possible designs is relatively limited. First, you need to consider the number of TNBC samples out of the 24 replicates. There are only 25 possibilities. For a given number of TNBC samples, the number of ER- / HER2+ samples is fixed. For example, if there are 13 TNBC then there are $24 - 13 = 11$ ER- / HER2+ replicates. Then, you just need to consider how many of these 13 TNBC samples are to be taken from the first study (there are only 14 possibilities) and how many of the 11 ER- / HER2+ samples are to be taken from the first study (there are only 12 possibilities). After that, the experimental design is perfectly defined. Thus, it is possible to explore all possible designs, with the help of a computer. This is what I did. For each possible number of TNBC replicates, I selected the design such that the sum of all variances associated to all comparisons was minimal. Results are represented in Figure 8.2. Based on this graph, we decided to pick 15 TNBC and 9 HER2+ samples corresponding to a good balance between the variance of all comparisons.

Experimental design of the last experiment

The design of the last experiment was developed to enable the simple correction of various known technical artifacts such as batch and hybridization effects. The correction of batch and hybridization effects is an important matter. Indeed, as can be seen on Figure 8.3 differences in samples of different batches can be much more important than what one would expect. To estimate and correct these effects, one can use a linear model.

The model

A few months after we selected the number of TNBC and ER- / HER2+ replicates needed, all the biological samples were retrieved. There were 132 samples including: 37 Luminal A tumors, 42 Luminal B tumors, 14 normal samples and 15 TNBC cell-lines, as well as 15 TNBC and 9 ER- / HER2+ tumors. At the time of the experimental design of this final experiment, there were two technical constraints to

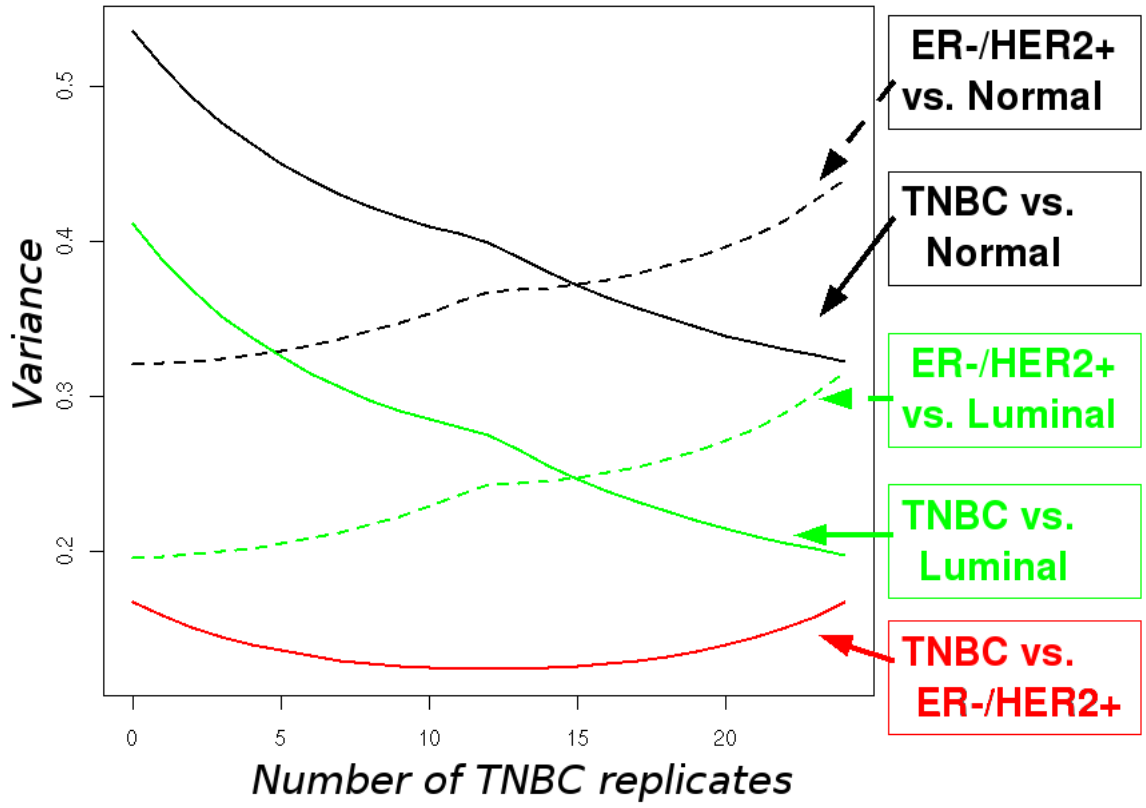


Figure 8.2: Variance of the TNBC vs. ER- / HER2+ (red solid line), TNBC vs. Luminal (green solid line), TNBC vs. Normal (black solid line), ER- / HER2+ vs. Luminal (green dashed line), ER- / HER2+ vs. Normal (black dashed line) as a function of the number of TNBC replicates in the design. Given that the variance of the ER- / HER2+ vs. Normal comparison is equal to that of the ER- / HER2+ vs. Cell-line comparison, the latter comparison is not plotted on the graph. Similarly, the variance of the TNBC vs. Normal comparison is equal to that of the TNBC vs. Cell-line comparison, the latter comparison is not plotted on the graph.

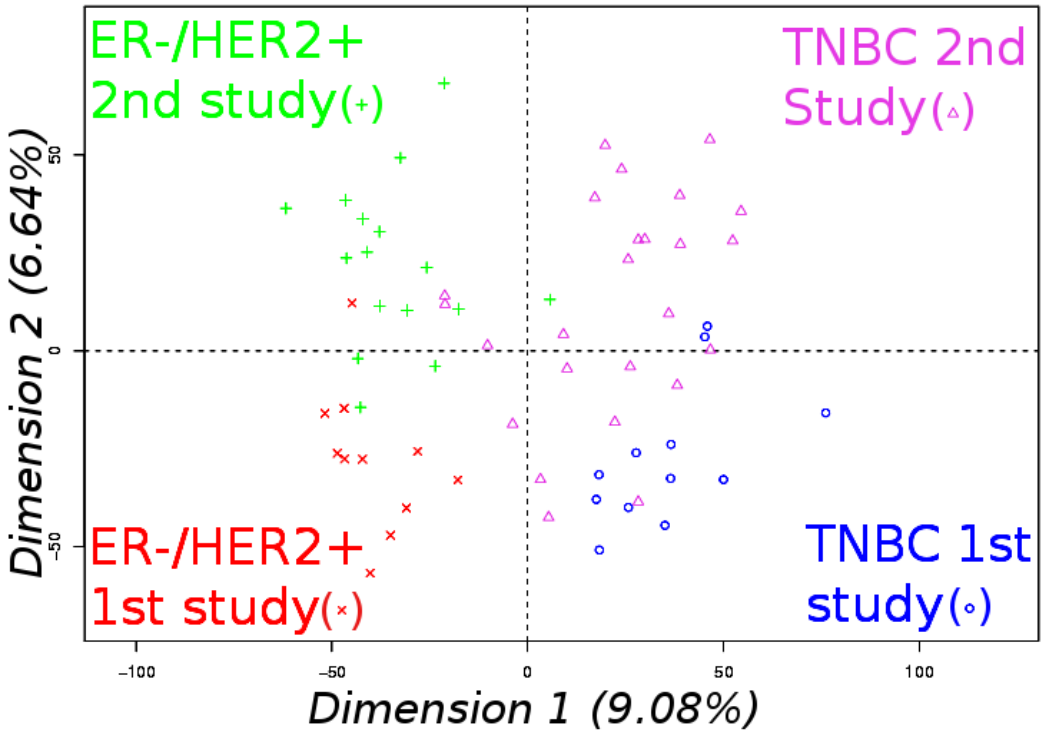


Figure 8.3: Example of batch effect. Each point represents a sample of the first or second experiment in the space of the two principal components. Samples of the first and second study were processed in two different batches. The first axis separates TNBC and ER- / HER2+ tumors. The second axis separates tumors from the first and second study. This is a batch effect that hampers our ability to detect true differences in between TNBC and ER- / HER2+. Importantly in terms of variability, this batch effect represents 6.64% of the variability, almost as much as the biological signal which accounts for 9.08% of the variability.

take into account for the design. First, the preparation of the mRNA samples before hybridization on the microarray is a difficult task and only 30 samples are processed in one batch. Furthermore only 45 samples are hybridized at the same time. Thus, the 132 samples were processed in 5 different batches and 3 different hybridization steps making for a total of $3 \times 5 = 15$ batch-hybridization configurations.

Having all this in mind, I proposed a linear model to explain for a given gene the differences observed between any two samples of this final experiment. The linear model takes into account 4 things:

Type The measured intensity depends on the sample types (t). Each sample type (TNBC, Luminal A, ...) has a different mean level. In the model, the effect of being of type t is associated to a fixed parameter α_t. While this is certainly not true for every gene, it is a practical hypothesis and it is precisely the existence of such a sample type effect that we want to test.

Batch The measured intensity is also explained by the batch (b). Each sample batch is associated to a different mean level. In the model, the effect of being in batch b is associated to a fixed parameter β_b. This is certainly not true for every batch. However, we want to be absolutely certain that differences observed between sample types are not mere artifacts of differences between batches.

Hybridization The measured intensity is also explained by the hybridization (h). Each hybridization step is associated to a different mean level. In the model, the effect of being in hybridization step h is associated to a fixed parameter δ_h. Again, it is a working hypothesis but we want to be absolutely certain that differences observed between sample types are not mere artifacts of differences between hybridization steps.

Residual error Each biological sample is associated to a given random variable ϵ_{tbhi} that takes into account the variability that exists even between samples of the same type, batch and hybridization step. ϵ_{tbhi} is the residual error of both the biological and technical variability.

All these effects are combined in an additive fashion. In a statistical framework, it can be written in the format that follows:

$$y_{tbhi} = \mu + \alpha_t + \beta_b + \delta_h + \epsilon_{tbhi}, \quad V(\epsilon_{tbhi}) = \sigma^2$$

Testing

The idea is now to recover genes that are overexpressed in TNBC compared to normal samples and other subtypes of tumors. In terms of our model, we want to assess the difference between any two subtypes. As we have seen in the previous subsection, it is possible to compute the variance associated to the difference between any two subtypes.

Exploring the set of possible designs

Having this model and a clear definition of which quantities we want to test, I explore the set of possible designs. Importantly, I restricted the search to balanced designs. Here, "balanced" means that if you consider one sample type, there should be little variation in the number of these samples per batch, per hybridization and per batch-hybridization configuration. To illustrate this point, I will take two examples. In the case of cell-lines, there are exactly 15 samples, matching the 15 batch-hybridization configurations, thus there is only one balanced design for this type: one cell-line sample per batch-hybridization configuration. In the case of Luminal A, it is a bit trickier. There are 37 samples. To obtain a balanced design, 3 batches should have 7 Luminal A samples and the two others batches should have 8. There should be 2 hybridization steps with 12 Luminal A and the third one with 13. Finally, there should be 8 batch-hybridization configurations with 2 Luminal A and 7 with 3 Luminal A. This greatly reduces the set of possible designs. In fact, the main point of this restriction is to have a design robust to problems. Indeed, suppose for example that for whatever reason one of the batches had failed to work properly, in the four remaining batches all sample types would have been well represented and it would still have been possible to compare all the types.

In the end, the number of balanced designs is relatively limited. Indeed, there are only four sample types to consider as there is exactly 15 TNBC and 15 TNBC cell-lines matching the 15 batch-hybridization configurations. For the remaining types (Luminal A, Luminal B, ER- / HER2+ and Normal), there are 7, 12, 9 and 14 samples to consider respectively. For each of these subtypes, it is possible to construct almost by hand the set of configurations that are balanced. For example, for Normal samples a balanced configuration is perfectly determined by the position of the one batch-

hybridization position that does not include a Normal sample. The ordering of the five batches and three hybridization steps is of no importance, which further reduces the number of possible designs. Thus, one can arbitrarily decide that the position that does not include a Normal sample is batch 1 and hybridization step 1. Overall, there are 930, 60, 180 and 15 balanced configurations for Luminal A, Luminal B, ER- / HER2+ and Normal samples respectively. As we have seen, we can consider only one configuration for the Normal sample and thus we obtain $930 \times 60 \times 180 = 10\ 044\ 000$ combinations.

Using a computer, it is possible to construct all the configurations for each type and combine them to recover all possible suitable experimental designs. Having these designs, it is possible to compare them in terms of their power to estimate some comparisons of interest as explained in the previous subsection. Overall, there was no great difference between all these designs. Thus, as long as the design was balanced, the choice was not critical and I picked a design with a slightly better ability to compare Normal and TNBC samples.

Chapter 9

Pre-processing

After the transcriptomic experiment, after the removal of bad-quality RNA samples and the incorporation of the two pilot studies, we were able to recover a total of 177 good-quality microarray experiments described in Figure 9.1. For each of these 177 experiments we retrieved a CEL files containing the measured intensity for each probes of the Affymetrix HGU133-plus2 microarray.

9.1 Probe annotation

Once the 177 CEL files were obtained, the first question was how to aggregate the information from the probes into probesets (the annotation problem). Each probeset hopefully groups probes corresponding to the same gene. Various solutions exist:

- The usual one is the annotation provided by Affymetrix. The problem with this method is that one gene can be found on several probesets while one probeset can be associated to several genes, this makes interpreting the data more difficult.

- Alternative annotations have been proposed (such as AffyProbeMiner (Liu et al., 2007) or Ballester et al. (2010)) that provide more robust annotations. But, with these annotations the information for some genes is lost.

	TNBC	ER-/ HER2+	Luminal A	Luminal B	TNBC cell-lines	Normal Tissue
ER/PR status	-	-	+	+	-	
overexpresion of HER2	-	+	-	+ / -	-	
CK14 or CK5/6 or EGFR	+	-	-	-		
Grade	III	III	I	III		
1st study	13	11				
2nd study	24	16				
Last study	15 (r)	9 (r)	31	31	14	11

Figure 9.1: Summary of good quality transcriptomic samples of the Curie-Servier dataset and their histological and immunohistochemical characterizations. 15 TNBC and 9 ER- / HER2+ samples of the first and second studies were replicated in the last study.

I choose to carry out the analysis both with the AffyProbeMiner annotation and the Affymetrix annotation. Indeed, the AffyProbeMiner annotation is a robust method which was publicly available at the beginning of the study while the Affymetrix annotation is useful for comparisons with other studies because it has been widely used.

9.2 Normalization

Once the probes are grouped into probesets, one needs to correct various non-relevant effects that hamper our ability to extract the biological signal. This is called normalization. As we have seen in the chapter 4 for DNA copy number, normalizing is a difficult issue. Here I chose to use a three-step method. The first step normalizes the measurements using the GC-RMA methodology (Wu et al., 2004). The second step removes low level measurements. Finally the third step corrects batch and hybridization effects using a mixed linear model. I chose to use the GC-RMA methodology which takes into account the probe sequence information to obtain more accurate measurements of the specific hybridization. This method has been shown to be efficient (Binder et al., 2010). For some of the analyses I also used the RMA methodology (Irizarry et al., 2003b) to compare the results with publicly available ones.

After the GC-RMA normalization of the data, we recover corrected measurements of the expression of all the genes present on the microarrays. From a biological point of view, it is likely that many of these genes are not expressed in the sample under consideration. These genes should have very low measurements for most probes. I thus decided to discard those genes that have very low measurements in most samples, since they do not inform us regarding differences between subtypes or with the normal samples. Removing these genes makes the remainder of the analysis easier and gives us a better chance to identify some relevant genes, if not all of them.

Using the GC-RMA normalization method, this discarding step is simple and well justified. Indeed, a typical histogram of log2 measurements of all genes across all tumors will always be bi-modal with a sharp peak around 2.5 log2-intensity and a flatter mode covering values between 3 and 15 log2-intensities (see Figure 9.2). The sharp peak can be interpreted as measurements that do not exceed

the background noise detection level while the flatter mode can be interpreted as measurements of genes for which expression is detected. Given the sharpness of the peak around 2.5 log2-intensity, it is quite easy with usual statistical techniques to identify a threshold to distinguish between the foreground and background level. Overall, in the case of the AffyProbeMiner annotation, approximately 10^4 probesets out of 2.10^4 were discarded. One can be surprised by the high number of probesets that are discarded (50%) but from a biological perspective it makes sense that only a fraction of the genes are expressed in a given cell type.

The experimental design I proposed allowed for the simple assessment and efficient correction of batch and hybridization effects. To correct these effects we used a mixed linear model. This model is similar to those described in chapter 8. It explains the differences in expression levels between the 177 samples for every gene by taking into account:

Type The measured intensity depends on the sample types (t). In the model the effect of being of type t is associated to a fixed parameter α_t. Note that this is certainly not true for every gene. Yet it is a practical hypothesis and it is precisely the existence of such a sample type effect that we want to test.

Batch The measured intensity is explained by the batch (b). In the model the effect of being in batch b is associated to a fixed parameter β_b.

Hybridization The measured intensity is also explained by the hybridization (h). In the model the effect of being in hybridization step h is associated to a fixed parameter δ_h.

Biological sample Some biological samples of the pilot studies are replicated in the last experiment. This is accounted for as random effect (as opposed to fixed). Each biological sample s of type t is associated to a given random variable A_{ts}.

Technical sample Finally, even in between replicate of the same biological sample, even after correction of the batch and hybridization step, there is some disparity. This is the residual error ϵ_{tsbh} which models the technical variability.

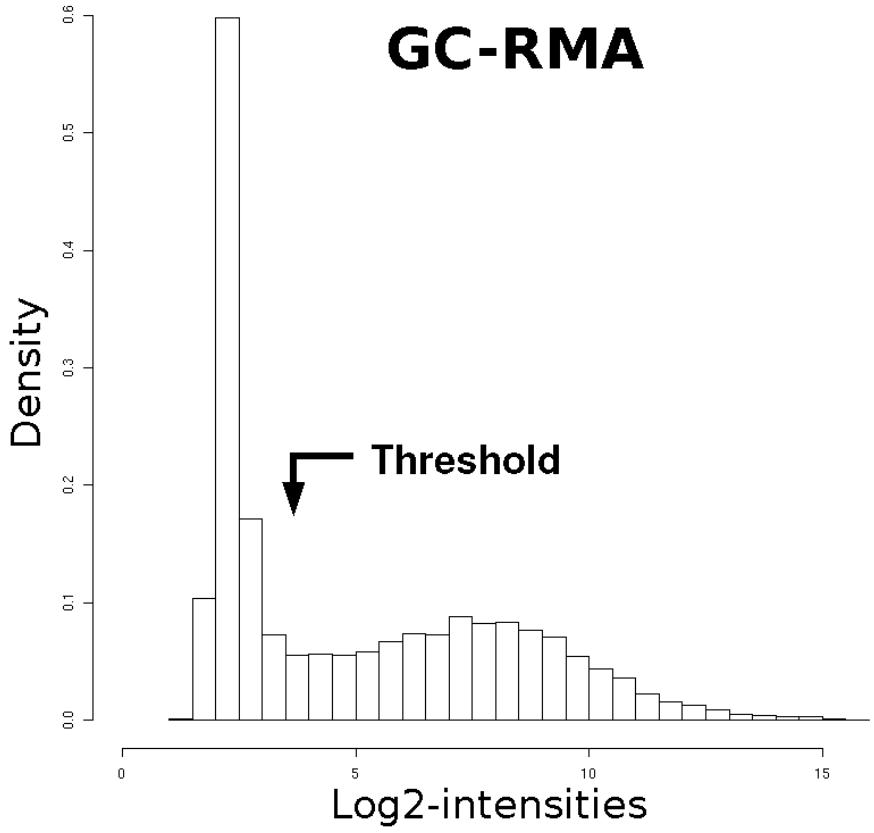


Figure 9.2: Histograms of log2-intensities after the GC-RMA normalization. The histogram is plotted so that the histogram has a total area of one. The height of a rectangle is proportional to the number of points falling into the cell. The histogram has a sharp peak around 2.5 log2-intensity.

All these effects are combined in an additive fashion without interactions. In a statistical framework all this can be written as follow:

$$Y_{tsbh} = \mu + \alpha_t + \beta_b + \delta_h + A_{ts} + \epsilon_{tsbh}, \quad V(A_{ts}) = \gamma^2 \quad and \quad V(\epsilon_{tsbh}) = \sigma^2$$

Having defined this model, it is possible to estimate from the data the influence of each molecular subtype, batch and hybridization set. Then using those estimations, it is straightforward to correct batch and hybridization effects. The result of this correction are shown in Figure 10.4 (in the next chapter) for the samples of the first two studies.

Chapter 10

Exploratory Analysis

10.1 Validation of the pre-processing step

Once I had carried out all these pre-processing steps, to validate them I checked three important points:

- replicates should be similar;
- there should be no visible differences between batches or hybridization steps;
- it should be relatively easy to identify the different molecular subtypes.

To acheive that, I used two exploratory methods, hierarchical clustering and Principal Component Analysis (PCA). In the following I will summarize the results obtained with both approaches. One difficulty when analyzing gene expression profiles is that you get measurements for thousands of genes. It is therefore impossible to visualize such an amount of data in a 2D or 3D plot. Thus one needs to reduce the dimension of the problem (dimension reduction). To get a good view one can use hierarchical clustering. In hierarchical clustering methods, samples are represented by a tree-like structure. The root of the tree corresponds to the whole dataset. The leaves of the tree correspond to samples. Starting from n samples or clusters a usual way to build such tree is to compute the distance between any two clusters and then aggregate the two closest clusters together. In the tree structure these two

clusters are branched together. This aggregation process is iterated until only one cluster is left. In the end, the recovered tree hopefully provides valuable information about the underlying structure of the data and it might be possible to identify well separated clusters.

A key issue with hierarchical clustering is to define the distance between two clusters. Most of the time one has to choose a linkage rule and a measure. The distance between two clusters can be the Euclidean distance (measure) between the two closest samples in the two clusters (simple linkage). One could also choose the City-block distance between the two furthest samples of the two clusters (complete linkage).

In all that follow, I have used the Ward's method. The Ward's method can be viewed as an analysis of variance approach to clustering. It considers the cost of a cluster as the sum of square distances between any sample of the cluster and the mean of the cluster. The distance between two clusters is defined as the cost of the two clusters taken as a whole minus the cost of the first and second cluster taken separately. In a more mathematical formulation, the distance between two clusters C_1 and C_2 is $D(C_1, C_2) = Cost(C_1 \cup C_2) - Cost(C_1) - Cost(C_2)$. The Ward's methods tends to minimize the loss of information associated to each grouping.

Using hierarchical clustering on our normalized data, I observed that, due to the correction of batch and hybridization effects, samples from similar batches did not cluster all together (data not shown, see results for PCA in Figure 10.4) and that replicates were always next to each other (see Figure 10.2). Moreover it was possible to identify visually six main branches in the structure corresponding to the TNBC, ER- / HER2+, Luminal A and Luminal B subtypes, the TNBC cell lines and the normal population. This means that there is a very good concordance between this classification and the immunohistochemistry-based classification (see Figure 10.1).

All these results were very comforting. However, there are three issues with hierarchical clustering that should not be forgotten. First, there is not a unique representation of the tree. The tree can be rotated around any of its branches. This means that, using tree-like representation, it is impossible to infer the geometric configuration of three different clusters. For example (in Figure 10.1), the TNBC cell line cluster appear in between the TNBC cluster and the ER- / HER2+ tumor cluster, but in fact

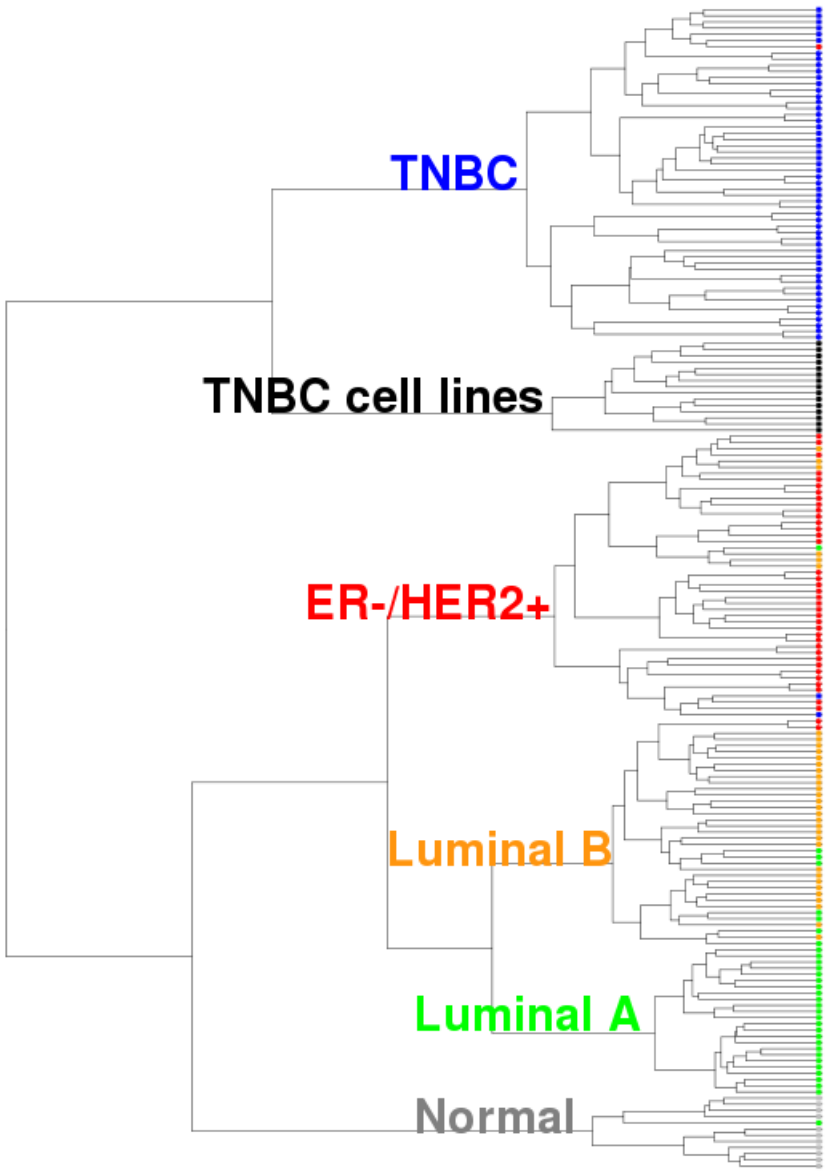


Figure 10.1: Validation of the pre-processing steps using hierarchical clustering of all 177 transcriptomic samples. The hierarchical clustering was done using the gene expression profiles of the thousand genes with the highest variance using the ward methods. The results are represented using a dendrogram.

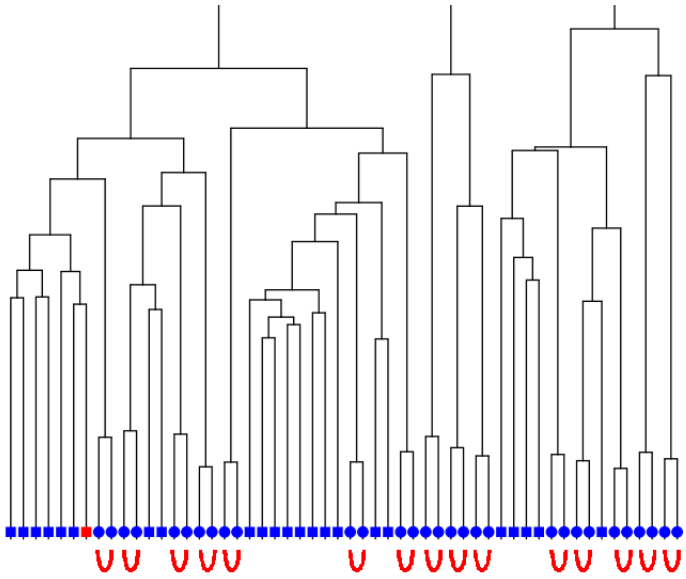


Figure 10.2: Good quality of the replicated samples after normalization. Zoom on TNBC on the clustering of all transcriptomic sample in Figure 10.1. TNBC replicates are joined by red U. All TNBC replicates are visibly close to each other in the dendrogram.

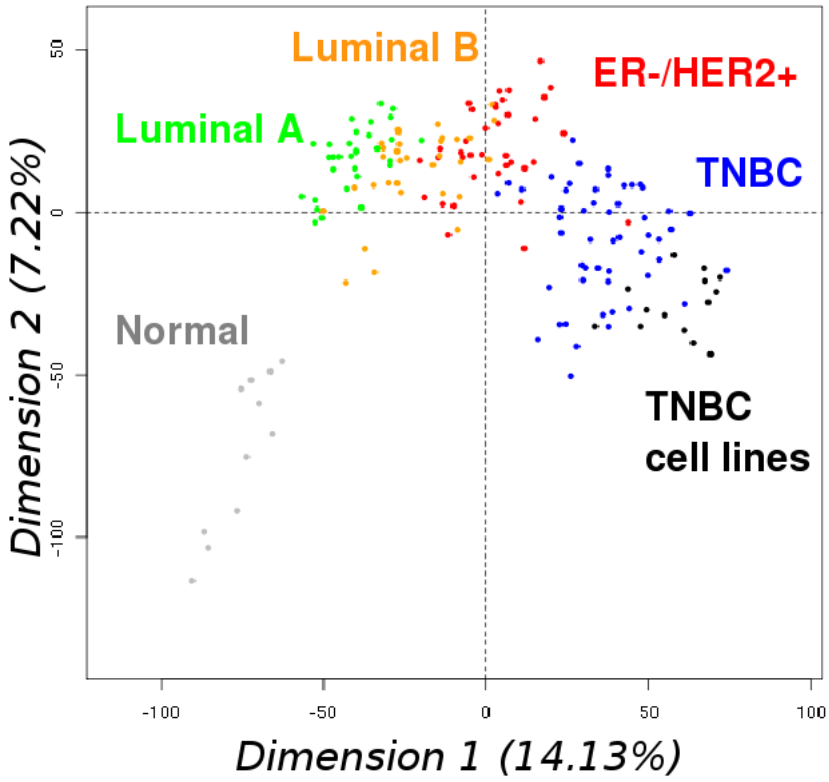


Figure 10.3: Validation of the pre-processing steps using principal component analysis of the 177 transcriptomic samples. Each point represents a sample in the two principal components.

this position is arbitrary and if one used this information, one would jump to false conclusions. In fact, the ER- / HER2+ tumor cluster is closer to the TNBC cluster than to the TNBC cell line cluster as suggested by the principal component analysis in Figure 10.3. The second issue is that hierarchical clustering is a heuristic. Indeed, the only guarantee is that each recursive step of the hierarchical clustering is optimal. As we have seen for the segmentation problem, this is not an overall optimal strategy. Third I have chosen a particular aggregation scheme (the Ward's method). There are many others. It is cautious to check these results with a completely different approach.

Here, I have chosen principal component analysis (PCA) as a complementary approach. Principal components aim at representing all samples in a given number of dimensions, usually 2 or 3. More precisely, it recovers the optimal representation of the dataset in a given number of dimensions. This representation is optimal in the sense that all the variability that it is possible to capture in so little dimensions is retrieved. This representation is unique and one can interpret the relative position of more than two points.

Using PCA on our normalized data, I confirmed the results of the hierarchical clustering that samples from similar batches do not appear to be close (see Figure 10.4) and that replicates were next to each other using the two principal components (data not shown). Moreover it was also possible to identify that the different tumor subtypes were relatively well separated even using only the first two principal components which represented 21% of the variability (see Figure 10.3). To conclude, that shows that the normalization process was efficient in removing non-relevant effects from our dataset.

10.2 A robust classification of breast tumors, but no intrinsic gene list?

Overall, conventional diagnostic techniques such as histopathology are still considered as the gold standard for tumor classification (Weigelt et al., 2010a). Indeed, as we have mentioned before, the quality of gene expression-based classification and prediction has been questioned and quite importantly it has been shown that some studies did not classify patients better than chance (Ein-Dor et al., 2005; Michiels et al., 2005). In the case of tumor subtypes the classification problem is (probably) easier yet the stability of each cluster is limited and many different list of genes have been proposed to classify tumors (Pusztai et al., 2006; Weigelt et al., 2010a). One might wonder why that is. A first possibility to explain unstable classification results might be the intrinsic complexity of the classification problem. Another possibility would be that the information to classify tumors exists but that looking for a given gene list to perform the classification and summarize the information may not be efficient. We have previously used hierarchical clustering for the analysis of the Curie-Servier dataset. This type

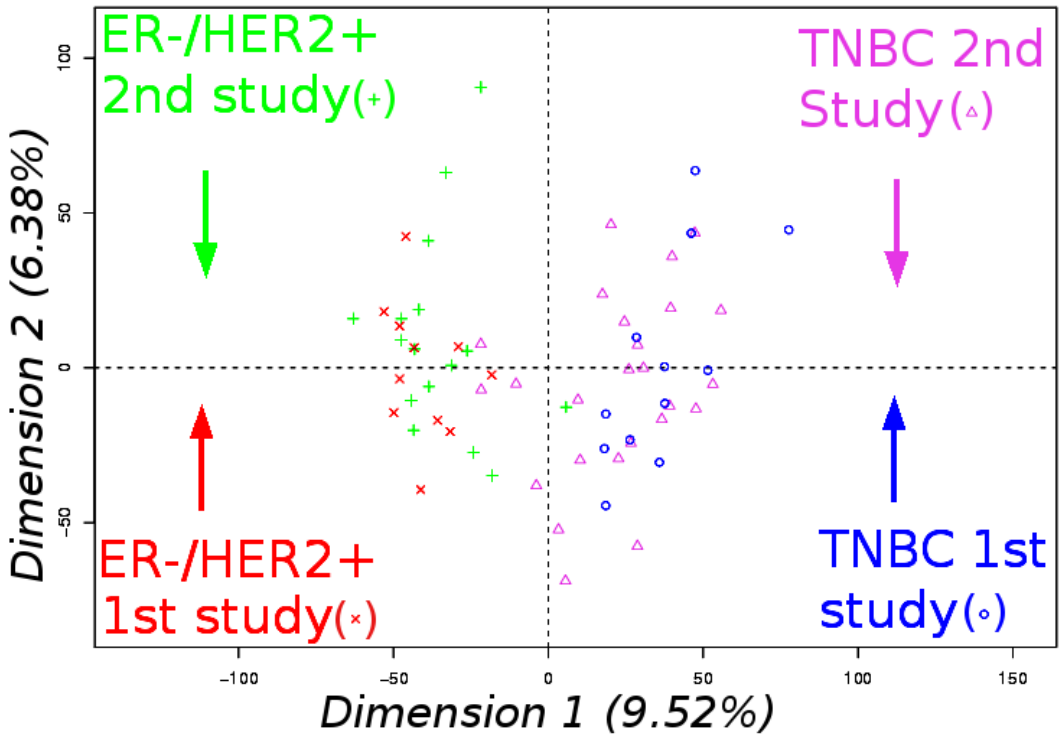


Figure 10.4: Correction of batch effects. This figure matches Figure 8.3 but after correction of batch and hybridization effects (using the mixed linear model described in the text). Each point represents a sample of the first and second study in the two principal components. Samples of the first and second study were processed in two different batches. The first axis still discriminates in between TNBC and ER- / HER2+ tumors. The second axis no longer discriminates between tumors of the first and second study (see Figure 8.3).

of classification is unsupervised, which means that it does not specifically try to recover the different subtypes. Nevertheless, we have found a very good correspondence between the unsupervised classification of our tumors and the immunohistochemistry (IHC) based classification. Thus the classification problem seems relatively simple. Looking for a specific gene list to perform the classification, might be responsible for the instability of the results. To further investigate these points, I set up the following simple experiment.

I assessed the ability of randomly selected lists of genes to segregate the four different tumor subtypes in an unsupervised setting. I used an unsupervised classification method because the goal was to assess the influence of the list of genes on the classification and to test whether the classification truly depends on the list. There was 685 genes in the list corresponding to the number of Affymetrix HGU133plus2 probeset matching the intrinsic gene list from Perou et al. (2000). For each of these lists, I classify all the tumors into four groups using an unsupervised model based clustering approach. I used a Gaussian mixture model for four groups. The Gaussian mixture model aims at identifying four groups based on the level of expression of the genes in the list. It returns a group number between 1 and 4 for each sample. Having identified those 4 groups, I compared them with the known groups identified by IHC. On average there was around 17% discrepancy between the reference IHC classification and the model-based classification (see Figure 10.5). This 17% discrepancy is much less than what would be expected from a completely random classification (66%).

The conclusion of this experiment is that there is, at least on our dataset, a very good concordance between the IHC information and gene expression information. The gene expression information is robust and not specific to a given gene list but is rather spread across many different genes. Finding a gene list that segregates correctly the different tumor subtypes is easy. Indeed, even randomly selected gene lists, without any specific training (unsupervised classification), have very good performances. In view of the nature of these results, it appears that looking for a unique and well-defined gene list that explains the differences between tumor subtypes is hazardous. These results and a detailed description of the IHC data on the Curie-Servier cohort are in preparation for publication in collaboration with Anne Vincent Salomon (MD/Ph.D., Institut Curie, Rigaill et al. (2010b)).

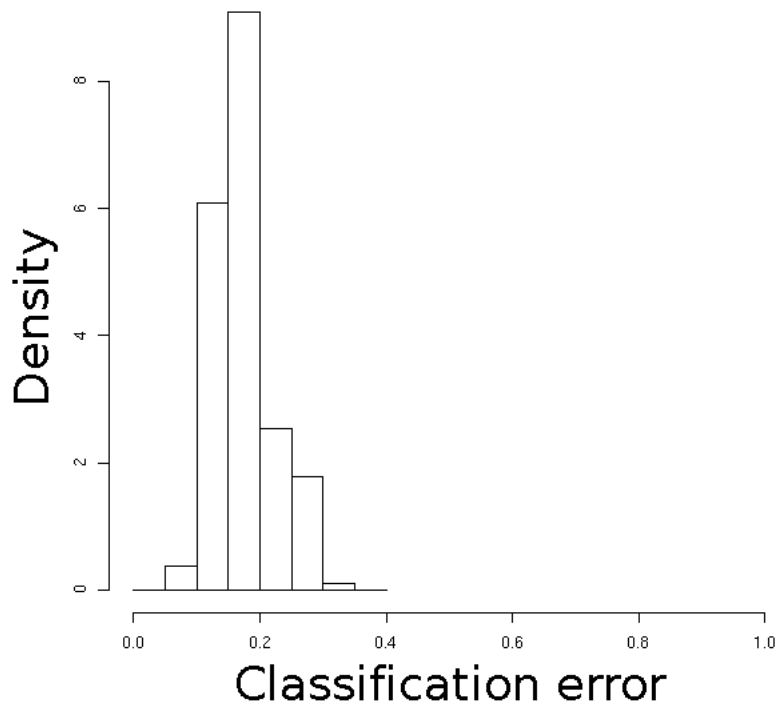


Figure 10.5: Correspondence between the IHC classification and an unsupervised classification in four groups using a Gaussian mixture model based on random lists of genes. The classification error is the proportion of samples that are not classified in the same group by the IHC-based and the Gaussian mixture classification. This classification error was measured for 300 000 randomly selected 685-long gene lists and represented as a histogram. The histogram is plotted so that the histogram has a total area of one. The height of a rectangle is proportional to the number of points falling into the cell. The classification error between the IHC classification and a completely random classification in four groups is on average 0.66.

Chapter 11

Comparison of TNBC with other tumor types

As we have seen in the experimental design chapter (8), the main question of the project was to identify key gene expression deregulations in TNBC that might explain the phenotype and be used as therapeutic targets. To this end, TNBC samples were compared to samples of other subtypes (class comparison). As we have seen in the previous section (10.2), it is extremely easy to find a gene list that explains the differences between subtypes. Thus, differential analysis alone will not reduce the list of candidate targets sufficiently and we need other filters. In the following, I will describe the strategy I have set up to propose candidate genes and identify pathways of interest.

11.1 Gene by gene differential analysis

11.1.1 Statistical testing

It was decided to look first for genes over-expressed in TNBC compared with either normal or Luminal A tumors. Many statistical methods have been proposed and some were specifically designed for microarray gene expression differential analysis (Tusher et al. (2001); Smyth (2004); Delmar et al. (2005)). These three methods were designed to accommodate very small sample sizes (less than 10

or even 5). Here I decided to use none of these 3 specific methodologies but rather the mixed linear model that I used for the estimation of batch and hybridization effects (see subsection 9.2). The reason for this choice is that the Curie-Servier dataset is a large one, with more than 30 samples for the four tumor subtypes considered. This is a rather favorable case and we will hopefully retrieve a good estimation of the variance using the 177 samples of the dataset simultaneously.

As we have seen previously, the mixed linear model leads to very good corrections of both batch and hybridization effects. The mixed linear model takes into account the existence of replicates and the fact that batch and hybridization effects have to be estimated. Thus, using this model, I computed a p-value for every gene. These p-values are close to 0 when there is a significant difference between the two types of sample compared.

In order to try to validate the use of a mixed linear model, it is usual to look at the empirical distribution of the p-values obtained for each gene. The results are shown in Figure 11.1. It can be seen on this figure that the distribution takes the shape we expected: there is a peak for small p-values corresponding to genes that are truly differently expressed in TNBC. The rest of the distribution is relatively flat and corresponds to genes that are not expressed differently in TNBC and thus have p-values distributed uniformly between 0 and 1. Looking at this distribution, it can be seen that there are lots of genes with a small p-value. It indicates a very high proportion of genes that are differentially expressed. This result is not surprising given that unsupervised classification clearly segregates TNBC from Luminal A and normal tissues (as we have seen earlier in section 10.1).

For any given gene, the computed p-value represents the chance or risk that we wrongly declared it as over-expressed in TNBC if, in fact, it is not. We can determine the difference between the mean level of a given gene in the TNBC samples and the mean level of the same gene in the normal samples. The p-value of this gene is the probability of observing such a difference for this gene if in fact there are no differences between the 2 types overall. Thus, p-values can be used as a decision-making tool to answer the question: “Is the gene up-regulated in TNBC?”. It is important to see that it is not because the p-value is close to 1 that the gene is not differentially expressed. This only means that there is no evidence in our data to support this possibility. Similarly, it is not because the p-value is

close to 0 that the gene is differentially expressed. It only means that the observed difference is highly unlikely if the gene is not differentially expressed.

Very often, a p-value or risk of less than 5% is considered to be acceptable. There is no particular reason for that choice. Here, we simultaneously look at thousands of genes. Taking a risk of 5% for each of them would necessarily lead to many errors. This is a major problem in statistics, which is known as the multiple-testing issue. A basic idea to reduce the overall risk is to take a much smaller risk for each gene, though various less drastic methodologies have been proposed. Here, we used the Benjamini-Hochberg FDR (False Discovery Rate) strategy. This strategy is very popular and straightforward. It orders genes by increasing p-values and for each successive gene it returns the expected percentage of false positives up to this point.

Based on this estimation, one selects the size of the list (and thus a set of genes) such that the FDR is sufficiently small, typically once again about 5% so that the expected number of false postives in the list is 5%. Using this FDR methodology, we determined that almost half of the tested genes were differentially expressed with an FDR of 5%. This is not surprising as we already know that many genes segregate TNBC from the other tumor types.

Overall, statistical testing is a valid and sound strategy to discriminate between genes for which over-expression in TNBC is supported by our data and genes for which no difference is detected in our data. In other words, it is used as a filter. Here, in the case of TNBC, even with a relatively small FDR, many genes are classified as over-expressed. It would not be realistic to try to look through a list of 1000 or 2000 genes. So, we need additional filters to narrow down the list of candidate targets.

11.1.2 Other filters

To narrow down our list of candidate targets, we considered additional filters. A filter discards genes for which the difference between the mean level in the TNBC samples and the mean level in the normal samples is too small. Such a filter is rather controversial, at least from a statistical point of view. Using a statistical test, we identified significantly over-expressed genes with small differences. In those cases, the difference is found to be significant because the variance of the signal is small within each

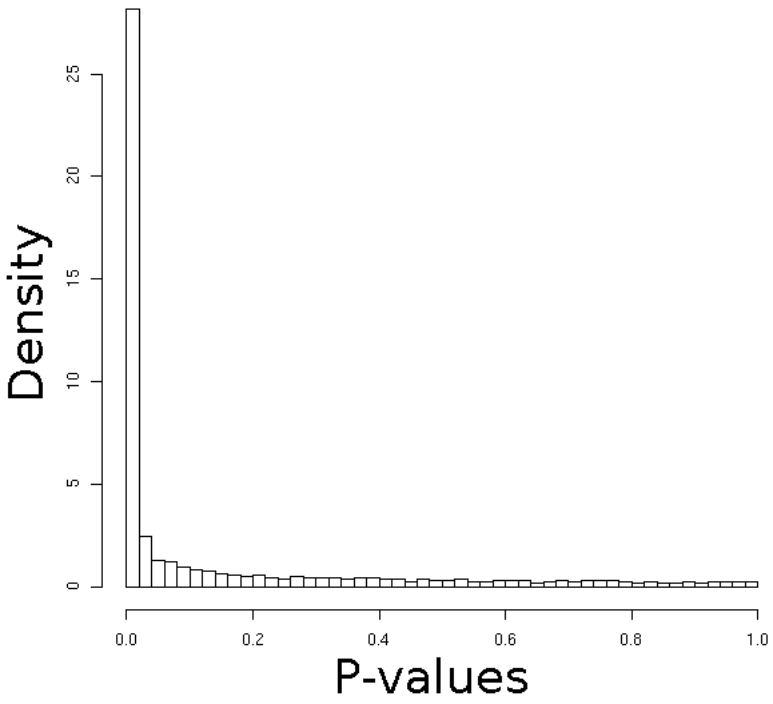


Figure 11.1: Histogram of p-values for all the genes tested in the TNBC vs. normal comparison. The histogram is plotted so that the histogram has a total area of one. The height of a rectangle is proportional to the number of points falling into the cell. There is a peak for very small p-values that indicates a high proportion of differentially expressed genes

type, which corresponds to a very high reproducibility. Nevertheless, those genes are usually discarded because it is thought that the difference is too small compared to the resolution of the technology. In fact, if presented with two genes found to be significantly over-expressed in TNBC with the same p-value and without any other information, one would probably choose the one with the highest mean difference. Usually, because the data uses the log2 scale, a threshold of 1 is used because it corresponds to a ratio between the mean levels of 2. It is an arbitrary choice which undoubtedly leads to the loss of interesting targets. For example, if I had to choose between a gene with a mean difference of 0.988 and a p-value of 1.2710^{-11} (*DNTB* in our dataset for the TNBC vs. normal comparison) compared to a gene with a mean difference of 1.02 and a p-value of 0.03 (*TNXIP* in our dataset), I would choose the first gene. Nonetheless, as part of the Curie-Servier collaboration, it was decided to keep the threshold at 1. A large proportion of genes are discarded with this threshold even though a number of them have very significant p-values (see Figure 11.2).

Not all genes are potentially interesting target genes and lists of drugable genes have been published (Hopkins and Groom, 2002), which limit potential targets to about 2000 genes. Within this list, one can also consider the list of kinases (see Figure 11.3) to further reduce the size of the list.

Finally, using the different comparisons (TNBC vs. normal, TNBC vs. Luminal A ...), it is possible to filter down those lists of genes as illustrated in Figure 11.4. In this example, we looked for genes that were specifically expressed in TNBC and retrieved 218 genes. Inside this list, it can be interesting to look for genes that are also over-expressed in TNBC cell-lines. Indeed, if genes are over-expressed both in TNBC biopsy samples and in TNBC cell-lines, it would be likely that these genes are over-expressed in the TNBC cells themselves and not in the stromal cells that were also sampled during the biopsy. Another advantage of these particular genes is that they are more amenable to experiments because cell-lines could be used as a model for the tumor.

From my transcriptomic and genomic analysis and also from some experimental data in TNBC cell lines, some candidate targets were subjected to a target assessment procedure. Its main objective was to classify the targets for further experimental validation programs. Target assessment is a manual procedure: it attributes an ad hoc score to each target in terms of drugability, competition with

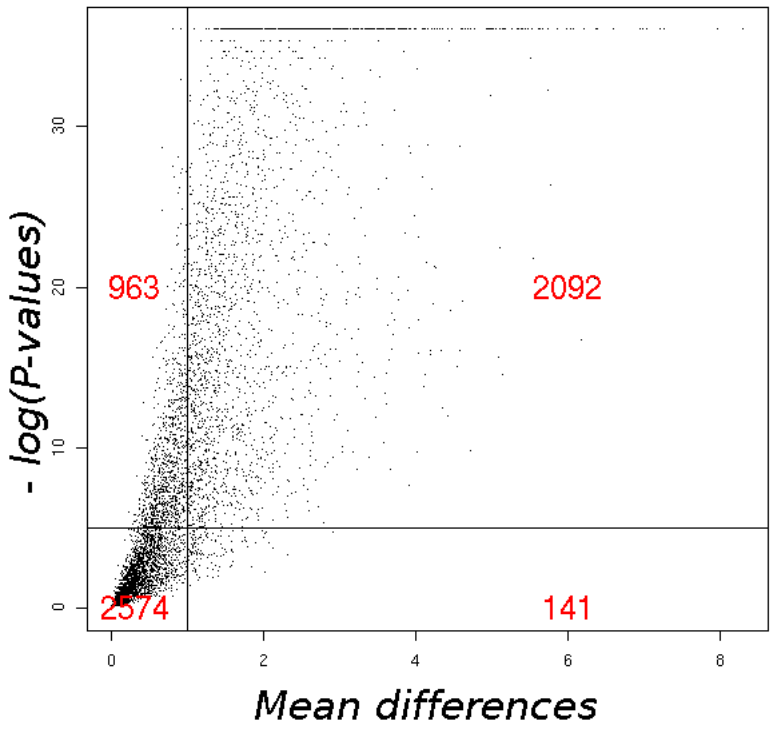


Figure 11.2: Volcano plot: $-log$(p-values) as a function of the mean differences. Only genes with a positive mean difference are considered. It can be seen that 963 genes with a p-value smaller than 10^{-5} are discarded because the mean difference is smaller than 1.

A

vs. Normal

symbol	ID	description	P-value
RIPK2	NM_003821	receptor-interacting	$< 10^{-7}$
ASK	NM_006716	activator of S phas	$< 10^{-7}$
MELK	NM_014791	maternal embryoni	$< 10^{-7}$
BUB1B	NM_001211	BUB1 budding uni	$< 10^{-7}$
PGK1	NM_000291	phosphoglycerate	$< 10^{-7}$
TOPK	NM_018492	T-LAK cell-originat	$< 10^{-7}$
CHEK1	NM_001274	CHK1 checkpoint	$< 10^{-7}$
TTK	NM_003318	TTK protein kinase	$< 10^{-7}$
LYN	NM_002350	v-yes-1 Yamaguch	$< 10^{-7}$
AURKB	NM_004217	aurora kinase B	$< 10^{-7}$

vs. Luminal A

symbol	ID	description	P-value
RIPK2	NM_003821	receptor-interacting	$< 10^{-7}$
ASK	NM_006716	activator of S phas	$< 10^{-7}$
MELK	NM_014791	maternal embryoni	$< 10^{-7}$
BUB1B	NM_001211	BUB1 budding uni	$< 10^{-7}$
SKP2	NM_005983	S-phase kinase-as	$< 10^{-7}$
TOPK	NM_018492	T-LAK cell-originat	$< 10^{-7}$
CHEK1	NM_001274	CHK1 checkpoint	$< 10^{-7}$
SRPK1	NM_003137	SFRS protein kina	$< 10^{-7}$
TTK	NM_003318	TTK protein kinas	$< 10^{-7}$
AURKB	NM_004217	aurora kinase B	$< 10^{-7}$

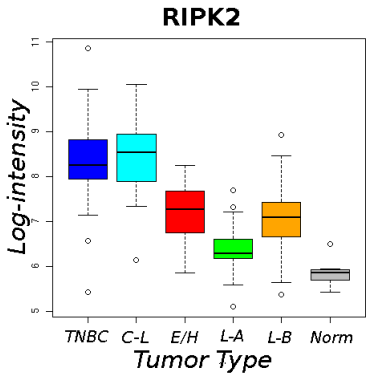
vs. Luminal B

symbol	ID	description	P-value
PRKCN	NM_005813	protein kinase C, n	$< 10^{-7}$
CHEK1	NM_001274	CHK1 checkpoint	$< 10^{-7}$
TTK	NM_003318	TTK protein kinase	$< 10^{-7}$
AURKB	NM_004217	aurora kinase B	$< 10^{-7}$
PLK1	NM_005030	polo-like kinase 1 (	$< 10^{-7}$
PRKX	NM_005044	protein kinase, X-li	$< 10^{-7}$
SKP2	NM_005983	S-phase kinase-as	$< 10^{-7}$
UMPK	NM_012474	uridine monophosp	$< 10^{-7}$
YES1	NM_005433	v-yes-1 Yamaguch	$< 10^{-7}$
MAGI1	NM_173515	membrane associa	$< 10^{-7}$

vs. ER-/HER2+

symbol	ID	description	P-value
PLK1	NM_005030	polo-like kinase 1 (	$< 10^{-7}$
MAGI1	NM_173515	membrane associa	$< 10^{-7}$
SKP2	NM_005983	S-phase kinase-as	$< 10^{-7}$
PRKX	NM_005044	protein kinase, X-li	$< 10^{-7}$
CHEK1	NM_001274	CHK1 checkpoint	$< 10^{-7}$
AURKB	NM_004217	aurora kinase B	$< 10^{-7}$
CKS1B	NM_001826	CDC28 protein kin	$< 10^{-7}$
TOPK	NM_018492	T-LAK cell-originat	$< 10^{-7}$
RIPK2	NM_003821	receptor-interacting	$< 10^{-7}$
UMPK	NM_012474	uridine monophosp	$< 10^{-7}$

B



C

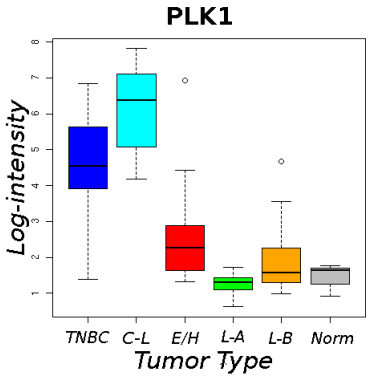
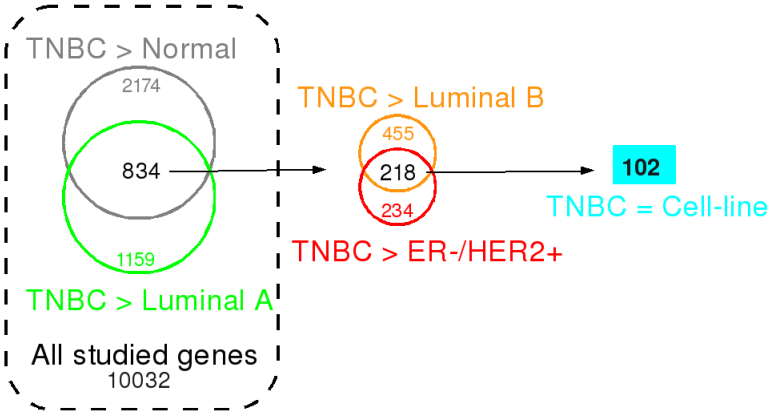


Figure 11.3: (A) Kinase Top 10 for the TNBC vs. normal, vs. Luminal A and vs. Luminal B comparisons. (B) Boxplot of the level of expression of the *RIPK2* gene (first hit of the TNBC vs. normal comparison). (C) Boxplot of the level of expression of the *PLK1* gene (first hit of the TNBC vs. ER- / HER2+ comparison). From left to right: TNBC are in blue, TNBC cell lines in cyan, ER- / HER2+ tumors in red, Luminal A tumors in green, Luminal B tumors in orange and normal samples in grey.

A



B

symbol		ID	P.Value
CDC20	#	NM_001255.	$< 10^{-7}$
TTK	#	NM_003318.	$< 10^{-7}$
ANP32E	#	NM_030920.	$< 10^{-7}$
CDCA7	#	NM_145810.	$< 10^{-7}$
CENPF	#	NM_016343.	$< 10^{-7}$
KIF4A	#	NM_012310.	$< 10^{-7}$
MAD2L1	#	NM_002358.	$< 10^{-7}$
NUF2	#	NM_145697.	$< 10^{-7}$
KIF23	#	NM_138555.	$< 10^{-7}$
SMC4	#	NM_005496.	$< 10^{-7}$
TYMS	#	NM_001071.	$< 10^{-7}$
KIF14	#	NM_014875.	$< 10^{-7}$
CTPS	#	NM_001905.	$< 10^{-7}$

C

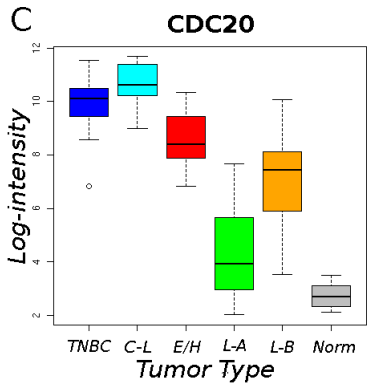


Figure 11.4: Successive filters. (A) Starting from the 10032 genes always above the background level, we first looked for genes over-expressed in TNBC vs. Luminal A and normal tissue and retrieved 834 genes. Out of the 834, 218 are over-expressed in TNBC vs. Luminal B and ER- / HER2+. Among those 218, we retrieved 102 that are also over-expressed in TNBC cell-lines. (B) The top 13 genes of this list of 102 genes. (C) Boxplot for the topmost gene of this list (*CDC20*). From left to right: TNBC are in blue, TNBC cell lines in cyan, ER- / HER2+ tumors in red, Luminal A tumors in green, Luminal B tumors in orange and normal samples in grey.

other pharmaceutical groups, easiness of validation. This scoring results from an intense search in the literature with the PubMed search engine. It sums up all the knowledge on the mechanism, structure, involvement in cancer and breast cancer of the considered potential target. The scoring is also based on the knowledge of registered patents that are accessible through various specialized search engines, such as Google patent. Once all this information is collected, targets are prioritized for further experimental validation in cell-lines before potential drug discovery programs.

Apart from target assessment, some results of this analysis have been studied more precisely in collaboration with various groups of the Institut Curie and some have already been published (Marty et al. (2008); Lizárraga et al. (2009), see subsection 11.1.3 and 11.1.4 respectively).

11.1.3 Paper: Frequent PTEN genomic alterations

We analyzed the oncogenic phosphatidylinositol 3-kinase (PI3K) pathway in 13 TNBC samples, and compared this pathway with a control series of 11 ER- / HER2+ carcinoma samples. These samples come from the first study of the Curie-Servier collaboration (see subsection 8.2). We analyzed the DNA copy number profiles, the gene expression profiles and the proteomic data generated using both Reverse Phase Protein Arrays (RPPA) and immunohistochemistry. The effect of the PI3K pathway inhibition on proliferation and apoptosis was further analyzed in several TNBC cell lines.

The PI3K pathway was found to be activated in TNBC and up-regulated compared to ER- / HER2+ tumors as shown by a significantly increased activation of the downstream targets Akt and mTOR (mammalian target of rapamycin). We linked this activation to a decrease in PTEN protein expression and the loss of the *PTEN* gene at the DNA copy number level. Interestingly, both PI3K and mTOR inhibitors led to TNBC cell growth arrest but only PI3K inhibition lead to cell death.

For this paper, I mainly took part in the transcriptomic and genomic analysis of the data. For the transcriptomic data, one of the questions was whether it was possible to distinguish between ER- / HER2+ and TNBC based on their transcription profiles. Using hierarchical clustering on the gene expression profiles of these 24 tumors, I showed that ER- / HER2+ tumors and TNBC were segregated in two different clusters (see Figure 1 (c) on page 4 of the paper). Thus, the two breast

cancer populations were accurately characterized and the subtypes identified by immunohistochemistry corresponded to the gene expression classification.

Using RPPA, it was shown that PTEN expression was low in TNBC compared to HER2+ carcinomas. We thus examined whether variations in PTEN protein expression could arise from genomic alterations in our TNBC samples. Genomic DNA isolated from tumors was analyzed on SNP arrays. Using those data and ITALICS (Rigaill et al., 2008), I showed that *PTEN* is lost in 50% of TNBC (see the paragraph "DNA and RNA microarray analysis" at page 5, paragraph "Genomic alterations at the PTEN tumor suppressor [..]" at page 6 and Figure 4 page 9 of the paper). Moreover, the measured copy number for *PTEN* correlated with PTEN protein level in a significant manner (with a p-value of 0.028, see page 7). These results suggest that genomic alterations at the *PTEN* locus are directly responsible for low PTEN protein expression in about 50% of TNBC (see Figure 4b page 9). Altogether, our data demonstrated a PTEN-dependent activation of Akt in TNBC.

Research article

Open Access

Frequent *PTEN* genomic alterations and activated phosphatidylinositol 3-kinase pathway in basal-like breast cancer cells

Bérengère Marty[1], Virginie Maire[1], Eléonore Gravier[1,2,3,6], Guillem Rigaill[1,7], Anne Vincent-Salomon[4], Marion Kappler[1], Ingrid Lebigot[4], Fathia Djelti[1], Audrey Tourdès[1], Pierre Gestraud[3,6], Philippe Hupé[3,5,6], Emmanuel Barillot[3,6], Francisco Cruzalegui[8], Gordon C Tucker[8], Marc-Henri Stern[9], Jean-Paul Thiery[1,10], John A Hickman[8] and Thierry Dubois[1]

[1]Département de Transfert, Institut Curie, 26 rue d'Ulm, 75005 Paris, France
[2]Département de Biostatistiques, Institut Curie, 26 rue d'Ulm, 75005 Paris, France
[3]INSERM U900, Institut Curie, 26 rue d'Ulm, 75005 Paris, France
[4]Service de Pathologie, Institut Curie, 26 rue d'Ulm, 75005 Paris, France
[5]CNRS UMR144, Institut Curie, 26 rue d'Ulm, 75005 Paris, France
[6]Ecole des Mines de Paris, 77300 Fontainebleau, France
[7]Unité de Mathématiques et Informatique Appliquées, UMR518, AgroParisTech/INRA, 75005 Paris, France
[8]Institut de Recherches Servier, 125 Chemin de Ronde, 78290 Croissy sur Seine, France
[9]INSERM U830, Institut Curie, 26 rue d'Ulm, 75005 Paris, France
[10]Current address: Institute of Molecular and Cell Biology, 61 Biopolis Drive (Proteos), 138673 Singapore

Corresponding author: Thierry Dubois, thierry.dubois@curie.fr

Received: 28 Aug 2008 Revisions requested: 9 Oct 2008 Revisions received: 22 Oct 2008 Accepted: 3 Dec 2008 Published: 3 Dec 2008

Breast Cancer Research 2008, **10**:R101 (doi:10.1186/bcr2204)
This article is online at: http://breast-cancer-research.com/content/10/6/R101

Abstract

Introduction Basal-like carcinomas (BLCs) and human epidermal growth factor receptor 2 overexpressing (HER2+) carcinomas are the subgroups of breast cancers that have the most aggressive clinical behaviour. In contrast to HER2+ carcinomas, no targeted therapy is currently available for the treatment of patients with BLCs. In order to discover potential therapeutic targets, we aimed to discover deregulated signalling pathways in human BLCs.

Methods In this study, we focused on the oncogenic phosphatidylinositol 3-kinase (PI3K) pathway in 13 BLCs, and compared it with a control series of 11 hormonal receptor negative- and grade III-matched HER2+ carcinomas. The two tumour populations were first characterised by immunohistochemistry and gene expression. The PI3K pathway was then investigated by gene copy-number analysis, gene expression profiling and at a proteomic level using reverse-phase protein array technology and tissue microarray. The effects of the PI3K inhibition pathway on proliferation and apoptosis was further analysed in three human basal-like cell lines.

Results The PI3K pathway was found to be activated in BLCs and up-regulated compared with HER2+ tumours as shown by a significantly increased activation of the downstream targets Akt and mTOR (mammalian target of rapamycin). BLCs expressed significantly lower levels of the tumour suppressor PTEN and PTEN levels were significantly negatively correlated with Akt activity within that population. PTEN protein expression correlated significantly with *PTEN* DNA copy number and more importantly, reduced *PTEN* DNA copy numbers were observed specifically in BLCs. Similar to human samples, basal-like cell lines exhibited an activation of PI3K/Akt pathway and low/lack PTEN expression. Both PI3K and mTOR inhibitors led to basal-like cell growth arrest. However, apoptosis was specifically observed after PI3K inhibition.

Conclusions These data provide insight into the molecular pathogenesis of BLCs and implicate the PTEN-dependent activated Akt signalling pathway as a potential therapeutic target for the management of patients with poor prognosis BLCs.

AFA: alcohol, formalin and acetic acid; BLC: basal-like breast carcinoma; BSA: bovine serum albumin; CK: cytokeratin; CN: DNA copy number; EGFR: epidermal growth factor receptor; ER: oestrogen receptor; HER2: human epidermal growth factor receptor 2; HER2+ carcinomas: HER2 overexpressing carcinomas; HES: haematoxylin-eosin-safran; IC_{50}: inhibition concentration 50%; mTOR: mammalian target of rapamycin; MTT: 3-(4,5-dimethylthiazol-2-yl)-2,5 diphenyltetrazolium bromide; PBS: phosphate buffered saline; PI3K: phosphatidylinositol 3-kinase; PIK3CA: PI3K p110 subunit alpha; PIP3: phosphatidylinositol-3,4,5-trisphosphate; PR: progesterone receptor; PTEN: phosphatase and tensin homolog deleted on chromosome 10; RPPA: reverse phase protein array; SNP: single nucleotide polymorphism; TBS: tris buffer saline; TBST: TBS + 0.1% tween-20; TBST-BSA: TBST + 5% BSA; TMA: tissue microarray.

Introduction

Gene expression profiling has enabled the identification of five subgroups of breast cancer characterised by different clinical outcomes and responses to therapy [1-10]. Among them, basal-like carcinomas (BLC) and human epidermal growth factor receptor 2 overexpressing (HER2+) carcinomas are associated with the worst prognosis [6,10,11]. BLCs are highly proliferative, genetically unstable, poorly differentiated, often grade III carcinomas [12,13] and preferentially metastase in the brain and lungs [14]. They are identified by immunohistochemistry as triple negative (lack of HER2 and oestrogen/progesterone receptor (ER/PR) expression) and positive for basal cytokeratins (CK5/6 and/or CK14) and/or epidermal growth factor receptor (EGFR) expression [8,15]. BLCs represent about 15% of cases of breast cancer and appear to be prevalent in pre-menopausal African American woman (39%) [16].

Patients with BLCs are treated exclusively with conventional therapy. Although they show high rates of objective initial response, the majority of patients do not have a complete, prolonged response, and they have a poorer prognosis than those within other breast tumour subgroups [12,13]. In contrast to HER2+ carcinomas treated with targeted therapy such as anti-HER2 [17], there is no available targeted therapy for BLCs. However, in patients with triple-negative breast cancer, some treatments are in preclinical trials, such as Dasatinib, a Src tyrosine kinase inhibitor, Cetuximab or Bevacizumab, which target EGFR and vascular endothelial growth factor, respectively [18]. Little is known about the pathogenesis of BLCs in spite of the recent genome and transcriptome microarray profiling [14,15,19,20]. Proteomics in tandem with genomic/transcriptomic analysis is essential to clarify the molecular pathology of BLCs and to discover druggable targets [21,22].

In order to identify such targets, we are exploring the phosphoproteome of BLCs to highlight deregulated signalling pathways. In this report, we have investigated the oncogenic phosphatidylinositol 3-kinase (PI3K) pathway in BLCs and compared it with that of HER2+ carcinomas in which it is known to be up-regulated [23-25]. Phosphatidylinositol-3,4,5-trisphosphate (PIP3) is an important lipid second messenger in tumourigenesis, in particular by activating Akt, which binds to membrane-associated PIP3 through its plekstrin homology domain, and other signalling molecules involved in a variety of cellular events, such as survival, proliferation, cell motility and invasion [26]. PI3K is activated downstream of extracellular signals and phosphorylates phosphatidylinositol-4,5-bisphosphate to generate PIP3. The tumour suppressor PTEN (phosphatase and tensin homologue deleted on chromosome 10) catalyses the opposite reaction, thereby reducing the pool of PIP3, inhibiting growth and survival signals, and suppressing tumour formation [27,28]. The PI3K signalling pathway is frequently deregulated in human solid tumours including breast cancers through Akt1 or PIK3CA (catalytic subunit of PI3K) mutations, HER2 overexpression and PTEN loss or mutation [24,25,29-34].

In this report, we demonstrate that the PI3K pathway is activated in BLCs. The PI3K pathway was up-regulated in BLCs compared with HER2+ carcinomas as shown by a significant increased activation of downstream targets such as Akt and mTOR (mammalian target of rapamycin). We also describe the molecular mechanism leading to this PI3K pathway activation, which occurs through a low PTEN protein expression that was found to be associated with genomic alterations at the *PTEN* locus, specifically in BLCs. In addition, we observed that basal-like cell lines exhibited an activation of Akt and a low/lack of PTEN expression. The exposure of basal-like cell lines to PI3K or mTOR inhibitors led to cell growth arrest. However, apoptosis was detected when PI3K, but not mTOR, was inhibited. Altogether, our data demonstrate a PTEN-dependent up-regulated PI3K pathway in BLCs and suggest this pathway as a therapeutic target for patients with poor prognosis BLCs.

Materials and methods

Immunohistochemistry

Twenty-four tumours were obtained from patients treated at the Curie Institute (Biological Resource Centre, Paris, France). Immunohistochemistry was performed as previously described [35]. Tumours contained between 50% and 90% tumour cells revealed by haematoxylin-eosin-safran (HES) staining.

For phospho-Akt (S473) staining, tissue microarrays (TMA) containing alcohol, formalin and acetic acid (AFA)-fixed paraffin-embedded tissue were made. For each biopsy, three representative tumour areas and one peritumoural tissue were carefully selected from a HES-stained section of a donor block. Using a specific arraying device (Manual Tissue Arrayer; Beecher Instruments, Sun Prairie, WI, USA) core cylinders of 1 mm in diameter were punched from each of those four areas and placed into recipient paraffin blocks. Sections of 3 μm were cut, placed onto positively charged slides (capillary gap microscope slides, Dako REAL, Dako, Trappes, France) and dried at 58°C for one hour. Sections were deparaffinised in toluene and hydrated in graded alcohol. Antigen retrieval was performed in 10 mM sodium citrate (pH 6.10) for 20 minutes at 95°C. Sections were then cooled for 20 minutes at room temperature. Endogenous biotins were blocked by Biotin blocking system (Dako, Trappes, France).

After washes in PBS-Tween buffer, endogenous peroxidase activity was quenched with 3% hydrogen peroxide for 5 minutes then rinsed in distilled water. Each tissue section was blocked with a solution of PBS (pH 7.4) containing 1% of BSA and 1.4% of normal horse serum for 5 minutes, followed by an overnight incubation at 4°C with primary antibody against phospho-Akt (S473). After washes, slides were incubated with rabbit biotinylated antibody (Jackson Immunoresearch,

Interchim, Clichy, France) for 30 minutes. Immunostaining was revealed using the Vectastain ABC peroxidase system (Vector Laboratories, Abcys, Paris, France) using diaminobenzidine as a chromogen. Slides were counter-stained with haematoxylin before mounting. The reactions were carried out using an automated stainer (LabVision, Thermo Scientific, Microm France, Francheville, France) except for the primary antibody. Omission of the primary antibody was used as a negative control. Immunohistochemistry conditions were first optimised using cell pellets from cell lines known to be positive or negative for phospho-Akt staining.

Positive nuclear staining for ER and PR were recorded in accordance with standardised guidelines, using 10% as the cut-off for ER- and PR-positive cells. For HER2, only staining of membranes was considered with a 30% cut-off as recommended [36]. The cut-off for CK5/6, CK14 and EGFR positivity was 10% of stained cells (weak or strong) for the results shown in Figure 1a.

EGFR (clone 31G7, 1:40 dilution, Zymed, Invitrogen, Cergy-Pontoise, France), CK5/6 (clone D5/16B4, 1:50 dilution, Dako, Trappes, France), CK14 (clone LL002, prediluted, Biogenex, San Ramon, CA, USA) and phospho-Akt (S473) (clone 736E11, 1:50 dilution, Cell Signaling Technology, Ozyme, Saint Quentin en Yveline, France) antibodies were used.

Tumour lysis

Frozen tumours were incubated with a lysis buffer containing 50 mM Tris (pH 6.8), 2% sodium dodecyl sulfate (SDS), 5% glycerol, 2 mM 1,4-dithio-DL-threitol (DTT), 2.5 mM ethylenediaminetetraacetic acid, 2.5 mM ethylene glycol tetraacetic acid, 2 mM sodium orthovanadate, 10 mM sodium fluoride and a cocktail of protease (Roche, Meylan, France) and phosphatase (Pierce, Perbio, Brebières, France) inhibitors. Homogenisation was obtained using a TissueLyser (Qiagen, Courtaboeuf, France) with stainless steel beads 5 mm in diameter (Qiagen, Courtaboeuf, France) for two to three minutes at 30 Hz. Lysates were boiled at 100°C for 10 minutes to inactivate proteases and phosphatases. Protein concentration was determined using the BCA Protein Assay Kit-Reducing Agent Compatible (Pierce, Perbio, Brebières, France). Lysates were then stored at -80°C.

Reverse phase protein array

We developed a robust reverse phase protein array (RPPA) technology allowing the printing of very small quantities of protein (about 1 ng per spot) convenient for the analysis of minimal quantities of biopsy material. This miniaturised dot-blot technology is based on robotic printing of a large number of different cell/tissue lysates onto nitrocellulose bound to histology slides and the analysis of proteins of interest with highly specific antibodies [37,38]. Five two-fold serial dilutions were made from each lysate in 96-well plates (conical bottom, 50 μl/well) and spotted in triplicates onto nitrocellulose-coated glass slides (FAST slides, Whatman, Schleicher & Schuell, Maidstone, Kent, UK) by using a MicroGrid Compact arrayer (BioRobotics, Dutscher Scientific Instrumentation, Brumath, France) with SMP3XB pins (Tip diameter = 75 μm, volume-spot = 1.2 nl; Telechem, Proteigene, Saint Marcel, France).

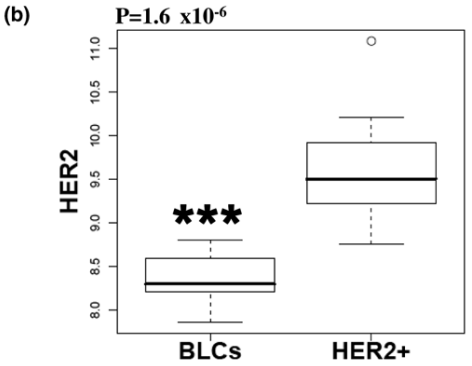
To avoid evaporation during spotting, the humidity was kept at about 50% to 60% in the array chamber with a humidification control unit. After printing, slides coated with two nitrocellulose pads were incubated with avidin, biotin and peroxydase blocking reagents (Dako, Trappes, France) before saturation with TBS containing 0.1% Tween-20 and 5% BSA (TBST-BSA). Each pad was then probed overnight at 4°C with primary antibodies (or without primary antibodies, for negative controls) at the appropriate dilution in TBST-BSA. After washes with TBST, arrays were probed with horseradish peroxidase secondary antibodies (Jackson ImmunoResearch Laboratories, Interchim, Clichy, France) diluted in TBST-BSA for one hour at room temperature. To amplify the signal, slides were incubated with Bio-Rad Amplification Reagent (BAR solution) supplied in the Western blot amplification module (Bio-Rad, Marnes la Coquette, France) for 10 minutes at room temperature. The arrays were washed with TBST containing 10% dimethyl sulfoxide (DMSO) for two minutes, then with TBST. To detect the bound biotin, slides were probed with Cy5-Streptavidin (Jackson ImmunoResearch Laboratories, Interchim, Clichy, France) diluted in TBST-BSA for one hour at room temperature. The processed slides were scanned using a GenePix 4000B microarray scanner (Molecular Devices, Saint Grégoire, France). Double staining was performed to quantify actin expression for the normalisation between samples using anti-beta-actin primary antibodies (Sigma-Aldrich, Saint Quentin Fallavier, France) and Cy3 secondary antibodies (Jackson ImmunoResearch Laboratories, Interchim, Clichy, France).

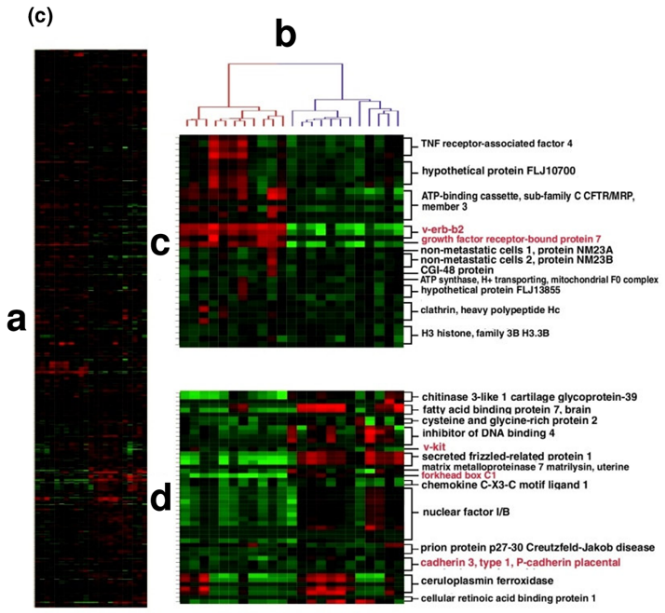
Specificity of each primary antibody used in this study was first validated by Western blotting on several cell and tumour lysates (not shown). Optimal dilution was determined for each antibody with different cell lysates using specific software developed at the Curie Institute with the following criteria: signal away from the negative control without saturation and correlation with Western blotting. Spot detection and quantification were determined with MicroVigene software (VigeneTech Inc, Carlisle, MA). Akt phospho-Akt (S473), PTEN and stathmin antibodies (Cell Signaling Technology, Ozyme, Saint Quentin en Yveline, France) were utilised at a dilution of 1:1000, 1:250, 1:200 and 1:100, respectively. HER2 antibodies (clone MS-432-P1, Ab11) used at 1:500 dilution were from Lab Vision (Interchim, Clichy, France). mTOR expression and phosphorylation was not examined by RPPA due to the poor specificity of mTOR antibodies.

Figure 1

(a)

Antibodies	IHC staining	BLCs (n=13)	HER2+ (n=11)
ER	-	13	11
	+	0	0
PR	-	13	11
	+	0	0
HER2	-	13	0
	+	0	11
CK5/6	-	5	10
	+	8	1
CK14	-	5	10
	+	8	1
EGFR	-	1	7
	+	12	4





Characterisation of basal-like carcinomas (BLCs) and human epidermal growth factor receptor overexpressing (HER2+) carcinomas from human biopsies. (a) Selection and characterisation of human samples by immunohistochemistry (IHC). Thirteen BLCs were first selected as grade III triple-negative ductal carcinomas (negative for oestrogen receptor (ER), progesterone receptor (PR) and HER2 expression) and then characterised for cytokeratin 5/6 (CK5/6), CK14 and epidermal growth factor receptor (EGFR) staining. The control series was composed of 11 hormone receptor negative- and grade III-matched HER2+ carcinomas. Tumours contained between 50% and 90% tumour cells revealed by haematoxylin-eosin-safran staining. (b) Higher HER2 protein expression in HER2+ carcinomas compared with BLCs. The box plot illustrates the expression of total HER2 protein expression measured by reverse phase protein array (RPPA) in human BLCs and HER2+ carcinomas. An outlier is present within the HER2+ population (open circle). The y axis represents logarithmic transformed HER2 relative quantification. The p value (*** $p < 0.001$) is represented (Mann-Whitney test). Data are representative of three separate RPPA experiments. (c) Tumours selected by immunohistochemistry clustered according to their gene expression signature. A hierarchical clustering was performed on the intrinsic/UNC genes as described [46]. (i) Overview of complete cluster diagram. Each row represents a gene and each column a human tumour. Black is for no change, red for up-regulation and green for down-regulation of gene expression. (ii) Experimental sample-associated dendrogram. Red dendrogram branch represents HER2+ carcinomas and blue designs basal-like carcinomas. (iii) HER2+ signature. A HER2+ expression cluster was observed and contained multiples genes from the 17q11 amplicon including HER2 and growth factor receptor-bound protein 7 (GRB7) (typed in red) as previously described [46]. (iv) Basal-like signature. A basal-like expression cluster was found and contained genes previously identified to be characteristic of basal epithelial cells such as v-kit, FOXC1 and P-cadherin (written in red) [46].

Western-blotting

Tissue lysates (10 μg/lane) were loaded onto 10% or four 12% Bis-Tris Criterion XT gels (Bio-Rad, Marnes la Coquette, France) and migration was performed using MOPS buffer (Bio-Rad, Marnes la Coquette, France). Proteins were then transferred to nitrocellulose (Bio-Rad, Marnes la Coquette, France). Membranes were saturated with TBST-BSA and then incubated overnight at 4°C with primary antibodies at the appropriate dilution in TBST-BSA. After washes, membranes were incubated with horseradish peroxidase secondary antibodies (Jackson ImmunoResearch Laboratories, Interchim, Clichy, France) for one hour at room temperature. Bound antibodies on immunoblots were visualised on membranes with a chemoluminescent detection system (ECL; Amersham Pharmacia Biotech, Orsay, France). Quantification was performed using a LAS-3000 Luminescent Image analyser and Image Gauge software (Fuji, FSVT, Courbevoie, France). Actin was detected for normalisation between samples using anti-beta-actin primary antibodies at the dilution of 1:5000 (Sigma-Aldrich, Saint Quentin Fallavier, France). Akt, phospho-Akt

(S473), mTOR, phospho-mTOR (S2448), PTEN and cleaved-PARP antibodies (Cell Signaling Technology, Ozyme, Saint Quentin en Yveline, France) were used at 1:1000 dilution. HER2 antibodies (clone MS-432-P1, Ab11) were used at a 1:500 dilution (Lab Vision, Interchim, Clichy, France).

DNA and RNA microarray analysis

DNA and RNA were purified as described [20]. For genomic arrays, Affymetrix GeneChip Human Mapping 100 K was normalised and analysed using ITALICS (ITerative and Alternative normaLIzation and Copy number calling for affymetrix Snp arrays) algorithm [39]. The segmentation of the genomic profile was performed using GLAD (Gain and Loss Analysis of DNA) software [40]. The forceGL parameter was set to 0.28. Single nucleotide polymorphisms with smoothing value lower and greater than 2 ± 0.28 were considered as loss and gain, respectively. After RNA quality control, 12 of the 13 BLCs and the 11 HER2+ carcinomas were hybridised onto U133 plus 2.0 Affymetrix chips. Transcriptomic data were normalised using GC-RMA [41]. Raw and normalised transcriptomic data are publically available at Gene Expression Omnibus (Accession number: [GSE13787]) and at the Curie Institute microarray dataset repositories [42].

Cell culture

The cell lines were obtained from the American Type Culture Collection (LGC Promochem, Molsheim, France) and from the European Collection of Animal Cell Cultures (Sigma-Aldrich, Saint Quentin Fallavier, France). HCC38 and HCC1937 were maintained in RPMI-1640 with 10% FBS, 1.5 g/L sodium bicarbonate, 10 mM Hepes and 1 mM sodium pyruvate. BT20 were cultured in Eagle's minimal essential medium containing 10% FBS, 1.5 g/L sodium bicarbonate, 0.1 mM non-essential amino acids and 1 mM sodium pyruvate. MDA-MB-468 were grown with RPMI with 10% FBS. MDA-MB-453 were cultured without carbon dioxide in Leibovitz's L-15 medium containing 10% FBS and 10 mM HEPES. SKBr3 were grown with McCoy5a containing 10% FBS and A431 with Eagle's minimal essential medium containing 10% FBS and 0.1 mM non-essential amino acids. A431 cells were either or not stimulated with 50 ng/ml EGF for five minutes after overnight serum starvation. Lysates were prepared at 60% to 90% cell confluency and analysed by Western blotting.

Cell proliferation assay

To test the effect of LY294002 and rapamycin on cell proliferation, cells were seeded into 96-well plates at a density determined on the basis of the growth characteristics of each cell line (750 cells/well for MDA-MB-468 and HCC1937; 1500 cells/well for BT20). Forty-eight hours later, cells (triplicate wells) were treated for seven days with varying concentration of LY294002 (Sigma-Aldrich, Saint Quentin Fallavier, France), rapamycin (Cell Signaling Technology, Ozyme, Saint Quentin en Yveline, France) or DMSO (Sigma-Aldrich, Saint Quentin Fallavier, France) as a control. LY094002 concentrations tested were 0.39, 0.78, 1.56, 3.12, 6.25, 12.5, 25 and 50 μM. Rapamycin concentrations analysed were 0.49, 0.98, 1.95, 3.91, 7.81, 15.62, 31.25, 62.5, 125 and 250 nM.

The relative percentages of metabolically active cells compared with untreated controls were determined on the basis of mitochondrial conversion of 3-(4,5-dimethylthiazol-2-yl)-2,5 diphenyltetrazolium bromide (MTT) to formazine using a MTT assay. To each well, 15 μL of MTT (5 mg/mL in PBS) was added. After four hours incubation at 37°C, floating plus adherent cells were lysed by the addition of 10% SDS in 10 mM hydrochloric acid. The absorbance was measured at the wavelength of 540 nm (Infinite 200, Tecan, Lyon, France) and results are presented as the percentage of control cell growth inhibition obtained from no treated cells grown in the same culture plate. The IC_{50}s were determined on the basis of the dose-response curves.

Apoptosis assays

Cells were harvested and seeded in 96-well plates (10 000 cells/well). After overnight growth, cells were treated in triplicate with various concentrations of LY294002, rapamycin or DMSO as a control. Twenty-four hours later, apoptosis was determined by caspase 3/7 activation and by the detection of PARP cleavage that serves as a marker of cells undergoing apoptosis. Caspase activity was determined using Caspase-Glo 3/7 luminescent assay (Promega, Charbonnières-les-bains, France) according to the manufacturer's instructions. Results are presented as caspase 3/7 activity normalised by caspase 3/7 activity from vehicle-treated cells. For PARP cleavage, Western blot was performed using whole protein lysates of floating plus adherent cells. Blots were incubated with a specific cleaved-PARP antibody (Cell Signaling Technology, Saint Quentin en Yveline, France).

Statistical analysis

As data did not display a normal distribution, a non-parametric test was performed. Mann-Whitney test was used to assess differential expression of a protein between the two groups (BLCs and HER2+). The R software v2.4.0 was used for statistical analyses [43]. A Spearman correlation test was performed to estimate a rank-based measure of association between two parameters. Values were log transformed. p values under 5% were considered significant. For the apoptosis assays, p values were calculated using Student's *t* test.

Results and discussion

Tumour selection and characterisation

The PI3K pathway was examined in two populations of highly proliferative, grade III, hormone receptor-negative invasive breast carcinomas. We chose this comparison, rather than that of BLCs with normal tissue, to compare two types of proliferating cells, avoiding a comparison with a largely differentiated, quiescent population. Thirteen BLCs were selected by immunohistochemistry as triple-negative ductal carcinomas

(lack of ER, PR and HER2 staining) that expressed CK5/6 and/or CK14 and/or EGFR (Figure 1a). The comparison series was composed of 11 patients with ER-negative/PR-negative and HER2+ tumours (Figure 1a). CK5/6 was expressed in 61.5% BLCs (8 of 13) and 9.1% HER2+ (1 of 11) (Figure 1a). Similarly, CK14 was expressed more in the same BLCs (61.5%) than in HER2+ (9.1%) (Figure 1a). EGFR was detected in 92.3% BLCs (12 of 13) and 36.4% HER2+ (4 of 11) (Figure 1a), in agreement with previous studies showing EGFR expression in most BLCs and in HER2+ carcinomas [8,44,45]. Expectedly, RPPA analysis confirmed a significantly higher HER2 protein expression in HER2+ carcinomas compared with BLCs ($p = 1.6 \times 10^{-6}$) (Figure 1b). Similar results were observed by Western blotting and significantly correlated with those obtained by RPPA [see Additional data file 1]. Of note, some BLCs carcinomas expressed HER2 protein but at lower levels than those observed in HER2+ carcinomas. In addition, these data indicated that RPPA technology could be useful to measure in a quantitative manner the expression of HER2 protein in human samples. Gene expression microarray analysis confirmed that the tumours clustered according to basal-like and HER2+ signatures [46] (Figure 1c). Therefore, the two breast cancer populations were accurately characterised and the subtypes identified by immunohistochemistry corresponded to the gene expression classification.

Activated PI3K pathway in basal-like breast cancer

Proteomic analysis was then continued by RPPA allowing analysis of a very limited amount of sample from biopsies [33,37,38]. Akt was expressed at similar levels in BLCs and HER2 carcinomas (Figure 2a) whereas the phosphorylated and active form of Akt (S473) tended to be expressed more in BLCs although not in a significant manner (Figure 2b). Akt activity, defined as the phospho/total ratio, was significantly increased in BLCs compared with HER2+ population ($p = 0.026$) (Figure 2c). Similar data, significantly correlated with RPPA data, were obtained by Western blotting [see Additional data file 2] and were in agreement with those showing an activation of Akt within a population of eight triple-negative carcinomas [47].

Our data further revealed that Akt was more active in BLCs compared with HER2+ carcinomas where Akt is known to be activated through HER2 overexpression [23-25]. We verified by immunohistochemistry of both BLCs and HER2+ carcinomas that the active form of Akt was expressed in tumour cells, with a plasma membrane localisation observed in tumours showing strong phospho-Akt immunoreactivity [see Additional data file 3]. We also examined the phosphorylation status of the target of rapamycin, mTOR, particularly at the S2448 residue known to be phosphorylated through PI3K/Akt signalling pathway activation. mTOR was expressed at similar levels in the two breast populations but was significantly more active (phospho/total ratio) in BLCs than in HER2+ carcinomas ($p = 0.015$) (Figure 2d,e,f), where mTOR has been shown to be activated [48]. The PI3K pathway was up-regulated in BLCs compared with HER2+ as shown by the significant activation of downstream targets such as Akt and mTOR.

Lower PTEN expression in basal-like breast cancer compared to HER2+ carcinomas

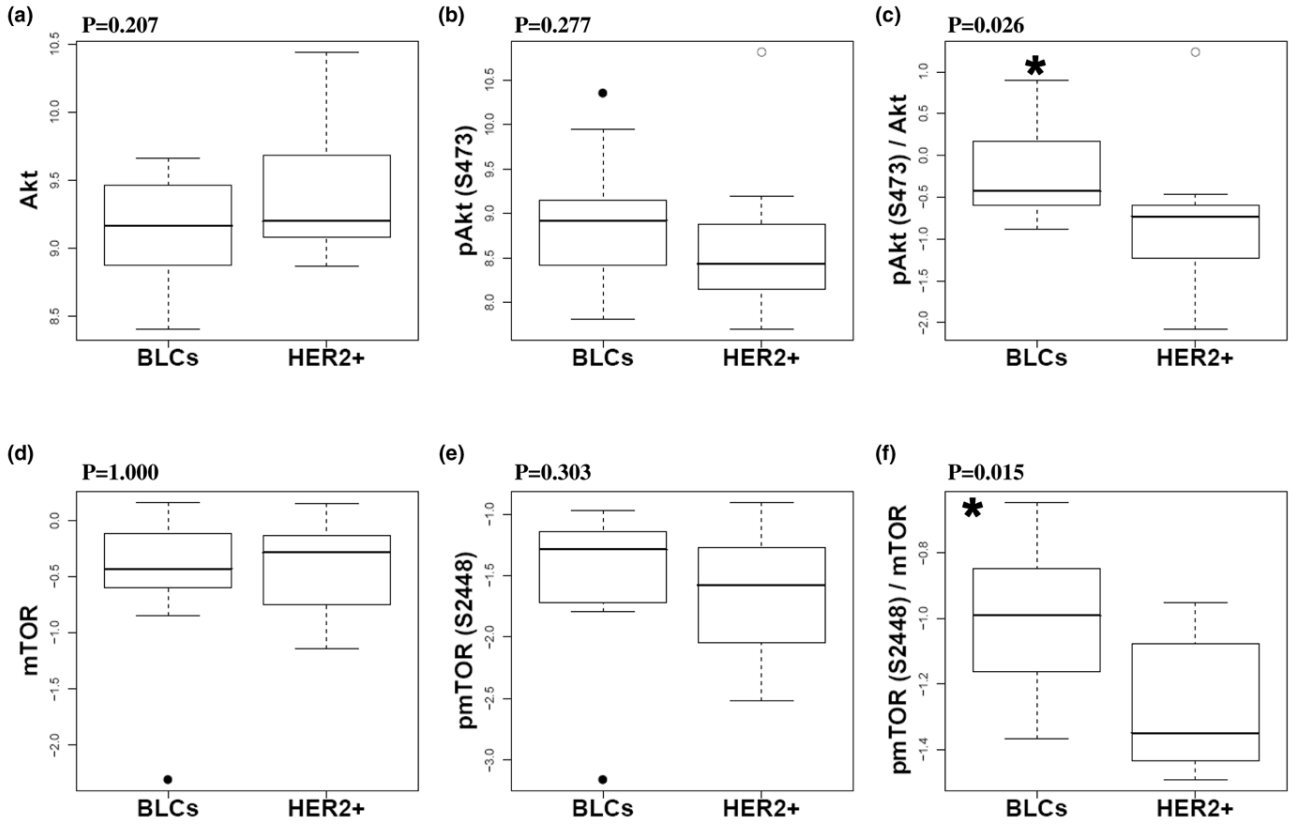
We then attempted to characterise the molecular mechanism(s) leading to Akt activation in BLCs. We evaluated PTEN expression because its loss has been associated with ER negative [24,49,50] and CK5/14-positive breast cancer [34]. RPPA analysis highlighted a lower expression of PTEN protein in BLCs compared with HER2+ carcinomas in a significant manner ($p = 0.002$) (Figure 3a). Similar data were obtained when PTEN was detected by Western blotting and significantly correlated with RPPA data [see Figures a and b in Additional data file 4]. So far, we failed to estimate PTEN level by immunohistochemistry, possibly because of the PTEN antibodies we tested and/or the AFA fixation of tissues. Lower PTEN expression in BLCs was also detected at the mRNA level ($p = 0.0002$) (Figure 3b).

In agreement with a previous report with PTEN protein levels measured by immunohistochemistry [49], PTEN mRNA and protein levels were well correlated ($p = 0.0002$) (Figure 3c) [see Figure c in Additional data file 4], indicating that we could estimate PTEN protein levels from transcriptomic analysis. Our analysis of published data [51] showed that lower PTEN mRNA levels in BLCs compared with normal samples ($p = 0.003$, Mann-Whitney test, data not shown), suggesting lower PTEN protein levels in BLCs compared with normal tissues. We examined the expression of stathmin, which has recently been shown to be overexpressed in low PTEN expressing breast cancers [49]. In accordance with these published observations, stathmin protein was overexpressed in BLCs compared with HER2+ carcinomas ($p = 0.018$) (Figure 3d). Stathmin therefore represents a potential marker for PTEN-dependent PI3K pathway activation [49]. Altogether, transcriptomic and proteomic analyses highlighted low expression of PTEN in BLCs.

Genomic alterations at the *PTEN* tumour suppressor gene in basal-like breast cancer

We then examined whether variations in PTEN protein expression could arise from genomic alterations in our BLC population. Genomic DNA isolated from tumours was analysed on SNP arrays. The two populations behaved differently for *PTEN* DNA copy-number (CN) in a significant manner ($p = 0.005$) (Figure 4a) [see Additional data file 5]. In contrast to the entire HER2+ population exhibiting normal *PTEN* CN, loss of *PTEN* CN was observed in 46.1% (6 of 13) BLCs (Figure 4a) [see Additional data file 5]. Of note is that our BLC population included one *BRCA1* tumour (c.2501delG *BRCA1* mutation) which also presented a loss of *PTEN* CN. We noticed that the only double deletion of the *PTEN* gene was observed in a BLC patient with a normal status of *BRCA1* with the exception of

Figure 2



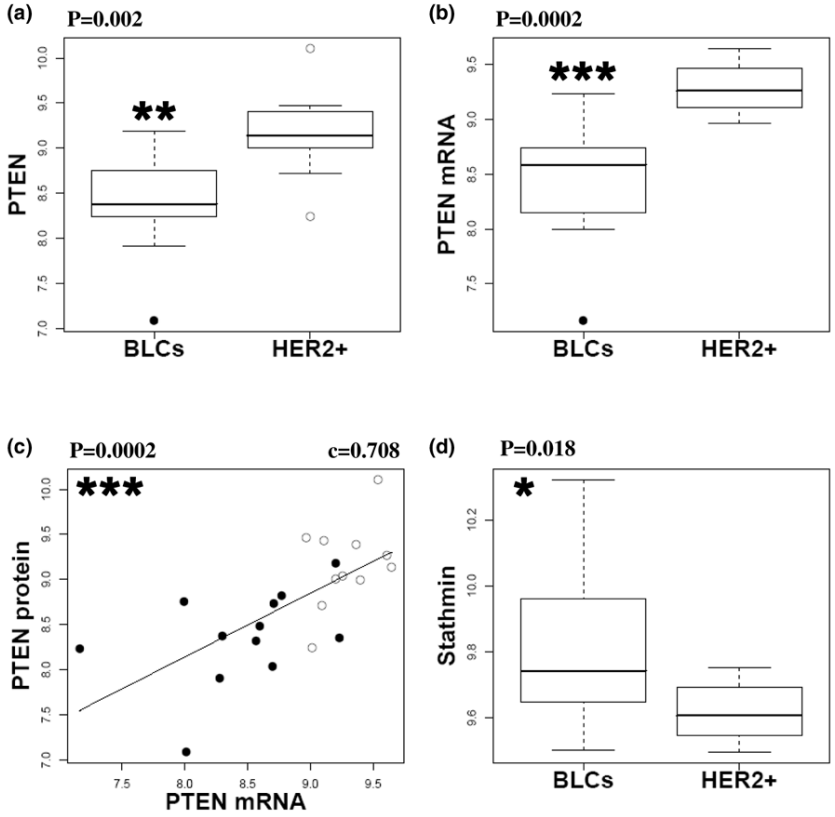
Up-regulated phosphatidylinositol 3-kinase (PI3K) signalling pathway in human basal-like breast cancers. Akt is activated in basal-like carcinomas (BLCs). The expression of (a) total Akt and (b) phosphorylated/active form of Akt (phospho-Akt (S473)) were measured by reverse phase protein array (RPPA) as well as Akt activity determined as the (c) ratio 'phospho/total' in human BLCs and human epidermal growth factor receptor overexpressing (HER2+) carcinomas. mTOR is activated in BLCs. Box plots show the (d) expression of mTOR and (e) its form phosphorylated via the PI3 kinase/Akt signalling pathway (phospho-mTOR (S2448)) determined by Western blotting as well as (f) mTOR activity (phospho/total ratio) in human BLCs and HER2+ carcinomas. Outliers are shown for BLCs (solid circles) and HER2+ carcinomas (open circles). The y axes represent logarithmic transformed relative quantifications. p values (* $p < 0.05$) are represented (Mann-Whitney test). Data are representative of four and two separate experiments for RPPA and Western-blot, respectively.

the c.4039A>G polymorphism. We also observed a gain of *PTEN* CN in 2 of 13 BLCs (15.4%) (Figure 4a) [see Additional data file 5] but these two tumours expressed PTEN protein at a level similar to that one in BLCs with normal *PTEN* CN (Figure 4b).

Importantly, *PTEN* CN correlated with PTEN protein level in a significant manner ($p = 0.028$) in the whole population (Figure 4b). These results suggest that genomic alterations at the *PTEN* locus are directly responsible for low PTEN protein expression in about 50% of BLCs (Figure 4b). Low PTEN protein expression in the other half of BLCs may result from *PTEN* promoter methylation and/or *PTEN* mutation. Although coding mutations of *PTEN* were thought to be rare in breast cancer, *PTEN* nucleotide sequence mutations have recently been detected exclusively in PTEN-null non-hereditary breast cancer [34]. However, we did not detect any *PTEN* mutation in our series of 13 BLCs (data not shown), in agreement with a recent report showing that the rare PTEN mutations observed in breast cancer (2.3%) were restricted to hormone receptor-positive carcinomas [33]. Therefore, low PTEN protein expression in the 50% BLCs with no *PTEN* CN loss may arise from epigenetic modifications.

In addition, by analysing a public data set generated from 42 BLCs and 32 hormone receptors-positive luminal A breast carcinomas [52], we also found a loss of *PTEN* CN, mainly in BLCs, and a correlation between *PTEN* CN and PTEN mRNA in the entire population ($p = 3.25 \times 10^{-7}$, $c = 0.614$, Spearman correlation, data not shown). In conclusion, we demonstrate the presence of genomic alterations at the *PTEN* locus specifically in BLCs. Our findings indicate that alteration of *PTEN* gene is not restricted to BRCA1-associated hereditary tumours (mostly corresponding to a specific basal-like sub-

Figure 3



Lower phosphatase and tensin homolog deleted on chromosome 10 (PTEN) expression in human basal-like cancers (BLCs) compared with human epidermal growth factor receptor overexpressing (HER2+) carcinomas. (a) Lower PTEN protein levels in BLCs compared with HER2+ carcinomas. PTEN protein level was quantified by reverse phase protein array (RPPA). Outliers are shown for BLCs (solid circles) and HER2+ carcinomas (open circles). Data are representative of four separate experiments. (b) Lower mRNA PTEN level (probeset 225363_at) in BLCs compared with HER2+ carcinomas. An outlier is present in BLC population (solid circles). (c) Correlation between PTEN protein measured by RPPA and PTEN messenger in the entire tumour population. Linear regression, Spearman correlation c and p value are presented. BLCs (solid circles) and HER2+ carcinomas (open circles) are shown. (d) Stathmin is overexpressed in BLCs compared with HER2+ carcinomas. Box plots indicate the levels of stathmin protein measured by RPPA within the two populations. Data are representative of four separate experiments. P values are shown (a,b,d: Mann-Whitney test). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Protein and mRNA relative quantifications were logarithmic transformed.

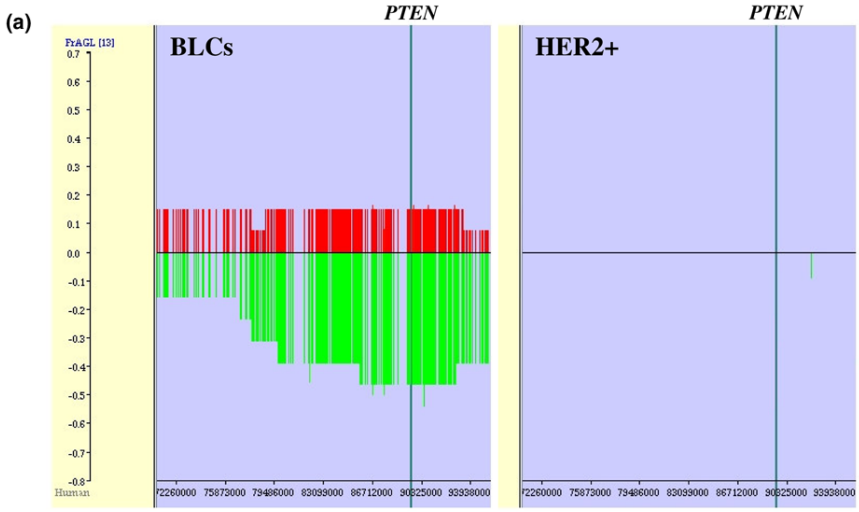
group) as recently suggested [34], but could be extended to the entire BLC population. These genetic modifications may drive to an aberrant PTEN-dependent signalling pathway in the whole BLC population.

PTEN-dependent activation of Akt in basal-like breast cancer

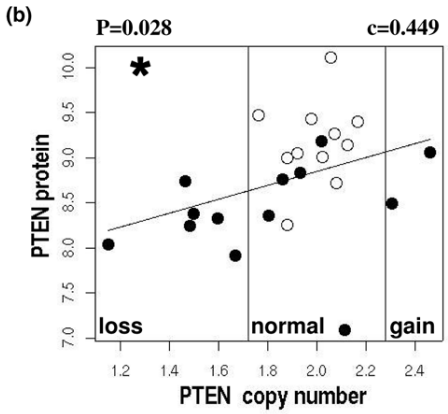
Low PTEN expression may therefore be responsible for Akt activation in BLCs. Indeed, data obtained by RPPA demonstrated that Akt activity correlated negatively with PTEN expression levels in BLCs ($p = 0.036$) (Figure 5) but not in HER2+ carcinomas (not shown). Similar conclusions arose from Western blot analysis [see Figure d in Additional data file 4]. Altogether, our data demonstrated a PTEN-dependent activation of Akt in BLCs, consistent with recent work showing higher phospho-Akt levels in PTEN-low compared with PTEN-high breast cancers [33].

We can not rule out the hypothesis that Akt could be activated through multiple mechanisms in BLCs, and not only through low PTEN expression. For example, transcriptomic microarray analysis revealed that the type II inositol polyphosphate-4-phosphatase mRNAs were expressed at significantly lower levels in BLCs compared with HER2+ human tumours (INPP4B reporter 205376_at from Affymetrix; $p = 0.0001$, data not shown). As INPP4B has been shown to negatively regulate Akt activity [53], its lower expression may represent an alternative pathway for Akt activation in BLCs. However, we could not test this hypothesis at a proteomic level because of the poor quality of the INPP4B antibody available. Mutations of *PIK3CA*, although more frequent in hormone receptor-positive tumours (34.5%) and HER2+ carcinomas (22.7%) occurs in BLCs (8.3%) and could represent another way to activate the PI3K signalling pathway in these tumours [33].

Figure 4

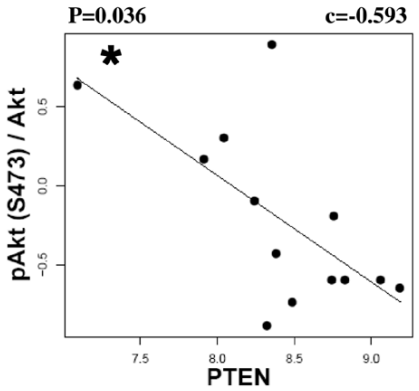


P=0.005 **	*PTEN* copy-number		
	Loss	Normal	Gain
HER2+ n=11	0	11 (100%)	0
BLCs n=13	6 (46.1%)	5 (38.5%)	2 (15.4%)



Loss of *phosphatase and tensin homolog deleted on chromosome 10 (PTEN)* DNA copy-number (CN) in human basal-like breast cancers. (a) Basal-like carcinomas (BLCs) and human epidermal growth factor receptor overexpressing (HER2+) carcinomas behaved differently for *PTEN* CN in a significant manner. Recurrent DNA CN alterations were observed around the *PTEN* gene (between 72,260,000 and 93,93,000 bp of chromosome 10) in BLCs compared with HER2+ carcinomas. Frequencies of genome CN gain (red) and loss (green) were calculated using the FrAGL (Frequency of Amplicon, Gain and Loss) option of VAMP software (Visualisation and Analysis of array-CGH, transcriptome and other Molecular Profiles) [63]. The vertical blue bar represents *PTEN* position from 89,613,175 to 89,718,511 bp. Percentages of tumours with loss, normal or gain of *PTEN* CN are presented within the two populations in the table. p value is shown (** $p < 0.01$, fisher exact test). (b) Correlation between PTEN protein level and *PTEN* DNA CN. PTEN protein level was quantified as in Figure 3a. Linear regression, Spearman correlation c and p value (* $p < 0.05$) are presented. BLCs (solid circles) and HER2+ carcinomas (open circles) are shown. The two vertical black lines ($X = 2 \pm 0.28$) separate loss/normal/gain *PTEN* CN (forceGL parameter: 0.28).

Figure 5



Phosphatase and tensin homolog deleted on chromosome 10 (PTEN)-dependent activation of Akt in human basal-like breast cancers. PTEN protein levels are correlated negatively with Akt activity in human basal-like cancer (BLC). Akt activity and PTEN protein levels were measured as in Figures 2c and 3a, respectively. BLCs (solid circles), linear regression, Spearman correlation c and p value (* $p < 0.05$) are shown.

PI3K but not mTOR inhibition induces apoptosis in basal-like cell lines

Akt activity was examined by Western blotting in four human basal-like cell lines (BT20, HCC38, HCC1937 and MDA-MB-468), one HER2+ (SKBr3) and one luminal (MDA-MB-453) human breast cell lines as well as in an epidermoid carcinoma cell line (A431) for a control (Figure 6a). Akt was phosphorylated indicating that PI3K pathway was activated in all breast cell lines analyzed (Figure 6a). PTEN was weakly expressed (BT20) or not detectable specifically in basal-like cell lines (Figure 6a). We noticed highest levels of Akt phosphorylation in MDA-MB-453 and BT20 (Figure 6a), and this may result from the mutation of the PI3K catalytic subunit (PIK3CA) reported in these two cell lines [24,33]. *PTEN* has been shown to be mutated in MDA-MB-468 [33]. Therefore, similar results were obtained from human biopsies and cell lines revealing an activation of Akt associated with a low/lack expression of PTEN in the basal-like population.

We then investigated whether the inhibition of the PI3K pathway altered proliferation and apoptosis of basal-like cell lines. First, we examined the growth inhibition response of three basal-like cell lines (BT20, HCC1937 and MDA-MB-468) treated with the PI3K inhibitor LY294002 or the mTOR inhibitor rapamycin. Exposure to LY294002 induced an inhibition of the proliferation for all three cell lines with a lower IC_{50} for MDA-MB-468 (IC_{50} = 7.6 ± 1.4 μM) compared with HCC1937 (IC_{50} = 14.5 ± 3.8 μM) and BT20 (IC_{50} = 13.3 ± 2.8 μM) (Figure 6b). The IC_{50} were in the same range than those obtained previously for MDA-MB-468 (IC_{50} = 9.5 μM) [54] and for other breast cell lines (2 to 20 μM LY294002) [55]. MDA-MB-468 cells were the most sensitive cells to LY294002 in agreement with the idea that PTEN mutation render cells more sensitive to growth inhibition by that inhibitor [33]. Exposure to rapamycin led to a growth inhibition that was not complete. The IC_{50} for rapamycin were not reached for HCC1937 and BT20 cell lines. MDA-MB-468 cells were the most sensitive cells to rapamycin with an IC_{50} = 1.2 ± 0.5 nM (Figure 6b). Similar data have been published previously for MDA-MB-468 cells (IC_{50} = 1 nM) [56,57].

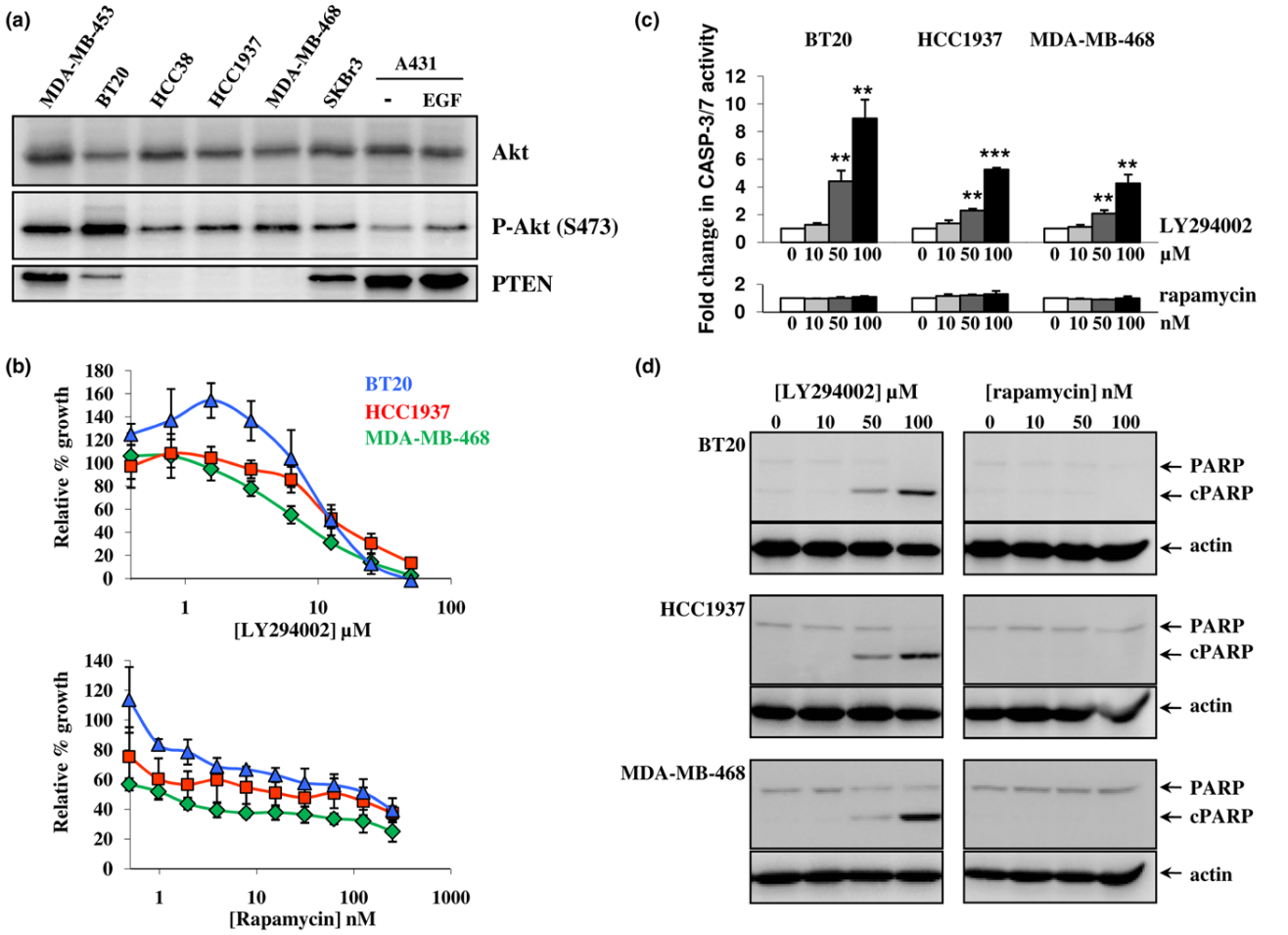
We next evaluated whether the growth inhibition resulted from apoptosis. Basal-like cell lines were treated with concentrations of inhibitors used to induce apoptosis, that is 50 to 100 μM LY294002 or 100 nM rapamycin [55-57]. Apoptosis was analysed 24 hours later by measuring casapase 3/7 activity (Figure 6c) and PARP cleavage (Figure 6d). In contrast to rapamycin, LY294002 treatment-induced apoptosis in all basal-like cell lines as judged by a rapamycin dose-dependent increased of caspase 3/7 activity (Figure 6c) and PARP cleavage (Figure 6d). These data are in agreement with a recent paper showing that LY294002 treatment, but not rapamycin, induced apoptosis in other breast cell lines [55]. It is likely that rapamycin inhibited basal-like cell proliferation by arresting the cell cycle in the G1 phase as reported for other breast cell lines [56].

In conclusion, exposure of basal-like cell lines to PI3K or mTOR inhibitors led to cell growth arrest but apoptosis was only observed in cells treated with LY294002. The inhibition of PI3K will directly affect Akt activity, which is involved in cell death and survival through several targets such as Bad, whereas the inhibition of mTOR, which acts downstream of Akt, is expected to inhibit proliferation but not apoptosis [28]. Moreover, the inhibition of mTOR may contribute to an unexpected activation of Akt through a negative feedback loop [58,59]. In order to bypass feedback loops, it may be more efficient to target PI3K or Akt than inhibiting mTOR. In contrast to LY294002, which broadly acts on the majority of PI3Ks and other related kinases [60], inhibitors of specific PI3K isoforms were recently identified [61]. In breast cell lines, PTEN loss was shown to sensitise to p110 beta inhibitors, a ubiquitously expressed class IA PI3K isoform [61]. Moreover, the inhibition of p110 beta was shown to block the tumourigenesis caused by *PTEN* loss in prostate [62]. Although further work is needed, these observations suggest that p110 beta may represent an attractive target for the treatment of patients with low PTEN expressing carcinomas such as BLCs.

Conclusion

Significant differences of protein expression patterns were observed between BLCs and HER2+ carcinomas, two types of highly proliferative breast cancers. Our data demonstrate that: the PI3K pathway is activated in BLCs and, to a higher extent than in HER2+ carcinomas, is known to have up-regulated Akt and mTOR activities; BLCs express less PTEN compared with HER2+ carcinomas and normal tissues; genomic

Figure 6



Phosphatidylinositol 3-kinase (PI3K) and mTOR inhibitors inhibit basal-like cell line proliferation whereas apoptosis is induced only by PI3K inhibition. (a) Akt activation is associated with low/lack of phosphatase and tensin homolog deleted on chromosome 10 (PTEN) expression in human basal-like cell lines. The expression of Akt, phospho-Akt (S473) and PTEN were analysed by Western blotting in four basal-like (BT20, HCC38, HCC1937 and MDA-MB-468), one human epidermal growth factor receptor overexpressing (HER2+) (SKBr3) and one luminal (MDA-MB-453) human breast cell lines as well as in epidermal growth factor stimulated (EGF) or not (-) A431 cells. (b) PI3K and mTOR inhibition induce cell growth arrest of basal-like cell lines. BT20 (blue triangle), HCC1937 (red square) and MDA-MB-468 (green diamond) cells were exposed continuously for seven days to increasing concentrations of LY294002 (upper panel) or rapamycin (lower panel). Growth was assessed by 3-(4,5-dimethylthiazol-2-yl)-2,5 diphenyltetrazolium bromide (MTT) dye conversion and presented as the percentage of control cell growth inhibition obtained from DMSO-treated cells. The x axes represent logarithmic transformed concentration of drugs. (c,d) The inhibition of PI3K, but not mTOR, induces apoptosis in basal-like cell lines. BT20, HCC1937 and MDA-MB-468 were exposed to varying concentrations of LY294002 (0 to 100 µM) or rapamycin (0 to 100 nM) for 24 hours and apoptosis was detected by measuring (c) caspase3/7 activity and the (d) cleavage of PARP (cPARP). (c) Caspase 3/7 activity was normalised by caspase 3/7 activity from vehicle-treated cells. (d) Actin was used as a loading control. The data represented the (b,c) average of three separate experiments performed in triplicates or are representative of (a,d) three separate experiments. Error bars represent standard deviation and p values (** $p < 0.01$, *** $p < 0.001$) were calculated by using Student's *t* test.

alterations at the *PTEN* locus are specifically found in BLCs; low PTEN expression in BLCs is associated with lost of *PTEN* DNA CN; Akt activity is dependent of PTEN expression in BLCs; similarly to human biopsies, basal-like breast cell lines exhibit low PTEN expression and activated Akt; PI3K or mTOR inhibition induced growth arrest in basal-like cell lines; PI3K inhibition, but not mTOR inhibition, induced apoptosis of basal-like cell lines; and finally that RPPA is a powerful quantitative tool for proteomic analysis and to examine signalling pathways in human tumours. Our study provides insight into the molecular pathology of BLCs with therapeutic implications and encourages the targeting of key players within the PI3K pathway, such as specific PI3K/Akt isoforms for the management of patients with poor prognosis BLC.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BM developed and performed the RPPA experiments and analyses. VM made the *in vitro* functional analyses and PTEN sequencing. EG, GR, PH, PG and EB contributed to statistical and bioinformatics analyses and development of informatics tools. AVS and IL collected, selected and provided human samples including DNA, RNA and biopsies. AVS and MK were involved in immunohistochemistry and TMA studies. VM, FD and AT performed the experiments in human breast cell lines. TD conceived the study design and performed the Western blot experiments from human samples. AVS, BM, FC, GCT, JPT and JAH participated in the design of the study. BM, VM, AVS, MK, FC, GCT, MHS, JAH and TD contributed to the biological interpretations of data. AVS, MHS and JPT provided expertise in clinical breast oncology. MHS and GR made the analysis of PTEN copy number from Affymetrix SNP data. BM and TD were responsible of manuscript editing. VM, EG, GR, AVS, MK, IL, MHS and JAH contributed to the preparation and corrections of the manuscript. All authors read and approved the final manuscript.

Additional files

The following Additional files are available online:

Additional file 1

A PDF containing figures showing the expression of HER2 measured by Western blotting and its correlation with RPPA data. Figure a illustrates the expression of total HER2 protein expression measured by Western blotting in human BLCs and HER2+ carcinomas. P value (*** $p < 0.001$) is represented (Mann-Whitney test). Figure b illustrates the correlation between RPPA and Western-blotting analysis for HER2 protein expression. Protein expressions are logarithmic transformed and illustrated by box plots with p values (Mann-Whitney test). Outliers are shown within the BLCs (solid circles) and HER2+ carcinomas (open circles) populations. The correlations are estimated using the Spearman correlation test (c) from logarithmic transformed values. Linear regression and p values are indicated. BLCs (solid circles) and HER2+ carcinomas (open circles) are shown. The significant p values are represented by stars (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).
See http://www.biomedcentral.com/content/supplementary/bcr2204-S1.pdf

Additional file 2

A PDF containing figures showing Akt activation in basal-like breast cancer measured by Western blotting and its correlation with RPPA data. Figures a, b and c illustrate the expression of total Akt, the phosphorylated/active form of Akt (phospho-Akt (S473)), and the activity of Akt determined as the 'phospho/total' ratio, respectively, in human BLCs and HER2+ carcinomas. Figures d and e show the correlation between RPPA and Western blotting data for Akt and phospho-Akt protein expressions, respectively. Protein expressions are logarithmic transformed and illustrated by box plots with p values (Mann-Whitney test). Outliers are shown within the BLCs (solid circles) and HER2+ carcinomas (open circles) populations. The correlations are estimated using the Spearman correlation test (c) from logarithmic transformed values. Linear regression and p values are indicated. BLCs (solid circles) and HER2+ carcinomas (open circles) are shown. The significant p values are represented by stars (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).
See http://www.biomedcentral.com/content/supplementary/bcr2204-S2.pdf

Additional file 3

A PDF containing a figure showing that the active form of Akt is detected in tumour cells within the biopsies by immunohistochemistry. Expression and localisation of phospho-Akt (S473), and hence activated Akt, was analysed on TMA in BLCs and HER2+ carcinomas. Phospho-Akt showed low, medium and high expression depending on the tumour samples. Phospho-Akt is preferentially expressed in tumour cells. It is located in the cytoplasm and at the plasma membrane particularly in tumour cells with strong staining. All photomicrographs are of the same 40× magnification (scale bar, 20 μm).
See http://www.biomedcentral.com/content/supplementary/bcr2204-S3.pdf

Additional file 4

A PDF containing a figure that validates the PTEN-dependent activation of Akt observed by RPPA in basal-like cancer with Western blot technology. Figure a shows PTEN protein level. Data are representative of two separate experiments. Figure b indicates the correlation between RPPA and Western blotting analysis for PTEN protein expression. Figure c exhibits the correlation between PTEN protein measured by Western blotting and PTEN messenger (probeset 225363_at). Figure d shows that PTEN protein levels correlates negatively with Akt activity within the BLC population (solid circles). Akt activity and PTEN protein levels were measured by Western blotting as in Figure a to c in Additional file 2 and as in Figure a in Additional file 4, respectively. Protein expressions are logarithmic transformed and illustrated by box plots with p values (Mann-Whitney test). Outliers are shown within the BLCs (solid circles) and HER2+ carcinomas (open circles) populations. The correlations are estimated using the Spearman correlation test (c) from logarithmic transformed values. Linear regression and p values are indicated. BLCs (solid circles) and HER2+ carcinomas (open circles) are shown. The significant p values are represented by stars (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).
See http://www.biomedcentral.com/content/supplementary/bcr2204-S4.pdf

Additional file 5

A PDF containing a figure that illustrates recurrent genomic alterations around PTEN locus in human basal-like breast cancers. Genomic analysis using the VAMP software (Visualization and Analysis of array-CGH, transcriptome and other Molecular Profiles) [60] around the PTEN gene (between 80,676,000 and 98,604,000 bp of chromosome 10) are shown for all the 13 BLCs (upper panel) and the 11 HER2+ tumours (lower panel). Each row represents one tumour profile. Gains are in red, losses in green and absence of alterations in yellow. The vertical blue bars represent the position of PTEN from 89,613,175 to 89,718,511 bp.
See http://www.biomedcentral.com/content/supplementary/bcr2204-S5.pdf

Acknowledgements

This work was supported by Institut de Recherches Servier and Institut Curie. GR was supported by a grant of Institut National du Cancer (INCa). We thank Paul Delgado, Sophie Duminil, Michèle Galut and Blandine Massemin for the purification of DNA and RNA from biopsies (Service de Pathologie, Institut Curie), Cécile Reyes, Audrey Rapinat, Benoit Albaud, David Gentien, Jean-Claude Haw-king-chon, Dr Charles Decraene and Dr Jean-Philippe Meyniel for genome and transcriptome microarrays (Département de Transfert, Institut Curie), Aurélie Cédenot for immunohistochemistry staining (Service de Pathologie, Institut Curie), Patricia Legoix-Né (Plateforme de Génomique, Institut Curie) and Elodie Manié (INSERM U830, Institut Curie) for advice regarding PTEN sequencing, and Yann de Rycke for statistical advice (Département de Biostatistiques, Institut Curie). We are indebted to Dr Sergio Roman-Roman (Département de Transfert, Institut Curie), Professor Dominique Stoppa-Lyonnet (INSERM U830, Service de Pathologie, Institut Curie) and Dr Xavier Sastre (Service de Pathologie, Institut Curie). We thank Séverine Lair (INSERM U900, Institut Curie) for loading the transcriptomic data into the Curie Institute microarray dataset repository. Finally, we are grateful Dr Minzi Ruan and Min Ma for their help for RPPA quantification (VigeneTech, Inc).

References

1. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET: **Breast cancer classification and prognosis based on gene expression profiles from a population-based study.** *Proc Natl Acad Sci USA* 2003, **100:**10393-10398.
2. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proc Natl Acad Sci USA* 2003, **100:**8418-8423.
3. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Rijn M van de, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98:**10869-10874.
4. Perou CM, Sorlie T, Eisen MB, Rijn M van de, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
5. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT: **Gene expression predictors of breast cancer outcomes.** *Lancet* 2003, **361:**1590-1596.
6. van 't Veer LJ, Dai H, Vijver MJ van de, He YD, Hart AA, Mao M, Peterse HL, Kooy K van der, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530-536.
7. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, Hess KR, Stec J, Ayers M, Wagner P, Morandi P, Fan C, Rabiul I, Ross JS, Hortobagyi GN, Pusztai L: **Breast cancer molecular subtypes respond differently to preoperative chemotherapy.** *Clin Cancer Res* 2005, **11:**5678-5685.
8. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, Hernandez-Boussard T, Livasy C, Cowan D, Dressler L, Akslen LA, Ragaz J, Gown AM, Gilks CB, Rijn M van de, Perou CM: **Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma.** *Clin Cancer Res* 2004, **10:**5367-5374.
9. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM, Esserman L, Albertson DG, Waldman FM, Gray JW: **Genomic and transcriptional aberrations linked to breast cancer pathophysiologies.** *Cancer Cell* 2006, **10:**529-541.
10. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schutz F, Goldstein DR, Piccart M, Delorenzi M: **Meta-analysis of gene-expression profiles in breast cancer: toward a unified understanding of breast cancer sub-typing and prognosis signatures.** *Breast Cancer Res* 2008, **10:**R65.
11. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365:**671-679.
12. Reis-Filho JS, Tutt AN: **Triple negative tumours: a critical review.** *Histopathology* 2008, **52:**108-118.

13. Fadare O, Tavassoli FA: **Clinical and pathologic aspects of basal-like breast cancers.** *Nat Clin Pract Oncol* 2008, **5:**149-159.
14. Smid M, Wang Y, Zhang Y, Sieuwerts AM, Yu J, Klijn JG, Foekens JA, Martens JW: **Subtypes of breast cancer show preferential site of relapse.** *Cancer Res* 2008, **68:**3108-3114.
15. Kreike B, van Kouwenhove M, Horlings H, Weigelt B, Peterse H, Bartelink H, Vijver MJ van de: **Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas.** *Breast Cancer Res* 2007, **9:**R65.
16. Couzin J: **Cancer research. Probing the roots of race and cancer.** *Science* 2007, **315:**592-594.
17. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, Gianni L, Baselga J, Bell R, Jackisch C, Cameron D, Dowsett M, Barrios CH, Steger G, Huang CS, Andersson M, Inbar M, Lichinitser M, Lang I, Nitz U, Iwata H, Thomssen C, Lohrisch C, Suter TM, Ruschoff J, Suto T, Greatorex V, Ward C, Straehle C, McFadden E, *et al.*: **Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer.** *N Engl J Med* 2005, **353:**1659-1672.
18. Rakha EA, Reis-Filho JS, Ellis IO: **Basal-like breast cancer: a critical review.** *J Clin Oncol* 2008, **26:**2568-2581.
19. Yehiely F, Moyano JV, Evans JR, Nielsen TO, Cryns VL: **Deconstructing the molecular portrait of basal-like breast cancer.** *Trends Mol Med* 2006, **12:**537-544.
20. Vincent-Salomon A, Gruel N, Lucchesi C, MacGrogan G, Dendale R, Sigal-Zafrani B, Longy M, Raynal V, Pierron G, de Mascarel I, Taris C, Stoppa-Lyonnet D, Pierga JY, Salmon R, Sastre-Garau X, Fourquet A, Delattre O, de Cremoux P, Aurias A: **Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity.** *Breast Cancer Res* 2007, **9:**R24.
21. Cleator S, Heller W, Coombes RC: **Triple-negative breast cancer: therapeutic options.** *Lancet Oncol* 2007, **8:**235-244.
22. Rodriguez-Pinilla SM, Sarrio D, Moreno-Bueno G, Rodriguez-Gil Y, Martinez MA, Hernandez L, Hardisson D, Reis-Filho JS, Palacios J: **Sox2: a possible driver of the basal-like phenotype in sporadic breast cancer.** *Mod Pathol* 2007, **20:**474-481.
23. Tokunaga E, Kimura Y, Mashino K, Oki E, Kataoka A, Ohno S, Morita M, Kakeji Y, Baba H, Maehara Y: **Activation of PI3K/Akt signaling and hormone resistance in breast cancer.** *Breast Cancer* 2006, **13:**137-144.
24. Saal LH, Holm K, Maurer M, Memeo L, Su T, Wang X, Yu JS, Malmstrom PO, Mansukhani M, Enoksson J, Hibshoosh H, Borg A, Parsons R: **PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma.** *Cancer Res* 2005, **65:**2554-2559.
25. Tokunaga E, Oki E, Kimura Y, Yamanaka T, Egashira A, Nishida K, Koga T, Morita M, Kakeji Y, Maehara Y: **Coexistence of the loss of heterozygosity at the PTEN locus and HER2 overexpression enhances the Akt activity thus leading to a negative progesterone receptor expression in breast carcinoma.** *Breast Cancer Res Treat* 2007, **101:**249-257.
26. Shaw RJ, Cantley LC: **Ras, PI(3)K and mTOR signalling controls tumour cell growth.** *Nature* 2006, **441:**424-430.
27. Engelman JA, Luo J, Cantley LC: **The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism.** *Nat Rev Genet* 2006, **7:**606-619.
28. Hennessy BT, Smith DL, Ram PT, Lu Y, Mills GB: **Exploiting the PI3K/AKT pathway for cancer drug discovery.** *Nat Rev Drug Discov* 2005, **4:**988-1004.
29. Liu W, Bagaitkar J, Watabe K: **Roles of AKT signal in breast cancer.** *Front Biosci* 2007, **12:**4011-4019.
30. Dillon RL, White DE, Muller WJ: **The phosphatidyl inositol 3-kinase signaling network: implications for human breast cancer.** *Oncogene* 2007, **26:**1338-1345.
31. Carpten JD, Faber AL, Horn C, Donoho GP, Briggs SL, Robbins CM, Hostetter G, Boguslawski S, Moses TY, Savage S, Uhlik M, Lin A, Du J, Qian YW, Zeckner DJ, Tucker-Kellogg G, Touchman J, Patel K, Mousses S, Bittner M, Schevitz R, Lai MH, Blanchard KL, Thomas JE: **A transforming mutation in the pleckstrin homology domain of AKT1 in cancer.** *Nature* 2007, **448:**439-444.
32. Kim MS, Jeong EG, Yoo NJ, Lee SH: **Mutational analysis of oncogenic AKT E17K mutation in common solid cancers and acute leukaemias.** *Br J Cancer* 2008, **98:**1533-1535.
33. Stemke-Hale K, Gonzalez-Angulo AM, Lluch A, Neve RM, Kuo WL, Davies M, Carey M, Hu Z, Guan Y, Sahin A, Symmans WF, Pusztai L, Nolden LK, Horlings H, Berns K, Hung MC, Vijver MJ van de, Valero V, Gray JW, Bernards R, Mills GB, Hennessy BT: **An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer.** *Cancer Res* 2008, **68:**6084-6091.
34. Saal LH, Gruvberger-Saal SK, Persson C, Lovgren K, Jumppanen M, Staaf J, Jonsson G, Pires MM, Maurer M, Holm K, Koujak S, Subramaniyam S, Vallon-Christersson J, Olsson H, Su T, Memeo L, Ludwig T, Ethier SP, Krogh M, Szabolcs M, Murty VV, Isola J, Hibshoosh H, Parsons R, Borg A: **Recurrent gross mutations of the PTEN tumor suppressor gene in breast cancers with deficient DSB repair.** *Nat Genet* 2008, **40:**102-107.
35. Azoulay S, Lae M, Freneaux P, Merle S, Al Ghuzlan A, Chnecker C, Rosty C, Klijanienko J, Sigal-Zafrani B, Salmon R, Fourquet A, Sastre-Garau X, Vincent-Salomon A: **KIT is highly expressed in adenoid cystic carcinoma of the breast, a basal-like carcinoma associated with a favorable outcome.** *Mod Pathol* 2005, **18:**1623-1631.
36. Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, Dowsett M, Fitzgibbons PL, Hanna WM, Langer A, McShane LM, Paik S, Pegram MD, Perez EA, Press MF, Rhodes A, Sturgeon C, Taube SE, Tubbs R, Vance GH, Vijver M van de, Wheeler TM, Hayes DF: **American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer.** *J Clin Oncol* 2007, **25:**118-145.
37. Gulmann C, Sheehan KM, Kay EW, Liotta LA, Petricoin EF 3rd: **Array-based proteomics: mapping of protein circuitries for diagnostics, prognostics, and therapy guidance in cancer.** *J Pathol* 2006, **208:**595-606.
38. Chan SM, Ermann J, Su L, Fathman CG, Utz PJ: **Protein microarrays for multiplex analysis of signal transduction pathways.** *Nat Med* 2004, **10:**1390-1396.
39. Rigaill G, Hupe P, Almeida A, La Rosa P, Meyniel JP, Decraene C, Barillot E: **ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays.** *Bioinformatics* 2008, **24:**768-774.
40. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E: **Analysis of array CGH data: from signal ratio to gain and loss of DNA regions.** *Bioinformatics* 2004, **20:**3413-3422.
41. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F: **A Model-Based Background Adjustment for Oligonucleotide Expression Arrays.** *J Am Stat Assoc* 2004, **99:**909-917.
42. **The Curie Institute microarray dataset repository** [http://microarrays.curie.fr/publications/transfert_signalisation/breast_transcriptome_P1/]
43. **The R Project for Statistical Computing** [http://www.r-project.org]
44. Livasy CA, Perou CM, Karaca G, Cowan DW, Maia D, Jackson S, Tse CK, Nyante S, Millikan RC: **Identification of a basal-like subtype of breast ductal carcinoma in situ.** *Hum Pathol* 2007, **38:**197-204.
45. Livasy CA, Karaca G, Nanda R, Tretiakova MS, Olopade OI, Moore DT, Perou CM: **Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma.** *Mod Pathol* 2006, **19:**264-271.
46. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, Nobel A, Parker J, Ewend MG, Sawyer LR, Wu J, Liu Y, Nanda R, Tretiakova M, Ruiz Orrico A, Dreher D, Palazzo JP, Perreard L, Nelson E, Mone M, Hansen H, Mullins M, Quackenbush JF, Ellis MJ, Olopade OI, Bernard PS, *et al.*: **The molecular portraits of breast tumors are conserved across microarray platforms.** *BMC Genomics* 2006, **7:**96.
47. Umemura S, Yoshida S, Ohta Y, Naito K, Osamura RY, Tokuda Y: **Increased phosphorylation of Akt in triple-negative breast cancers.** *Cancer Sci* 2007, **98:**1889-1892.
48. Zhou X, Tan M, Stone Hawthorne V, Klos KS, Lan KH, Yang Y, Yang W, Smith TL, Shi D, Yu D: **Activation of the Akt/mammalian target of rapamycin/4E-BP1 pathway by ErbB2 overexpression predicts tumor progression in breast cancers.** *Clin Cancer Res* 2004, **10:**6779-6788.
49. Saal LH, Johansson P, Holm K, Gruvberger-Saal SK, She QB, Maurer M, Koujak S, Ferrando AA, Malmstrom P, Memeo L, Isola J, Bendahl PO, Rosen N, Hibshoosh H, Ringner M, Borg A, Parsons R: **Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor**

pathway activity. *Proc Natl Acad Sci USA* 2007, **104:**7564-7569.
50. Depowski PL, Rosenthal SI, Ross JS: **Loss of expression of the PTEN gene protein product is associated with poor outcome in breast cancer.** *Mod Pathol* 2001, **14:**672-676.
51. Richardson AL, Wang ZC, De Nicolo A, Lu X, Brown M, Miron A, Liao X, Iglehart JD, Livingston DM, Ganesan S: **X chromosomal abnormalities in basal-like human breast cancer.** *Cancer Cell* 2006, **9:**121-132.
52. Adelaide J, Finetti P, Bekhouche I, Repellini L, Geneix J, Sircoulomb F, Charafe-Jauffret E, Cervera N, Desplans J, Parzy D, Schoenmakers E, Viens P, Jacquemier J, Birnbaum D, Bertucci F, Chaffanet M: **Integrated profiling of basal and luminal breast cancers.** *Cancer Res* 2007, **67:**11565-11575.
53. Barnache S, Le Scolan E, Kosmider O, Denis N, Moreau-Gachelin F: **Phosphatidylinositol 4-phosphatase type II is an erythropoietin-responsive gene.** *Oncogene* 2006, **25:**1420-1423.
54. DeGraffenried LA, Fulcher L, Friedrichs WE, Grunwald V, Ray RB, Hidalgo M: **Reduced PTEN expression in breast cancer cells confers susceptibility to inhibitors of the PI3 kinase/Akt pathway.** *Ann Oncol* 2004, **15:**1510-1516.
55. Heinonen H, Nieminen A, Saarela M, Kallioniemi A, Klefstrom J, Hautaniemi S, Monni O: **Deciphering downstream gene targets of PI3K/mTOR/p70S6K pathway in breast cancer.** *BMC Genomics* 2008, **9:**348.
56. Noh WC, Mondesire WH, Peng J, Jian W, Zhang H, Dong J, Mills GB, Hung MC, Meric-Bernstam F: **Determinants of rapamycin sensitivity in breast cancer cells.** *Clin Cancer Res* 2004, **10:**1013-1023.
57. Mondesire WH, Jian W, Zhang H, Ensor J, Hung MC, Mills GB, Meric-Bernstam F: **Targeting mammalian target of rapamycin synergistically enhances chemotherapy-induced cytotoxicity in breast cancer cells.** *Clin Cancer Res* 2004, **10:**7031-7042.
58. Wan X, Harkavy B, Shen N, Grohar P, Helman LJ: **Rapamycin induces feedback activation of Akt signaling through an IGF-1R-dependent mechanism.** *Oncogene* 2007, **26:**1932-1940.
59. O'Reilly KE, Rojo F, She QB, Solit D, Mills GB, Smith D, Lane H, Hofmann F, Hicklin DJ, Ludwig DL, Baselga J, Rosen N: **mTOR inhibition induces upstream receptor tyrosine kinase signaling and activates Akt.** *Cancer Res* 2006, **66:**1500-1508.
60. Gharbi SI, Zvelebil MJ, Shuttleworth SJ, Hancox T, Saghir N, Timms JF, Waterfield MD: **Exploring the specificity of the PI3K family inhibitor LY294002.** *Biochem J* 2007, **404:**15-21.
61. Torbett NE, Luna-Moran A, Knight ZA, Houk A, Moasser M, Weiss W, Shokat KM, Stokoe D: **A chemical screen in diverse breast cancer cell lines reveals genetic enhancers and suppressors of sensitivity to PI3K isotype-selective inhibition.** *Biochem J* 2008, **415:**97-110.
62. Jia S, Liu Z, Zhang S, Liu P, Zhang L, Lee SH, Zhang J, Signoretti S, Loda M, Roberts TM, Zhao JJ: **Essential roles of PI(3)K-p110beta in cell growth, metabolism and tumorigenesis.** *Nature* 2008, **454:**776-779.
63. La Rosa P, Viara E, Hupe P, Pierron G, Liva S, Neuvial P, Brito I, Lair S, Servant N, Robine N, Manie E, Brennetot C, Janoueix-Lerosey I, Raynal V, Gruel N, Rouveirol C, Stransky N, Stern MH, Delattre O, Aurias A, Radvanyi F, Barillot E: **VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles.** *Bioinformatics* 2006, **22:**2066-2073.

11.1.4 Paper: Formins regulate tumor cell invasion

This work was done in collaboration with the group of Philippe Chavrier (Ph.D., Institut Curie). We investigated the role of Diaphanous-related formins (DRF) in invadopodia formation and breast tumor cell invasion. Using small interfering RNA (siRNA) on highly-invasive MDA-MB-231 TNBC cell-line, it was shown that some members of the DRF family are required for invadopodia formation and two-dimensional matrix proteolysis. It was also shown that invasion of a three-dimensional Matrigel matrix involves filopodia-like protrusions and the formation of these filopodia depends on DRF. Overall, we show that DRF are critical components of the invasive apparatus of tumor cells in two-dimensional and three-dimensional matrices.

I participated in the transcriptomic analysis of the data. An important question was whether DRF are overexpressed in TNBC. My statistical analysis of the Curie-Servier dataset and another public dataset revealed a 2 to 2.5-fold over-expression of DRF2 and DRF3 transcripts in TNBC compared with normal human breast tissues (see page 9 of the paper). This result suggests that DRF do play a role in the invasion of real TNBC.

Diaphanous-Related Formins Are Required for Invadopodia Formation and Invasion of Breast Tumor Cells

Floria Lizárraga,[1,3] Renaud Poincloux,[1,3] Maryse Romao,[1,4] Guillaume Montagnac,[1,3] Gaëlle Le Dez,[1,3] Isabelle Bonne,[5] Guillem Rigaill,[2,6] Graça Raposo,[1,4] and Philippe Chavrier[1,3]

[1]Research Center and [2]Department of Translational Research, Institut Curie; [3]Membrane and Cytoskeleton Dynamics and [4]Structure and Membrane Compartments, Centre National de la Recherche Scientifique (CNRS), UMR144; [5]Institut Pasteur, Ultrastructural Microscopy Platform; and [6]UMR AgroParisTech/INRA 518, Statistics and Genome Team, Paris, France

Abstract

Proteolytic degradation of the extracellular matrix by metastatic tumor cells is initiated by the formation of invadopodia, i.e., actin-driven filopodia-like membrane protrusions endowed with matrix-degradative activity. A signaling cascade involving neural Wiskott-Aldrich syndrome protein and the Arp2/3 actin nucleating complex is involved in actin assembly at invadopodia. Yet, the mechanism of invadopodia formation is poorly understood. Based on their role as actin nucleators in cytoskeletal rearrangements, including filopodia formation, we examined the function of Diaphanous-related formins (DRF) in invadopodia formation and invasion by breast tumor cells. Using small interfering RNA silencing of protein expression in highly invasive MDA-MB-231 breast adenocarcinoma cells, we show that three members of the DRF family (DRF1–DRF3) are required for invadopodia formation and two-dimensional matrix proteolysis. We also report that invasion of a three-dimensional Matrigel matrix involves filopodia-like protrusions enriched for invadopodial proteins, including membrane type 1 matrix metalloproteinase, which depend on DRFs for their formation. These data identify DRFs as critical components of the invasive apparatus of tumor cells in two-dimensional and three-dimensional matrices and suggest that different types of actin nucleators cooperate during the formation of invadopodia. [Cancer Res 2009;69(7):2792–800]

Introduction

Tumor cell invasion across tissue boundaries and metastasis are dependent on the capacity of invasive cancer cells to breach the basement membrane (BM) and migrate through the three-dimensional interstitial collagen network (1, 2). One major route of invasion requires tumor cells to proteolytically cleave extracellular matrix (ECM) components via a mechanism involving matrix-degrading proteases (3). In particular, extracellular proteases belonging to the matrix metalloproteinase (MMP) family, including transmembrane membrane type 1 MMP (MT1-MMP), play a crucial role in cancer dissemination by degrading and remodeling ECM components (4–8).

Intriguingly, when invasive cancer cells are grown on a two-dimensional ECM substratum layered on glass, matrix proteolytic activity is restricted to invadopodia, which correspond to actin-rich finger-like structures protruding into the matrix and enriched in MT1-MMP (9–13). Neural Wiskott-Aldrich syndrome protein (N-WASP) and the Arp2/3 complex, both components of the actin polymerization machinery, are required for invadopodia formation and thus have been proposed to assemble actin filaments at invadopodia (14, 15). The cytoskeletal protein cortactin is also enriched at invadopodia and is critical for the formation and activity of these structures possibly through stabilization of the actin network and, as recently suggested, by controlling delivery/recruitment of MMPs at invadopodia (10, 11, 16–18). Furthermore, members of the Rho family of small GTPases are required for invadopodia formation (12, 15, 19). In particular, Cdc42 was shown to control the formation of invadopodia in human melanoma and rat mammary adenocarcinoma tumor cell lines through activation of the N-WASP/Arp2/3 complex cascade (14, 15), whereas we recently implicated Cdc42 and RhoA in the mechanism of invadopodia formation and MT1-MMP delivery in breast adenocarcinoma MDA-MB-231 cells (12). However, the complete machinery of invadopodia formation in cancer cells remains poorly understood. Based on the filopodia-like morphology of invadopodia (16, 20), it was postulated that formins might elongate actin filaments in invadopodia (14, 21).

Formins are filamentous actin (F-actin) nucleators that polymerize linear filaments through conserved formin homology domains (22). Among the formin family, Diaphanous-related formins (DRF) produce linear actin filaments that are the hallmarks of stress fibers and filopodia (23–27). In addition, DRFs are downstream effectors of active Rho GTPases, RhoA, and Cdc42 (22). Roles for DRF1 during formation of membrane protrusions by tumor cells and invasion have been recently reported (28, 29). However, the function of DRF proteins in invadopodia formation and in the acquisition of invasive phenotypes by cancer cells has not been thoroughly explored.

In this study, we assessed the contribution of DRF1, DRF2, and DRF3 to the invasion capacity of human MDA-MB-231 cells, a highly invasive cell line of basal-like breast tumor phenotype, the most aggressive form of breast cancers (30). Our data show a pivotal role of DRF proteins during invadopodia formation in two-dimensional and three-dimensional matrices for BM degradation and invasion by breast cancer cells.

Materials and Methods

Antibodies. Mouse monoclonal antibody for DRF1 was obtained from BD Biosciences. Goat polyclonal anti-DRF2 was purchased from Santa Cruz Biotechnology. Mouse MT1-MMP monoclonal antibody was a gift from

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

F. Lizárraga and R. Poincloux contributed equally to this work.

Requests for reprints: Philippe Chavrier, Institut Curie-Section Recherche, 26 rue d'Ulm, Paris, 75248, France. Phone: 33-156-246-359; Fax: 33-156-246-349; E-mail: Philippe.Chavrier@curie.fr.

doi:10.1158/0008-5472.CAN-08-3709

Dr. M.C. Rio (Institut de Génétique et de Biologie Moléculaire et Cellulaire). Monoclonal mouse anti–β-actin antibody (clone AC15) was from Sigma-Aldrich. Monoclonal anti-RhoA was provided by Dr. J. Bertoglio (Institut National de la Sante et de la Recherche Medicale U461). Monoclonal anti-cortactin (Clone 4F11) and anti-phosphotyrosine (pY; Clone 4G10) antibodies were obtained from Millipore. AlexaFluor-phalloidin, rabbit polyclonal anti-pY antibody, and antimouse IgG Alexa488 antibody were from Invitrogen. Horseradish peroxidase–conjugated and fluorescently conjugated secondary antibodies were from Jackson ImmunoResearch Laboratories.

Constructs. EGFP-mDia2 and cSRC-Y527F expression constructs were kind gifts of Drs. T. Svitkina (University of Pennsylvania) and M. Arpin (Institut Curie), respectively. MT1-MMP with internal pHLuorin was generated by PCR resulting in the insertion of pHLuorin (super ecliptic variant, a gift from Dr. T. Galli) between amino acids 534 and 535, NH_2 terminally to the transmembrane region. The sequence of Lifeact (31), kindly provided by Dr. R. Wedlich-Söldner (MPI Biochem), was fused to the aminoterminus of mCherry by PCR (mCh-Lifeact) with a GDPPVAT spacer and subcloned in pIRESpuro3 (Clontech).

Cell culture, transfections, and stable cell lines. Human breast adenocarcinoma cells MDA-MB-231 (American Type Culture Collection) were maintained in L-15 culture medium (Sigma-Aldrich) with 2 mmol/L glutamine and 15% FCS at 37°C in 1% CO_2.

For small interfering RNA (siRNA) treatment, MDA-MB-231 cells were treated with 10 to 100 nmol/L of specific siRNA (see Supplementary Table S1) with Oligofectamine (Invitrogen). Cells were analyzed after 72 h of treatment. MDA-MB-231 cells were transfected with expression constructs using Lipofectamine (Invitrogen). Cells were analyzed after 24 h of transfection. Stable lines of MDA-MB-231 cells expressing mCh-Lifeact alone or together with pHLuorin-MT1-MMP were selected with 1 μg/mL puromycin or 1 μg/mL puromycin, together with 0.5 mg/mL G418, respectively.

Reverse transcription–PCR. Total RNA was obtained using the RNeasy Mini kit from QIAGEN (Hilden). cDNA synthesis was carried out using SuperScript III Reverse Transcriptase enzyme (Invitrogen). PCR reactions were performed using Platinum Taq DNA Polymerase (Invitrogen). Primers are listed in Supplementary Table S2.

Fluorescent-gelatin degradation assay and quantification of invadopodia. MDA-MB-231 cells were incubated for 5 h on FITC-conjugated or AlexaFluor 594–conjugated gelatin (Invitrogen) to quantify gelatin degradation and were stained for F-actin and cortactin to identify invadopodia positive cells as described (12, 13). Statistical analysis was carried out using SigmaStat 3.5.

Indirect immunofluorescence analysis. MDA-MB-231 cells were cultured on gelatin-coated coverslip (Figs. 1 and 2) or on top of Matrigel (10 mg/mL, BD Biosciences; Fig. 4 and Supplementary Figs. S4 and S5). Cells were preextracted with 0.3% Triton-X100 in 4% PFA and processed for immunofluorescence analysis as described (12, 13). Cells were imaged with a DM6000 B/M microscope (Leica Microsystems; Figs. 1 and 2), or with a Leica DMRA2 microscope with 100× PL APO HCX, 1.4 NA objective equipped with a piezoelectric driver (0.2-μm increment, Physik Instrumente; Fig. 4 and Supplementary Figs. S4 and S5). Microscopes were equipped with a CoolSnapHQ camera (Roper Scientific) and steered by Metamorph 6 (Molecular Devices Corporation).

Live cell imaging. MDA-MB-231 cells expressing mCh-Lifeact were plated on FITC-gelatin coated glass-base dishes (Iwaki) and kept in a humidified atmosphere at 37°C and 1% CO_2. For Supplementary Videos S1 and S6, images were recorded using the 100× objective of a Leica DMIRE2 microscope equipped with a Cascade II camera (Roper Scientific). For Supplementary Videos S2 to S4, images of mCh-Lifeact and FITC-gelatin were recorded with the 60× objective of an automated Nikon TE2000-E microscope equipped with a CoolSnapHQ camera. To allow representative sampling of mock and siRNA-treated cell populations, six fields per condition were recorded simultaneously.

Scanning and transmission electron microscopy. The upper chamber of a Transwell cell culture insert (BD Biosciences) was filled with 100 μL of Matrigel, and cells were added in serum-free L15 medium. The lower chamber contained L15 medium with 15% FCS. For scanning electron microscopy, cells were prefixed in 2.5% glutaraldehyde/0.1 mol/L sodium cacodylate (pH 7.4). After postfixation in 1% osmium tetraoxide (in 0.2 mol/L cacodylate buffer), cells were dehydrated in a series of increasing ethanol concentrations and critical point dried using carbon dioxide. After coating with gold, cells were examined with a JEOL JSM-6700F scanning electron microscope. For transmission electron microscopy, 5 h after contact with Matrigel, cells were fixed overnight in 2.5% glutaraldehyde and 2% paraformaldehyde in 0.1 mol/L cacodylate buffer, postfixed with 2% OsO_4, dehydrated in ethanol, and embedded in Epon. Ultrathin sections were prepared with a Reichert Ultracut-E Microtome (Leica), counterstained with 2% uranyl acetate in 70% methanol, and viewed with a Philips CM120 Transmission Electron Microscope (FEI Company) equipped with a KeenView camera (Olympus).

Results

DRFs are required for invadopodia formation and matrix degradation. Western blotting and semiquantitative reverse transcription–PCR analysis showed that MDA-MB-231 cells express all three members of the DRF family (Fig. 1*A*). To characterize the function of DRF proteins in matrix degradation and invadopodia activity in MDA-MB-231 cells, we silenced each DRF protein individually by using two independent siRNAs (Fig. 1*A*).

Mock and siRNA-treated cells were plated on a thin layer of fluorescently labeled gelatin for 5 hours in an assay monitoring matrix degradation (10, 12). We observed that ~20% of mock-treated cells degraded the fluorescent matrix (Fig. 1*B*; data not shown) and silencing of MT1-MMP reduced matrix proteolysis of MDA-MB-231 cells to ~10% of control (Fig. 1*D*), confirming recent reports (10, 12, 17). Next, we analyzed matrix proteolysis by DRF-depleted cells. Silencing of DRF1 to DRF3 with two independent siRNAs resulted in 70% to 80% reduction of degradation compared with mock-treated cells without significant perturbation of cell and actin cytoskeleton morphology or spreading on gelatin (Fig. 1*B* and *D*).

As described (9, 10, 16), proteolysis of the matrix by MDA-MB-231 cells was mainly focal and coincided with the presence of F-actin/cortactin-positive invadopodia at the ventral cell surface (Fig. 1*C*). Consistent with the proportion of cells able to degrade the matrix, invadopodia were observed in ~20% of mock-treated MDA-MB-231 cells plated on gelatin (6.0 ± 0.7 invadopodia per cell, n = 95 cells; Fig. 1*D*). Interestingly, this proportion dropped to 8% to 11% in cells depleted for each individual DRF (Fig. 1*D*). When cells were incubated on gelatin for a longer time (15 h), inhibition of matrix degradation and invadopodia formation was still observed in DRF-depleted cells compared with controls, indicating that loss of DRF function does not delay but rather inhibits invadopodia formation (not shown). No additional effect of simultaneously knocking down two DRF proteins was observed, suggesting that the availability of each DRF is limiting and that they are functionally linked (not shown). Of note, triple knockdown was inefficient for individual protein suppression (not shown). Altogether, these data show that DRF proteins are required for invadopodia formation and subsequent matrix degradation in MDA-MB-231 breast cancer cells.

To specify at which step of invadopodia formation DRF proteins are implicated, we performed live cell imaging of mock-depleted, DRF3-depleted, and MT1-MMP–depleted MDA-MB-231 cells stably expressing mCh-Lifeact (F-actin–binding peptide of *Saccharomyces cerevisiae* Abp140p fused to mCherry; ref. 31). MDA-MB-231 cells plated on gelatin displayed lamellipodial and membrane ruffling activities and exhibited random motility irrespective of siRNA

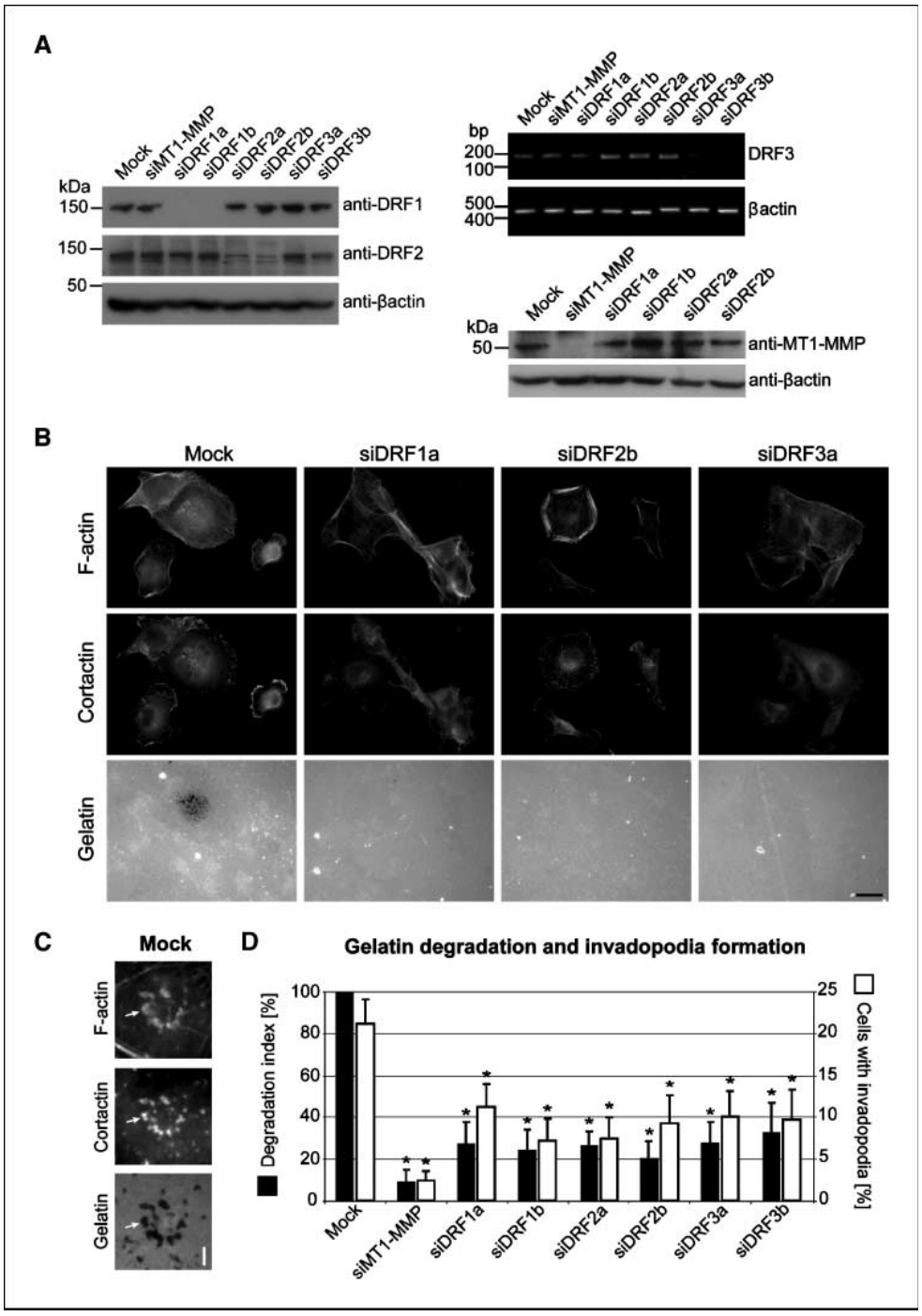


Figure 1. DRF knockdown results in decreased gelatin degradation and invadopodia formation. *A,* protein expression levels of DRF1, DRF2, and mRNA expression levels of *DRF3* in MDA-MB-231 cells treated with two independent siRNAs as indicated. MT1-MMP expression was analyzed by immunoblotting. β-Actin immunoblotting or mRNA levels were used as a loading control. Molecular weight markers are indicated in kDa and bp. *B,* cells transfected with the indicated siRNA were incubated on fluorescent gelatin for 5 h, fixed, and stained for F-actin and cortactin. *Scale bar,* 20 μm. *C,* higher magnification of MDA-MB-231 cell showing F-actin and cortactin-positive invadopodia lying on degraded gelatin (*arrows*). *Scale bar,* 2 μm. *D,* graph depicting gelatin degradation (*black columns*) and the presence of invadopodia (*white columns*) in MDA-MB-231 cells treated with the indicated siRNA. *Black columns,* mean degradation area setting mock to 100; *white columns,* percentage of cells with invadopodia as defined in *C* in the different cell populations; *bars,* SE. Quantifications were obtained from five independent experiments. *, siRNA-treated cell populations are significantly different compared with mock-treated cells (see Supplementary Tables S3 and S4).

treatment. In mock-treated degrading cells, invadopodia appeared as bright, static, and long-lived F-actin puncta, some being stable for at least 2 hours (Supplementary Fig. S1*A*; Supplementary Videos S1 and S2). Newly formed invadopodia appeared either as single isolated puncta (Supplementary Fig. S1*A* and *B*; Supplementary Videos S1 and S2) or formed collectively as a group of puncta originating from a wave of actin assembly (Supplementary Fig. S1*A*; Supplementary Video S1). Highly dynamic small actin dots could also be observed at the rear of extending lamellipodia that did not coincide nor precede matrix degradation and were not related to invadopodia (Supplementary Fig. S1*B*; Supplementary Video S2). These small actin dots were also present in cells depleted for DRF3 (Supplementary Fig. S1*C*; Supplementary Video S3). In contrast, long-lived degradative invadopodia were rarely observed in DRF3-ablated cells (Supplementary Video S3). Finally, F-actin recruitment or aggregation was virtually absent in MT1-MMP depleted cells (Supplementary Video S4). Altogether, these observations suggest that DRF3, as well as other DRF-family members, are important for the early stage of invadopodia formation.

DRF localization at invadopodia. Endogenous DRF1 was distributed throughout the cytoplasm of MDA-MB-231 cells plated on gelatin as observed in many different cell types (29, 32), with some dotty accumulations at the cell edge but no obvious localization at invadopodia (not shown). This distribution was

lost in cells silenced for DRF1 (not shown). GFP-mDia2, the mouse orthologue of human DRF3, was similarly diffuse (not shown). As invadopodia are generally present at the center of cells where cytoplasmic signal was strongest, we analyzed DRF localization in MDA-MB-231 cells transiently expressing an active form of c-Src (Y527F) that triggers appearance of large peripheral invadopodia (10, 12). In these cells, endogenous DRF1 was detected into small dots. Some DRF1 dots were closely apposed to and surrounded F-actin–positive and pY-positive invadopodia (Fig. 2*A* and *B*). In addition, GFP-mDia2 also colocalized with F-actin and pY at invadopodia under these conditions (Fig. 2*C* and *D*). This association of DRF1 and mDia2/DRF3 to invadopodia argues for a direct role for DRFs in the formation of these structures.

DRFs are required for invasion through a three-dimensional reconstituted BM. As MT1-MMP–mediated proteolysis is critical for breaching of the BM by carcinoma cells (3, 5), we evaluated the role of DRFs during remodeling and invasion of Matrigel, a reconstituted matrix mimicking BM. For this purpose, MDA-MB-231 cells were plated on top of a thick layer of Matrigel (~3.5 mm) and were observed by scanning electron microscopy (SEM) after 2 to 14 hours on the matrix. MDA-MB-231 cells rapidly adhered to Matrigel and adopted a rounded morphology (Fig. 3*A*, 2 hours). Cells progressively invaded the matrix and were completely embedded in Matrigel after 12 to 14 hours (Fig. 3*A* and *C*). Of note, gel retraction during sample preparation created domes of Matrigel around invading cells (Fig. 3*A*). Plasma membrane folds, filopodia-like protrusions, and some membrane blebs were observed at the free surface of invading cells (see Figs. 3*A* and 5*A*). In contrast to mock-treated cells that invaded into the BM, cells silenced for MT1-MMP

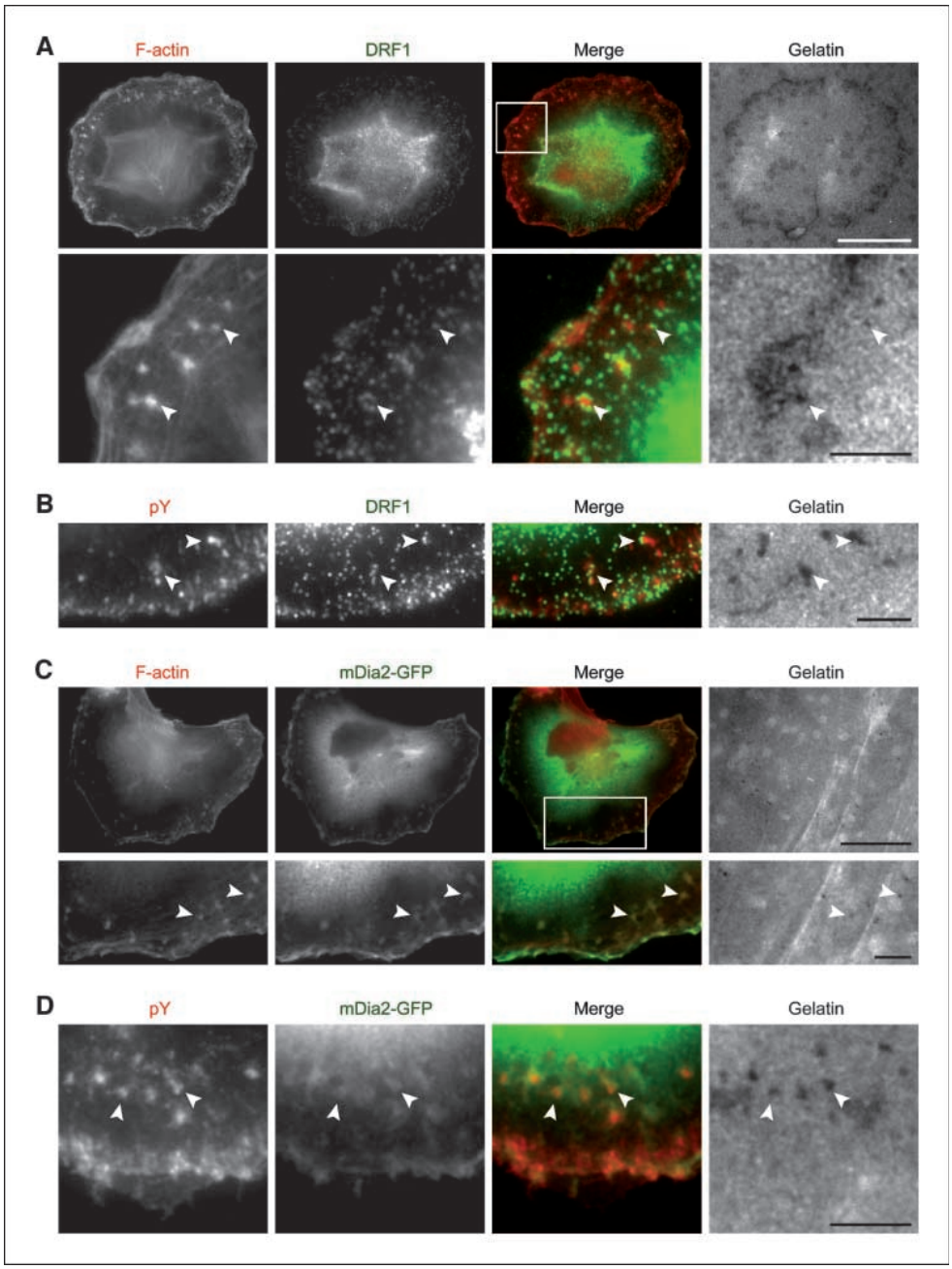


Figure 2. DRF formins localize at invadopodia. *A* and *B,* MDA-MB-231 cells transiently transfected with Y527F-cSrc were incubated 5 h on fluorescent gelatin. After fixation, cells were stained for F-actin (*A*), pY (*B*), and DRF1 (*A* and *B*). Merged images show DRF1 accumulations (*green*) adjacent to F-actin (*red*) or pY-positive invadopodia (*red*). *Bottom,* enlargement of boxed region in *A*. *C* and *D*, MDA-MB-231 cells transiently transfected with Y527F-cSrc and mDia2-GFP constructs incubated for 5 h on fluorescent-gelatin and stained for F-actin (*C*) or pY (*D*). Merged images show mDia2-GFP (*green*) colocalization with F-actin (*red*) or pY (*red*) with underlying gelatin degradation. *Bottom,* enlargement of boxed region in *C*. Arrowheads point to colocalization of DRF1 or mDia2-GFP with F-actin or pY and degradation of the gelatin. *Scale bars,* 20 μm (*A, C*) and 5 μm (*B, D*, and *insets in A* and *C*).

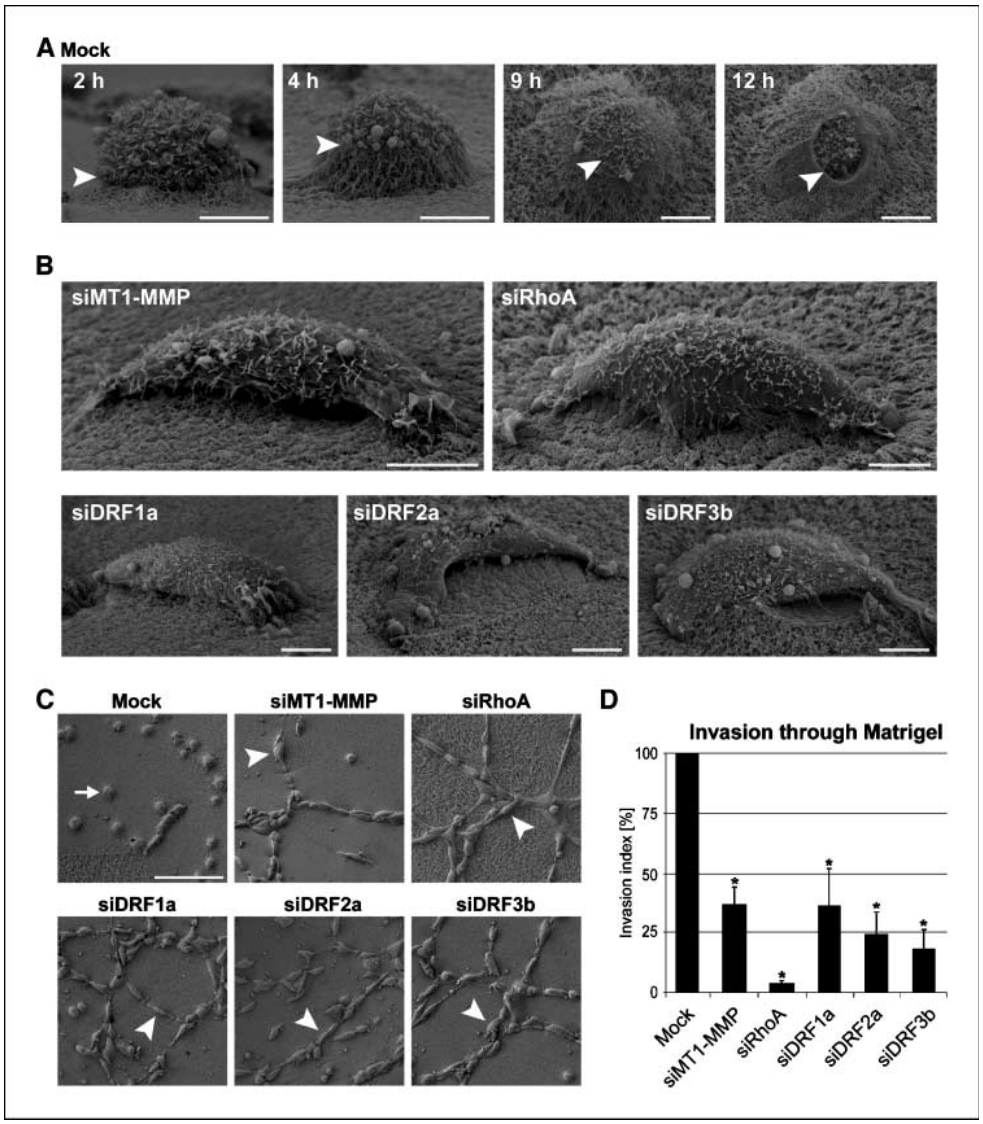


Figure 3. Depletion of RhoA and DRF impairs cell invasion through Matrigel. *A-C,* scanning electron micrographs of MDA-MB-231 cells invading a thick layer of Matrigel. *A,* mock MDA-MB-231 cells were plated on top of Matrigel and fixed at the indicated time points. Arrowheads point to the limit between cells and the matrix. *B,* MDA-MB-231 cells treated with the indicated siRNA were fixed after 4 h on Matrigel. *C,* low-magnification micrographs of siRNA-treated cells fixed 14 h after plating. Mock-treated cells display a rounded morphology and are partially or completely embedded within the matrix (*arrow*), whereas cells depleted for MT1-MMP, RhoA, or DRF proteins are spread and often form chains at the surface of Matrigel (*arrowheads*). *Scale bar,* 5 μm (*A* and *B*) and 100 μm (*C*). *D,* quantification of cell invasion through Matrigel from low-magnification SEM micrographs. *Columns,* mean invasion setting mock to 100%; *bars,* SE. *, all siRNA-treated cell populations are significantly different compared with mock-treated cells (see Supplementary Table S5).

remained spread and formed chains on the surface of Matrigel (Fig. 3*B* and *C*).

SEM micrographs (Fig. 3*C*) allowed a precise quantification of cells completely buried within Matrigel after 14 hours and, hence, of the invasion capacity of the different cell populations (Fig. 3*D*). Strikingly, the invasion capacity of MT1-MMP–depleted and DRF1/DRF3–depleted cells in Matrigel dropped to 18% to 37% compared with mock-treated cells (Fig. 3*D*). Invasion capacity of DRF-depleted or MT1-MMP–depleted cells was also significantly reduced compared with mock when assessed with commercial Matrigel invasion chambers (47–66% of control value; Supplementary Fig. S2). Interestingly, silencing of RhoA, a common regulator of the three DRF proteins also known to be necessary for gelatin degradation (12), phenocopied the effect of MT1-MMP and DRF depletion (Fig. 3*B–D*). Together, these data indicate that, similar to the effect observed on two-dimensional gelatin (Fig. 1), RhoA and each one of the DRF family members are required for invasion through Matrigel.

Invadopodia-like membrane protrusions are present during invasion of three-dimensional BM. Whether invadopodia are present in tumor cells invading three-dimensional ECM and share characteristics with those defined on a two-dimensional rigid matrix are critical issues still awaiting more detailed analysis (7, 16, 20). As an attempt to characterize mechanisms underlying three-dimensional matrix invasion and, in particular, requirement of DRFs, MDA-MB-231 cells either mock-treated or treated with siRNAs specific for DRF proteins were plated on Matrigel for 6 hours and the distribution of key invadopodial markers, i.e., F-actin, cortactin, and pY, was analyzed by indirect immunofluorescence and three-dimensional microscopy.

MDA-MB-231 cells labeled for F-actin and cortactin displayed the rounded morphology typical of invading cells in Matrigel (Supplementary Video S5). Focusing at the invasive surface of cells in contact with the matrix revealed a higher number of F-actin–rich protrusions compared with invadopodia in cells grown on two-dimensional gelatin (55.4 $\pm$ 7.0 protrusions per cell, n = 35 cells; see Fig. 4*A*; Supplementary Video S5). These structures were positive for cortactin and pY, with visible enrichment of both markers at the basis of the protrusions (Fig. 4*B*). In addition, examination of SEM micrographs of partially invading MDA-MB-231 cells revealed the presence of thin filopodia-like protrusions at the free dorsal side of cells partially embedded in Matrigel

(Fig. 5*A* and Supplementary Fig. S3). Some of these structures breaching through Matrigel likely represent proteolytically active structures. Consistent with this assumption, a fusion of MT1-MMP with pHLuorin, a pH-sensitive GFP-emitting fluorescence only in the external milieu, colocalized with F-actin–positive protrusions labeled with mCh-Lifeact at the surface of MDA-MB-231 cells embedded in Matrigel (Supplementary Video S6).

The invasive interface of MDA-MB-231 cells with Matrigel was analyzed by transmission electron microscopy (TEM) on cross-sections perpendicular to the matrix. At an early time point (5 hours), cells that partially entered the matrix presented numerous short protrusions heterogeneous both in length and diameter (Fig. 5*B*). Furthermore, protrusions visible at the cell-matrix interface extending into Matrigel (Fig. 5*B*) were often surrounded by an electroluscent zone devoid of Matrigel, probably as a result of matrix degradation (Fig. 5*B*). At higher magnification, electron-dense regions of the cytoplasm devoid of organelles were visible at the cell cortex in contact with Matrigel (Fig. 5*B*). This cortical material consisting of fibrous structures extending within the protrusions (see Fig. 5*B*) is likely to correspond to cortical actin filament bundles. Altogether, these findings indicate that invadopodia-like structures are present at the invasive surface of MDA-MB-231 cells, invading a three-dimensional reconstituted BM and likely represent the sites of matrix proteolysis.

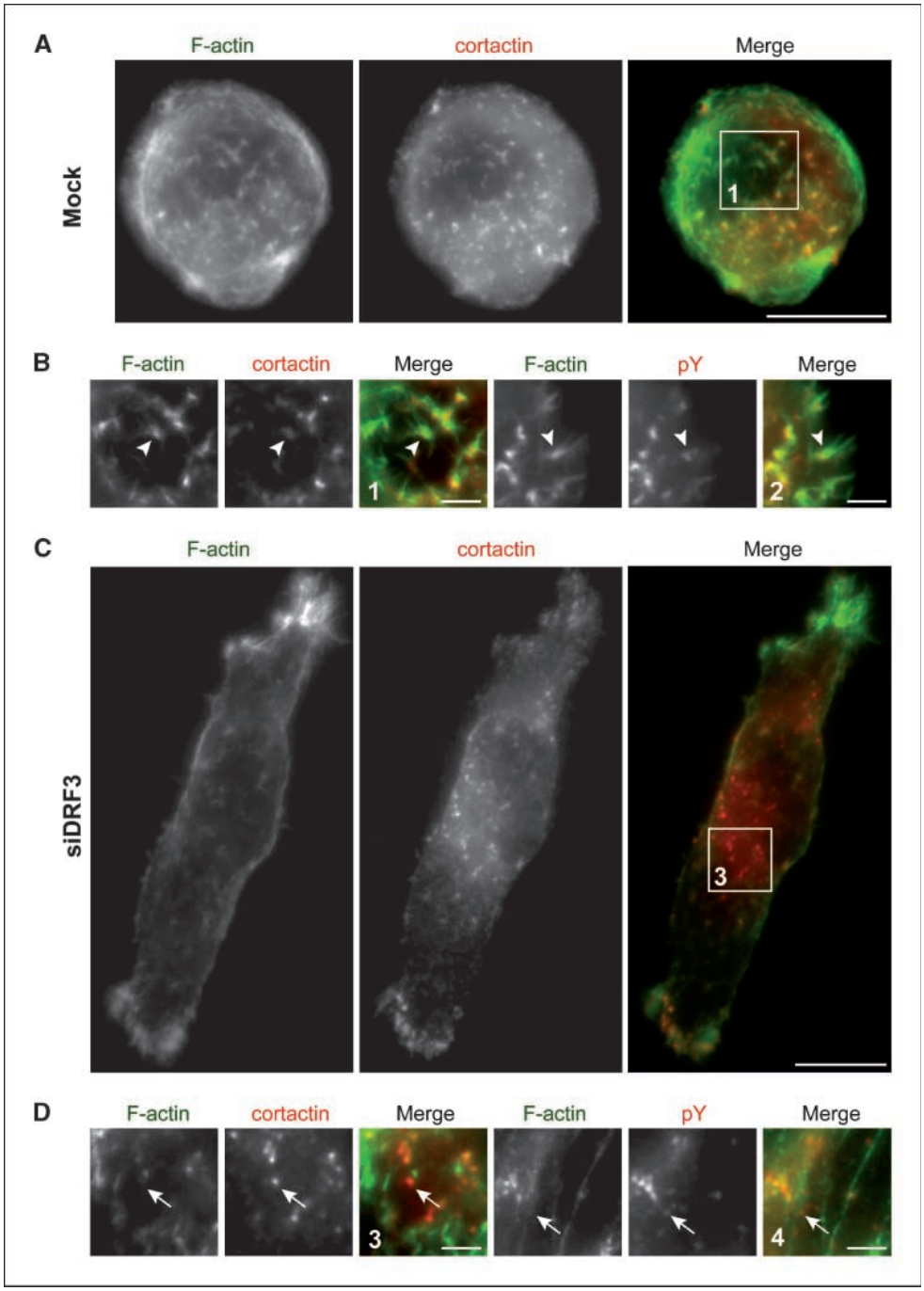


Figure 4. MDA-MB-231 cells display invadopodia-like protrusions in Matrigel that require DRF proteins for their formation. MDA-MB-231 cells either mock-treated (*A* and *B*) or treated with siRNA specific for DRF3 (*C* and *D*) were added on top of Matrigel, fixed after 6 h, and labeled for F-actin and cortactin or pY as indicated. Pictures are three-dimensional reconstructions from several planes corresponding to the ventral half of cells in contact with the matrix. Enlargements of boxed regions in *A* and *C* are shown in *B* (*inset 1*) and *D* (*inset 3*), respectively. For F-actin/pY double-labeling, only enlarged insets are shown (*B, inset 2* and *D, inset 4*), corresponding to full-size pictures in Supplementary Fig. S4*A* (*box 2*) and *B* (*box 4*). In overlaid images, F-actin is pseudocolored in green and cortactin or pY is in red. Arrowheads in insets point to F-actin–rich invadopodia-like protrusions at the surface of mock-treated cells with their base enriched for cortactin (*B, inset 1*) or pY (*B, inset 2*). Arrows in *D* point to cytoplasmic aggregates of cortactin (*inset 3*) or pY (*inset 4*), with weak or no F-actin enrichment in cells silenced for DRF3. *Scale bar,* 10 μm (*A* and *C*) and 2 μm (*B* and *D*).

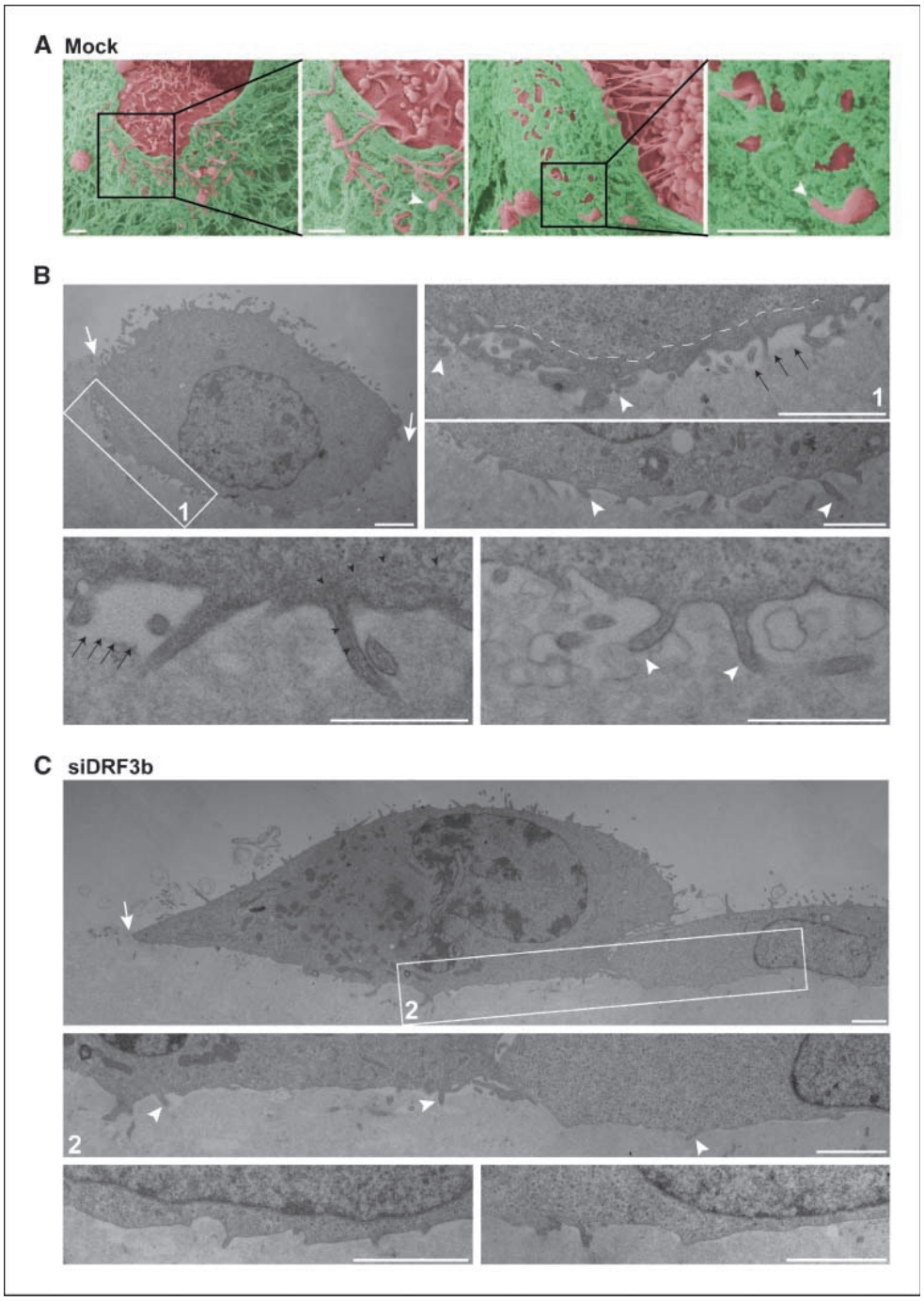


Figure 5. Ultrastructural analysis of invadopodia in Matrigel. *A,* SEM micrographs of MDA-MB-231 cells invading through Matrigel. *Red,* cells; *green,* Matrigel. Arrowheads point to finger-like cellular protrusions going through the matrix. *B* and *C,* TEM micrographs of thin sections across mock-treated (*B*) or DRF3-depleted cells. *C,* assembly of several individual micrographs. Higher magnifications of regions boxed in 1 and 2 are shown in adjacent panels. White arrows in *B* and *C* point to the limit between cells and Matrigel. White arrowheads point to cell protrusions extending within the matrix. Black arrows in *B* point to regions of matrix degradation. *Dashed line in B, inset 1,* limit of an organelle-free zone of the cytoplasm at the invasive edge of the cell. Black arrowheads in *B* point to some fibrous material underneath the invasive surface extending within cell protrusions. *Scale bar,* 1 μm (*A* and *B, bottom*) and 2 μm (*B, top* and *C*).

Three-dimensional invadopodia require DRFs. Cells depleted for each one of the DRF proteins remained elongated on Matrigel with some sparsely distributed F-actin and cortactin–positive or pY-positive protrusions at their surface (Fig. 4*C* and *D* and Supplementary Figs. S4*B* and S5*B–D*; Supplementary Video S7), corroborating results of SEM (Fig. 3). Numerous bright cortactin aggregates with little or no F-actin were observed in DRF-depleted cells. These seemed to be cytoplasmic on three-dimensional reconstruction of Z-stack of images (Supplementary Video S7; Fig. 4*D, inset 3* and Supplementary Fig. S4*B, inset 4*). Similar observations were made in MDA-MB-231 cells silenced for RhoA or MT1-MMP (Supplementary Figs. S4*C* and S5*A*). In addition, the flat morphology of DRF3-depleted cells was also clearly visible on TEM analysis of cross-sections (Fig. 5*C*). More strikingly, although these cells developed some microvilli-like extensions on their free dorsal surface, they were almost devoid of membrane protrusions at the interface with Matrigel (Fig. 5*C*). DRF3 knocked-down cells remained closely apposed to the matrix, this being indicative of the absence of Matrigel proteolysis and remodeling, and a dense cortex was hardly visible underneath the plasma membrane (Fig. 5*C*). Therefore, as shown previously for two-dimensional invadopodia, three-dimensional invadopodia did not form in the absence of DRF1 to DRF3, further establishing the critical role of DRFs for invadopodia formation.

Discussion

We investigated the contributions of three members of the DRF family (DRF1–DRF3) to the mechanism of invadopodia formation and invasion of human breast cancer cells. Each of the three DRF proteins is required for the formation and activity of invadopodia when MDA-MB-231 cells are plated on a two-dimensional matrix. Similarly, DRF1 to DRF3 are required for invasion of breast tumor cells through a thick three-dimensional layer of Matrigel with a composition resembling BM matrix. Remodeling of the three-dimensional matrix occurs at the level of submicrometric finger-like actin-based protrusions possessing a composition similar to invadopodia, i.e., enriched in F-actin, cortactin, pY, and MT1-MMP, and depending on DRFs for their formation. Hence, our work adds DRFs to the list of proteins required for invasion by tumor cells and brings new insight into the mechanism of invasion in a three-dimensional matrix environment.

Invadopodia are viewed as dynamic filopodia-like extensions of the plasma membrane, wherein signaling components and cellular machineries involved in actin-driven membrane protrusion and exocytosis cooperate for delivering and concentrating active MMPs at sites of matrix degradation (10–13, 16, 20, 33, 34). The present data clearly implicate DRF proteins in the regulation of actin assembly during invadopodia formation as silencing of each DRF led to (*a*) a drastic reduction in cells with F-actin/cortactin–rich invadopodia in two-dimensional and three-dimensional matrices and (*b*) decreased matrix degradation and invasion capacity. Of note, no additional effect of knocking down simultaneously two DRF proteins was observed. One possibility is that DRF proteins may be functionally linked in the mechanism of invadopodia formation, which is supported by the observation that DRFs can form heterodimers (35, 36). Alternatively, each formin may be individually required for the induction of actin nucleation at invadopodia or for other individual roles. In this respect, DRF proteins are implicated in various cellular functions, including regulation of endosome motility, which could contribute to the delivery of MT1-MMP at invadopodia (13, 17). We did not find colocalization of GFP-mDia2 with endocytic markers at invadopodia. However, some association of mDia2 with transferrin-positive early endosomes was visible (not shown). Therefore, a role for DRF in membrane trafficking events related to invadopodia function cannot be excluded (37, 38). In addition, the known function of DRF proteins in the regulation of microtubule stability may be potentially relevant for the mechanism of invadopodia formation (37–39).

A signaling cascade based on N-WASP/Arp2/3 complex activation downstream of the Rho-GTP-binding protein Cdc42 is required for actin assembly during invadopodia formation in metastatic rat mammary adenocarcinoma cells (15). We confirmed that the Arp2/3 complex is required for invadopodia formation in two-dimensional and three-dimensional matrices in MDA-MB-231 cells.[7] The Arp2/3 complex nucleates actin filaments and forms branched dendritic filament arrays, whereas formins produce unbranched actin filaments (22). Therefore, two types of actin-nucleating machineries cooperate during invadopodia formation in invasive cells. Using a FRET biosensor, N-WASP activation was visualized at the base of invadopodia, suggesting that Arp2/3–mediated actin nucleation is constrained to the base of invadopodia (14). In addition, the F-actin binding protein cortactin, which is recruited early at invadopodia concomitantly with F-actin assembly, is also essential for invadopodia formation and may act by stabilizing newly formed branches within the dendritic filament network (10, 16–18). Noticeably, cortactin and pY are enriched at the base of invadopodia that protrude from the invasive surface of MDA-MB-231 cells in a thick layer of Matrigel (see Fig. 4). Cells silenced for the Arp2/3 complex (15), cortactin (10, 17, 18), or DRFs are similarly impaired for invadopodia formation, indicating that both pathways of actin nucleation are required for invadopodia formation and cannot compensate for each other. Convergent extension of filopodia from an Arp2/3 complex–induced lamellipodial actin meshwork has been proposed as a mechanism for filopodia emergence (40), although this model is debated (41). In addition, transition from Arp2/3 to formin-mediated actin assembly may occur during actin dynamics associated with integrin-based adhesion sites (42). Invadopodia, which are enriched in adhesion proteins, including integrins and focal adhesion components, also correspond to cell-matrix adhesion sites (16, 43). Therefore, it is plausible that DRFs take over from N-WASP/Arp2/3/cortactin dendritic array at the base of invadopodia and elongate actin filaments into an invadopodial protrusion. DRF3/mDia2 and DRF1/mDia1 localize to filopodial tips and are involved in actin filament elongation during filopodia and membrane protrusion formation in mammalian cells including invasive tumor cells (23, 25, 27, 29, 44). In accordance, we observed endogenous DRF1 and mDia2-GFP at invadopodia in MDA-MB-231 cells. Although more work will be necessary to understand the cooperation between Arp2/3 complex and DRFs and unravel the ultrastructural architecture of invadopodial organization, the present study clearly identifies DRFs as important components involved in breast cancer cell invasion.

Rho GTP-binding proteins act as regulators of actin organization and membrane trafficking events in cells under physiologic conditions and have also been shown to contribute to various aspects of tumorigenesis including invasion of carcinoma cells (45). Several groups, including ours, found that Rho proteins control the formation of invadopodia in tumor-derived cell lines of diverse origins (15, 19). In particular, we recently reported that silencing of RhoA or Cdc42 abolishes invadopodia formation and matrix degradation by MDA-MB-231 cells cultured on a two-dimensional matrix (12). DRF proteins are downstream effectors of Rho GTPases (22). Our finding that RhoA and Cdc42 are required for invadopodia formation in both two-dimensional and three-dimensional matrices (ref. 12, and this study), suggests that beside the aforementioned regulation of Arp2/3 complex by Cdc42 (15), Cdc42/RhoA GTPases may regulate the function of DRFs in actin polymerization at invadopodia in invasive MDA-MB-231 cells. Along the same line, the formation of cellular protrusions associated with migration of rat mammary carcinoma cells involves a RhoA/DRF1 pathway acting in a coordinated network together with WASP family proteins and Arp2/3 complex (29). Interestingly, invasion of MDA-MB-435 human cancer cells in Matrigel is dependent on a RhoA/DRF1 pathway, but not DRF2 (DRF3 was not tested; ref. 28). The reason for this discrepancy with our findings is unclear and may involve the different origin of these two cell lines (30). It is worth noticing that, at a mechanistic level, MDA-MB-435 cells in Matrigel use a bleb-associated mode of invasion and lose the ability to form membrane blebbing when silenced for DRF1 (28). In contrast, invasion of MDA-MB-231 cells involves filopodia-like membrane protrusions that are enriched for MT1-MMP (see Supplementary Video S6) and require RhoA and DRFs for their formation.

[7] Our unpublished observations.

In conclusion, this study highlights structural composition and mechanistic similarities between classic invadopodia, degradative structures of invasive cells cultured on two-dimensional matrix, and filopodial-like protrusions forming at sites of matrix degradation at the surface of breast tumor cells invading through three-dimensional Matrigel. It also reveals a new role for DRF proteins as essential components of the invasive machinery of metastatic cells through the regulation of actin assembly underlying the mechanism of invadopodia formation. Along this line, it is quite remarkable that analysis of gene expression array data (46) revealed a significant 2-fold to 2.5-fold overexpression of DRF2-encoding and DRF3-encoding transcripts in highly invasive basal-like breast tumors compared with normal human breast tissues (see Supplementary Materials). Understanding whether and how the multiple activities of DRF proteins contribute to the acquisition of specific invadopodial functions and to the invasive phenotype of cancer cells will be a challenge for future studies.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

Received 9/24/08; revised 12/22/08; accepted 1/5/09; published OnlineFirst 3/10/09.

Grant support: Institut Curie, CNRS, Ligue Nationale contre le Cancer "Equipe Labellisée," and Fondation BNP-Paribas grants (P. Chavrier); Association pour la Recherche contre le Cancer and Institut National du Cancer grants (R. Poincloux); CNRS and Fondation pour la Recherche Médicale postdoctoral fellowships (F. Lizárraga); and Institut National du Cancer (G. Rigaill).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Dr. M-C. Prévost for the access to the Ultrastructural Microscopy Platform of Institut Pasteur and the help with image acquisition and processing at the Cell and Tissue Imaging facility of CNRS/Institut Curie, Dr. P. Hupé (Institut Curie) for his help with statistical analysis, Drs. A. Steffen and K. Rottner for critical reading of the manuscript, and Dr. T. Dubois for drawing our attention to the overexpression of *DIAPH* genes in human basal-like breast tumors.

References

1. Yamaguchi H, Wyckoff J, Condeelis J. Cell migration in tumors. Curr Opin Cell Biol 2005;17:559–64.
2. Sahai E. Mechanisms of cancer cell invasion. Curr Opin Genet Dev 2005;15:87–96.
3. Friedl P, Wolf K. Tumour-cell invasion and migration: diversity and escape mechanisms. Nat Rev Cancer 2003; 3:362–74.
4. Sabeh F, Ota I, Holmbeck K, et al. Tumor cell traffic through the extracellular matrix is controlled by the membrane-anchored collagenase MT1-MMP. J Cell Biol 2004;167:769–81.
5. Hotary K, Li XY, Allen E, Stevens SL, Weiss SJ. A cancer cell metalloprotease triad regulates the basement membrane transmigration program. Genes Dev 2006; 20:2673–86.
6. Zaman MH, Trapani LM, Sieminski AL, et al. Migration of tumor cells in 3D matrices is governed by matrix stiffness along with cell-matrix adhesion and proteolysis. Proc Natl Acad Sci U S A 2006;103:10889–94.
7. Wolf K, Wu YI, Liu Y, et al. Multi-step pericellular proteolysis controls the transition from individual to collective cancer cell invasion. Nat Cell Biol 2007;9:893–904.
8. Itoh Y, Seiki M. MT1-MMP: a potent modifier of pericellular microenvironment. J Cell Physiol 2006;206:1–8.
9. Chen WT, Olden K, Bernard BA, Chu FF. Expression of transformation-associated protease(s) that degrade fibronectin at cell contact sites. J Cell Biol 1984;98:1546–55.
10. Artym VV, Zhang Y, Seillier-Moiseiwitsch F, Yamada KM, Mueller SC. Dynamic interactions of cortactin and membrane type 1 matrix metalloproteinase at invadopodia: defining the stages of invadopodia formation and function. Cancer Res 2006;66:3034–43.
11. Clark ES, Weaver AM. A new role for cortactin in invadopodia: regulation of protease secretion. Eur J Cell Biol 2008;87:8–9.
12. Sakurai-Yageta M, Recchi C, Le Dez G, et al. The interaction of IQGAP1 with the exocyst complex is required for tumor cell invasion downstream of Cdc42 and RhoA. J Cell Biol 2008;181:985–98.
13. Steffen A, Le Dez G, Poincloux R, et al. MT1-MMP-dependent invasion is regulated by TI-VAMP/VAMP7. Curr Biol 2008;18:926–31.
14. Lorenz M, Yamaguchi H, Wang Y, Singer RH, Condeelis J. Imaging sites of N-wasp activity in lamellipodia and invadopodia of carcinoma cells. Curr Biol 2004;14:697–703.
15. Yamaguchi H, Lorenz M, Kempiak S, et al. Molecular mechanisms of invadopodium formation: the role of the N-WASP-Arp2/3 complex pathway and cofilin. J Cell Biol 2005;168:441–52.
16. Bowden ET, Barth M, Thomas D, Glazer RI, Mueller SC. An invasion-related complex of cortactin, paxillin and PKCmu associates with invadopodia at sites of extracellular matrix degradation. Oncogene 1999;18: 4440–9.
17. Clark ES, Whigham AS, Yarbrough WG, Weaver AM. Cortactin is an essential regulator of matrix metalloproteinase secretion and extracellular matrix degradation in invadopodia. Cancer Res 2007;67:4227–35.
18. Ayala I, Baldassarre M, Giacchetti G, et al. Multiple regulatory inputs converge on cortactin to control invadopodia biogenesis and extracellular matrix degradation. J Cell Sci 2008;121:369–78.
19. Nakahara H, Otani T, Sasaki T, Miura Y, Takai Y, Kogo M. Involvement of Cdc42 and Rac small G proteins in invadopodia formation of RPMI7951 cells. Genes Cells 2003;8:1019–27.
20. Bowden ET, Onikoyi E, Slack R, et al. Co-localization of cortactin and phosphotyrosine identifies active invadopodia in human breast cancer cells. Exp Cell Res 2006;312:1240–53.
21. Linder S. The matrix corroded: podosomes and invadopodia in extracellular matrix degradation. Trends Cell Biol 2007;17:107–17.
22. Higgs HN. Formin proteins: a domain-based approach. Trends Biochem Sci 2005;30:342–53.
23. Peng J, Wallar BJ, Flanders A, Swiatek PJ, Alberts AS. Disruption of the Diaphanous-related formin Drf1 gene encoding mDia1 reveals a role for Drf3 as an effector for Cdc42. Curr Biol 2003;13:534–45.
24. Schirenbeck A, Bretschneider T, Arasada R, Schleicher M, Faix J. The Diaphanous-related formin dDia2 is required for the formation and maintenance of filopodia. Nat Cell Biol 2005;7:619–25.
25. Pellegrin S, Mellor H. The Rho family GTPase Rif induces filopodia through mDia2. Curr Biol 2005;15: 129–33.
26. Hotulainen P, Lappalainen P. Stress fibers are generated by two distinct actin assembly mechanisms in motile cells. J Cell Biol 2006;173:383–94.
27. Yang C, Czech L, Gerboth S, Kojima S, Scita G, Svitkina T. Novel roles of formin mDia2 in lamellipodia and filopodia formation in motile cells. PLoS Biol 2007; 5:e317.
28. Kitzing TM, Sahadevan AS, Brandt DT, et al. Positive feedback between Dia1, LARG, and RhoA regulates cell morphology and invasion. Genes Dev 2007;21:1478–83.
29. Sarmiento C, Wang W, Dovas A, et al. WASP family members and formin proteins coordinate regulation of cell protrusions in carcinoma cells. J Cell Biol 2008;180: 1245–60.
30. Neve RM, Chin K, Fridlyand J, et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell 2006;10:515–27.
31. Riedl J, Crevenna AH, Kessenbrock K, et al. Lifeact: a versatile marker to visualize F-actin. Nat Methods 2008; 5:605–7.
32. Gupton SL, Eisenmann K, Alberts AS, Waterman-Storer CM. mDia2 regulates actin and focal adhesion dynamics and organization in the lamella for efficient epithelial cell migration. J Cell Sci 2007;120:3475–87.
33. Mueller SC, Yeh Y, Chen WT. Tyrosine phosphorylation of membrane proteins mediates cellular invasion by transformed cells. J Cell Biol 1992;119:1309–25.
34. Chen WT, Wang JY. Specialized surface protrusions of invasive cells, invadopodia and lamellipodia, have differential MT1-MMP, MMP-2, and TIMP-2 localization. Ann N Y Acad Sci 1999;878:361–71.
35. Copeland SJ, Green BJ, Burchat S, Papalia GA, Banner D, Copeland JW. The Diaphanous inhibitory domain/Diaphanous autoregulatory domain interaction is able to mediate heterodimerization between mDia1 and mDia2. J Biol Chem 2007;282:30120–30.
36. Gavard J, Patel V, Gutkind JS. Angiopoietin-1 prevents VEGF-induced endothelial permeability by sequestering Src through mDia. Dev Cell 2008;14:25–36.
37. Tominaga T, Sahai E, Chardin P, McCormick F, Courtneidge SA, Alberts AS. Diaphanous-related formins bridge Rho GTPase and Src tyrosine kinase signaling. Mol Cell 2000;5:13–25.
38. Gasman S, Kalaidzidis Y, Zerial M. RhoD regulates endosome dynamics through Diaphanous-related Formin and Src tyrosine kinase. Nat Cell Biol 2003;5: 195–204.
39. Palazzo AF, Cook TA, Alberts AS, Gundersen GG. mDia mediates Rho-regulated formation and orientation of stable microtubules. Nat Cell Biol 2001;3:723–9.
40. Svitkina TM, Bulanova EA, Chaga OY, et al. Mechanism of filopodia initiation by reorganization of a dendritic network. J Cell Biol 2003;160:409–21.
41. Ladwein M, Rottner K. On the Rho'd: the regulation of membrane protrusions by Rho-GTPases. FEBS Lett 2008;582:2066–74.
42. Butler B, Gao C, Mersich AT, Blystone SD. Purified integrin adhesion complexes exhibit actin-polymerization activity. Curr Biol 2006;16:242–51.
43. Mueller SC, Chen WT. Cellular invasion into matrix beads: localization of β 1 integrins and fibronectin to the invadopodia. J Cell Sci 1991;99:213–25.
44. Higashida C, Miyoshi T, Fujita A, et al. Actin polymerization-driven molecular movement of mDia1 in living cells. Science 2004;303:2007–10.
45. Sahai E, Marshall CJ. RHO-GTPases and cancer. Nat Rev Cancer 2002;2:133–42.
46. Richardson AL, Wang ZC, De Nicolo A, et al. X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell 2006;9:121–32.

11.2 Pathway by pathway differential analysis

As we have seen in the previous subsection, the gene-by-gene analysis revealed many over-expressed genes in TNBC. It may be easier to understand the disease at the pathway level and one would hope that only a few pathways are deregulated. An idea is to consider sets of genes which are somehow related (a "geneset"). For example, one can consider the set of all genes involved in proliferation (which is a very large set). Of course, this definition fails to capture the fact that in some cases we know how the different proteins or genes of the pathway interact with each other. Various pathway databases are available, such as the GO (Ashburner et al., 2000) and KEGG databases (Kanehisa and Goto, 2000) (see Figure 11.5).

Various methods have been proposed to identify links between pathways and a biological or clinical variable based on gene expression levels (see for example Goeman et al. (2004); Subramanian et al. (2005); Efron and Tibshirani (2007); Beissbarth and Speed (2004)). Here, I have decided to use the global test method (Goeman et al., 2004) rather than the GSEA, GSA or Gostat methods (respectively described in Subramanian et al. (2005); Efron and Tibshirani (2007); Beissbarth and Speed (2004)). The main reason for my choice is a modeling one. The three last methods work after the gene-by-gene differential analysis and model the resulting p-values. These methods look for pathways which have many genes with small p-values. On the contrary, the global test directly describes the relationship between the gene expression levels and the tumor subtypes. More precisely, the dependence between the tumor subtype and the gene expression levels is modeled using the framework of generalized linear model. Using this model, it is possible to detect pathways with gene levels related to the tumor subtypes. From a biological perspective, the global test model is rather natural, whereas modeling a pathway through gene p-values appears less intuitive.

Using the global test methodology, I retrieved a list of pathways sorted by p-values (see Figure 11.6). As we have seen at the gene level, identifying targets is a difficult issue. One would hope that looking at the pathway level would simplify the problem and that identifying candidate target pathways would be easier. But it is not the case and actually, it is probably the opposite. Indeed, many genes are differentially expressed between TNBC and other subtypes. Moreover, as we have seen

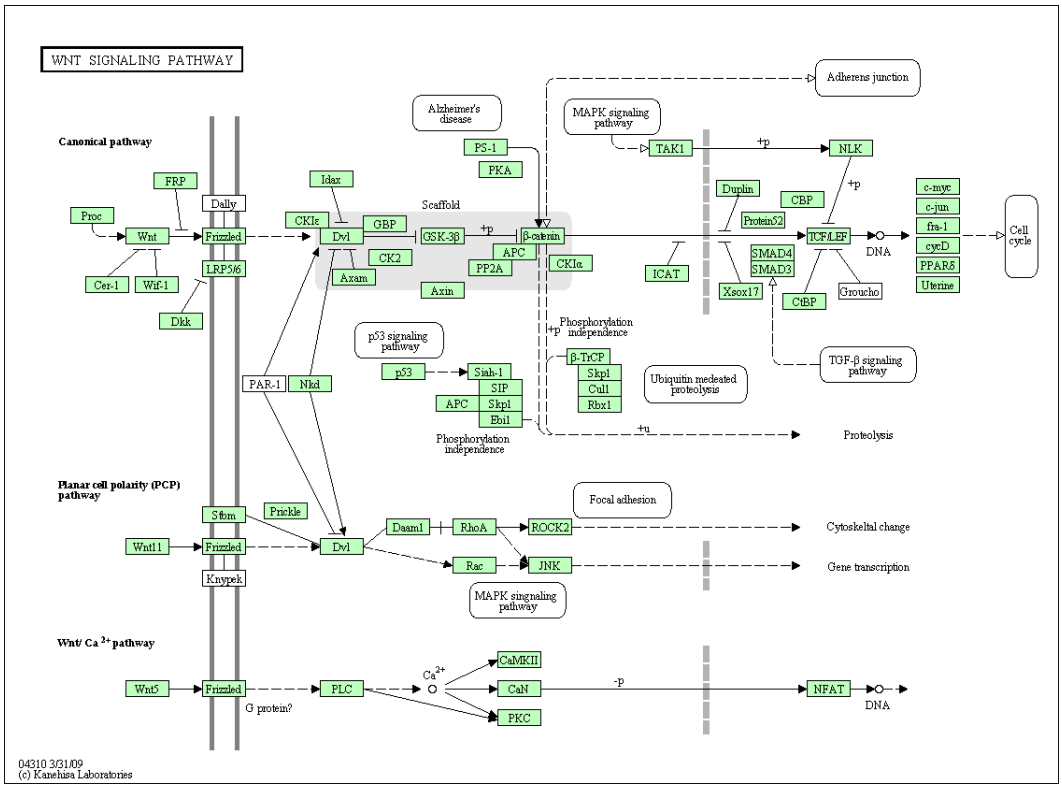


Figure 11.5: Example of a KEGG pathway: the Wnt pathway. Each box represents a protein.

NUM	KEGG_ID	KEGG NAME	p.value
1	05215	Prostate cancer	0
2	05212	Pancreatic cancer	0
3	00230	Purine metabolism	0
4	04520	Adherens junction	0
5	05210	Colorectal cancer	0
6	05213	Endometrial cancer	0
7	05223	Non-small cell lung cancer	0
8	04012	ErbB signaling pathway	0
9	04010	MAPK signaling pathway	0
10	01030	Glycan structures - biosynthesis 1	0
11	04115	p53 signaling pathway	0
12	05220	Chronic myeloid leukemia	0
13	04910	Insulin signaling pathway	0
14	04310	Wnt signaling pathway	0
15	04110	Cell cycle	0
16	05219	Bladder cancer	0
17	00512	O-Glycan biosynthesis	0
18	04020	Calcium signaling pathway	0
19	04120	Ubiquitin mediated proteolysis	0
20	00030	Pentose phosphate pathway	0
21	00600	Sphingolipid metabolism	0

Figure 11.6: Top 21 deregulated KEGG pathways between TNBC and normal tissue according to the global test (Goeman et al., 2004). The global test tries to detect pathways with gene levels related to the tumor subtypes. All p-values are smaller than 2.10^{-16}.

previously, even randomly selected sets of genes show differences between different tumor subtypes. Overall, almost all possible pathways should be detected as related to all tumor subtypes. Indeed, it would be very unlikely to recover a pathway without any genes differentially expressed between TNBC and normal samples. In fact, this is what we observed and most pathways are found to have small p-values: out of the 205 KEGG pathways we tested, only five had a p-value bigger than 5% in the TNBC versus normal comparison. Similar results were obtained for the comparison between TNBC and other tumor subtypes.

Moreover, pathways are not sufficient and one needs to go back to the gene level in the pathway in

order to better interpret the results. In a collaboration with Fatima Mechta-Grigoriou (Ph.D., Institut Curie), we have studied more in depth the oxidative stress pathways in ER- / HER2+ tumors (as published in Toullec et al. (2010)).

11.2.1 Paper: Reactive oxygen species (ROS) control myofibroblast and metastases

This work was done in collaboration with the group of Fatima Mechta-Grigoriou (Ph.D., Institut Curie). In a model of mammary carcinogenesis (junD-/- mice were crossed with an MMTV v-Ha-ras transgenic strain), it was shown that junD inactivation increased tumor incidence. Moreover, junD inactivation in the stroma was sufficient to shorten tumor-free survival rate and enhance metastatic spread. ROS promoted conversion of fibroblasts into highly migrating myofibroblasts through accumulation of the Hypoxia-Inducible Factor (HIF)-1a and the CXCL12 chemokine. It was demonstrated that CXCL12 accumulated in the stroma of ER- / HER2+ human breast tumors. Moreover, ER- / HER2+ tumors exhibited a high proportion of myofibroblasts, a significant nuclear exclusion of JunD and an associated oxido-reduction signature. Collectively, these data uncover a new mechanism by which oxidative stress increases the migratory properties of stromal fibroblasts and these properties in turn favor tumor dissemination.

For this paper, I participated in the transcriptomic analysis. Our data indicate that accumulation of CXCL12 in the stroma of ER- / HER2+ tumors is associated with high myofibroblast content, which impacts tumor spreading in lymph nodes. We then wondered whether genes regulating oxidative stress could be abnormally expressed in this set of tumors. To address this question, I used the Curie-Servier transcritpomic dataset, the GO Ontology terms and the global test methods. I showed that the GO terms "oxidation-reduction" ($GO:0055114$) and "oxido-reductase" activity ($GO:0016491$) appeared among the 100 most significantly deregulated pathways (at the 36th and 83th positions, respectively) with p-values smaller than 2.10^{-16} (numerical precision). I confirmed this difference using hierarchical clustering and principal component analysis (PCA). Both techniques showed that using the gene expression level of these 2 pathways it is possible to discriminate ER- / HER2+, TNBC and Luminal

A tumors. This result suggests that ER- / HER2+ tumors have a gene expression profile characteristic of an oxidative-stress response.

EMBO Molecular Medicine

Oxidative stress promotes myofibroblast differentiation and tumour spreading

Aurore Toullec[1], Damien Gerald[1†], Gilles Despouy[1†], Brigitte Bourachot[1], Melissa Cardon[1], Sylvain Lefort[1], Marion Richardson[2], Guillem Rigaill[2], Maria-Carla Parrini[1,3], Carlo Lucchesi[1,3], Dorine Bellanger[3], Marc-Henri Stern[3], Thierry Dubois[2], Xavier Sastre-Garau[4], Olivier Delattre[3], Anne Vincent-Salomon[4], Fatima Mechta-Grigoriou[1]*

Keywords: AP-1; SDF-1; HIF-1; stroma; metastasis

DOI 10.1002/emmm.201000073

Received November 06, 2009
Revised March 10, 2010
Accepted April 28, 2010

JunD regulates genes involved in antioxidant defence. We took advantage of the chronic oxidative stress resulting from *junD* deletion to examine the role of reactive oxygen species (ROS) in tumour development. In a model of mammary carcinogenesis, *junD* inactivation increased tumour incidence and revealed an associated reactive stroma. *junD*-inactivation in the stroma was sufficient to shorten tumour-free survival rate and enhance metastatic spread. ROS promoted conversion of fibroblasts into highly migrating myofibroblasts through accumulation of the hypoxia-inducible factor (HIF)-1α transcription factor and the CXCL12 chemokine. Accordingly, treatment with an antioxidant reduced the levels of HIF and CXCL12 and numerous myofibroblast features. CXCL12 accumulated in the stroma of HER2-human breast adenocarcinomas. Moreover, HER2 tumours exhibited a high proportion of myofibroblasts, which was significantly correlated to nodal metastases. Interestingly, this subset of tumours exhibited a significant nuclear exclusion of JunD and revealed an associated oxido-reduction signature, further demonstrating the relevance of our findings in human cancers. Collectively, our data uncover a new mechanism by which oxidative stress increases the migratory properties of stromal fibroblasts, which in turn potentiate tumour dissemination.

INTRODUCTION

Carcinomas are highly complex tissues composed of neoplastic and stromal cells, including mesenchymal cells, fibroblasts or myofibroblasts, endothelial cells, pericytes and inflammatory cells (Bissell & Radisky, 2001; Mueller & Fusenig, 2004). In past decades, the major focus of cancer research has been the transformed cell itself. However, new clinical data have shown that the stroma contributes significantly to the development of a wide variety of tumours. Tissues exhibiting chronically inflamed stroma or those suffering from repetitive wound healings display a higher incidence of tumour formation (Joyce & Pollard, 2009; Tlsty & Coussens, 2006). Fibroblasts are the most common type of stromal cells in various human carcinomas, yet their specific contributions to tumour growth have only recently been clarified (Erez et al, 2010; Orimo et al, 2005). Stromal fibroblasts, named carcinoma-associated fibroblasts (CAFs), have been extracted from a number of invasive human breast carcinomas. CAFs are more competent in promoting growth of mammary carcinoma cells and enhancing tumour angiogenesis than fibroblasts derived from outside tumour masses (Olumi et al, 1999; Orimo et al, 2005). CAFs also mediate tumour-enhancing inflammation (Erez et al, 2010). CAFs isolated from the stroma of invasive human breast cancers include large populations of myofibroblasts (Eyden et al, 2008). Myofibroblasts are often referred to as activated fibroblasts that play key

(1) Laboratory of "Stress and Cancer", Inserm U830, Institut Curie, 26 rue d'Ulm, 75248 Paris Cedex 05, France.
(2) Institut Curie, Département de Transfert, Laboratoire de signalisation, Paris, France.
(3) Inserm, Génétique et Biologie des cancers, Paris, France.
(4) Institut Curie, Service de Pathologie, Paris, Cedex, France.
***Corresponding author:** Tel: +33-1-56-24-66-53;
Fax: +33-1-56-24-66-50
E-mail: fatima.mechta-grigoriou@curie.fr
†These authors contributed equally to this study.

roles in wound repair (Hinz et al, 2007). The myofibroblastic properties of CAFs are believed to increase tumour growth and enhance vascular remodelling. Myofibroblasts are characterized by high *de novo* expression of smooth muscle α-actin (SM-α-actin), the actin isoform typically found in vascular smooth muscle cells, and possess greatly enhanced contractile ability. Recent work has shown that CAFs secrete elevated levels of CXCL12, also called stromal cell-derived factor 1 (SDF-1) (Orimo et al, 2005). CXCL12 is a homeostatic chemokine that mediates homing of stem cells to bone marrow by binding to its receptor (CXCR4) on circulating cells (Rossi & Zlotnik, 2000). CXCL12 not only stimulates carcinoma cell growth but also helps in recruiting endothelial progenitor cells to tumours, thereby furthering neo-angiogenesis (Orimo et al, 2005). The importance of this CXCL12–CXCR4 signalling pathway in the tumour microenvironment has already been addressed but, to our knowledge, its role in CAFs remains unexplored.

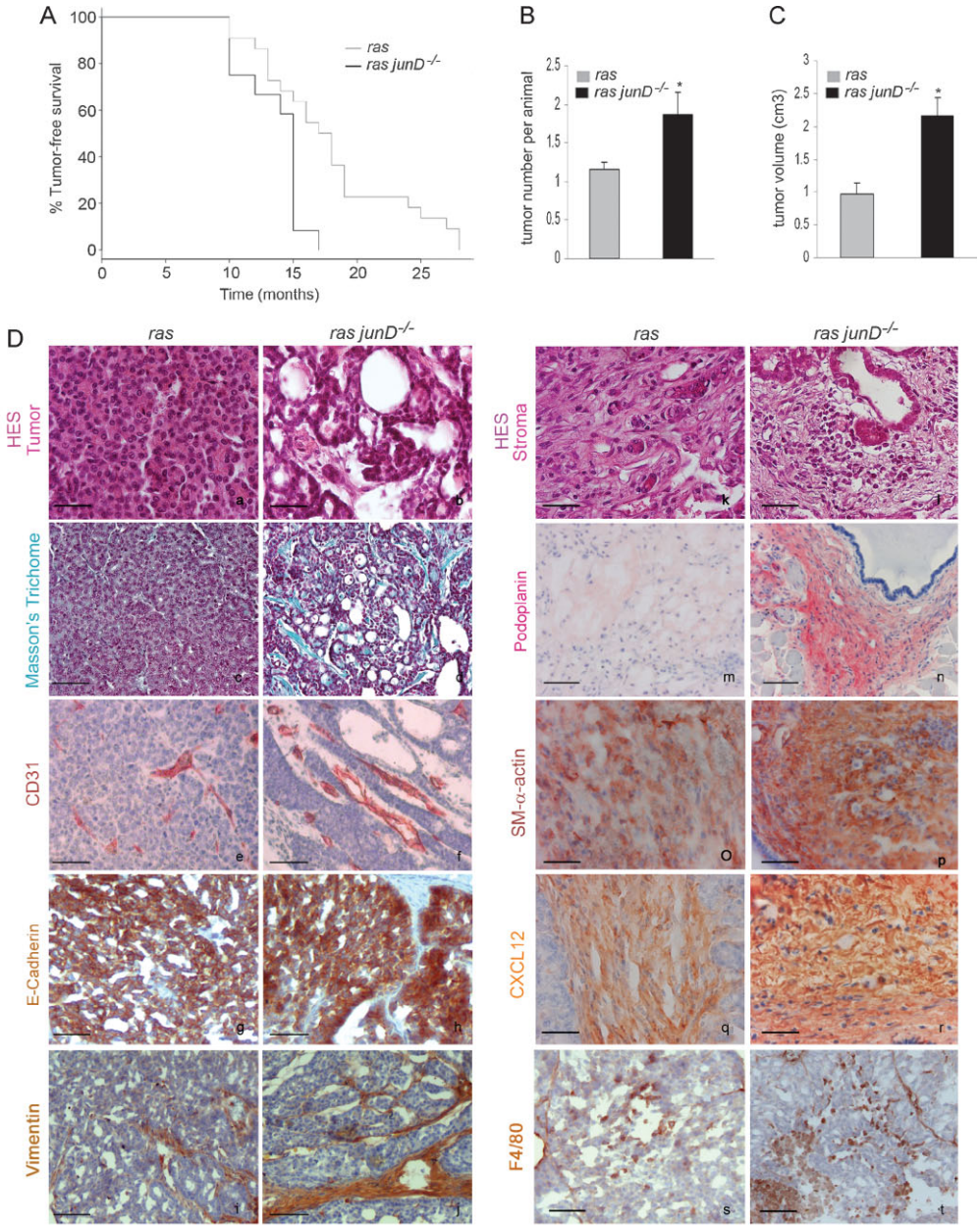
The AP-1 (activator protein-1) transcription factor plays a critical role in regulating environmental stress responses (Mechta-Grigoriou et al, 2001). Recently, we discovered a new function of JunD, a member of the AP-1 family, in controlling oxidative stress and angiogenic switch (Gerald et al, 2004; Laurent et al, 2008). JunD protects cells against oxidative stress by regulating genes involved in antioxidant defence and H_2O_2 production. Subsequently, inactivation of *junD* leads to a persistent accumulation of reactive oxygen species (ROS) in cells and tissues. Thus, *junD*-deficient mice and $junD^{-/-}$ derived-fibroblasts constitute good models for investigating the physiological consequences of chronic oxidative stress. Using these systems, we uncovered a molecular mechanism linking oxidative stress to angiogenesis and ageing (Gerald et al, 2004; Laurent et al, 2008). Accumulation of H_2O_2 reduces the activity of hypoxia-inducible factor (HIF)-prolyl-hydroxylases (PHDs), which signal HIF-α subunits for proteosomal degradation. In consequence, HIF-α proteins accumulate and enhance transcription of specific target genes such as *VEGF-A* (Pouyssegur & Mechta-Grigoriou, 2006).

In this paper, we take advantage of the persistent oxidative stress due to *junD* inactivation to examine the role of ROS in tumour development. In a model of *ras*-mediated mammary carcinogenesis, *junD* deletion increases tumour growth and revealed extensively modified stroma. Moreover, dissemination of $junD^{+/+}$ neoplastic cells was enhanced when grafted into $junD^{-/-}$ mice. Since JunD expression is detected in stromal fibroblasts, these data suggest that inactivation of *junD* in these cells, and consecutive oxidative stress, may affect the fibroblastic properties and potentiate tumour spread. Using *junD*-deficient fibroblasts, we demonstrated that oxidative stress promotes conversion of fibroblasts into myofibroblasts in an HIF-1α and CXCL12-dependent pathway. Conversely, long-term antioxidant treatment partially reverses myofibroblast differentiation. In agreement with our observations on mice, HER2-amplified tumours exhibit the highest expression levels of CXCL12 in the stroma and the highest correlated proportion of myofibroblasts, when compared to aggressive basal-like breast cancers (BLC) or to good prognosis luminal-A (Lum-A) breast carcinomas. Interestingly, HER2-subtype of tumours display a molecular signature characteristic of stress–response and a nuclear exclusion of the JunD protein. Since breast tumours which overexpress HER2 exhibit one of the poorest prognosis of all molecular classes of breast carcinomas and show a high rate of axillary lymph node metastases (Bartlett et al, 2007), these observations underline the role of oxidative stress and myofibroblasts in cancer metastases.

RESULTS

junD inactivation results in a reactive stroma and promotes tumour metastasis

We have previously shown that inactivation of the *junD* gene leads to constitutive oxidative stress (Gerald et al, 2004; Laurent et al, 2008). To further investigate the role of such stress in tumour development, we crossed $junD^{-/-}$ mice with an *MMTV-v-Ha-ras* transgenic strain; a model for breast cancer (Sinn et al, 1987). Tumour-free survival rates were significantly lower in *junD*-deficient females compared to *ras* $junD^{+/+}$ (referred to as *ras*) ones (Fig 1A). Moreover, the number and volume of tumours were higher in *ras* $junD^{-/-}$ mice compared to control animals (Fig 1B and C). To better understand the underlying mechanism, we compared histological properties of *ras* $junD^{-/-}$ and *ras* tumours when they reached the same size (Fig 1D; Fig S1). Inactivation of *junD* generally affected characteristics of the tumours, as well as features of the associated stroma. When compared to *ras* tumours (Fig 1Da,c; Fig S1Aa), deletion of *junD* in *ras*-mediated tumours resulted in an increased proportion of polycystic carcinomas (Fig 1Db; Fig S1Ac) and massive fibrosis (Fig 1Dd), indicated by the accumulation of various forms of collagen. Moreover, the number and the size of blood vessels increased in *ras* $junD^{-/-}$ tumours compared to *ras* (Fig 1De,f), confirming our previous results that *junD* deletion increased angiogenesis *in vivo* (Gerald et al, 2004; Laurent et al, 2008). We quantified each cell type composing the tumours by specific immunohistochemistry staining. Epithelial cells, fibroblasts, myofibroblasts, macrophages and haematopoietic cells were specifically stained using E-cadherin, vimentin, SM-α-actin, F4/80 and CD45-specific antibodies, respectively (Fig 1D and E; Fig S1B). In both genotypes, epithelial cells remained highly differentiated, as evaluated by expression of E-cadherin at cellular surface (Fig 1Dg,h; Fig S1B), further indicating that *junD* inactivation did not promote massive epithelial to mesenchymal transition (EMT). In contrast, the tumour surrounding stroma was quantitatively and qualitatively modified by *junD* deletion. The proportion of CAF was significantly higher in *ras* $junD^{-/-}$ tumours compared to *ras* ones (Fig 1Di,j and E; Fig S1B). Moreover, *ras* $junD^{-/-}$ fibroblasts expressed higher levels of vimentin, a type III intermediate filament, than controls (Fig 1Di,j; Fig S1B) and accumulated podoplanin, a glycoprotein characteristic of reactive stroma (Fig 1Dm,n). Furthermore, *ras* $junD^{-/-}$ tumours exhibited significant increase in myofibroblasts content, evaluated by the number of SM-α-actin-positive fibroblasts (Fig 1Do,p). Finally, the stroma of *ras* $junD^{-/-}$ tumours overproduced the CXCL12 chemokine (Fig 1Dq,r; Fig S1D) and exhibited increased



E

		Tumor	Stroma	Epithelial cells E-cadherin			Fibroblasts Vimentin			Myofibroblasts SM-α-actin			Macrophages F4/80	Hematopoietic CD45
		%		Int	%	H score	Int	%	H score	Int	%	H score	Nb / µm²	Nb / µm²
Mean	*ras*	84,5±4	15,5±4	1,9±0,3	100	189±25	2,9±0,5	100	290±5	1,2±0,3	50±11	61,5±39	0,012±1,4E-3	0,0614±1,2E-3
	ras junD$^{-/-}$	67,8±5	32,2±5	2,1±0,2	100	212±32	4,9±0,3	100	490±3	3,3±0,4	96,7±3	319±26	0,055±2,8E-3	0,089±6,2E-3
pvalue	*ras* vs *ras junD*$^{-/-}$	0,032	0,01	NS	NS	NS	0,05	NS	0,05	0,044	0,047	0,01	0,048	0,05

Figure 1. *junD* inactivation promotes appearance of a reactive stroma and tumour progression.

A. Kaplan–Meyer tumour-free survival curve of *ras junD*$^{+/+}$ (referred to as *ras*) animals ($n = 12$) and *ras junD*$^{-/-}$mice ($n = 12$) ($p = 0.0092$, log-rank test).

B. Number of tumours per animal in *ras* ($n = 12$) and *ras junD*$^{-/-}$ ($n = 12$) mice.

C. Tumour volumes in *ras* ($n = 10$) and *ras junD*$^{-/-}$ ($n = 9$) mice.

D. Sections and histological analysis of epithelial tumours (a–h) and immediate adjacent stroma (i–t) from *ras* or *ras junD*$^{-/-}$ animals. Sections have been coloured with HES (haematoxylin-eosin-saffranin) (a,b,k,l), Masson's trichrome (c,d) or immunostained with specific antibodies, as indicated (e–j,m–t).

E. Percentage of epithelial and fibroblastic compartments in the tumours has been evaluated using E-cadherin and Vimentin-specific staining, respectively. Are also indicated intensity (Int), percentage of positive cells (%) and H scoring (Int × %) for E-cadherin, Vimentin and SM-α-actin-staining as well as the number of F4/80- or CD45-positive cells per µm^2. *n* represents the number of animals analysed per genotype; *n* represents the number of tumours analysed per genotype. Data are means ± SEM. $^{*}p < 0.05$ by student's test. Scale bars = 20 µm in (Da,b,k,l) and 40 µm in (Dc–j,Dm–t).

recruitment of inflammatory cells (Fig 1Dk,l). Staining with specific markers for macrophages (Fig 1Ds,t and E) and haematopoietic cells (data not shown) indicated that both cell types were recruited more efficiently in *ras junD*$^{-/-}$ tumours compared to *ras*. All these observations reveal highly vascularized tumours with reactive stroma in *junD*-deficient *ras*-mediated tumours, features that are not seen in *ras* tumours alone strongly arguing that JunD is involved in these processes.

To further define the function of JunD, we monitored its pattern of expression in *ras*-derived tumours. JunD expression was detected in neoplastic epithelial cells (Fig S1Ca,e) and in stromal fibroblasts (Fig S1Cb,f), suggesting that its deletion can directly impact both compartments. In order to explore the role of JunD only in stromal fibroblasts, we performed transplant experiments by injecting B16F10, a transformed cell line of the same immunotype as our immunocompetent *wt* and *junD*$^{-/-}$ mice. Resulting tumours developed in either a *wt* or a *junD*$^{-/-}$ host. Although *junD*-deficient mice showed earlier tumour onset than *wt* animals (Fig 2A), both types of tumours reached the same mean volume (data not shown) and did not display obvious accumulation of inflammatory cells (Fig S1E). In contrast, tumours developed in *junD*-deficient environment accumulated significantly both SM-α-actin and CXCL-12 (Fig 2B and C). In addition, stromal inactivation of *junD* notably increased the incidence and size of metastases in lungs (Fig 2D and E). To confirm the role of CXCL12 in *junD*-mediated tumourigenesis, we have treated daily grafted *junD*$^{-/-}$ mice by specific CXCL12 siRNA (Fig 2F). Interestingly, silencing of CXCL12 decreased significantly tumour size and prevented lung metastases. Taken together, these results indicate that inactivation of *junD* in the tumour environment is sufficient to modify tumour properties, increase the content in SM-α-actin-expressing cells and promote tumour growth and spread, in a CXCL12-dependent manner.

junD-deficient fibroblasts exhibit features of CAFs

Since *junD* expression has been detected in tumour-associated fibroblasts, we next investigated whether inactivation of *junD*, followed by oxidative stress, was sufficient to alter the properties of fibroblasts. We investigated the gene expression profile and morphology of immortalized *junD*$^{-/-}$ fibroblasts in a tumour-free context and compared them with the already reported characteristics of CAFs (Fig 3). We first identified a subset of 1934 genes that were significantly up-regulated ($p < 0.05$) in *junD*-deficient fibroblasts compared to *wt* cells. We next compared this list (referred to as *junD*$^{-/-}$) with two partially overlapping lists of CAF-specific genes (Allinen et al, 2004; Farmer et al, 2009), set of genes that is also up-regulated in desmoid-type fibromatosis, further underscoring their tumour-associated fibroblastic molecular signature (West et al, 2005) (Fig 3A). The Allinen' and Farmer's lists were composed of 201 and 161 genes, respectively. Among these genes, 44 from Allinen's list and 17 from Farmer's were up-regulated in *junD*$^{-/-}$ *versus wt* fibroblasts (Table S1). This is significantly more than would be expected by chance (namely, $p = 10^{-20}$ for Allinen *versus junD*$^{-/-}$ and $p = 10^{-4}$ for Farmer *versus junD*$^{-/-}$, using Fisher Exact test). Genes encoding extracellular matrix proteins (including collagens I, III, IV, fibronectin, sparc), components of the cytoskeleton (myosin) and matrix metalloproteases (MMP2, MMP14) were up-regulated in the three lists (Fig 3A). These data argue that expression profiles of *junD*$^{-/-}$ fibroblasts are related to the expression signature of CAFs. *junD* inactivation is thus sufficient to confer CAF properties, even in a tumour-free context.

Since CAFs contain a high proportion of myofibroblasts, we analysed whether *junD* deletion caused fibroblasts to adopt myofibroblastic features. Compared to *wt* cells, *junD*$^{-/-}$ fibroblasts exhibited significant accumulation of SM-α-actin (Fig 3Ba,b), increased assembly of F-actin containing stress fibres (Fig 3Bc,d) and recruitment of SM-α-actin into those stress fibres (Fig S2A). Moreover, just as in differentiated myofibroblasts, *junD*$^{-/-}$ fibroblasts differed from *wt* cells in having a significantly higher number of adherens junctions (AJ) (Fig 3Be,f) and an enhanced assembly of mature focal adhesions (FA), characterized by an increase in vinculin, tensin and focal adhesion kinase (FAK) content (Fig 3Bg-l, Fig S2B for quantitative analyses). SM-α-actin protein (Fig 3C) and mRNA (Fig S2C) were increased in *junD*$^{-/-}$ cells compared to *wt*. In contrast, total amounts of all other tested proteins remained similar between the two cell types (Fig 3C), further suggesting that inactivation of *junD* modulated the polymerization of F-actin, the recruitment of N-cadherin to AJ and the association of vinculin or tensin to FA but had only a marginal effect on their total levels. Finally, cellular migration assessed by transwell assays was increased in *junD*$^{-/-}$ fibroblasts as compared to *wt* cells (Fig 3D). Hence, these data show that inactivation of *junD* in fibroblasts converts them into myofibroblasts and increases their cellular migration potential.

CXCL12 plays a key role in acquisition of myofibroblast properties

Since expression of SM-α-actin in myofibroblasts is coordinately regulated by transforming growth factor β1 (TGF-β1) (Hinz et al, 2007; Ronnov-Jessen et al, 1995), we analysed the role of TGF-β1 in acquisition of *junD*$^{-/-}$ myofibroblast properties. Treatment of *junD*$^{-/-}$ cells with an inhibitory drug targeting the TGF-β1 pathway did not alter their myofibroblast properties (such as accumulation of SM-α-actin containing stress fibres), despite clearly decreasing phosphorylation of key TGF-β effector Smad3 (Fig S2D). These observations indicate that the myofibroblast properties of *junD*$^{-/-}$ cells do not result from activation of the TGF-β pathway and strongly suggest that JunD regulates another process, which contributes to the phenotype.

wt fibroblasts incubated with conditioned medium from *junD*$^{-/-}$ cells exhibited accumulation of SM-α-actin containing stress fibres (Fig 4A), further arguing for the role of a secreted factor. The expression of the chemokine CXCL12 was increased in *junD*$^{-/-}$ fibroblasts (Fig S2E). Since we observed that *junD*-dependent CXCL12 accumulation in the stroma was critical for tumour growth and spread, we next investigated if the myofibroblastic phenotype detected in *junD*-deficient cells may be dependent upon CXCL12. Addition of exogenous CXCL12 into the culture medium of *wt* cells was sufficient to

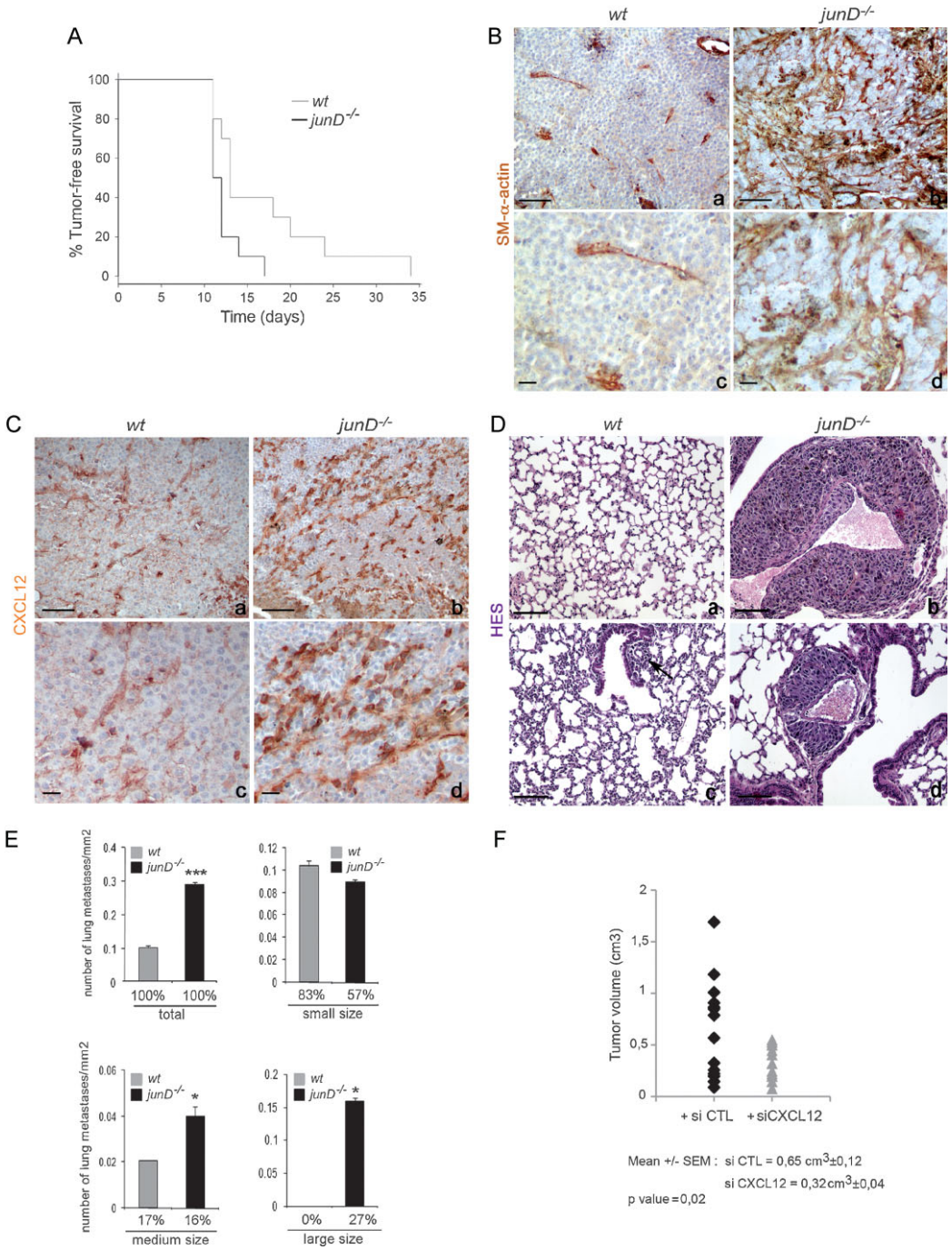


Figure 2. *junD* inactivation in the stroma potentiates tumour metastasis.

A. Kaplan–Meyer tumour-free survival curve of *wt* ($n = 10$) and $junD^{-/-}$ mice ($n = 10$) in graft experiments using B16F10 cells ($p = 0.041$, log-rank test).

B, C. Representative immunohistochemistry of tumours from injected *wt* (a,c) and $junD^{-/-}$ mice (b,d) using SM-α-actin and CXCL12-specific antibodies.

D. Typical HES views of lungs from injected mice. Sections show large (b) and medium (d) sizes of metastatic nodules in $junD^{-/-}$ mice compared to *wt* animals (a,c).

E. Number of total, small-sized (<10 cells), medium-sized (10 cells $< \times <$ 50 cells) and large-sized (>50 cells) metastasis in *wt* and $junD^{-/-}$ mice. Numbers below indicate the percentage of large-, medium- and small-sized metastasis in the respective populations.

F. Tumour volumes in $junD^{-/-}$ mice treated daily either with control siRNA (black) or with specific CXCL12-directed siRNA. Data are means ± SEM. $^{*}p < 0.05$ and $^{***}p < 0.005$ by student's test. *n* represents the number of tumours analysed per genotype. Scale bars = 40 μm.

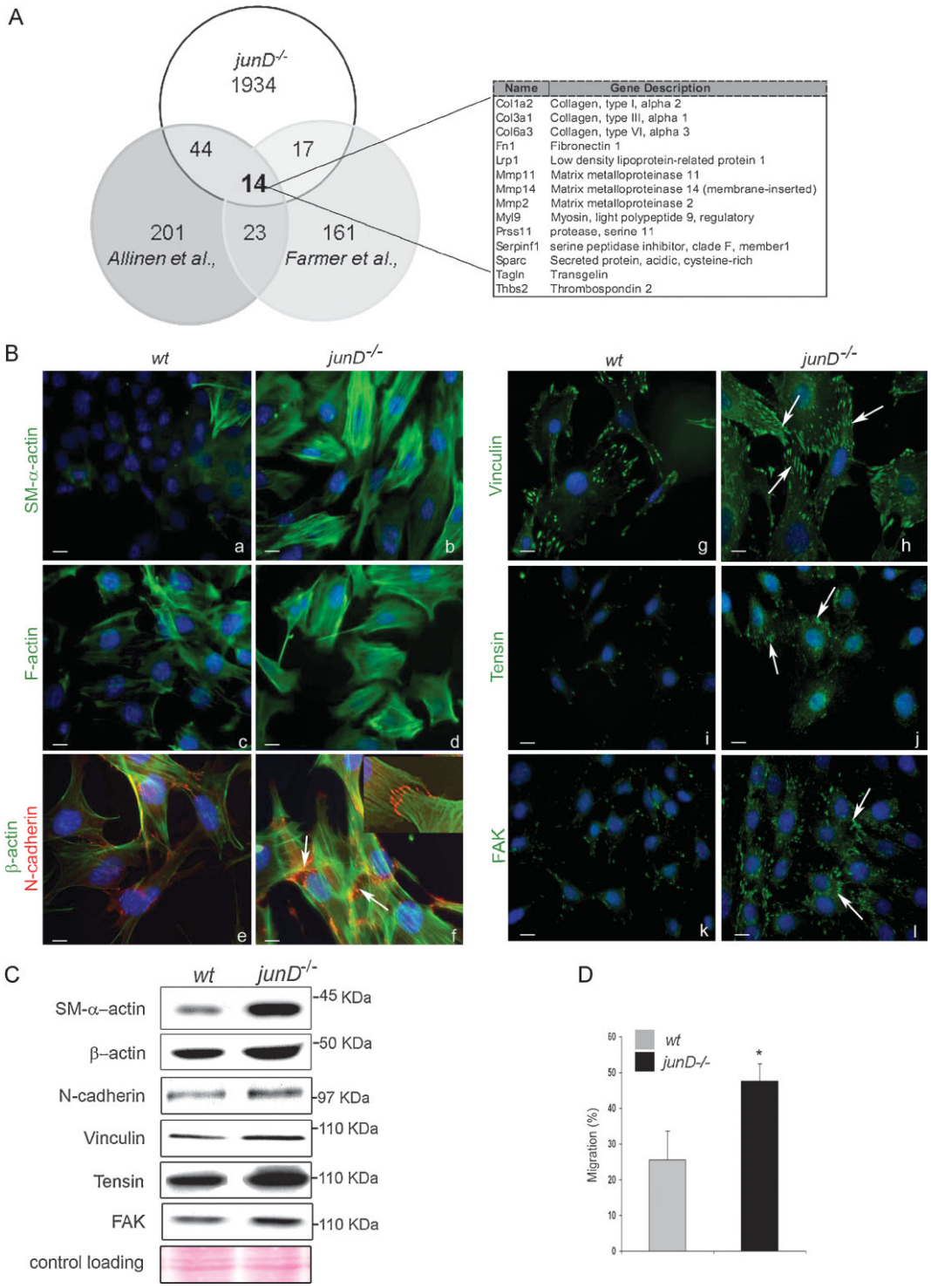


Figure 3. *junD*$^{-/-}$ fibroblasts exhibit features of carcinoma-associated myofibroblasts.

A. Venn's diagram showing the number of common up-regulated genes in CAF (from Allinen's and Farmer's studies) and *junD*$^{-/-}$ fibroblasts. On the right panel, the 14 genes common to the three lists are listed.

B. Representative immunofluorescence staining from *wt* and *junD*$^{-/-}$ cells using specific antibodies, as indicated. Arrows indicate typical staining. Inserted section in (f) shows a higher magnification (100×) image of a representative AJ co-stained with SM-α-actin (in green) and N-cadherin (in red).

C. Western blots of whole cell extracts from *wt* and *junD*$^{-/-}$ fibroblasts. Ponceau colouration was used as an internal control for each protein loading; a representative gel is shown.

D. Migration assay of *wt* compare with *junD*$^{-/-}$ fibroblasts. $^{*}p < 0.05$ by student test. Scale bars = 10 μm (Ba–d,g–l) and 5 μm in (Be,f).

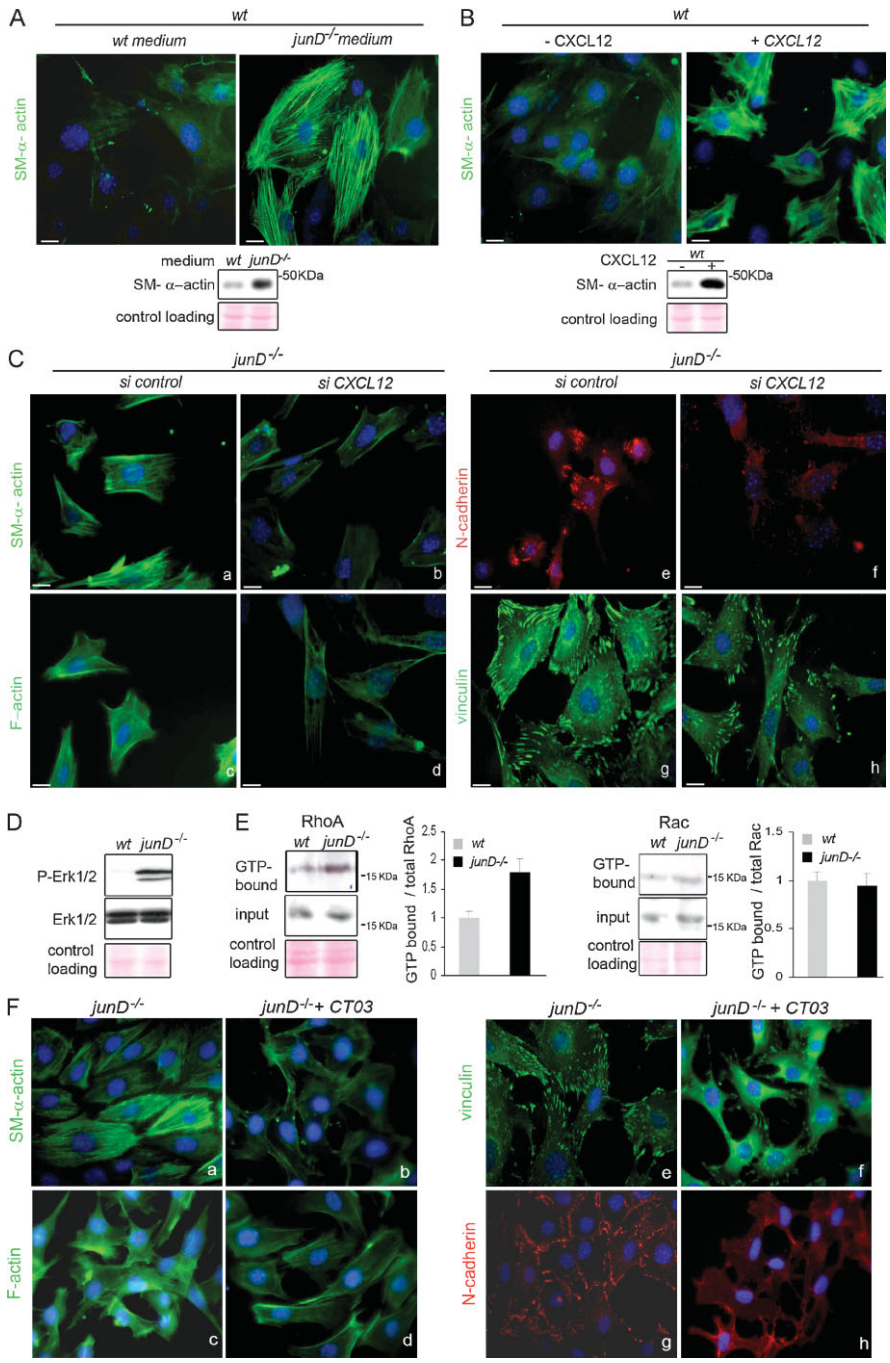


Figure 4. CXCL12 is necessary and sufficient for promoting myofibroblast properties.

A. SM-α-actin staining in *wt* fibroblasts after incubation in *wt* or *junD*$^{-/-}$-conditioned medium. At the bottom is the corresponding Western blot.

B. SM-α-actin staining in *wt* fibroblasts after addition of exogenous CXCL12 protein. At the bottom is the corresponding Western blot.

C. Representative immunofluorescence of myofibroblast markers in *junD*$^{-/-}$ fibroblasts after transfection with a scramble siRNA (si control) (a,c,e,g) or with a CXCL12-directed siRNA (si CXCL12) (b,d,f,h).

D. Western blots from *wt* and *junD*$^{-/-}$ whole cell extracts showing p44 and p42 MAPK (Erk1/2) and their phosphorylated forms.

E. Representative GST pull-down assays on *wt* and *junD*$^{-/-}$ fibroblasts for RhoA (left panel) and Rac (right panel). GTP-bound form and total amount (input) of each protein are shown. Histograms show relative Rho- or Rac-GTP levels normalized to their respective total protein amounts.

F. Representative immunofluorescence of myofibroblast markers in *junD*$^{-/-}$ fibroblasts either untreated (a,c,e,g) or incubated with exoenzyme C3 transferase (b,d,f,h). Scale bars = 10 μm.

increase the proportion of SM-α-actin containing stress fibres (Fig 4B). Conversely, silencing of CXCL12 decreased the level of SM-α-actin, the formation of stress fibres, the number of AJ and the proportion of mature FA in *junD*$^{-/-}$ cells, while the control siRNA had no effect (Fig 4C; Fig S2F). To further validate the CXCL12-dependent autocrine loop in fibroblasts, we confirmed that the CXCR4 receptor was expressed in fibroblasts and detected at the cellular surface (Fig S2G). Moreover, as expected from elevated CXCR4 activity, *junD*-deficient fibroblasts accumulated phosphorylated forms of ERK1/2, a typical response elicited by this G-protein coupled receptor (Fig 4D). Because small Guanosine triphosphate (GTP)-binding proteins of the Rho family play a central role in regulation of the actin-based cytoskeleton and cell movement, we looked at their activation status by pull-down assays (Fig 4E). Although RhoA and Rac protein levels were comparable between *wt* and *junD*$^{-/-}$ fibroblasts (input, Fig 4E and data not shown), *junD*-deficient cells exhibited higher levels of the GTP-bound form of RhoA as compared to *wt* fibroblasts (Fig 4E, left part). However, in parallel experiments, we did not detect accumulation of GTP-Rac in *junD*$^{-/-}$ cells (Fig 4E, right part). Furthermore, treatment of *junD*$^{-/-}$ fibroblasts with exoenzyme C3 transferase (CT03), a drug that inhibits RhoA Guanosine diphosphate (GDP)/GTP exchange activity, severely affected SM-α-actin polymerization, the formation of F-actin containing stress fibres, the number of AJ and the assembly of FA (Fig 4F). Taken as a whole, these data strongly suggest that the CXCL12/CXCR4 pathway activates the RhoA-GTPase and in turn promotes myofibroblastic properties.

HIF-1 is necessary and sufficient for converting fibroblasts into myofibroblasts

It has been shown that hypoxic gradients regulate CXCL12 through HIF induction (Ceradini et al, 2004). Since we demonstrated that HIF-1α protein accumulates in *junD*$^{-/-}$ fibroblasts and mice (Gerald et al, 2004; Laurent et al, 2008), the up-regulation of CXCL12 in *junD*$^{-/-}$ fibroblasts could be mediated, at least partly, through HIF. Specific inhibition of HIF-1α by siRNA strongly reduced HIF-1α mRNA levels (Fig S3A) and decreased the expression of its target gene, CXCL12 (Fig S2E). Moreover, HIF-1α inhibition reduced the proportion of SM-α actin- and F-actin-containing stress fibres in *junD*$^{-/-}$ fibroblasts, as well as the number and size of AJ and mature FA (Fig 5A; Fig S3B). These observations indicate that HIF-1α is a key regulator of the contractile features of *junD*$^{-/-}$ fibroblasts. To further establish whether HIF was sufficient to establish myofibroblast properties, we treated *wt* fibroblasts with desferrioxamine (DFO), an iron chelator that mimicked hypoxia and promoted accumulation of HIF-1α (Fig S3C). Accumulation of HIF-1α in *wt* cells stimulated polymerization of SM-α-actin and F-actin-containing stress fibres, as well as formation of AJ and mature FA (Fig 5B; Fig S3D). Moreover, DFO treatment also increased RhoA activity in *wt* cells (Fig 5C, left part), whilst the same treatment had no effect on Rac activity (Fig 5C, right part). Therefore, treatment of fibroblasts with hypoxia-mimetic DFO was sufficient for activation of RhoA and differentiation into myofibroblasts.

Finally, to investigate if myofibroblast properties detected in *junD*$^{-/-}$ cells were dependent upon chronic oxidative stress, we subjected them to long-term antioxidant treatment. Culturing *junD*$^{-/-}$ fibroblasts with *N*-acetylcysteine (NAC) has been shown to decrease ROS content and HIF protein levels (Gerald et al, 2004). This treatment collectively decreased the contractile features of *junD*-deficient cells (Fig 5D; Fig S3E), further demonstrating the role of ROS in myofibroblastic differentiation. These results suggest redox-dependent accumulation of HIF stimulates the CXCL12/CXCR4 signalling pathway, triggers activation of RhoA and thereby elicits myofibroblast features.

Human HER2-amplified tumours accumulate CXCL12 and myofibroblasts in their stroma

Having established that myofibroblast content correlated with an increased risk of tumour cell dissemination and that stress-induced CXCL12-dependent signalling played a key role in acquisition of myofibroblast properties, we next investigated the potential relevance of these findings in humans. In that purpose, we analysed the pattern of expression of CXCL12/CXCR4, the myofibroblast content and the possible link with oxidative stress in three classes of human breast cancers, chosen according to their distinct invasive properties and clinical outcomes. We compared the stromal properties of (1) Lum-A breast carcinomas, a subtype associated with a good prognosis, (2) HER2-amplified adenocarcinomas, a subset of aggressive tumours characterized by amplification of the HER2/ERBB2 oncogene and high rate of nodal metastases and (3) basal-like cancers (BLC), another type of aggressive and highly proliferative tumours, *albeit* less prone to lymph node metastases than HER2 ones. We first investigated the expression pattern of CXCL12/CXCR4 in both stromal and tumour compartments by performing immunohistochemical staining using tissue microarrays (TMA) from HER2, BLC and Lum-A primary tumours (Fig 6; Table S2). Expression of CXCL12 was significantly increased in HER2-neoplasic cells (Fig 6Aa–c) compared to BLC (Fig 6Ad–f) or Lum-A (Fig 6Ag–i) tumour cells. CXCR4 was also strongly expressed in HER2 (Fig 6Ba–c) and BLC (Fig 6Bd–f) epithelial compartment but to a lesser extent in Lum-A (Fig 6Bg–i), as it has been previously reported (Li et al, 2004; Muller et al, 2001). Interestingly, HER2-amplified tumours exhibited the highest expression levels of both CXCL12 and CXCR4 in the fibroblastic compartment (Fig 6Aa–c and Ba–c), compared to BLC (Fig 6Ad–f and Bd–f) or Lum-A (Fig 6Ag–i and Bg–i; Table S2).

We next evaluated the proliferation rate and myofibroblastic content in these three classes of human breast cancers. In agreement with the previously known characteristics of these tumours, Ki67 nuclear staining showed that the proliferation rate—as detected in both tumour and stromal compartments—was higher in BLC (Fig S4Ad–f), than HER2 (Fig S4Aa–c), itself significantly higher than in Lum-A (Fig S4Ag–i). These observations suggest that the high rate of lymph node metastasis in HER2-derived tumours compared to BLC is not strictly correlated to the proliferation rate. We also evaluated the recruitment of macrophages in each type of breast cancer using CD68-specific marker (Fig S4B). Both forms of aggressive breast

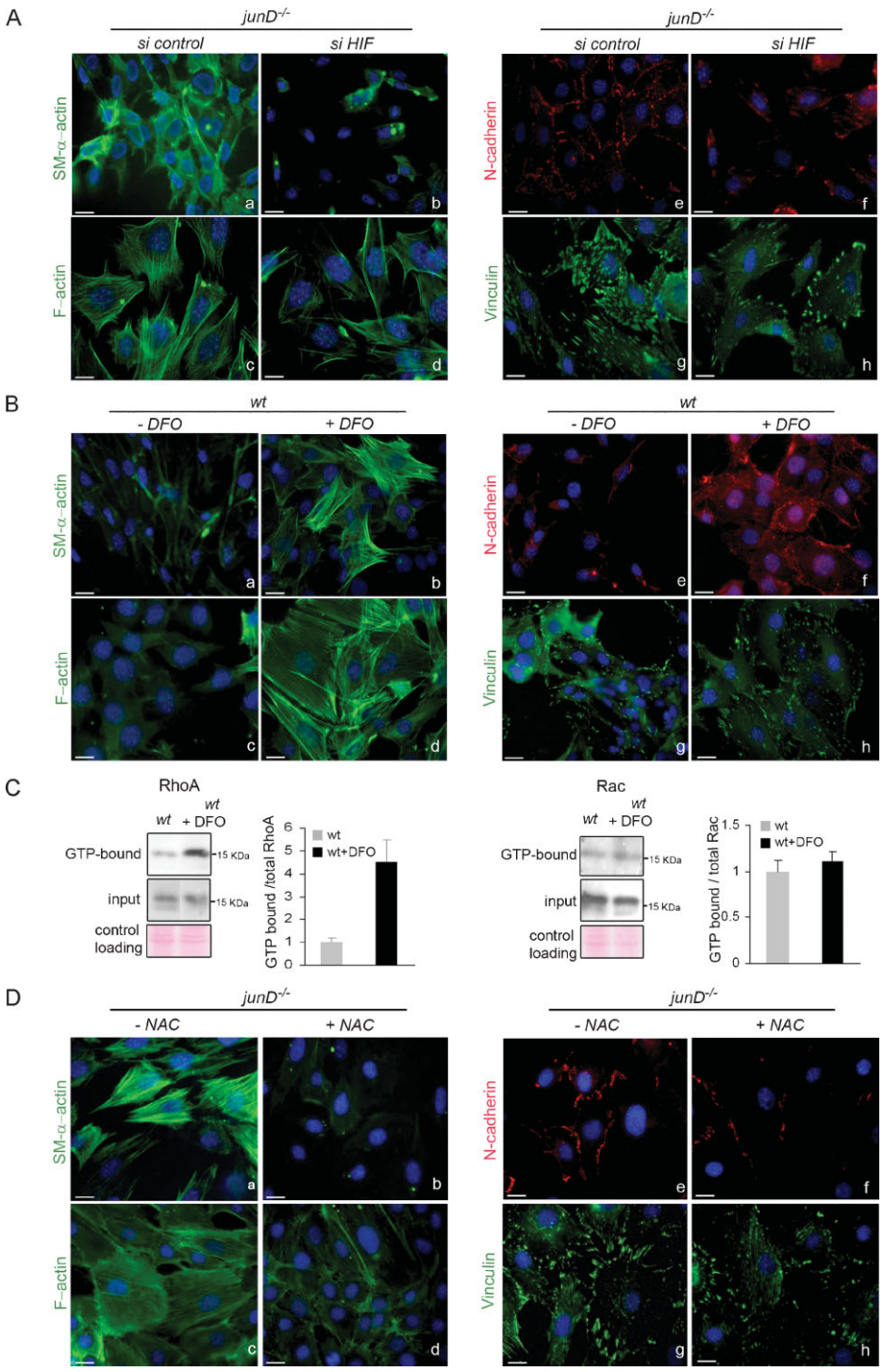


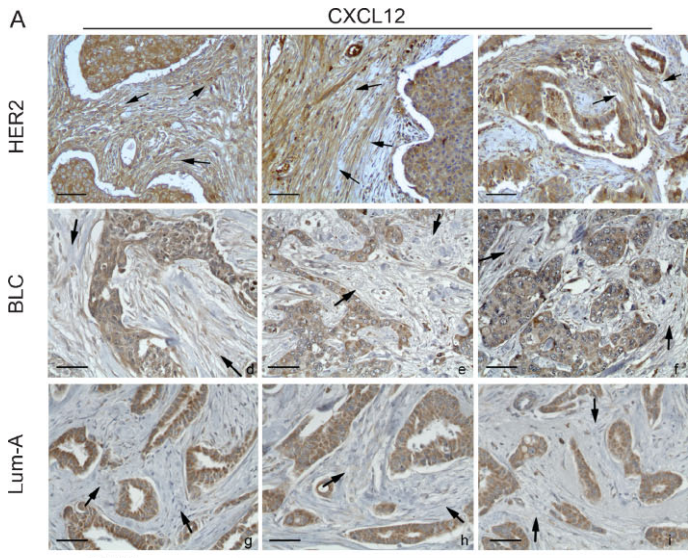
Figure 5. Oxidative stress-mediated HIF-1α accumulation promotes myofibroblast properties.

A. Representative immunofluorescence of myofibroblast markers in *junD*$^{-/-}$ fibroblasts after transfection with a scramble siRNA (si control) (a,c,e,g) or with an HIF-1α-directed siRNA (si HIF) (b,d,f,h).

B. Representative immunofluorescence of myofibroblast markers in *wt* fibroblasts either untreated (a,c,e,g) or incubated with DFO (b,d,f,h).

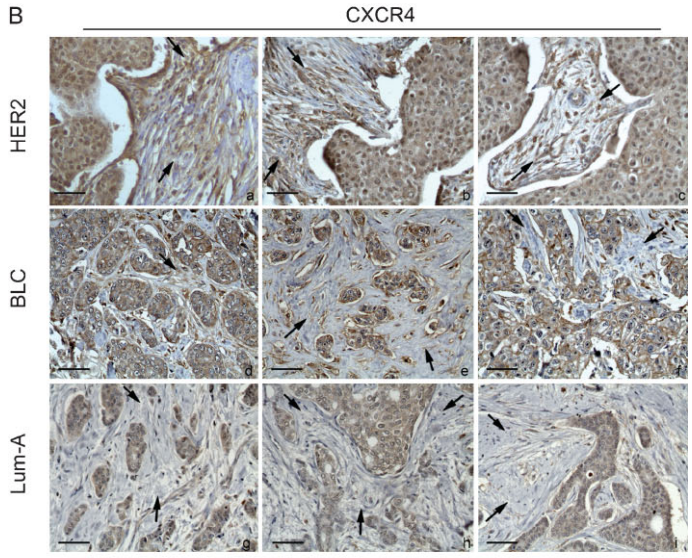
C. Representative GST pull-down assays for RhoA (left panel) and Rac (right panel) on *wt* fibroblasts with or without DFO.

D. Representative immunofluorescence of myofibroblast markers in *junD*$^{-/-}$ fibroblasts either untreated (a,c,e,g) or incubated with NAC (b,d,f,h). Scale bars = 10 μm.



CXCL12

		Epithelium			Stroma		
		Intensity	% cells	H score	Intensity	% cells	H score
Mean	HER2	1,86	98,2	182,43	1,82	68,9	137,6
	BLC	1,06	98,4	104,35	1,07	49,4	64,7
	Lum-A	1,39	96,1	129,13	0,38	24,7	10,7
p value	HER2 vs BLC	0,0005	NS	0,0006	0,0007	0,0026	0,0008
	HER2 vs LUM	0,05	NS	0,05	1,47E-09	3E-08	6,3E-09
	BLC vs LUM	NS	NS	NS	0,001	0,0008	0,0002



CXCR4

		Epithelium			Stroma		
		Intensity	% cells	H score	Intensity	% cells	H score
Mean	HER2	1,15	99,2	120,0	2,20	71,0	159,5
	BLC	1,14	100,0	113,9	1,39	61,8	100,9
	Lum-A	0,41	95,6	40,8	0,20	15,4	5,3
p value	HER2 vs BLC	NS	NS	NS	0,0005	NS	0,02645
	HER2 vs LUM	7,6E-05	NS	8E-05	7,71E-14	2E-10	5,1E-09
	BLC vs LUM	2,4E-06	NS	1E-06	1,44E-10	8,5E-09	1,4E-07

Figure 6. CXCL12 and CXCR4 levels are increased in the stroma of HER2-driven breast cancer.

A. Sections and histological analysis of HER2 (a–c), BLC (d.f) and Lum-A (g–i) human breast tumours using $CXCL_{12}$-specific antibody.

B. Sections and histological analysis of HER2 (a–c), BLC (d.f) and Lum-A (g–i) human breast tumours using $CXCL_4$-specific antibody. Scale bars = 40 μm.

cancers (HER2 and BLC) exhibited a clear enhanced immune response, when compared to the non-aggressive one (Lum-A) (Fig S4B). In contrast, there was no significant difference in macrophages recruitment between HER2 and BLC, suggesting that immune response does not drive increased metastatic potential into lymph nodes in HER2 tumours. To further evaluate the influence of the fibroblastic component of the stroma in each tumour type, we analysed the proportion of myofibroblasts using an SM-α-actin marker (Fig 7A). Among the total population of CAFs, the percentage of SM-α-actin-positive

cells and the intensity of SM-α-actin staining were significantly higher in HER2- and BLC-derived stroma (Fig 7Aa–c; Fig 7Ad–f) than in Lum-A (Fig 7Ag–i). Moreover, we detected a clear difference in the general fibroblasts content in these three subtypes of tumours (Fig 7B). For each tumour, analysis was based on percentage of fibroblasts compared to total cells forming the tumour. The maximum proportion of stromal fibroblasts in BLC was evaluated at 40–45%, while it reached 60–65% in HER2 or Lum-A tumours. Thus, there were qualitative and quantitative differences in CAFs properties in HER2, BLC and Lum-A tumours. Lum-A tumours exhibited high fibroblasts content but low expression of SM-α-actin; BLC tumours showed high expression of SM-α-actin but low fibroblasts content; HER2-amplified tumours exhibited both high fibroblasts content and high expression of SM-α-actin, indicating that this subtype of breast cancers accumulate the highest proportion of myofibroblasts.

Interestingly, in HER2-driven tumours, the myofibroblastic content was significantly correlated with the stromal expression levels of CXCL12 (Pearson correlation coefficient: $r = 0.56$; $p = 5 \times 10^{-4}$). Even more importantly, the myofibroblastic content of HER2-amplified tumours was also significantly correlated with the metastatic rate in lymph nodes ($r = 0.76$; $p = 0.01$), further indicating that myofibroblasts promote migration of tumour cells in lymph nodes. Taken together, these data indicate that accumulation of CXCL12 in the stroma of HER2-amplified tumours is associated with high myofibroblasts content, which impacts tumour spreading in lymph nodes.

HER2-amplified tumours display a gene expression profile involved in oxido-reduction

Since HER2-tumours demonstrated correlated metastatic rate, myofibroblasts content and CXCL12 staining, we next wondered whether genes regulating oxidative stress could be abnormally expressed in this set of tumours. We first performed unsupervised analysis and pathway enrichment studies using all Gene Ontology (GO) terms and the global test methods to examine the predominant signatures in HER2, BLC and Lum-A tumours. The GO terms oxidation–reduction (GO: 0055114) (Fig S5A) and oxido-reductase activity (GO: 0016491) (Fig S6A) appeared among the 100 most significantly deregulated pathways (at the 36th and 83th positions, respectively) with *p*-value smaller than 2×10^{-16}. We confirmed this difference using hierarchical clustering and principal component analysis (Fig S5B–D; Fig S6B–D), further highlighting that these pathways were able to discriminate HER2, BLC and Lum-A tumours. In order to better characterize HER2-specific expression pattern, we next performed supervised clustering according to the tumour subtypes. We defined significantly up-regulated genes in HER2 *versus* Lum-A, BLC *versus* Lum-A and HER2 *versus* BLC and submitted the identified set of genes to GO analysis (Fig 7C). As expected according to the clinical outcomes of the tumours, the up-regulated genes in BLC *versus* Lum-A or HER2 *versus* Lum-A were involved in cell cycle regulation or associated with specific signatures known to denote estrogen-negative tumours or poor prognosis. Interestingly, when comparing up-regulated genes in HER2 *versus* BLC, the first identified statistically relevant GO signature was involved in oxido-reduction (Fig 7C). This confirmed the global test analysis and indicated that one of the major differences identified between these two aggressive breast cancer subtypes — HER2 and BLC — was dependent upon oxido-reduction. Genes that are up-regulated in HER2 tumours when compared to BLC have been directly linked to H_2O_2 generation (NADPH oxidase, Nox4), production of fatty acid hydroxyperoxides (lipoxygenases), oxidative deamination of collagens and elastin (lysyl oxidases), metabolism of xenobiotics (cytochromes P450) (Table S3). Moreover, genes known to be induced upon oxidative stress (such as NQO1) or hypoxia (LOX) were up-regulated in HER2 *versus* BLC tumours. Thus, these data indicate that HER2-amplifying tumours exhibit an expression profile characteristic of a stress response, when compared to BLC. Finally, in order to reconcile parts of this study based on mouse and human analyses, we have tested JunD expression pattern in breast cancers. We did not observe any difference in *junD* mRNA levels in Lum-A, BLC and HER2 tumours. In contrast, we detected significant variations in the subcellular localization of the protein (Fig 8A). Although high levels of JunD were detected in the nucleus of Lum-A tumours, this nuclear localization was reduced in BLC and almost undetectable in HER2-amplified tumours (Fig 8A). Although the involved mechanism remains unknown, exclusion of JunD from the nucleus can efficiently inactivate it and suggests that JunD is far less active in HER2 than in BLC or Lum-A, further correlating with the oxidative stress signature of this breast cancer subtype.

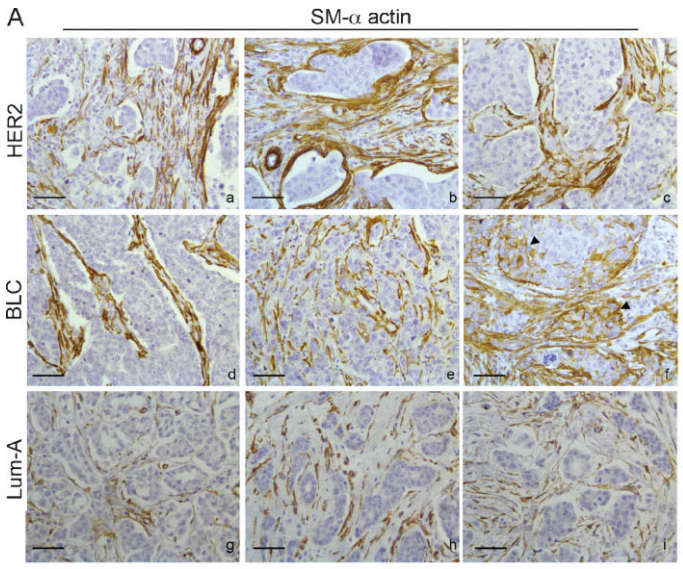
In conclusion, HER2-amplified tumours were characterized by high expression of CXCL12 and accumulation of myofibroblasts and revealed an associated stress response signature, all features that may correlate with their high tendency to metastasize in lymph nodes. Taking both mouse and human studies, our data underline the role of persistent oxidative stress on metastatic spread through the conversion of fibroblasts into myofibroblasts. We have used our data to establish a proposed model, as described in Fig 8B.

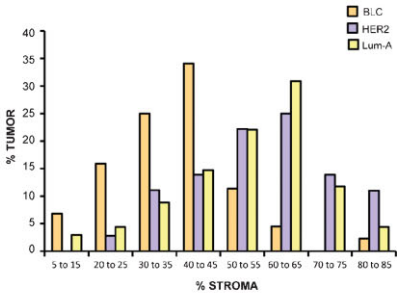
DISCUSSION

An initial step in understanding the mechanisms of stromal reaction in tumour progression is to fully define the reactive stroma phenotype and its formation. In the present study, by combining mouse models and studies on human materials, we uncover a new ROS-dependent mechanism that impacts on tumourigenesis. Collectively, our data indicate that the oxidative stress-mediated accumulation of HIF-1α stimulates the CXCL12/CXCR4 signalling axis that converts fibroblasts into myofibroblasts and is associated with a high rate of metastases in both mouse and human adenocarcinomas.

Role of myofibroblasts in metastatic spread

The current evidence considers the surrounding tumour microenvironment as an important determinant in the final outcome of cancer. It has been clearly established that CAFs, one of the most abundant stromal components, promote tumour cell proliferation and angiogenesis (Allinen et al, 2004; Bhowmick





p value (χ^2)
HER2 vs BLC : p=0,001
HER2 vs Lum-A : NS
BLC vs Lum-A : p=0,001

SM-actin

		Epithelium			Stroma		
		Intensity	% cells	H score	Intensity	% cells	H score
Mean	HER2	0	0	0	3,19	54,9	178,8
	BLC	0,56	8,49	0,30	3,03	61,5	193,0
	Lum-A	0	0	0	2,63	45,4	128,0
p value	HER2 vs BLC	0,008	0,016	0,020	NS	NS	NS
	HER2 vs LUM	NS	NS	NS	0,048	NS	0,019
	BLC vs LUM	0,008	0,016	0,020	NS	0,022	0,005

C

	GO.ID	Observed in selection	Observed on array	BH Fisher pValue	GO.Term
HER2 vs LUM-A					
	Gene Ontology Biological Process				
	GO:0051301	64	191	8,7E-14	**cell division**
	GO:0007067	52	141	9,9E-13	**mitosis**
	GO:0007049	91	352	3,8E-13	**cell cycle**
	Gene Ontology Cellular Component				
	GO:0005819	18	42	3,0E-06	**spindle**
	GO:0005654	72	338	3,2E-08	**nucleoplasm**
	GO:0000775	16	37	9,6E-06	**chromosome, centromeric region**
	Broad/MIT Pathways				
Prognosis	BRCA_ER_NEG	317	900	1,9E-84	Higher expression is associated with **ER-negative** tumors
	BRCA_PROGNOSIS_NEG	56	91	6,4E-24	Higher expression is associated with **metastasis** and **poor prognosis**
	BREAST_DUCTAL_CARCINOMA_GENES	16	19	3,4E-09	Genes upregulated in **high tumor grade** breast cancers (IDC vs DCIS)
Metastases	CANCER_UNDIFFERENTIATED_META_UP	42	67	1,7E-18	Up-regulated in **undifferentiated** cancer (OncoMine database)
	CANCER_NEOPLASTIC_META_UP	24	64	9,7E-08	Up-regulated in **cancer** relative to normal tissue (OncoMine database)
	GILDEA_BLADDER_UP	9	27	2,0E-03	Differentially expressed in **metastatic** and nonmetastatic human bladder cancer
	ZUOCHI_EPITHELIAL_UP	11	42	3,3E-03	Up-regulated genes in primary **IDC** isolated from lymph nodes vs normal tissue
	EMT_UP	14	60	2,5E-03	Up-regulated in **TGFbeta-induced EMT** of Ras-mammary epithelial cells
Serum response	SERUM_FIBROBLAST_CELLCYCLE	82	131	7,7E-35	Cell-cycle dependent genes regulated by **serum** in human fibroblasts
	SERUM_FIBROBLAST_CORE_UP	53	189	2,7E-11	Up-regulated genes by **serum** in human fibroblasts
	CHANG_SERUM_RESPONSE_UP	38	144	6,5E-08	Serum **Response** signature for activated genes (Stanford)
BLC vs LUM-A					
	Gene Ontology Biological Process				
	GO:0007049	125	352	6,8E-18	**cell cycle**
	GO:0007067	65	141	2,6E-14	**mitosis**
	GO:0051301	80	191	6,5E-15	**cell division**
	Gene Ontology Cellular Component				
	GO:0005654	98	338	5,5E-11	nucleoplasm
	GO:0000775	20	37	2,9E-06	chromosome, centromeric region
	GO:0000777	15	26	2,8E-05	condensed chromosome kinetochore
	Broad/MIT Pathways				
Prognosis	BRCA_ER_NEG	569	900	2E-193	Higher expression is associated with **ER-negative** tumors
	BRCA_PROGNOSIS_NEG	75	91	9E-30	Higher expression is associated with **metastasis** and **poor prognosis**
	VANTVEER_BREAST_CUTCOME_GOOD_VS_POOR	52	65	1E-20	Marker genes in Breast Cancer - **Poor prognosis**
	BRCA_BRCA1_POS	55	104	7E-16	Positively correlated with **brca1** germline status in breast cancer - sporadic tumors
Serum response	SERUM_FIBROBLAST_CELLCYCLE	98	131	4E-36	Cell-cycle dependent genes regulated by **serum** in human fibroblasts
	SERUM_FIBROBLAST_CORE_UP	75	189	1E-15	Up-regulated genes by **serum** in human fibroblasts
	CHANG_SERUM_RESPONSE_UP	58	144	1E-12	Serum **Response** signature for activated genes (Stanford)
HER2 vs BLC					
	Gene Ontology Biological Process				
	GO:0055114	61	503	1,6E-07	**oxidation reduction**
	Gene Ontology Molecular Function				
	GO:0016491	52	433	1,3E-06	**oxidoreductase activity**
	Gene Ontology Cellular Component				
	GO:0005581	7	20	0,0005	**collagen**
	Broad/MIT Pathways				
Prognosis	BRCA_BRCA1_NEG	61	147	1,3E-30	Higher expression is associated with BRCA1 tumors
	BRCA_PROGNOSIS_POS	13	39	1,1E-06	Higher expression is associated with good prognosis
	VANTVEER_BREAST_CUTCOME_GOOD_VS_POOR	9	26	3,9E-05	Marker genes in Breast Cancer - Good prognosis
	BREAST_CANCER_ESTROGEN_SIGNALING	15	95	1,2E-04	Estrogen-receptor-dependent signal transduction
Oxidoreduction	OXIDOREDUCTASE_ACTIVITY	40	276	5,5E-08	**Catalysis of an oxidation-reduction reaction**
	TRANSITION_METAL_ION_TRANSPORTER_ACTIVITY	3	10	4,0E-04	**Catalysis of transition metal ion transporter activity**
Metabolism	CARBOXYLIC_ACID_METABOLIC_PROCESS	21	176	1,3E-03	Chemical reactions involving carboxylic acids
	ORGANIC_ACID_METABOLIC_PROCESS	21	178	1,4E-03	Chemical reactions involving organic acids
Hypoxia	MENSE_HYPOXIA_UP	14	97	1,4E-03	**List of Hypoxia-induced genes** found in both Astrocytes and HeLa Cell

Figure 7. HER2 tumours exhibit high content of myofibroblasts and display a stress response signature.

A. Representative staining of SM-α-actin in HER2 (a–c), BLC (d–f) and Lum-A (g–i) human breast tumours. Arrowheads indicate SM-α-actin staining in epithelial cells in BLC (f). p-values are as in Fig 6.

B. Representative graph of the percentage of fibroblasts compared to total cells forming the tumour in each breast cancer subtype. *p*-values by χ^2 test are highly significant between BLC and HER2 or Lum-A ($\leq$0.001) and non-significant between HER2 and Lum-A.

C. Gene Ontology pathways significantly at least 2-fold up-regulated in HER2 *versus* Lum-A, BLC *versus* Lum-A and HER2 *versus* BLC, as indicated. *p*-values by Fisher Exact test adjusted using the Benjamini–Hochberg correction are indicated.

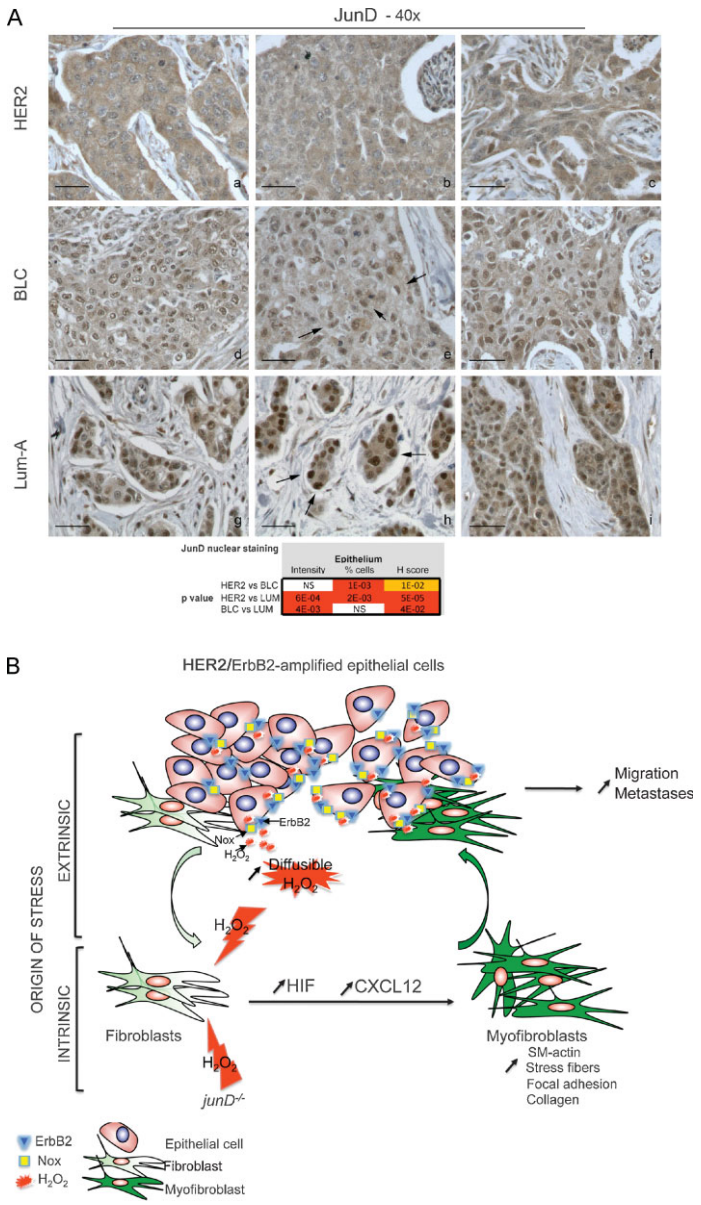


Figure 8. JunD nuclear exclusion in HER2-tumours and model describing ROS, as key players in the reciprocal cross-talk between tumour cells and surrounding fibroblasts.

A. Representative staining of JunD in HER2 (a–c), BLC (d–f) and Lum-A (g–i) human breast tumours. *p*-values by χ^2 test are highly significant between BLC and HER2 or Lum-A.

B. Model: In carcinoma, chronic oxidative stress promotes the conversion of fibroblasts into myofibroblasts, contractile cells with high migration properties capacity. Pro-invasive myofibroblast properties result from ROS-mediated accumulation of the pro-angiogenic HIF-1α factor and the CXCL12 chemokine. The origin of oxidative stress can be either intrinsic or extrinsic to fibroblasts. Indeed, fibroblasts may acquire genetic alteration, such as *junD* inactivation, which will intrinsically increase ROS contents. Since genetic alteration in stromal fibroblasts may be rare events, stress may also more often originate from tumour cells themselves. Accordingly, we identified a signature characteristic of stress–response in HER2-amplified tumours. In this set of tumours, stress can originate from non-exclusive mechanisms, such as JunD nuclear exclusion, up-regulation of Nox4 or HER2/ERBB2 amplification *per se*, since it has been previously demonstrated that growth factor-stimulated RTKs enhance H_2O_2 levels through activation of Nox. H_2O_2 is highly diffusible and can easily cross-cellular membranes to act on surrounding fibroblasts. Acute stress in fibroblasts increases levels of HIF and CXCL12 that, in turn, convert fibroblasts into myofibroblasts. These highly contractile SM-α-actin-expressing cells would subsequently promote migration and dissemination of neoplastic cells.

et al, 2004; Kalluri & Zeisberg, 2006; Littlepage et al, 2005; Olumi et al, 1999; Orimo et al, 2005; Tlsty & Coussens, 2006). Moreover, CAFs have been recently shown to orchestrate tumour inflammation (Erez et al, 2010), further highlighting a new interaction between fibroblasts and immune cells. The link between inflammation and cancer has also been well demonstrated. Indeed, inflammatory immune cells are highly recruited in carcinomas and promote tumour growth and metastases

(Condeelis & Pollard, 2006; Coussens & Werb, 2002; de Visser et al, 2006; Erez et al, 2010; Grivennikov & Karin, 2010; Joyce & Pollard, 2009; Karin & Greten, 2005; Mantovani et al, 2008; Radisky & Radisky, 2007; van Kempen et al, 2006). Our work complements these views and assigns to CAFs an essential role in increasing the risk of metastatic dissemination by modulating SM-α-actin expression in a CXCL12-dependent manner. Indeed, the myofibroblast-enriched *junD*-deficient stroma accelerates tumour onset and increases the number and the size of metastases. Moreover, the HER2 class of human breast cancers, which displays a high rate of lymph node metastases, exhibits a significantly correlated high proportion of SM-α-actin-expressing fibroblasts. In contrast, increased rate of metastasis was not correlated with enhanced immune response, suggesting that other mechanism exists. Similar to our findings, experimental and clinical data support the notion that stromal myofibroblasts participate in tumour development. Moreover, recent data show that genetic inactivation of *Pten* in stromal fibroblasts of mouse mammary glands accelerates tumour initiation and progression by massive remodelling of extracellular matrix (ECM) and increased angiogenesis (Trimboli et al, 2009). In xenograft mouse models, myofibroblasts were shown to stimulate growth of human breast cancer cells (Olumi et al, 1999; Orimo et al, 2005). Moreover, *in vitro* co-cultures of myofibroblasts and tumour epithelial cells demonstrated that myofibroblasts promote invasion of breast, colon, pancreas and squamous carcinoma cells (Casey et al, 2008; De Wever et al, 2004; Hwang et al, 2008; Lewis et al, 2004). Similarly, progression of *in situ* to invasive breast carcinoma is promoted by fibroblasts and inhibited by normal myoepithelial cells (Hu et al, 2008). Furthermore, mesenchymal stem cells and derivatives within tumours promote breast cancer metastasis (Karnoub et al, 2007). In clinical studies, the abundance of stromal myofibroblasts predicts human disease recurrence. Tumours with abundant myofibroblasts are associated with significantly shorter event-free survival rates for stages II and III human colorectal cancers (Tsujino et al, 2007). In addition, in lung and breast adenocarcinomas, myofibroblast content is significantly correlated with lymph node metastasis and shortened patient survival (Tokunou et al, 2001; Yazhou et al, 2004). Accordingly, we show here that HER2-amplifying invasive adenocarcinomas, human breast cancers that display high rate of lymph node metastasis, are associated with the highest proportion of myofibroblasts, when compared to BLC or Lum-A. Taken together, these data emphasize that tumour spreading could be facilitated by the myofibroblastic component of the stroma.

Myofibroblasts express genes that encode invasion associated-secreted factors, ECM proteins and ECM remodelling proteases, which may facilitate tumour cell dissemination. The transcriptome of CAFs, as well as *junD*-deficient fibroblasts, shows abundant expression of collagens, cytoskeleton components, cell adhesion molecules and MMPs. Moreover, we observe that HER2-breast tumours reveal a high content of various collagens, when compared to BLC. Interestingly, imaging of invading co-cultures of squamous cell carcinoma with stromal fibroblasts revealed that the leading cell is always a fibroblast (Gaggioli et al, 2007). In this study, tumour cells migrate within tracks of ECM molecules, secreted by the leading fibroblasts, whose migration is dependent on Rho-GTPase. Thus, myofibroblasts trigger both deposition and proteolysis of ECM molecules that promote migration of cancer cells. Accordingly, collagen density in mammary tissue significantly increases tumour formation and results in an invasive phenotype with high numbers of lung metastases (Provenzano et al, 2008). In accordance with the prevalent role of tumour microenvironment in cancer cell dissemination, a stromal gene expression signature, similar to that observed during wound healing, predicts human breast and prostate cancer progression and patient survival (Bacac et al, 2006; Chang et al, 2004, 2005; West et al, 2005). All these data strongly suggest that stromal myofibroblasts impact tumour aggressiveness.

Oxidative stress-dependent origin of myofibroblasts

Although the importance of CAFs in tumour development is becoming clear, their origin is still controversial and the basis of their myofibroblast characteristics debated (Haviv et al, 2009; Hinz et al, 2007; Ostman & Augsten, 2009). In adenocarcinomas, it has been suggested that myofibroblasts derive from epithelial cells throughout EMT (Kalluri & Weinberg, 2009; Neilson, 2006; Radisky et al, 2007; Zavadil et al, 2008). Although some data indicate that the stromal compartment, when microdissected from human breast cancers, exhibits genetic alterations (Eng et al, 2009), the proportion of karyotypic alterations in fibroblasts remain less frequent than in cancer cells (Qiu et al, 2008). In addition to EMT, recent data from human breast cancer and animal models established that tumour-associated myofibroblasts can also derive from haematopoietic or mesenchymal stem cells from the bone marrow (Direkze et al, 2004; Ishii et al, 2003; LaRue et al, 2006; Mishra et al, 2008; Mori et al, 2005; Studeny et al, 2004). Local resident fibroblasts or fibroblasts stimulated by members of the TGF-β family have also been considered as a major source of CAFs (Kalluri & Zeisberg, 2006; Mueller et al, 2007; Ostman & Augsten, 2009). Consistently, we demonstrate here that fibroblasts can differentiate into myofibroblasts upon stress *in vitro*, effect which is reversed by long-term antioxidant treatment. Moreover, our study shows different patho-physiological conditions (*junD*-deficient animals, HER2-amplified breast adenocarcinomas), in which stress response is associated with myofibroblast accumulation. Thus, our data are consistent with previous published data and further argue that myofibroblasts can originate from stress-exposed fibroblasts.

Many distinct biological circumstances can stimulate oxidative stress in tumours. The origin of oxidative stress in tumours can be either intrinsic or extrinsic to fibroblasts. Acquisition of genetic alterations (*e.g.* p53 loss) in the stroma, either induced by cancer cells or by chemotherapy, may allow a local production of ROS that would further promote the appearance of myofibroblasts. In addition, MMPs, such as MMP3, a remodelling enzyme that is up-regulated in the earliest stage of human breast cancer, have been shown to modulate activity of the mitochondrial respiratory chain, subsequently enhancing ROS content and stimulating tumour progression (Radisky et al,

2005). MMP3 was also defined as a major factor secreted from senescent fibroblasts (Parrinello et al, 2005), which constitute an inflammatory environment that compromises the structure and the functions of the surrounding tissue (Coppe et al, 2008). Interestingly, genotoxic stress and persistent DNA damage signalling promote the development of senescent fibroblasts that resemble reactive stroma and stimulate tumour growth (Rodier et al, 2009). Thus, stress-signalling originating from tumour epithelial cells may modulate local environment. In that respect, the stress–response signature that we detect in HER2-amplified tumours, expressing high levels of ERBB2 receptor, is interesting. It has been shown that stimulation of receptor tyrosine kinase (RTK) by growth factors, such as epidermal growth factor (EGF), is associated with ROS generation (mainly H_2O_2) through activation of non-phagocytic Nox (Aslan & Ozben, 2003). ROS production is crucial for RTK-downstream signalling pathways and stimulation of proliferation. Application of an antioxidant treatment to HER2-transformed cells demonstrates that ROS are important for their proliferation and survival (Preston et al, 2001; Wang et al, 2005). In agreement with these data, the present study shows that genes that are directly linked to H_2O_2 synthesis, such as Nox4, are significantly up-regulated in HER2-amplifying tumours. Interestingly, HER2-amplified tumours display significant exclusion of JunD from the nuclear compartment, when compared to BLC and Lum-A. Nox4 has been already shown to be up-regulated in *junD*-deficient cells (Gerald et al, 2004). Taken together, these data suggest that amplification and constitutive activation of ERBB2 is associated with reduced JunD activity and massive ROS production, potentially through Nox activation. Since H_2O_2, despite short half-life, can easily diffuse among cellular membranes, H_2O_2 production by HER2-neoplasic cells may have an impact on surrounding fibroblasts and promote their conversion into myofibroblasts. Accordingly, HER2-amplified tumours exhibit high content in myofibroblasts, which is significantly correlated with lymph node metastases. Thus, oncogene-mediated hyperplasia may trigger a stress-mediated stromal reaction that initiates a vicious cycle promoting tumour invasion.

Role of HIF and CXCR4/CXCL12 signalling pathways in myofibroblast differentiation and metastases

We uncover a novel cell autonomous HIF-mediated mechanism that regulates differentiation of fibroblasts into myofibroblasts and may enhance tumour spreading. Previously, a stroma-derived predictor of poor prognosis, consisting of angiogenic and hypoxic gene expression, was discovered (Finak et al, 2008). Similarly, it has been suggested that a hypoxia-induced HIF-mediated response reflects metastatic potential in soft tissue sarcomas (Francis et al, 2007). Finally, HIF-1α expression was correlated to aggressiveness in breast cancers, especially in node positive HER2-driven carcinomas (Giatromanolaki et al, 2004; Gruber et al, 2004; Vleugel et al, 2005; Yamamoto et al, 2008). Consistently, we identified a hypoxia-related signature among the genes that are significantly up-regulated in HER2 tumours, when compared to BLC. Thus, HIF-dependent signature has been closely linked to aggressive phenotype in human cancers, further arguing for the deleterious effect of chronic stress.

In addition to HIF, we identified the chemokine CXCL12 as a new mediator for the formation of contractile features in myofibroblasts. CXCL12 has already been involved in tumour growth and metastatic spread, mostly through its role in chemoattraction (Muller et al, 2001; Ostman & Augsten, 2009). Indeed, CXCL12 stimulates carcinoma cell proliferation and recruitment of endothelial precursor cells (Littlepage et al, 2005). Moreover, CXCR4-positive tumour cells are attracted to CXCL12-expressing metastatic organs, through a chemotactic gradient (Zlotnik, 2008). Furthermore, CXCL12 has been defined as a master regulator of trafficking of haematopoietic- and cancer-stem cells (Gelmini et al, 2008). Finally, CXCL12-dependent tumour cell migration has also been associated with macrophages and their cross-talk with tumour cells (Joyce & Pollard, 2009). In agreement with this previously identified chemoattractive function, we detected a significant increase in recruitment of inflammatory cells in our mouse model of mammary adenocarcinomas. In contrast, this chemoattractive effect was not detected in the transplanted tumour model, despite clear accumulation of CXCL12. Similarly, HER2-human breast cancers did not show enhanced rate of CD68-positive cells when compared to BLC, while CXCL12 accumulated. In contrast, SM-α-actin-expressing fibroblasts accumulated in the stroma of the two mouse models used in this study. Thus, our study complements previous published data about CXCL12 by deciphering a new function for this chemokine in stromal fibroblasts and acquisition of contractile properties. Indeed, we show here that CXCL12 is necessary and sufficient to convert normal fibroblasts into myofibroblasts. Accordingly, CXCL12 has been previously shown to induce intracellular actin polymerization in lymphocytes (Bleul et al, 1996). We show here also that HER2 human breast cancers, which are associated with increased risk of nodal metastasis, exhibit a statistically significant increase in CXCL12 and CXCR4 expression in the stroma. Similarly, CXCR4 expression in breast cancers has been correlated with survival and metastatic development (Kang et al, 2005; Li et al, 2004; Muller et al, 2001). This suggests that there may be some value in developing CXCR4-blocking antibodies for lymph node-positive HER2 patients, in addition to the already existing treatments (Baselga & Swain, 2009). Many reports have associated stromal CXCL12-positive staining with increased aggressive metastatic foci in human cancers (Kryczek et al, 2007). Fibroblast-derived CXCL12 stimulates the invasion of oral squamous cell carcinoma and CXCL12 blocking antibodies reduce the level of invasion (Daly et al, 2008). In our model of fibroblasts, CXCL12 triggers activation of the Rho family of GTPases and increases migratory properties, as it has been shown in melanoma cells and inflammatory cells (Bartolome et al, 2004; Tan et al, 2006). Moreover, previous studies on the pro-invasive capacity of myofibroblasts revealed an essential role of the RhoA/Rho-kinase axis in cancer cell invasion (De Wever et al, 2004; Nguyen et al, 2005). Altogether, these data suggest that activation of RhoA triggered by CXCL12/CXCR4 signalling in stressed fibroblasts could contribute to dissemination of tumour cells.

In conclusion, our data support the conclusion that production of CXCL12 by stromal fibroblasts is tightly regulated by oxidative stress in an HIF-dependent manner. Most probably, this enhances metastatic spread by increasing the migratory potential of both tumour cells and their associated stroma. These data provides new insights into the contribution of oxidative stress to the tumour-associated microenvironment and may contribute to a better knowledge of mechanisms stimulating cell dissemination.

MATERIALS AND METHODS

Cell culture and siRNA

Independent immortalized cell lines derived from *wt* or *junD*$^{-/-}$ embryos were generated using a conventional 3T3 protocol as previously described in Gerald et al (2004). Experiments were performed at least in triplicate on three independent cell lines of each genotype. Treatments of cells were performed for 16 h by addition of exogenous CXCL12 (100 nM), a kind gift of Arenzana-Seistedos, for 4 h with DFO (100 μM) (D9533-Sigma) and 3 h with exoenzyme C3 transferase (1 μg/ml) (CT03A-cytoskeleton). Long-term antioxidant treatment, using NAC (500 μM) (A9165-Sigma), was applied for 20 days, with addition of the product every 2 days. For siRNA experiments, cells were transfected with 25–50 nM siRNA using Dharmafect1 reagent (Dharmacon). siRNA sequences targeting mouse HIF-1α and CXCL12 were, respectively, 5′-CCCUAUAUCCCAAUGGAUG-3′ and 5′-CAACG-UCAAGCAUCUGAAA-3′.

Mouse strains and graft experiments

Due to male sterility, the *junD*-deficient mice were maintained through the breeding of heterozygous animals. *Ras junD*$^{+/+}$ and *Ras junD*$^{-/-}$ mice have been obtained by crossing *junD*$^{+/-}$ mice with *MMTV-Ha-Ras* mouse mammary tumour model (Sinn et al, 1987). Mice were checked weekly for tumour growth. On average, tumours appeared in 15-month-old animals. Tumour volume was determined by the use of a calliper and the following equation: $0.5 \times [\text{length} \times (\text{width})^2]$. For immunohistochemistry, tumours were fixed in 4% paraformaldehyde (PFA) for 1 h 30 min at room temperature and then embedded in gelatin 15%/sucrose 7.5%. Ten micrometre sections were incubated with antibody against SM-α-actin (A2547, clone1A4, Sigma), CXCL12 (MAB350, R&D system), CD31 (7388-50, Abcam), E-cadherin (4065, Cell Signaling), Vimentin (RV202, Abcam), Podoplanin (Abcam, ab11936), F4/80 (ab6640, Abcam) or coloured with Masson's trichrome or haematoxyline-eosin-saffranin (HES). Fibroblasts, myofibroblasts, epithelial cells, macrophages and haematopoietic cells were specifically stained using vimentin, SM-actin, E-cadherin, F4/80 and CD45-specific antibodies, respectively. For quantification, three different tumours of each genotype were analysed and three sections from distinct areas of each tumour were evaluated. Staining intensity and percentage of labelled cells were scored for SM-actin, vimentin and E-cadherin positive cells. Numbers of CD45- or F4/80-positive cells per tumour surface were also quantified. Statistical analysis were done using student test. Graft experiments were performed using 10-month-old mice. Single cell suspensions containing 2×10^6 B16F10 cells in 200 μl were injected subcutaneously. The mice were checked daily for tumour growth and tumours were measured using callipers. Tumours and lungs were collected when tumours reach appreciatively 2 cm^3. Removals were fixed in 10% formol, sectioned in paraffin (5 μm) and coloured. When required, mice have been treated daily with 3 μg of control- or CXCL12-specific siRNA (sequence above), previously validated on cells. The Institut Curie ethical committee approved all experiments.

Gene expression analysis

Whole genome expression profiling of *wt* and *junD*$^{-/-}$ fibroblasts were performed using mouse expression beadchip (Sentrix Mouse-6 v1.1) from Illumina (see Supporting Information). Three RNA samples were pooled according to their genotype. Pooled RNA were used to synthesize cRNA and hybridized to Illumina mouse-6 expression arrays (version 1). Detected probe sets were selected ($n = 46{,}673$) and further analysed using beadstudio software. The background was subtracted and intensities normalized using cubic spline algorithm. Only at least 2-fold up-regulated genes with a significant *p*-value ($p < 0.05$) ($n = 1934$) in *junD*$^{-/-}$ *versus wt* cells were taken into account. The comparison between lists of genes obtained in independent gene expression analyses was carried out using the hypergeometric law via Fisher's Exact test in R (The R development Core Team, R: A Language and Environment for Statistical Computing, Version 2.8.1, 2008).

Gene expression profiling and pathway enrichment analysis in human breast cancers

Only human tumours with a high content in epithelial tissue (at least 65%) have been used. Total RNA were extracted from frozen tumours with TRIzol reagent (Life Technology, Inc.) and purified using the RNeasy MinElute Cleanup kit (Qiagen). RNA quality was checked on an Agilent 2100 bioanalyser. Samples were analysed on Human Genome U133 Plus 2.0 array (Affymetrix), according to manufacturer's procedures. Log-intensity values were normalized using the GC-RMA algorithm. Probes, with log-intensity value smaller than 3.5, were discarded. A linear model was then fitted to detect and correct any batch and hybridization effects. All GO pathways were retrieved using the R software and Bioconductor. We first applied the global test method proposed by Goeman et al (2004), investigating whether the expression pattern of a group of genes is significantly related to a clinical outcome of interest. We also performed supervised comparative analysis using Welch test and adjusted *p*-values using Benjamini–Hochberg procedure (R-Multitest package).

Immunofluorescence and immunoblotting

Fluorescence microscopy was performed as previously described with few modifications (Mechta et al, 1997). In brief, cells were fixed in 4% PFA for 30 min, permeabilized in 0.01% sodium dodecyl sulphate (SDS) for 10 min, rinsed twice in phosphate buffered saline (PBS) solutions and blocked for 30 min in 10% foetal calf serum (FCS). Cells were stained with 4,6-diamidino-2-phenylindole (DAPI) (50 μg/ml, Roche) for DNA detection, together with specific antibody recognizing SM-α-actin (A2547 clone 1A4, Sigma 1/400), vinculin (V9131 clone hVIN-1, Sigma 1/1000), N-cadherin (SC-7939,

The paper explained

PROBLEM:

While the tumour microenvironment is known to contribute to tumour progression, the role of carcinoma-associated fibroblasts (CAFs) remains controversial and their origin unclear. This study addresses the hypothesis that chronic oxidative stress can modulate tumour growth and spread by modulating surrounding tumour fibroblasts.

RESULTS:

We took advantage of the chronic oxidative stress resulting from *junD* deletion to examine the role of reactive oxygen species (ROS) in tumour development. In this model, CAFs derive from stress-exposed fibroblasts and promote metastatic dissemination of neoplastic cells. Pro-invasive myofibroblast properties resulted from ROS-mediated accumulation of the pro-angiogenic HIF-1α and the pro-inflammatory chemokine CXCL12 that activated the RhoA-GTPase. Invasive HER2-human breast adenocarcinomas, characterized by high rate of lymph node metastases, exhibit a correlated stromal accumulation of both CXCL12 and myofibroblasts and display an associated oxido-reduction signature, indicating the relevance of our findings in human cancers.

IMPACT:

HER2-amplifying human breast adenocarcinomas, a breast cancer molecular subtype associated with a very poor prognosis and lymph node metastases, express high levels of CXCL12 in the stroma. Our study raises the intriguing possibility that, in addition to current treatments, CXCR4-blocking antibodies may be effective in combating tumour metastasis in lymph node-positive HER2 patients.

Santa Cruz 1/250), tensin (610064, BD Biosciences 1/500) or FAK (F2918, Sigma 1/1000), followed by either fluorescein isothiocyanate (FITC)-coupled or texas red-coupled secondary antibody (Amersham). F-actin was visualized with FITC-phalloidine (P5282, Sigma 1/1000). Slides were examined using a Zeiss Axioplan 2 and images were acquired with identical exposure times and settings using a digital camera (Photometrix Quantix). Fluorescence image analysis was performed using the ImageJ software (Rasband, WS., ImageJ, U.S National Institutes of Health, Bethesda, Maryland, USA, http://rsb.info.nih.gov/ij/, 1997–2008). After background subtraction, the mean fluorescence intensity of SM-α-actin and F-actin was measured considering all cells per field per condition ($n \geq 20$ cells), from at least three independent experiments. In order to count the FA and the adherent junctions and measure their size, we used the 'analysis particle' tool of ImageJ ($n \geq 50$ per condition). The FA size was measured as the length of the FA in the direction perpendicular to the cell boundary. For immunoblotting analysis, whole cell extracts and Western blotting was performed as in Gerald et al (2004) using antibodies described above. Blots were incubated with horseradish peroxidase-conjugated secondary antibody (Amersham) followed by detection with enhanced chemoluminescence and exposed to autoradiography.

Immunohistochemistry on human breast carcinomas

Sections of paraffin-embedded tissue (3 μm) were stained using streptavidin-peroxidase protocol, immunostainer Benchmark, Ventana, Illkirch, France with specific antibodies recognizing CXCR4 (1/50; ab2074, Abcam), CXCL12 (1/100; ab9797, Abcam), Ki67 (1/200; MIB-1, Dakocytomation), SM-α actin (1/400; A2547, Sigma), CD68 (clone KP1 M081401-2, Dako) and JunD (1/100; sc-74, Santa Cruz) (see also Fig S7). TMA from 36 HER2, 44 BLC and 23 Lum-A tumours were composed using three cores of tumour tissue per case and one core of normal tissue (1 mm of diameter each) and hybridized simultaneously. Invasive HER2-amplified carcinomas have been defined according to ERBB2 immunostaining using ASCO's guideline. Among invasive ductal carcinomas, the BLC immunophenotype was defined as follows: ERPR ERBB2 with the expression of at least one of the following markers: KRT5/6$^+$, EGF-R$^+$, kit$^+$. Lum-A tumours were ER$^+$. For quantification, three sections from distinct areas of each tumour were evaluated independently by at least two different investigators. A score, associated with a colour code, was given as a function of the percentage of positive cells and the staining intensity. The colour code is as the following: white = no or weak signal; yellow = moderate; orange = high; red = intense. Experiments were approved by the ethics committee of the Institut Curie and informed consent was obtained from all included patients prior to inclusion in the study.

Migration assay

Migration assays were performed by using Corning polycarbonate Transwell 24-well plates. Cells (7×10^5) were seeded to the upper chamber of each well (6.5 mm in diameter, 8 μm pore size). Medium containing 7% FCS was placed in the lower compartment of the chamber. After 24 h at 37°C, any remaining cells on the upper membrane surface were removed by careful wiping with a cotton swab, and the filters were fixed and stained with 0.2% crystal violet solution in 20% methanol for 15 min. The colouration was removed using acetic acid (10%) and absorbance at 530 nm was used to measure the proportion of migrating cells adhering to the under surface of the filter. The number of total cells was evaluated by using unwrapped chambers.

RhoA- and Rac-pull-down assay

RhoA and Rac activity were quantified by measuring the amounts of RhoA-GTP and Rac-GTP precipitated in a pull-down assay from cell lysates, using the GTPase-binding domain of Rhotekin (RBD) or p21-activated kinase, PAK, (PBD), fused to the glutathione-*S*-transferase (GST-RBD or GST-PBD, respectively). Briefly, 10^7 cells were lysed in lysis buffer (200 mM NaCl, 0.5% NP-40, 1 mM ethylenedia-

minetetraacetic acid (EDTA) pH 8.0, 20 mM Tris–HCl pH 8.0). Whole cell extracts (500 μg) were added to GST-RBD or GST-PBD beads. The pull-down reaction mixtures were incubated for 45 min at 4°C with gentle agitation. The supernatants were removed by brief centrifugation and the precipitated proteins were subjected to immunoblot analysis using monoclonal antibody to RhoA (2117, Cell Signaling) or to Rac1/2/3 (2465, Cell Signaling). The intensities of the bands of GTP-bound RhoA and Rac were quantified by ImageJ and normalized to the total amount of the corresponding protein in whole lysates.

Statistical analysis

All experiments were performed at least three times. Differences were considered to be statistically significant at values of $p \leq 0.05$ by Student's *t*-test and Mann Whitney test. Graphs show mean and standard error of mean using Student's *t*-test. Single, double and triple asterisks indicate statistically significant differences: $^{*}p \leq 0.05$; $^{**}p \leq 0.01$; $^{***}p \leq 0.005$. All survival analyses were carried out using Kaplan–Meier method and log-rank test in R (refer above).

Author contributions

FMG and AT participated in the conception and design of the experiments. AT, DG, BB, MC, MCP and DB performed the experiments on mouse models and derived cells. AT, GD, SL, MR, MHS, TD and OD participated in the studies of human samples. XSG and AVS provided all human samples used in this study. CL and GR contributed to the statistical analyses of the data. FMG directed the work and wrote the paper with suggestions and comments from all authors.

Acknowledgements

We thank M. Yaniv, A. Aurias, D. Williamson and D. Lallemand for critical reading and interesting comments of the manuscript. We are grateful to Arenzana-Seistedos and H. Lortat-Jacob for fruitful discussions about this work. We thank G. Laurent for assistance in graft experiments, T. Gruosso for help in FACS experiments, A. Sadou for GST-pull down assays and F. Assayag for the gift of B16F10 cells. We acknowledge the Institut Curie Breast Cancer Study Group, headed by B. Sigal-Zafrani, A. Fourquet and R. Salmon for providing human breast tumours. We are grateful to all members of the animal facilities of Curie Institute for their helpful expertise. The experimental work was supported by grants from the Institut National de la Santé et de la Recherche Médicale (Inserm), the Institut Curie, the Fondation de France, the Agence Nationale de la Recherche, the French National Institut of Cancer (INCa) and the Association pour la Recherche contre le Cancer.

Supporting information is available at EMBO Molecular Medicine online.

The authors declare that they have no conflict of interest.

For more information

Oncomine data base:
https://www.oncomine.org/

Breast cancer information:
http://www.breastcancer.org

F. Mechta-Grigoriou's laboratory:
http://www.curie.fr/recherche/themes/detail_equipe.cfm/lang/_gb/id_equipe/318.htm

References

Allinen M, Beroukhim R, Cai L, Brennan C, Lahti-Domenici J, Huang H, Porter D, Hu M, Chin L, Richardson A, *et al* (2004) Molecular characterization of the tumor microenvironment in breast cancer. Cancer Cell 6: 17-32

Aslan M, Ozben T (2003) Oxidants in receptor tyrosine kinase signal transduction pathways. Antioxid Redox Signal 5: 781-788

Bacac M, Provero P, Mayran N, Stehle JC, Fusco C, Stamenkovic I (2006) A mouse stromal response to tumor invasion predicts prostate and breast cancer patient survival. PLoS ONE 1: e32

Bartlett JM, Ellis IO, Dowsett M, Mallon EA, Cameron DA, Johnston S, Hall E, A'Hern R, Peckitt C, Bliss JM, *et al* (2007) Human epidermal growth factor receptor 2 status correlates with lymph node involvement in patients with estrogen receptor (ER) negative, but with grade in those with ER-positive early-stage breast cancer suitable for cytotoxic chemotherapy. J Clin Oncol 25: 4423-4430

Bartolome RA, Galvez BG, Longo N, Baleux F, Van Muijen GN, Sanchez-Mateos P, Arroyo AG, Teixido J (2004) Stromal cell-derived factor-1alpha promotes melanoma cell invasion across basement membranes involving stimulation of membrane-type 1 matrix metalloproteinase and Rho GTPase activities. Cancer Res 64: 2534-2543

Baselga J, Swain SM (2009) Novel anticancer targets: revisiting ERBB2 and discovering ERBB3. Nat Rev Cancer 9: 463-475

Bhowmick NA, Neilson EG, Moses HL (2004) Stromal fibroblasts in cancer initiation and progression. Nature 432: 332-337

Bissell MJ, Radisky D (2001) Putting tumours in context. Nat Rev Cancer 1: 46-54

Bleul CC, Fuhlbrigge RC, Casasnovas JM, Aiuti A, Springer TA (1996) A highly efficacious lymphocyte chemoattractant, stromal cell-derived factor 1 (SDF-1). J Exp Med 184: 1101-1109

Casey TM, Eneman J, Crocker A, White J, Tessitore J, Stanley M, Harlow S, Bunn JY, Weaver D, Muss H, *et al* (2008) Cancer associated fibroblasts stimulated by transforming growth factor beta1 (TGF-beta 1) increase invasion rate of tumor cells: a population study. Breast Cancer Res Treat 110: 39-49

Ceradini DJ, Kulkarni AR, Callaghan MJ, Tepper OM, Bastidas N, Kleinman ME, Capla JM, Galiano RD, Levine JP, Gurtner GC (2004) Progenitor cell trafficking is regulated by hypoxic gradients through HIF-1 induction of SDF-1. Nat Med 10: 858-864

Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, Montgomery K, Chi JT, van de Rijn M, Botstein D, Brown PO (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol 2: E7

Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H, *et al* (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci USA 102: 3738-3743

Condeelis J, Pollard JW (2006) Macrophages: obligate partners for tumor cell migration, invasion, and metastasis. Cell 124: 263-266

Coppe JP, Patil CK, Rodier F, Sun Y, Munoz DP, Goldstein J, Nelson PS, Desprez PY, Campisi J (2008) Senescence-associated secretory phenotypes reveal cell-nonautonomous functions of oncogenic RAS and the p53 tumor suppressor. PLoS Biol 6: 2853-2868

Coussens LM, Werb Z (2002) Inflammation and cancer. Nature 420: 860-867

Daly AJ, McIlreavey L, Irwin CR (2008) Regulation of HGF and SDF-1 expression by oral fibroblasts—implications for invasion of oral cancer. Oral Oncol 44: 646-651

de Visser KE, Eichten A, Coussens LM (2006) Paradoxical roles of the immune system during cancer development. Nat Rev Cancer 6: 24-37

De Wever O, Nguyen QD, Van Hoorde L, Bracke M, Bruyneel E, Gespach C, Mareel M (2004) Tenascin-C and SF/HGF produced by myofibroblasts in vitro provide convergent pro-invasive signals to human colon cancer cells through RhoA and Rac. FASEB J 18: 1016-1018

Direkze NC, Hodivala-Dilke K, Jeffery R, Hunt T, Poulsom R, Oukrif D, Alison MR, Wright NA (2004) Bone marrow contribution to tumor-associated myofibroblasts and fibroblasts. Cancer Res 64: 8492-8495

Eng C, Leone G, Orloff MS, Ostrowski MC (2009) Genomic alterations in tumor stroma. Cancer Res 69: 6759-6764

Erez N, Truitt M, Olson P, Hanahan D (2010) Cancer-associated fibroblasts are activated in incipient neoplasia to orchestrate tumor-promoting inflammation in an NF-kappaB-dependent manner. Cancer Cell 17: 135-147

Eyden B, Banerjee SS, Shenjere P, Fisher C (2009) The myofibroblast and its tumours: a review. J Clin Pathol 62: 236-249

Farmer P, Bonnefoi H, Anderle P, Cameron D, Wirapati P, Becette V, Andre S, Piccart M, Campone M, Brain E, *et al* (2009) A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. Nat Med 15: 68-74

Finak G, Bertos N, Pepin F, Sadekova S, Souleimanova M, Zhao H, Chen H, Omeroglu G, Meterissian S, Omeroglu A, *et al* (2008) Stromal gene expression predicts clinical outcome in breast cancer. Nat Med 14: 518-527

Francis P, Namlos HM, Muller C, Eden P, Fernebro J, Berner JM, Bjerkehagen B, Akerman M, Bendahl PO, Isinger A, *et al* (2007) Diagnostic and prognostic gene expression signatures in 177 soft tissue sarcomas: hypoxia-induced transcription profile signifies metastatic potential. BMC Genomics 8: 73

Gaggioli C, Hooper S, Hidalgo-Carcedo C, Grosse R, Marshall JF, Harrington K, Sahai E (2007) Fibroblast-led collective invasion of carcinoma cells with differing roles for RhoGTPases in leading and following cells. Nat Cell Biol 9: 1392-1400

Gelmini S, Mangoni M, Serio M, Romagnani P, Lazzeri E (2008) The critical role of SDF-1/CXCR4 axis in cancer and cancer stem cells metastasis. J Endocrinol Invest 31: 809-819

Gerald D, Berra E, Frapart YM, Chan DA, Giaccia AJ, Mansuy D, Pouyssegur J, Yaniv M, Mechta-Grigoriou F (2004) JunD reduces tumor angiogenesis by protecting cells from oxidative stress. Cell 118: 781-794

Giatromanolaki A, Koukourakis MI, Simopoulos C, Polychronidis A, Gatter KC, Harris AL, Sivridis E (2004) c-erbB-2 related aggressiveness in breast cancer is hypoxia inducible factor-1alpha dependent. Clin Cancer Res 10: 7972-7977

Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20: 93-99

Grivennikov SI, Karin M (2010) Inflammation and oncogenesis: a vicious connection. Curr Opin Genet Dev 20: 65-71

Gruber G, Greiner RH, Hlushchuk R, Aebersold DM, Altermatt HJ, Berclaz G, Djonov V (2004) Hypoxia-inducible factor 1 alpha in high-risk breast cancer: An independent prognostic parameter? Breast Cancer Res 6: R191-R198

Haviv I, Polyak K, Qiu W, Hu M, Campbell I (2009) Origin of carcinoma associated fibroblasts. Cell Cycle 8: 383-395

Hinz B, Phan SH, Thannickal VJ, Galli A, Bochaton-Piallat ML, Gabbiani G (2007) The myofibroblast: one function, multiple origins. Am J Pathol 170: 1807-1816

Hu M, Yao J, Carroll DK, Weremowicz S, Chen H, Carrasco D, Richardson A, Violette S, Nikolskaya T, Nikolsky Y, *et al* (2008) Regulation of in situ to invasive breast carcinoma transition. Cancer Cell 13: 394-406

Hwang RF, Moore T, Arumugam T, Ramachandran V, Amos KD, Rivera A, Ji B, Evans DB, Logsdon CD (2008) Cancer-associated stromal fibroblasts promote pancreatic tumor progression. Cancer Res 68: 918-926

Ishii G, Sangai T, Oda T, Aoyagi Y, Hasebe T, Kanomata N, Endoh Y, Okumura C, Okuhara Y, Magae J, *et al* (2003) Bone-marrow-derived myofibroblasts contribute to the cancer-induced stromal reaction. Biochem Biophys Res Commun 309: 232-240

Joyce JA, Pollard JW (2009) Microenvironmental regulation of metastasis. Nat Rev Cancer 9: 239-252

Kalluri R, Weinberg RA (2009) The basics of epithelial-mesenchymal transition. J Clin Invest 119: 1420-1428

Kalluri R, Zeisberg M (2006) Fibroblasts in cancer. Nat Rev Cancer 6: 392-401

Kang H, Watkins G, Douglas-Jones A, Mansel RE, Jiang WG (2005) The elevated level of CXCR4 is correlated with nodal metastasis of human breast cancer. Breast 14: 360-367

Karin M, Greten FR (2005) NF-kappaB: linking inflammation and immunity to cancer development and progression. Nat Rev Immunol 5: 749-759

Karnoub AE, Dash AB, Vo AP, Sullivan A, Brooks MW, Bell GW, Richardson AL, Polyak K, Tubo R, Weinberg RA (2007) Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. Nature 449: 557-563

Kryczek I, Wei S, Keller E, Liu R, Zou W (2007) Stroma-derived factor (SDF-1/CXCL12) and human tumor pathogenesis. Am J Physiol Cell Physiol 292: C987-C995

LaRue AC, Masuya M, Ebihara Y, Fleming PA, Visconti RP, Minamiguchi H, Ogawa M, Drake CJ (2006) Hematopoietic origins of fibroblasts. I. In vivo studies of fibroblasts associated with solid tumors. Exp Hematol 34: 208-218

Laurent G, Solari F, Mateescu B, Karaca M, Castel J, Bourachot B, Magnan C, Billaud M, Mechta-Grigoriou F (2008) Oxidative stress contributes to aging by enhancing pancreatic angiogenesis and insulin signaling. Cell Metab 7: 113-124

Lewis MP, Lygoe KA, Nystrom ML, Anderson WP, Speight PM, Marshall JF, Thomas GJ (2004) Tumour-derived TGF-beta1 modulates myofibroblast differentiation and promotes HGF/SF-dependent invasion of squamous carcinoma cells. Br J Cancer 90: 822-832

Li YM, Pan Y, Wei Y, Cheng X, Zhou BP, Tan M, Zhou X, Xia W, Hortobagyi GN, Yu D, *et al* (2004) Upregulation of CXCR4 is essential for HER2-mediated tumor metastasis. Cancer Cell 6: 459-469

Littlepage LE, Egeblad M, Werb Z (2005) Coevolution of cancer and stromal cellular responses. Cancer Cell 7: 499-500

Mantovani A, Allavena P, Sica A, Balkwill F (2008) Cancer-related inflammation. Nature 454: 436-444

Mechta F, Lallemand D, Pfarr CM, Yaniv M (1997) Transformation by ras modifies AP1 composition and activity. Oncogene 14: 837-847

Mechta-Grigoriou F, Gerald D, Yaniv M (2001) The mammalian Jun proteins: redundancy and specificity. Oncogene 20: 2378-2389

Mishra PJ, Humeniuk R, Medina DJ, Alexe G, Mesirov JP, Ganesan S, Glod JW, Banerjee D (2008) Carcinoma-associated fibroblast-like differentiation of human mesenchymal stem cells. Cancer Res 68: 4331-4339

Mori L, Bellini A, Stacey MA, Schmidt M, Mattoli S (2005) Fibrocytes contribute to the myofibroblast population in wounded skin and originate from the bone marrow. Exp Cell Res 304: 81-90

Mueller MM, Fusenig NE (2004) Friends or foes—bipolar effects of the tumour stroma in cancer. Nat Rev Cancer 4: 839-849

Mueller L, Goumas FA, Affeldt M, Sandtner S, Gehling UM, Brilloff S, Walter J, Karnatz N, Lamszus K, Rogiers X, *et al* (2007) Stromal fibroblasts in colorectal liver metastases originate from resident fibroblasts and generate an inflammatory microenvironment. Am J Pathol 171: 1608-1618

Muller A, Homey B, Soto H, Ge N, Catron D, Buchanan ME, McClanahan T, Murphy E, Yuan W, Wagner SN, *et al* (2001) Involvement of chemokine receptors in breast cancer metastasis. Nature 410: 50-56

Neilson EG (2006) Mechanisms of disease: fibroblasts—a new look at an old problem. Nat Clin Pract Nephrol 2: 101-108

Nguyen QD, De Wever O, Bruyneel E, Hendrix A, Xie WZ, Lombet A, Leibl M, Mareel M, Gieseler F, Bracke M, *et al* (2005) Commutators of PAR-1 signaling in cancer cell invasion reveal an essential role of the Rho-Rho kinase axis and tumor microenvironment. Oncogene 24: 8240-8251

Olumi AF, Grossfeld GD, Hayward SW, Carroll PR, Tlsty TD, Cunha GR (1999) Carcinoma-associated fibroblasts direct tumor progression of initiated human prostatic epithelium. Cancer Res 59: 5002-5011

Orimo A, Gupta PB, Sgroi DC, Arenzana-Seisdedos F, Delaunay T, Naeem R, Carey VJ, Richardson AL, Weinberg RA (2005) Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion. Cell 121: 335-348

Ostman A, Augsten M (2009) Cancer-associated fibroblasts and tumor growth—bystanders turning into key players. Curr Opin Genet Dev 19: 67-73

Parrinello S, Coppe JP, Krtolica A, Campisi J (2005) Stromal-epithelial interactions in aging and cancer: senescent fibroblasts alter epithelial cell differentiation. J Cell Sci 118: 485-496

Pouyssegur J, Mechta-Grigoriou F (2006) Redox regulation of the hypoxia-inducible factor. Biol Chem 387: 1337-1346

Preston TJ, Muller WJ, Singh G (2001) Scavenging of extracellular H_2O_2 by catalase inhibits the proliferation of HER-2/Neu-transformed rat-1 fibroblasts through the induction of a stress response. J Biol Chem 276: 9558-9564

Provenzano PP, Inman DR, Eliceiri KW, Knittel JG, Yan L, Rueden CT, White JG, Keely PJ (2008) Collagen density promotes mammary tumor initiation and progression. BMC Med 6: 11

Qiu W, Hu M, Sridhar A, Opeskin K, Fox S, Shipitsin M, Trivett M, Thompson ER, Ramakrishna M, Gorringe KL, *et al* (2008) No evidence of clonal somatic genetic alterations in cancer-associated fibroblasts from human breast and ovarian carcinomas. Nat Genet 40: 650-655

Radisky ES, Radisky DC (2007) Stromal induction of breast cancer: inflammation and invasion. Rev Endocr Metab Disord 8: 279-287

Radisky DC, Levy DD, Littlepage LE, Liu H, Nelson CM, Fata JE, Leake D, Godden EL, Albertson DG, Nieto MA, *et al* (2005) Rac1b and reactive oxygen species mediate MMP-3-induced EMT and genomic instability. Nature 436: 123-127

Radisky DC, Kenny PA, Bissell MJ (2007) Fibrosis and cancer: Do myofibroblasts come also from epithelial cells via EMT? J Cell Biochem 101: 830-839

Rodier F, Coppe JP, Patil CK, Hoeijmakers WA, Munoz DP, Raza SR, Freund A, Campeau E, Davalos AR, Campisi J (2009) Persistent DNA damage signalling triggers senescence-associated inflammatory cytokine secretion. Nat Cell Biol 11: 973-979

Ronnov-Jessen L, Petersen OW, Koteliansky VE, Bissell MJ (1995) The origin of the myofibroblasts in breast cancer. Recapitulation of tumor environment in culture unravels diversity and implicates converted fibroblasts and recruited smooth muscle cells. J Clin Invest 95: 859-873

Rossi D, Zlotnik A (2000) The biology of chemokines and their receptors. Annu Rev Immunol 18: 217-242

Sinn E, Muller W, Pattengale P, Tepler I, Wallace R, Leder P (1987) Coexpression of MMTV/v-Ha-ras and MMTV/c-myc genes in transgenic mice: synergistic action of oncogenes in vivo. Cell 49: 465-475

Studeny M, Marini FC, Dembinski JL, Zompetta C, Cabreira-Hansen M, Bekele BN, Champlin RE, Andreeff M (2004) Mesenchymal stem cells: potential precursors for tumor stroma and targeted-delivery vehicles for anticancer agents. J Natl Cancer Inst 96: 1593-1603

Tan W, Martin D, Gutkind JS (2006) The Galpha13-Rho signaling axis is required for SDF-1-induced migration through CXCR4. J Biol Chem 281: 39542-39549

Tlsty TD, Coussens LM (2006) Tumor stroma and regulation of cancer development. Annu Rev Pathol 1: 119-150

Tokunou M, Niki T, Eguchi K, Iba S, Tsuda H, Yamada T, Matsuno Y, Kondo H, Saitoh Y, Imamura H, *et al* (2001) c-MET expression in myofibroblasts: role in autocrine activation and prognostic significance in lung adenocarcinoma. Am J Pathol 158: 1451-1463

Trimboli AJ, Cantemir-Stone CZ, Li F, Wallace JA, Merchant A, Creasap N, Thompson JC, Caserta E, Wang H, Chong JL, *et al* (2009) Pten in stromal fibroblasts suppresses mammary epithelial tumours. Nature 461: 1084-1091

Tsujino T, Seshimo I, Yamamoto H, Ngan CY, Ezumi K, Takemasa I, Ikeda M, Sekimoto M, Matsuura N, Monden M (2007) Stromal myofibroblasts predict disease recurrence for colorectal cancer. Clin Cancer Res 13: 2082-2090

van Kempen LC, de Visser KE, Coussens LM (2006) Inflammation, proteases and cancer. Eur J Cancer 42: 728-734

Vleugel MM, Greijer AE, Shvarts A, van der Groep P, van Berkel M, Aarbodem Y, van Tinteren H, Harris AL, van Diest PJ, van der Wall E (2005) Differential prognostic impact of hypoxia induced and diffuse HIF-1alpha expression in invasive breast cancer. J Clin Pathol 58: 172-177

Wang XF, Witting PK, Salvatore BA, Neuzil J (2005) Vitamin E analogs trigger apoptosis in HER2/erbB2-overexpressing breast cancer cells by signaling via the mitochondrial pathway. Biochem Biophys Res Commun 326: 282-289

West RB, Nuyten DS, Subramanian S, Nielsen TO, Corless CL, Rubin BP, Montgomery K, Zhu S, Patel R, Hernandez-Boussard T, *et al* (2005) Determination of stromal signatures in breast carcinoma. PLoS Biol 3: e187

Yamamoto Y, Ibusuki M, Okumura Y, Kawasoe T, Kai K, Iyama K, Iwase H (2008) Hypoxia-inducible factor 1alpha is closely linked to an aggressive phenotype in breast cancer. Breast Cancer Res Treat 110: 465-475

Yazhou C, Wenlv S, Weidong Z, Licun W (2004) Clinicopathological significance of stromal myofibroblasts in invasive ductal carcinoma of the breast. Tumour Biol 25: 290-295

Zavadil J, Haley J, Kalluri R, Muthuswamy SK, Thompson E (2008) Epithelial-mesenchymal transition. Cancer Res 68: 9574-9577

Zlotnik A (2008) New insights on the role of CXCR4 in cancer metastasis. J Pathol 215: 211-213

11.2.2 An overview of the Wnt pathway in breast cancers

In our global test analysis the Wnt pathway appears as the 14th most significantly associated pathway (see Figure 11.6) with a p-value smaller than 2.10^{-16}. We performed an in-depth analysis of the Wnt pathway: looking at the gene expression patterns for the different Wnt genes and identifying genes that are robustly over-expressed in several datasets (Rigaill et al. (2010a)). This paper is still in preparation and a first draft is proposed in the following pages.

The Wnt signaling pathway is highly conserved in evolution and is essential during the embryogenesis and morphogenesis of many organs (Clevers, 2006) in particular the morphogenesis of the breast (see Boras-Granic and Wysolmerski (2008) for a review). Moreover, it remains critical in regenerating adult tissues, such as colon, skin, hair follicles, lymphoid tissues and bone (Clevers, 2006). This regeneration relies on a tight regulation of the Wnt pathway to maintain a balance between proliferation and differentiation. The Wnt signaling pathway is composed of two distinct signaling arms: the canonical and the non-canonical pathways. The non-canonical pathway modulates cytoskeletal organization, controls cell movement and tissue polarity, and can directly antagonize the canonical signaling pathway. The canonical pathway promotes cell proliferation, fate determination and survival. In the canonical pathway, the secreted Wnt proteins bind the cell-surface receptor Frizzled and the LRP5/6 co-receptors. This interaction can be inhibited by secreted Frizzled-related proteins (SFRP), Dickkopfs (DKK) and Wnt inhibitory factor 1 (WIF1). On the first hand, in the absence of Wnt ligands, cytoplasmic β-catenin is recruited into a destruction complex. In this complex, it interacts with adenomatosis polyposis coli (APC) and axins, and is phosphorylated by casein kinase I alpha (CKI-α) and GSK3-β. Following its phosphorylation, β-catenin is targeted for proteosome-dependent degradation. On the other hand, in the presence of canonical Wnt ligands, LRP5 and 6 are phosphorylated by CKI-γ and GSK3-β. Dishevelled is recruited to the plasma membrane and interacts with Frizzled receptors. LRP5/6 phosphorylation and the formation of Dishevelled complexes mediate the translocation of axin to the plasma membrane and lead to the inactivation of the destruction complex. It induces the stabilization of cytoplasmic β-catenin and its translocation to the nucleus. Once in the nucleus, β-catenin forms a transcriptionally active complex with the LEF and TCF transcription

factors, leading to the expression of a plethora of so-called Wnt target genes (indicative of Wnt activation). Overall, the Wnt network is very complex: there are 19 Wnt ligands, 10 Frizzled receptors, two LRP co-receptors (LRP5 and LRP6) and several TCF/LEF DNA-binding proteins. This large amount of diverse proteins enables many signaling interactions and the transduction of complex signals (see Clevers (2006) for a review).

Mutations in genes or deregulated expression of components leading to Wnt pathway hyperactivity have been shown to be involved in cancer progression, for example in the colon cancer (Polakis, 2007). There is also some evidence that deregulations of the Wnt signaling pathway leads to mammary carcinomas (Howe and Brown, 2004; Zardawi et al., 2009). Analysis from mouse model systems strongly involves Wnt signaling in both mammary development and tumorigenesis (Nusse and Varmus, 1982; Turashvili et al., 2006). In addition, an elevation of cytoplasmic and/or nuclear β-catenin has been observed in human breast carcinomas suggesting that the Wnt pathway was activated (Lin et al., 2000; Ryo et al., 2001; Johannsdottir et al., 2006) and more recently (Khramtsov et al., 2010). Moreover, many Wnt ligands were shown to be often over-expressed in breast tumors (Lejeune et al., 1995; Huguet et al., 1994; Bui et al., 1997; Kirikoshi et al., 2001) and many Wnt pathway actors were found to be often deregulated in breast tumors, such as WIF-1, sFRP1, DVL-1, DKK-3 (Cowling et al., 2007; Ai et al., 2006; Nagahata et al., 2003; Veeck et al., 2008). Some of these were also shown to be repressed by epigenetic silencing (Suzuki et al., 2008; Veeck et al., 2008; Klarmann et al., 2008). However, the way the Wnt signaling pathway is deregulated in breast cancer remains to be elucidated.

Several recent publications link TNBC and Wnt pathway deregulation (Smid et al., 2008; Matsuda et al., 2009; DiMeo et al., 2009; Liu et al., 2010; Khramtsov et al., 2010). Using TNBC cell-line models, both Matsuda et al. (2009) and DiMeo et al. (2009) have thoroughly shown that the Wnt pathway was activated and involved in tumor migratory ability and metastasis formation. In humans, at the transcriptomic level, it seems that Wnt pathway genes or Wnt-related genes are more specifically deregulated in TNBC (Smid et al., 2008; DiMeo et al., 2009). Moreover, recent publications have more directly shown that the Wnt pathway is activated in TNBC (Liu et al., 2010; Khramtsov et al., 2010). Liu et al. (2010) showed that LRP6 was over-expressed in TNBC, and that this over-expression could

be responsible for the activation of the Wnt pathway.

In light of all these results, there is little doubt that the Wnt pathway is involved in breast cancer development and more specifically in TNBC. However, as highlighted by Collu et al. (2009), contrary to colorectal cancer, the mechanisms of activation and the role of the Wnt pathway in breast cancer remain to be elucidated. In colorectal cancer, the Wnt signaling pathway is known to be activated by mutations in APC, axin or β-catenin and tumor progression is driven by β-catenin stabilization.

In this study, our initial goal was to validate the deregulation of the Wnt pathway in TNBC compared to other subtypes. We aimed to better understand the Wnt pathway gene pattern in TNBC and identify the key gene deregulations characteristic of this pattern. The reproducibility of such differential analysis is very often limited (Subramanian et al., 2005), so I chose to use three different datasets simultaneously (further described below). One of our hypotheses was that if the Wnt pathway is indeed involved in the TNBC phenotype then these tumors should exhibit a reproducible Wnt pathway gene expression pattern.

11.2.3 Transcriptomic statistical analysis of the Wnt pathway

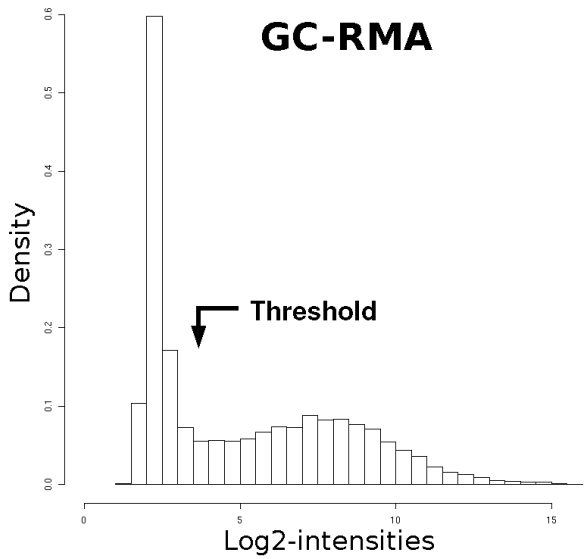
For this analysis, three different datasets were used (Own, Chin and Adelaïde described below). All datasets used the Affymetrix HGU133plus2 technology to enable an easy comparison of the results. The expression profiling data of Chin et al. (2006) were collected from ArrayExpress. This dataset comprises Luminal A and B, HER2+, TNBC and normal-like tumors. The data from Adélaïde et al. (2007) were retrieved from their website. This dataset comprises Luminal A and TNBC. Data from the Adelaïde dataset were already normalized with RMA (Irizarry et al., 2003a) and they used the Affymetrix annotation (see section 9.1). Thus, specifically for this study, I used the Affymetrix annotation and the RMA normalization for both the Chin and our own datasets to allow an easy comparison of the three datasets. As previously described in section 9.2, a threshold was implemented to discard probes with intensities in the background level. In the case of the RMA normalization, this threshold is much more arbitrary than with the GC-RMA analysis. Indeed, the histogram of log2 measurements is not bi-modal (as described in section 9.2) and there is no sharp peak corresponding to the background

noise (see Figure 11.7). For all datasets, probesets were discarded if at least 95% of all samples had a probeset log-intensity smaller than 5.

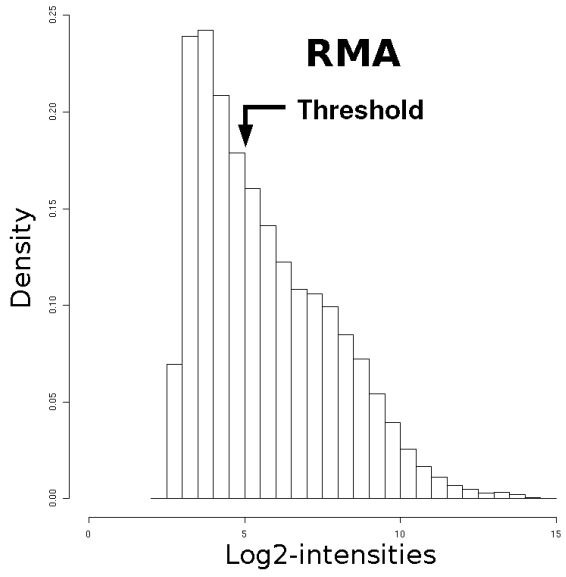
As the link between the Wnt gene mRNA expression and the tumor subtypes is still unclear, I first tried to confirm the deregulation of the Wnt pathway shown by Smid et al. (2008) and DiMeo et al. (2009) on other datasets (see Figure 11.8). Using the global test method (Goeman et al., 2004) on our own data set, I tested for an association between tumor subtypes and KEGG pathways (Kanehisa and Goto, 2000). To be more specific, it can be determined using this global test whether the global expression pattern of a group of genes or genesets is significantly related to tumor subtype. Moreover, the test allows genesets of different sizes to be compared, and gives one p-value per geneset. The Wnt pathway appears as the 14th most significantly associated pathway (see Figure 11.6) with a p-value smaller than 2.10^{-16}.

This association was further validated using the Chin dataset and the Adelaïde datasets. For both datasets, the Wnt pathway was significantly associated with tumor subtype with a p-value smaller than 2.10^{-16}.

To further confirm that the Wnt pathway mRNA levels segregate the different tumor subtypes at least to a certain extent, we sought to better characterize the Wnt molecular pattern of breast tumors. Because the number of genes of the Wnt pathway (as defined by KEGG) is important, I used three exploratory methods to describe the Wnt pathway expression profile: Hierarchical clustering, Principal Component Analysis (previously described in section 10.1) and model-based clustering (Yeung et al., 2001). Model-based clustering is an unsupervised clustering method that aims at determining both the number of clusters and the structure of these clusters. To be more specific, I used a Gaussian mixture model to identify groups of tumors based on the gene expression pattern of the Wnt pathway. This second classification methodology was used to confirm the partitioning of the data obtained with hierarchical clustering. Figure 11.9 shows that both hierarchical clustering and model-based clustering segregate all tumor subtypes into two very distinct groups, one of which clearly corresponds to TNBC. The robustness of this classification is assessed by the tight agreement between hierarchical clustering and model-based clustering (see Figure 11.9) and the clear difference between these two



A



B

Figure 11.7: Histograms of log2 intensities after (A) GC-RMA normalization and (B) RMA-normalization for the Curie-Servier transcriptomic dataset. The histograms are plotted so that they have a total area of one. The height of a rectangle is proportional to the number of points falling into the cell. The GC-RMA histogram has a sharp peak around 2.5 log2 intensity. The RMA histogram does not. The threshold used to discard probes in the background level is shown: 3.5 for GC-RMA and 5 for RMA.

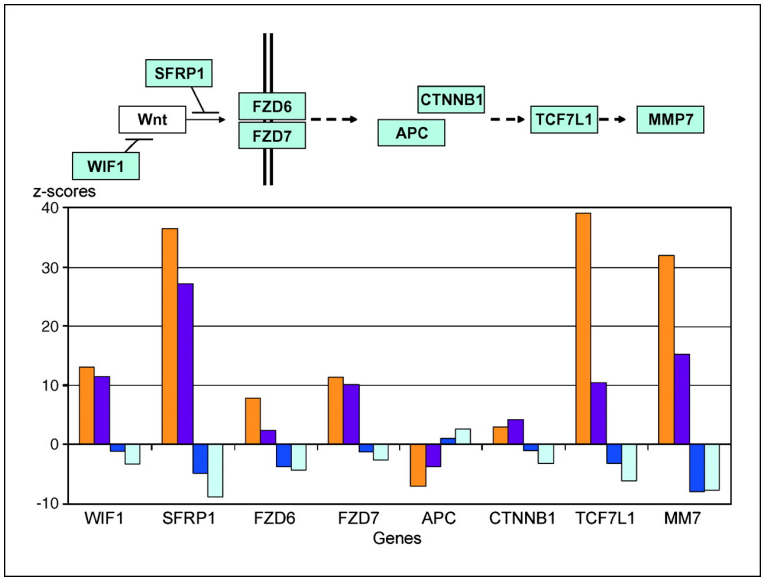


Figure 11.8: Selected genes of WNT signaling. Important molecules of the WNT signaling pathway that are involved in the TNBC and Luminal B subtypes and in tumors from bone and brain relapse patients. Capped lines, inhibitory effect on the protein-protein interaction. The bottom graph depicts z-score values. The z-scores is a measure of the unlikelihood of the null hypothesis that the gene is not associated with the clinical attribute. Positive and negative z-values indicate genes expressed significantly higher or lower in the corresponding group, respectively. Orange bar, TNBC subtype; maroon, brain relapse patient; blue, Luminal B subtype; light blue, bone relapse patient. (Figure and legend from Smid et al. (2008).)

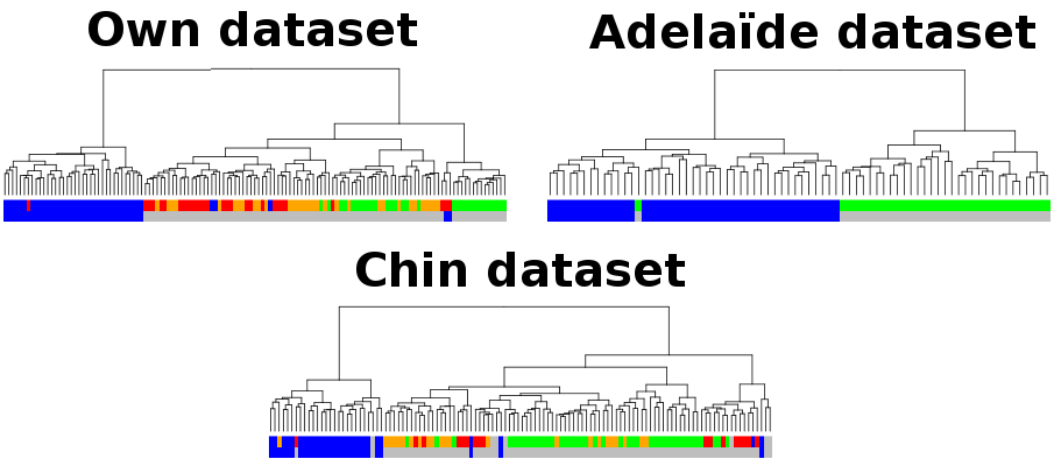


Figure 11.9: Classification of tumors based on the gene expression levels of the KEGG Wnt pathway for our own, the Adelaïde and the Chin datasets. On top is the hierarchical clustering, below is the corresponding immunohistological classification of the tumors in subtypes (blue: TNBC, red: ER- / HER2+, green: Luminal A, orange: Luminal B, green: normal-like). The last lane is the model-based classification in two groups (blue and grey groups).

groups on the first two principal components in the PCA (see Figure 11.10). In conclusion, the Wnt mRNA levels clearly segregate TNBC from all other subtypes. Once again, many genesets are able to segregate between the different subtypes. Thus, this unsupervised analysis is not by itself a proof of the importance of the Wnt pathway in TNBC.

Having shown that TNBC have a very distinct Wnt pathway gene mRNA pattern, we wanted to identify which of these genes best segregated TNBC from other tumors and are therefore characteristic of the TNBC Wnt pathway pattern. Indeed, if the Wnt pathway is involved in the TNBC phenotype, the expression pattern in TNBC of the Wnt pathway genes should be reproducible. Thus I looked for genes of the Wnt pathway (as defined by KEGG) over-expressed in TNBC. To be more precise, I compared TNBC to all other tumors using a two sample t-test. Probesets with a mean difference between TNBC less than log2(1.5) corresponding to a 150% over-expression were discarded. Multiple-testing issues were accounted for using the Bonferonni correction. In the end, we obtained a list of 28

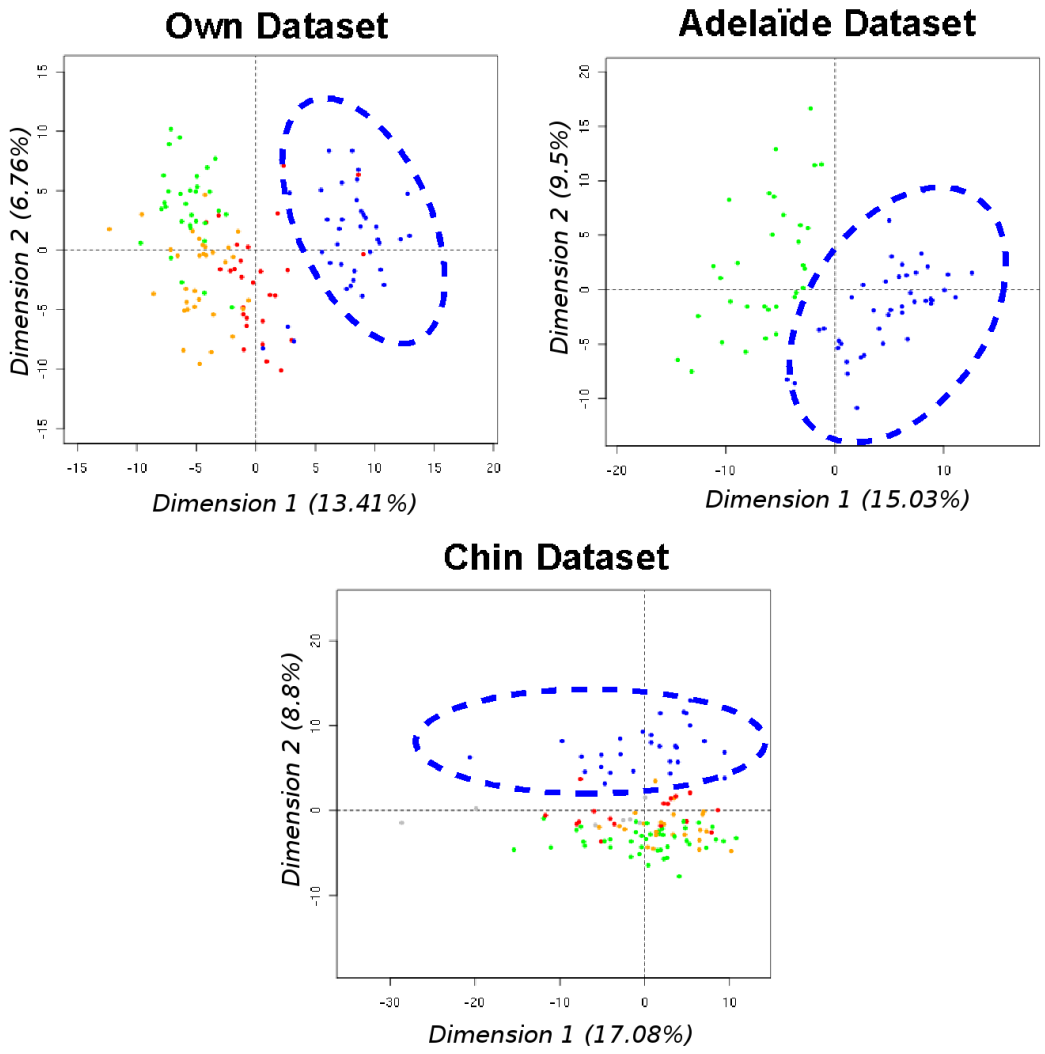


Figure 11.10: PCA individual plot based on the gene expression levels of the KEGG Wnt pathway for our own, the Adelaïde and the Chin datasets. Representation of the tumors using the first two principal components (blue: TNBC, red: ER- / HER2+, green: Luminal A, orange: Luminal B, green: normal-like). Individuals in the blue group (model-based clustering) are circled in blue.

probesets corresponding to 23 genes significantly over-expressed in TNBC compared to other subtypes.

The reproducibility of such a differential analysis is very often limited (Subramanian et al., 2005). Thus, to validate this list, I replicated the same analysis on the Adelaïde and the Chin datasets. To ensure a valid comparison, we kept only the 211 probesets that were present in the three datasets. First, TNBC samples were compared to Luminal A samples, as Luminal A was the only subtype present in all datasets. Using the previously described t-test and Bonferonni correction, I obtained a restricted list of respectively 30, 20 and 28 probesets for the Adelaïde, Chin and our own datasets (see supplementary table) corresponding to respectively 21, 14 and 23 genes. The intersection of those three lists of genes is a list of 9 Wnt pathway genes over-expressed in TNBC compared to Luminal A: MMP7, SFRP1, MYC , FZD7, EN1, TCF7L1, PRKX, PRKCA, PLCB4. To check that those genes were not specific of the TNBC/Luminal A comparison, I then compared TNBC samples to all subtypes available in the datasets. Restricted lists of respectively 30, 15 and 21 probesets were obtained for the Adelaïde, Chin and our own datasets respectively. The intersection was the same except for PLCB4. We decided to call this list of 8 genes (MMP7, SFRP1, MYC , FZD7, EN1, TCF7L1, PRKX, PRKCA) "WOTNBC" for Wnt pathway genes Over-expressed in TNBC.

I then assessed how significant it was to recover an intersection of more than 8 genes to validate the reproducibility of this WOTNBC gene expression pattern. To be precise, we quantified how easy it is to obtain an intersection of 8 genes by picking at random 30, 15 and 21 probesets out of the 211 of the Wnt pathway present on the microarrays. If there was a one-to-one relation between genes and probesets, a hypergeometric test with three classes could have been used. But here some genes correspond to several probesets (Affymetrix annotation). I used a re-sampling strategy and computationally picked at random 30, 15 and 21 probesets out of the 211, and then intersected those three probeset lists and counted how many genes were retrieved. This process was repeated 10 million times. I recovered more than 8 genes only 7 times out of 10^7. This result comforts the fact that it is very unlikely to obtain such a large intersection at random and thus that these 8 core WOTNBC genes are most probably linked to TNBC. In conclusion, we identified a reproducible and robust set of over-expressed Wnt pathway genes in TNBC. Interestingly, a good proportion of these genes are

known to be up-regulated by Wnt activation, in particular MMP-7, SFRP1, FZD7 and MYC (see the Wnt home page: http://www.stanford.edu/ rnusse/wntwindow.html). These over-expressed Wnt target genes may reflect the activation of the Wnt pathway in TNBC.

To validate the importance of this over-expression between TNBC and other tumor types, we compared the expression of the WOTNBC genes in tumors with their expression in normal samples. Figure 11.11 shows that TCF7L1, FZD7 and PRKCA have levels similar to normal (they are not significantly different, p-values of 0.22, 0.51 and 0.72 respectively). For all other WOTNBC genes, the difference between normal samples and TNBC is significant. Interestingly, sFRP1 is down-regulated in TNBC compared to normal (see Figure 11.11) and it is well documented that sFRP1 down-regulation can lead to Wnt activation. All these results, over-expression in TNBC and presence of known Wnt target genes, strongly suggest that these WOTNBC genes are co-regulated and lead us to believe that the Wnt pathway is activated in TNBC.

To conclude, using gene expression profiling of breast tumors and bioinformatic tools, we demonstrated that the Wnt pathway mRNA levels clearly segregate TNBC from all other subtypes. Moreover, using our own and several publicly available datasets, we identified a consistent and reproducible set of Wnt genes over-expressed in TNBC. The reproducibility of this set of genes is remarkable given that in general reproducibility is very often limited (Subramanian et al., 2005). It indicates that the Wnt pathway might indeed be involved in the TNBC phenotype as it exhibits a reproducible gene expression pattern. Moreover, many of the identified genes are known to be Wnt pathway target genes, suggesting an activation of the Wnt pathway.

Following these statistical analyses, the activation of the Wnt pathway in TNBC was further investigated by our group, in collaboration with Anne Vincent-Salomon (MD/Ph.D.) and Marion Richardson (M.Sc.). They looked for the localization of the β-catenin in TNBC using immunohistochemistry (IHC). They detected nuclear β-catenin (indicative of Wnt activation) in only 2 TNBC samples of the Curie-Servier cohort in contrast to a recent report (Khramtsov et al., 2010). In addition, we look for potential causes leading to Wnt pathway activation in TNBC. One possibility is the activation of the PI3K/AKT pathway in TNBC (Stemke-Hale et al., 2008; Marty et al., 2008). Indeed, AKT can phos-

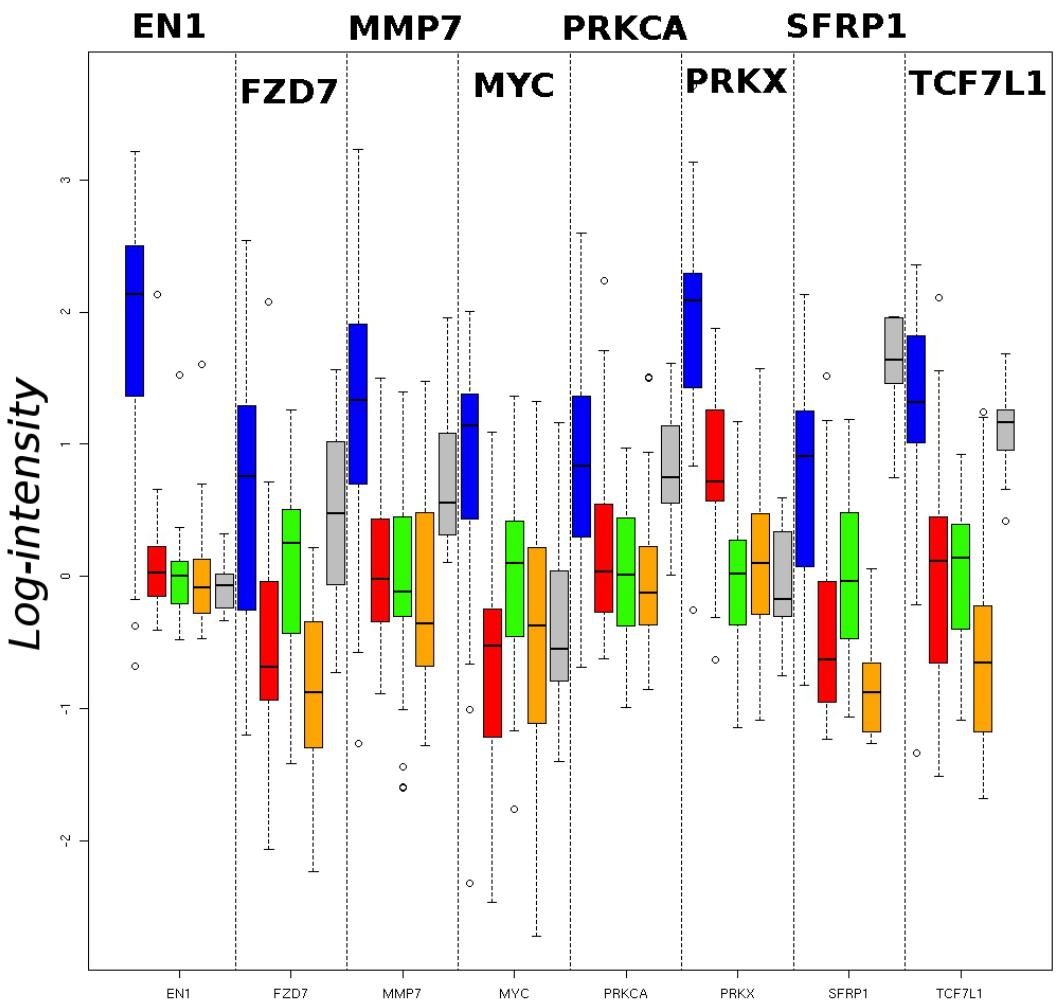


Figure 11.11: Gene expression of the different TNBC and normal samples for the Wnt pathway genes over-expressed in TNBC (WOTNBC) in our own dataset. Log-intensities on this graph are centered on the mean log-intensity of Luminal A samples. TNBC are in blue, ER- / HER2+ in red, Luminal A in green, Luminal B in orange and normal samples in grey.

phorylate and inhibit GSK3. Another possibility is the frequent loss in TNBC of *APC* on chromosome 5q (see Chapter 6). A third possibility is the over-expression of the co-receptor LRP6. Indeed, we found using RPPA that LRP6 was over-expressed in TNBC compared to other subtypes (data not shown), in agreement with the recent study of Liu et al. (2010) showing by immunohistochemistry a higher level of LRP6 in TNBC compared to other subtypes. It has been shown in cell-lines that the over-expression of LRP6 is sufficient to induce the activation of the Wnt pathway (Li et al., 2004; Liu et al., 2010). Of course, this does not exclude other causes leading to Wnt pathway activation. The experimental analyses of this pathway are still under investigation in our laboratory.

Part IV

Conclusion

My PhD project was part of a collaboration between the Institut Curie and the Servier pharmaceutical group. The goal of this collaboration is to discover deregulated genes and signaling pathways in human TNBC to identify new therapeutic targets. In this manuscript, I present my work on the biostatistical analysis of the Curie-Servier transcriptomic and genomic datasets. I developed statistical tools and algorithms for the analysis of DNA copy number profiles. I developed ITALICS to normalize Affymetrix SNP 50K and 250K arrays (Rigaill et al., 2008). ITALICS identified a 50% loss of PTEN in TNBC (Marty et al., 2008) and it helped to define true recurrence among ipsilateral breast cancers (Bollet et al., 2008). I also worked on the segmentation of DNA copy number profiles. Most segmentation methods return a single segmentation, characterized by a set of breakpoints. The quality of this segmentation is rarely questioned. To answer this problem, I proposed algorithms to explore the segmentation space and derived from this exploration new statistical criteria to assess the stability of the segmentation and select the number of breakpoints (Rigaill et al., 2010c,d). Moreover, most segmentation methods rely on heuristics and computation time is an issue for the analysis of large SNP profiles with more than 10^5 probes. I proposed an algorithm to recover the best segmentation of very large DNA copy number profiles with respect to the quadratic loss (Rigaill, 2010b). This algorithm is able to process Affymetrix SNP 6.0 profiles (containing a million points) in a matter of minutes compared to several days for other optimal computational schemes. To my knowledge this algorithm is by far the fastest optimal computational scheme available to recover the best segmentation (with respect to the MSE) of large profiles. I applied this algorithm to the Curie-Servier genomic dataset.

The analysis of the Curie-Servier transcriptomic dataset was carried out using already available biostatistical and bioinformatical tools. The analyses I made were a multi-step process, which are often viewed as a simple pipeline. The validity and importance of each step is rarely questioned. Yet, the choice of a given methodology is often a subtle decision and there rarely is a definitive answer. In my opinion the work of the biostatistician is to justify these choices with at least three things in mind, first and foremost the biological question, second the statistical inference and third the computational strategy. In my analysis of the transcriptomic data, I justified (as much as possible) my choices and overall this analysis answered precise and well-defined biological questions. I recovered lists of drugable

genes that are significantly overexpressed in TNBC. In collaboration with groups of the Institut Curie I identified interesting genes (Lizárraga et al., 2009) and pathway deregulations (Toullec et al., 2010).

Importantly, my work is part of a much larger project between Institut Curie and Institut Servier. The goal of the project is to discover deregulated genes and signaling pathways in human TNBC to identify new therapeutic targets. To do so breast, tumors of Luminal A, Luminal B, HER2 and TNBC were selected and characterized by a pathologist from Institut Curie. Transcriptomic, genomic, miRNA and proteomic microarray profiles were generated at the Institut Curie translational department. This huge amount of data has been analyzed in the Institut Curie bioinformatics team and the biostatistics team of Agroparistech. I have been responsible for the analysis of the genomic and transcriptomic data. This analysis led to some interesting potential targets. Some of these potential targets have been further characterized and functionally validated in cell lines by a team of biologists at the Institut Curie translational department. The analysis of the RPPA and omic data will continue. Indeed the analysis of the miRNA and proteomic data will provide valuable information about the biology of these tumors and hopefully lead to new therapeutic targets.

In the end, the Curie-Servier dataset will include well-characterized tumors at many different levels (clinical, histology, DNA, mRNA, miRNA, proteomic). It will be necessary to integrate or at least to compare these different levels of information to distinguish between what is common to these different levels and what is specific to one or several of these. On top of that, it might be interesting to integrate the already-acquired biological knowledge, such as how proteins interact (interaction network) or how proteins regulate the transcription of other genes (regulation network). Using this prior information, one would hope to simplify the problem at hand. As part of the Curie-Servier collaboration the integration of these different levels of information will be done and, in fact, a system biology integration of these data is ongoing.

Biostatistics and bioinformatics certainly aim to integrate all these different layers of information. But, this integration raises a number of questions. What is the purpose of this integration? Is it possible to integrate this information with the data we have? We are looking at thousands of genes, millions of sequences along the genome, hundreds of proteins and hundreds of miRNA. But we look

at all these on a very limited number of biological or clinical samples. Moreover, a cell is much more than a simple collection of gene transcription levels, DNA alterations and protein expression levels. In the case of cancer study, we are looking at several cells that interact with one another and these cells are part of a whole organism made of billions of cells. This is of course a very pessimistic view of the problem. But it also shows that a lot of work and numerous investigations need to be done. If we truly want to know whether it is possible to model "biology", we have to try.

To conclude, I feel that the true challenge of biostatistics and bioinformatics is to understand what are the biological or clinical questions of interest and out of these questions to distinguish between those that can be answered using the following: the biotechnology at our disposal, an efficient experimental design, and our statistical and computational expertise.

Appendix A

A few more papers

A.1 DNA Breakpoints to Define True Recurrences Among Ipsilateral Breast Cancers

High-Resolution Mapping of DNA Breakpoints to Define True Recurrences Among Ipsilateral Breast Cancers

Marc A. Bollet, Nicolas Servant, Pierre Neuvial, Charles Decraene, Ingrid Lebigot, Jean-Philippe Meyniel, Yann De Rycke, Alexia Savignoni, Guillem Rigaill, Philippe Hupé, Alain Fourquet, Brigitte Sigal-Zafrani, Emmanuel Barillot, Jean-Paul Thiery

Background To distinguish new primary breast cancers from true recurrences, pangenomic analyses of DNA copy number alterations (CNAs) using single-nucleotide polymorphism arrays have proven useful.

Methods The pangenomic profiles of 22 pairs of primary breast carcinoma (ductal or lobular) and ipsilateral breast cancers from the same patients were analyzed. Hierarchical clustering was performed using CNAs and DNA breakpoint information. A partial identity score developed using DNA breakpoint information was used to quantify partial identities between two tumors. The nature of ipsilateral breast cancers (true recurrence vs new primary tumor) as defined using the clustering methods and the partial identity score was compared with that based on clinical characteristics. Metastasis-free survival was compared among patients with primary tumors and true recurrences as defined using the partial identity score and by clinical characteristics. All statistical tests were two-sided.

Results All methods agreed on the nature of ipsilateral breast cancers for 14 pairs of samples. For five pairs, the clinical definition disagreed with both clustering methods. For three pairs, the two clustering methods were discordant and the one using DNA breakpoints agreed with the clinical definition. The partial identity score confirmed the nature of ipsilateral breast cancers as defined by clustering of DNA breakpoints in 21 of 22 pairs. The difference in metastasis-free survival of patients with new primary tumors and those with true recurrences was not statistically significant when tumors were defined based on clinical and histologic characteristics (5-year metastasis-free survival: 76%, 95% confidence interval [CI] = 52% to 100% for new primary tumors and 38%, 95% CI = 17% to 83% for true recurrences; $P = .18$; new primary tumor vs true recurrence, hazard ratio = 2.8, 95% CI = 0.6 to 13.7), but the difference was statistically significant when tumors were defined using the partial identity score (5-year metastasis-free survival: 100% for new primary tumors and 29%, 95% CI = 11% to 78% for true recurrences; $P = .01$).

Conclusions DNA breakpoint information more often agreed with the clinical determination than CNAs in this population. The partial identity score, which was calculated based on DNA breakpoints, allows statistical discrimination between new primary tumors and true recurrences that could outperform the clinical determination in terms of prognosis.

J Natl Cancer Inst 2008;100:48–58

Breast-conserving therapy is the preferred treatment for patients with early-stage breast cancer (1). It offers equal local control and overall survival (2) and superior psychosocial outcomes (3,4) compared with modified radical mastectomy. However, an ipsilateral breast cancer recurrence can be traumatizing and can lead to death (2).

When an ipsilateral breast cancer develops, the new tumor can either be a true recurrence—that is, a regrowth of clonogenic cells that were not removed by surgery or killed by radiotherapy—or a new primary tumor that arises from the remaining breast tissue (5). Several definitions have been used to distinguish true recurrences from new primary tumors. Initially, these distinctions were based

Affiliations of authors: Département d'oncologie radiothérapie (MAB, AF), Service de Bio-informatique (NS, PN, GR, PH, EB), Département de Transfert (CD, JPM, JPT), Département de Biologie des tumeurs (IL, BSZ), Service de Biostatistiques (YDR, AS), and Centre National de la Recherche Scientifique, Unité Mixtes de Recherche 144 (CD, PH), Institut Curie, Paris, France; Institute of Molecular and Cell Biology Biopolis A*STAR, Singapore (JPT).

Correspondence to: Marc A. Bollet, MD, Département d'oncologie radiothérapie, Institut Curie, 26, rue d'Ulm, 75248 Paris cedex 05, France (e-mail: marc.bollet@curie.net).

See "Funding" and "Notes" following "References."

DOI: 10.1093/jnci/djm266

on the location of the ipsilateral breast cancer (ie, the farther from the initial primary tumor, the more likely it is to be a new primary tumor) and on shared common histopathologic criteria (eg, type, grade, and hormone receptor status) (6–10). In the quest for additional ways to distinguish new primary breast tumors from true breast cancer recurrences, biologic studies of clonal relationships between the new and original tumor have also been performed. These studies have relied on ploidy (5,11), loss of heterozygosity (12–14), p53 analysis (15), or X chromosome inactivation (16) or have been based on DNA copy number alterations (CNAs) (17–19). CNA data can be obtained by high-resolution techniques, such as array-based comparative genomic hybridization or single-nucleotide polymorphism (SNP) arrays (20). One of the most commonly used ways to look at clonal relatedness using pangenomic data is to perform an unsupervised hierarchical clustering that organizes primary breast tumors and ipsilateral breast cancers on the basis of their overall genomic similarity (18,19). These measures of similarity are summarized in a dendrogram, in which the pattern and length of the branches reflect the relatedness of the samples in terms of DNA CNAs.

Changes in DNA copy numbers occur at chromosomal locations called breakpoints. We hypothesized that the precise locations of these breakpoints could serve as markers for clonal relatedness and that we could distinguish true recurrences from new primary tumors by the number of common breakpoints in the ipsilateral breast cancer and the primary tumor. In this study, we first aimed to test the added value of examining the clustering of breakpoints (over CNAs) when determining the nature of the ipsilateral breast cancer. Second, we aimed to develop a score to quantify the partial identity between two tumors according to their clonal relatedness (determination of the partial identity score). Third, we examined prognosis in terms of metastasis-free survival. In each case, these methods were compared with the clinical determination of the nature of the ipsilateral breast cancer.

CONTEXT AND CAVEATS

Prior knowledge

Detecting changes in DNA copy number using single nucleotide polymorphism arrays has been a useful tool in distinguishing new primary breast tumors from recurrences.

Study design

Comparison of hierarchical clustering of DNA copy number and DNA breakpoints, an identity score based on the DNA breakpoint information, and clinical characteristics to accurately designate ipsilateral breast tumors as new primary tumors or true recurrences in breast tumor pairs from 22 patients.

Contributions

For 14 of the pairs, all methods agreed on the designation of the ipsilateral breast cancer as a new primary tumor or a true recurrence; however, for five pairs and three pairs, both clustering methods and clustering by DNA breakpoints, respectively, agreed with the clinical definition. For 21 pairs, the partial identity score confirmed the designation of the tumor as defined by both clustering methods. Patients with recurrences had poorer metastasis-free survival than patients with new primary tumors, according to the partial identity score, but this difference was not statistically significant using the clinical definition.

Implications

The partial identity score may outperform clinical determination for the prognosis of ipsilateral breast cancers.

Limitations

Freshly frozen tissue samples that contain a large number of cells from both the initial primary tumor and the ipsilateral tumor are needed to perform the DNA breakpoint analyses.

Subjects and Methods

Selection of Patients

Specimens from patients with primary breast cancers and ipsilateral breast cancers were selected from freshly frozen samples of the Institut Curie tissue bank according to the following criteria: the primary tumor was either ductal or lobular invasive breast carcinoma; the patient was 49 years or younger at diagnosis of the initial tumor; all patients were premenopausal; and there was no previous history of cancer, except for one nonmelanoma skin cancer. All patients had been treated at the Institut Curie by breast-conserving surgery, including dissection of the axillary lymph nodes in most patients, followed by radiotherapy to the breast with or without a boost to the tumor bed (external beam radiotherapy or brachytherapy) and/or to the regional lymph node–bearing areas if indicated and, when required, systemic treatment as part of their initial management. For all tumors, histopathologic characteristics were reviewed by one pathologist (B. Sigal-Zafrani).

To ensure that the data would be informative, we restricted genomic analyses to tumors (primary and recurrences) in which at least 50% of cancer cells had been assessed by hematoxylin, eosin, and saffron staining of sections from snap-frozen samples. This study reports a series of 22 patients with assessable pairs of primary breast tumors and ipsilateral breast cancers.

To evaluate the genomic features of a population with similar breast cancers, 44 control patients from the pool of patients with primary tumors who met the above selection criteria were matched to the case patients in accordance with their age at diagnosis and adjuvant treatment. The control patients had not experienced an ipsilateral breast recurrence within the time span of the local recurrence of the index patient.

This research was approved by the institutional review boards of the Institut Curie. No patient refused the use of her tumor specimens for research purposes.

Clinical and Histologic Studies

The histologic/biologic properties of the breast cancers were determined by subjecting tissue sections to immunohistochemical analysis for the estrogen receptor (clone 6F11, 1:200 dilution; Novocastra, Newcastle Upon Tyne, England) and progesterone receptor (clone 1A6, 1:200 dilution; Novocastra) antibodies. Tumors were considered to be positive for these receptors if at least 10% of the invasive tumor cells in a section showed nuclear staining (21).

In accordance with theories of the clonal evolution of tumor cell populations, ipsilateral breast cancers were clinically defined as true recurrences if they had the same histologic subtype (ductal or

lobular) and a similar or increased growth rate, similar or loss of dependence on either estradiol or progesterone, and similar or decreased differentiation as the initial tumor (22).True recurrences also had to share with their primary tumors the same breast quadrant. Thus, new primary tumors were clinically defined as such when the ipsilateral breast cancer had occurred in a different location, had a distinct histologic type, or had less aggressiveness features (lower grade, appearance of hormonal receptors) than the initial tumor.

Genomic Studies

Total genomic DNA was extracted from tissue samples using a variation of the standard phenol : chloroform protocol (23). Genomic DNA was quantified by spectrophotometry using a ND-1000 Spectrophotometer (NanoDrop, Wilmington, DE), and quality was assessed by 0.8% agarose gel electrophoresis.

Genomic DNA from each sample was prepared for microarray hybridization using the GeneChips Mapping 50K Xba Assay Kit (Affymetrix Inc., Santa Clara, CA). Briefly, 250 ng of total genomic DNA was digested with the restriction enzyme XbaI and ligated to an adaptor sequence (XbaI adaptator: 5′-ATTATGAGCACGAC AGACGCCTGATCT-3′ and 5′-CTAGAGATCAGGCGTCTG TCGTGCTCATAA-3′) that recognizes the cohesive four base pair (bp) region (3′-GATC-5′). A generic primer (5′-ATT ATG AGC ACG ACA GAC GCC TGA TCT-3′) that recognizes the adaptor sequence was used to preferentially amplify adaptor-ligated DNA fragments 250–2000 bp in size by the optimized polymerase chain reaction (PCR) conditions, according to the manufacturer's instructions. The amplified DNA was then fragmented by DNase treatment and hybridized to the Affymetrix GeneChips Human Mapping 50K array Xba 240 (Affymetrix), according to the manufacturer's instructions. Washing, staining, and scanning of chips were performed using materials and methods provided by the manufacturer. The pangenomic profiles of the 22 pairs of primary tumors/ipsilateral breast cancers are available on ACTuDB (24) (http://bioinfo.curie.fr/actudb/). Human mapping 50K array Xba 240 annotations and sequence files are available on the Affymetrix website (http://www.affymetrix.com/support/technical/byproduct.affx?product=100k).

Metastasis-Free Survival

Metastasis-free survival was estimated by the Kaplan–Meier method (25) and compared between the groups of patients defined as having been diagnosed with either a true recurrence or a new primary tumor using the log-rank test. The confidence interval (CI) of the hazard ratio was obtained using a semiparametric Cox model (26).

Statistical Methods

Copy Number Alteration Determination. SNP data were gathered from the pangenomic profile and analyzed using the iterative and alternative normalization of copy number SNP array (ITALICS) algorithm with default parameters, which simultaneously normalizes the genomic profile and detects the biologic signal. Briefly, ITALICS alternatively estimates the biologic signal (ie, the DNA copy number at each SNP locus) with the gain and loss analysis of DNA algorithm (27) and normalizes the data to correct the nonrelevant effects (CG content and fragment length of PCR products, oligonucleotide CG content, and SNP effect). These two steps are repeated iteratively to improve the biologic signal estimation until no more improvement is seen. ITALICS outperforms other methods of normalization. The result of this process is a segmented genomic profile that consists of regions of homogeneous DNA and information on their corresponding copy numbers. Each region is given a smoothing value (ie, the median of the SNP copy numbers within the region) and a status (ie, gain, normal, or loss).

We defined a breakpoint as 1) a SNP locus located at a change of status (eg, normal/gain or gain/loss) or as 2) a SNP locus located at a change of smoothing value that occurred within a region of gain or loss, thus defining different levels of gain or loss among these regions. Additional breakpoints were also added at the extremities of the chromosome to take into account their gain or loss whenever applicable. Because some breakpoints could be due to copy number variations that occur in healthy individuals, breakpoints arising in the copy number variable regions in the HapMap collection (28) were excluded. The visualization and further analysis of the data was performed through a graphic user interface, Visualization and analysis of array CGH, transcriptome and other molecular profiles (29).

Hierarchical Clustering. *Similarity between genomic profiles.* We considered two measures of similarity among the genomic profiles of a primary tumor and ipsilateral breast cancer. First, we used the Pearson correlation between their CNA profiles. Second, we used a measure M that is derived from the percent concordance proposed by Waldman et al. (18) and adapted from Dice's formula (30) and corresponds to the number of common breakpoints divided by the mean number of breakpoints in either a primary tumor or an ipsilateral breast cancer. M is computed as follows, for a (i,j) pair.

$$M_{i,j} = \frac{\#(S_i \cap S_j)}{1/2 \times (\#S_i + \#S_j)},$$

in which S_i and S_j are the subsets of breakpoints present in the SNP arrays of the primary tumor, i, and ispilateral breast cancer, j. An example of M is given in Supplementary Fig. 1 (available online).

Two tumors had common breakpoints if the following conditions were fulfilled: 1) the changes in copy number occurred at the exact same locus and 2) the changes in copy number were of the same nature (ie, either an increase or a decrease in numbers) in the two tumors.

Assessing clonal relatedness from a dendrogram. We assumed that clonal unrelatedness was revealed by the clustering apart of the two tumors (primary tumor and ipsilateral breast tumor) from the same patient, reflecting that they were more similar to carcinomas of other patients than to each other. In contrast, the clustering together of two tumors from the same patient indicated clonal relatedness among them. For both measures of similarity (Pearson coefficient and M measure), we used Ward's criteria (31) as an agglomerative method in the hierarchical clustering.

Partial Identity Score. *Score definition.* To distinguish true recurrences from new primary tumors, we developed a partial identity score

Table 1. Patient and tumor characteristics of the 22 patients whose tumors (both PT and IBC) had exploitable SNP arrays*

Pair	Age, y	Family	Prob	BRCA1	BRCA2	pT	pN	Surgical margin, mm	Radiotherapy dose, Gy		No. of cycles of chemotherapy†
									Whole breast	Tumorectomy bed	
P1	23.1	0	20	0	2	1	0	≥4	54	54	4
P2	42.1	1	NA	NA	NA	1	0	≥4	50	50	0
P3	42.6	0	NA	NA	NA	1	0	≥4	54	54	0
P4	48.2	1	44	0	0	1	0	≥4	50	50	0
P5	45.5	0	NA	NA	NA	1	1	≥4	50	60	4
P6	35.7	0	8	0	0	2	0	≥4	51	66	4
P10	46.2	0	NA	NA	NA	2	0	0–3	50	70	0
P11	49.0	1	95	0	1	2	0	≥4	50	64	0
P12	48.9	1	NA	NA	NA	1	0	≥4	52	52	0
P13	45.0	0	NA	NA	NA	2	0	≥4	51	67	6
P14	43.6	0	NA	NA	NA	1	0	0–3	50	50	0
P15	46.1	0	NA	NA	NA	1	0	≥4	50	65	0
P16	48.4	0	NA	NA	NA	1	0	≥4	50	66	0
P18	27.9	1	82	0	0	2	0	0–3	50	70	4
P19	49.1	0	NA	NA	NA	2	0	0–3	51	65	4
P20	47.1	0	NA	NA	NA	2	1	0–3	45	65	4
P21	46.3	0	NA	NA	NA	1	0	DCIS	50	70	0‡
P22	35.0	0	NA	NA	NA	2	2	≥4	50	75	6‡
P23	30.8	0	NA	NA	NA	2	0	≥4	50	66	4
P24	47.7	0	NA	NA	NA	1	1	≥4	50	60	6
P25	43.0	0	NA	NA	NA	1	0	0–3	45	60	0‡
P26	30.5	0	NA	NA	NA	NA	1	≥4	52	70	4‡

* PT = primary tumor; IBC = ipsilateral breast cancer; SNP = single nucleotide polymorphism; Family = family history of breast cancer in the first two degrees (0 = no, 1 = yes); Prob = age-specific risk estimates of breast cancer according to the Claus Model (32); BRCA1 and BRCA2 = mutation found in BRCA1 and BRCA2 (0 = not found, 1 = deleterious, 2 = possibly deleterious, NA = not available); pT = histologic tumor classification according to Union Internationale Contre le Cancer (UICC) (33); pN = histologic lymph node classification according to UICC; DCIS = ductal carcinoma in situ.

† Chemotherapy consisted of 5-fluorouracil, anthracyclines, and cyclophosphamide.

‡ Patients were treated with tamoxifen for 5 years.

that is based on the *M* measure of similarity described above. The score reflects the number of common breakpoints among the ipsilateral breast cancer and the primary tumor. In addition, because very frequent breakpoints may be less informative than frequent ones in estimating the clonal relatedness between two tumors, the added value of each breakpoint was weighted according to its frequency among the samples of 44 control patients. The partial identity score (PS) was thus

$$\mathrm{PS}_{i,j} = \frac{\sum_{k\in(S_i\cap S_j)} (1-F_k)^2}{1/2\times[\sum_{k\in S_i}(1-F_k)+\sum_{k\in S_j}(1-F_k)]},$$

in which F_k represents the frequency of appearance of the breakpoint k calculated in the series of the 44 control breast cancers. An example of a partial identity score is given in Supplementary Fig. 1 (available online).

Statistical testing for partial identity. The partial identity score was calculated for all 462 possible "artificial pairs" (462 = 22 × 21, because each of the 22 primary tumors could be artificially paired with the ipsilateral breast cancer of the 21 other patients, *see* Table 3 notes). The distribution under the null hypothesis, H0, of no partial identity between the two tumors was estimated using all 462 possible artificial pairs. We rejected H0 with a type I error fixed at 5%, that is, we considered that a local recurrence shared partial identity with a primary tumor when the score was higher than the upper 5th percentile in the distribution of artificial pairs. The score was then calculated for the "natural pairs," that is, a primary tumor and its ipsilateral breast cancer occurring in the same patients (*see* Table 3 notes). Ipsilateral breast cancers from pairs with scores higher than this cutoff, that is, with shared partial identity, were considered to be true recurrences.

Robustness of the score. The robustness of the partial identity score was assessed by randomly selecting two subgroups of 15 and 7 patients from the population of 22 breast cancer patients. The first subgroup of 15 patients was used to compute the scores of the artificial pairs and to record the cutoff score corresponding to the 95th percentile. This score was then used to determine the status of each of the natural pairs in the seven patients of the other subgroup. To make the comparison statistically sound, each process was repeated 1000 times. The variation of the cutoff scores was assessed by box plot representation. The consistency of the ipsilateral breast cancer status was calculated as the percentage of extractions when the status of this pair was respectively a true recurrence or a new primary tumor.

All statistical tests were two-sided. *P* values less than .05 were considered to be statistically significant.

Results

Clinical and Histologic Features

The clinical and tumor characteristics of 22 patients whose tumors had exploitable SNP arrays were analyzed (Tables 1 and 2). According to clinical and histologic criteria (Table 2), nine of the 22 ipsilateral breast cancers were new primary tumors and the other

Table 2. Histologic characteristics of the primary tumors and their ipsilateral breast cancers: distinctions between new primary tumors and true recurrences according to clinical criteria or clustering methods*

	Primary tumors					Ipsilateral breast cancers					New primary tumors or true recurrences				
Pair	Type	Grade	ER	PR	Time, y	Location	Type	Grade	ER	PR	CNA	BKP	Clinical	Divergence	Score
P1	D	3	0	40	6.5	1	D	2	90	15	TR	NP‡	NP	CNA	0.020
P2	D	2	90	40	5.3	1	L	1	90	70	TR	NP‡	NP	CNA	0.000
P3	D	3	30	80	3.1	1	D	3	60	90	TR	TR‡	TR	No	0.465
P4	L	1	90	80	3.5	1	L	2	90	80	TR	TR‡	TR	No	0.278
P5	D	2	90	40	2.0	1	D	2	80	90	TR	TR‡	TR	No	0.555
P6	L	1	90	100	3.1	1	L	2	70	70	NP	NP‡	TR	Clinical	0.104
P10	L	3	80	95	5.0	0	D	2	70	40	NP	NP‡	NP	No	0.059
P11	L	3	0	0	6.3	1	D	3	0	0	NP	NP‡	NP	No	0.029
P12	L	2	90	50	2.9	0	L	2	90	0	TR	TR†	NP	Clinical	0.116
P13	D	2	20	85	4.6	1	D	2	95	20	TR	TR‡	TR	No	0.240
P14	L	2	90	60	2.5	1	L	2	0	100	TR	TR‡	TR	No	0.310
P15	D	2	100	80	3.3	1	D	2	70	100	NP	TR‡	TR	CNA	0.127
P16	D	2	80	30	3.8	1	D	1	20	70	TR	TR‡	NP	Clinical	0.317
P18	D	3	0	0	2.2	1	D	2	80	50	NP	NP‡	NP	No	0.004
P19§	D	3	0	0	3.0	1	D	3	0	0	TR	TR‡	TR	No	0.325
P20	D	3	0	0	1.4	0	D	3	0	0	TR	TR‡	NP	Clinical	0.139
P21	D	2	80	0	4.2	1	D	2	70		TR	TR‡	TR	No	0.360
P22§	D	2	20	50	3.5	1	M	3	15	0	TR	TR‡	NP	Clinical	0.394
P23	D	3	0	0	0.8	1	D	3	0	0	TR	TR‡	TR	No	0.341
P24§	D	3	0	0	1.0	1	D	3	0	0	TR	TR‡	TR	No	0.311
P25§	D	3	75	70	2.2	1	D	3	70	15	TR	TR‡	TR	No	0.375
P26	D	3	0	0	1.8	1	D	3	0	0	TR	TR‡	TR	No	0.519

* Type = histologic type (D = ductal, L = lobular, M = micropapillary); Grade = histologic grade; ER = estrogen receptor; PR = progesterone receptor; Location (1 = IBC at the index quadrant, 0 = IBC at a different quadrant); CNA = cluster according to copy number alterations; BKP = cluster according to breakpoints; Clinical = definition according to clinical criteria; NP = new primary tumor; TR = true recurrence.

† NP according to the partial identity score.

‡ Agreement with the definition by the partial identity score.

§ The ipsilateral breast cancers of these pairs received chemotherapy before surgery.

13 were true recurrences. Ipsilateral breast cancers occurred at a median time of 3.1 years after the initial breast cancer diagnosis (range = 0.8–6.5 years). In three of 22 (14%) patients, ipsilateral breast cancers occurred in a different quadrant than the initial tumor; all of these were defined clinically as new primary tumors.

Genomic Studies

The pangenomic profiles of a primary tumor and its ipsilateral breast cancer revealed common breakpoints, with a precision within a SNP that can be used as markers of their clonal relatedness. Pair 5 is given as an illustration (Fig. 1).

The median number of breakpoints per array was statistically significantly higher for ipsilateral breast cancers (median = 71, range = 21–433) than for primary tumors (median = 52, range = 4–646) ($P = .001$) (Table 3). The mean number of common breakpoints per pair was also statistically significantly higher for natural pairs (mean = 18.8, SD = 18.8) than for artificial pairs (mean = 4.1, SD = 3.1) ($P = 0.5 \times 10^{-6}$).

Clustering by Copy Number Alterations or Breakpoints

According to hierarchical clustering by DNA CNAs (Fig. 2) and by breakpoints (Fig. 3), five and six ipsilateral breast cancers, respectively, were new primary tumors. The two clustering methods and the clinical definition agreed for 14 pairs (Table 2). However, for five pairs (P6, P12, P16, P20, P22), the clinical definition disagreed with both clustering methods and, for three others (P1, P2, P15), the clustering by breakpoints disagreed with that by CNAs but agreed with the clinical definition. The recurrences in pairs 1 and 2 were identified as true recurrences by the CNA clustering but as new primary tumors by the clinical definitions because of the reappearance of estrogen receptors in the pair 1 ipsilateral breast cancer and different histologic type (ductal instead of lobular carcinoma) in pair 2. In pair 15, CNA clustering did not find a true recurrence, whereas the clinical definition did. No statistically significant differences in clinical and histologic characteristics between the patients diagnosed with new primary tumors or true recurrences were observed by breakpoint information, apart from a suggestion for patients with new primary tumors to be younger and to have a more frequent family history of breast cancer (Supplementary Table 1, available online).

Partial Identity Score

According to the partial identity score reported for each pair in Table 2, 15 ipsilateral breast cancers were true recurrences and seven were new primary tumors (Fig. 4). With a type I error set at 5%, the partial identity score disagreed with clustering by breakpoints in pair 12 only; the clinical definition was new primary tumor because of a change in tumor location. When the score was determined according to Waldman's percent of concordance without either weighing the influence of the coexistence of breakpoints according to their frequency in a similar population or excluding

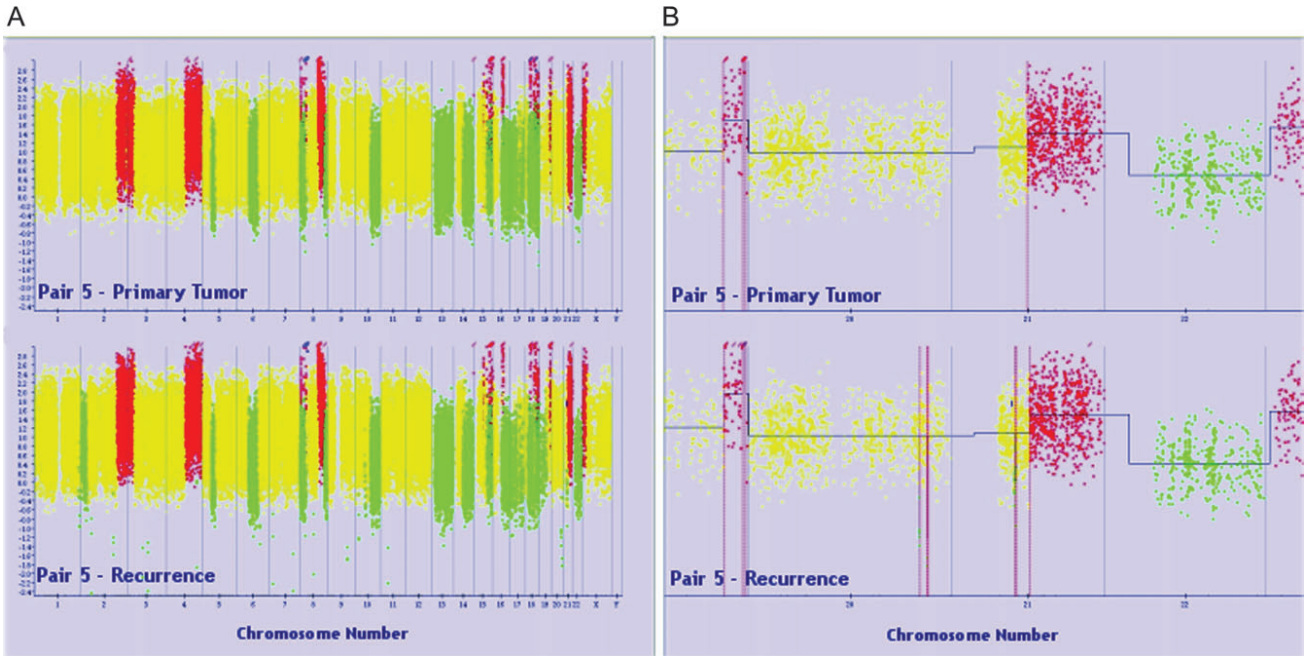


Fig. 1. Genomic profiles of tumors of pair 5 to illustrate the finding of common breakpoints within a single nucleotide polymorphism (SNP). A genomic profile represents the ordered values of the DNA copy numbers obtained as described in "Subjects and Methods". Each **dot** represents the number of DNA copies at each SNP position. Regions with gains are in **red**, with losses in **green**, with no DNA copy number alterations in **yellow**. **A**) Pangenomic profiles. **B**) Profiles of chromosomes 20, 21, and 22. Top primary tumor of pair 5; bottom, ipsilateral breast cancer of pair 5. The **blue horizontal line** represents the smoothing line and the **dotted vertical line** the breakpoint position.

the breakpoints that occur in the copy number variable regions in the HapMap collection, the attribution of the status of three pairs (20 changed from a true recurrence to a new primary, whereas 6 and 12 became true recurrences) and two pairs (10 and 12 changed from new primaries to true recurrences) changed, respectively.

The status of all pairs was confirmed by the 1000 random extractions (Supplementary Table 2, available online). The mean cutoff value was 0.1203 (SD = 0.0102) (Supplementary Fig. 2, available online). The cutoff used to determine the status of the 22 ipsilateral breast cancers, which was defined using all 462 artificial pairs, was 0.1212.

Prognostic Value of the Determination of the Nature of the Ipsilateral Breast Cancer

Patients who were diagnosed with true recurrences had lower metastasis-free survival than those diagnosed with new primary tumors (Supplementary Fig. 3, available online). The difference in metastasis-free survival in the two groups was not statistically significant when they were defined based on clinical and histologic characteristics (5-year metastasis-free survival: 76%, 95% CI = 52% to 100% for new primary tumors and 38%, 95% CI = 17% to 83% for true recurrences; $P = .18$; primary tumors vs true recurrences, hazard ratio = 2.8, 95% CI = 0.6 to 13.7). However, metastasis-free survival was different when the groups were defined according to the partial identity score (5-year metastasis-free survival: 100% for new primary tumors and 29%, 95% CI = 11% to 78% for true recurrences; $P = .01$).

Discussion

DNA breakpoint information was more often in agreement with the clinical definition than that from CNAs to define true recurrences among ipsilateral breast cancers in this population. We developed a partial identity score that is based on DNA breakpoints, which allowed statistical discrimination between new primary tumors and true recurrences. This score outperformed the clinical prognosis determination in terms of metastasis-free survival.

We chose to base our study on a series of young (<50 years old) premenopausal women not only because young age is recognized as one of the most important independent prognostic factors for ipsilateral breast recurrence (34–40) but also to ensure a very high level of homogeneity. In addition, all patients had undergone breast-conserving surgery followed by whole-breast radiotherapy for their initial breast cancers, which were selected as either ductal or lobular invasive carcinomas, and were treated at the same cancer center.

Our results show that some ipsilateral breast cancers share with their primary tumors many DNA CNA breakpoints at the same locations (precision to within a SNP, as illustrated in Fig. 1). From these observations, we produced a method of determining true recurrences that relies on a number of assumptions. The first and most obvious is that the vast majority of breast cancers are of clonal origin. The second is that a tumor retains a substantial number of genomic alterations throughout its evolution. The third assumption, which is key to the method that we have developed, is that the exact locations of the breakpoints that are on the edge of a given change in DNA copy numbers are better hallmarks of a given tumor than the magnitude or width of the genomic alteration itself. For example, because the deletion that causes the loss of Phosphatase and TENsin homolog (PTEN) alters regulatory pathways that lead to precocious development and neoplasia in the mammary gland (41), it can be found in many breast cancers (42–44); however, the exact location of the breakpoints bordering this deletion can be specific to a given tumor. We provide as an

Table 3. Number of common breakpoints in natural (same patient) and artificial (two different patients) pairs of primary tumors (vertically) and ipsilateral breast cancers (horizontally)

No. of BKPs in		No. of BKPs in PT*																					
		77	11	46	16	94	8	22	4	31	55	12	11	58	646	89	69	127	49	60	57	41	72
IBC*	Pair	P1	P2	P3	P4	P5	P6	P10	P11	P12	P13	P14	P15	P16	P18	P19	P20	P21	P22	P23	P24	P25	P26
433	P1	6†	3	12‡	3	8	5§	5	1	4	5	6	1	1	7§	8	6	7	3	8	8	5	12‡
25	P2	0	0†	1	0	1	0	0	0	3‡	0	1	0	0	0	1	0	2	1	0	1	0	0
43	P3	3	2	23†‡§	5	5	2	10§	2§	4	6	5	4	3	4	11	5	7	6	4	8	4	9
26	P4	5	3	7	9†‡§	5	2	7	0	6§	4	4	3	2	0	9‡	3	4	5	3	6	3	5
128	P5	3	3	11	4	64†‡§	1	7	0	4	4	5	2	2	2	8	4	3	8	3	2	3	10
21	P6	3	3	4‡	3	3	3†	4‡	0	4‡	1	4‡	2	0	0	3	1	2	1	1	2	4‡	2
23	P10	3	2	4	3	3	1	3†	1	2	2	1	1	1	3	5‡	1	1	2	1	1	5‡	3
97	P11	5	2	19‡	6	9	1	9	2†§	6§	9	7	6§	5	7§	14	7	10	9	4	12	4	13
35	P12	6‡	3	4	5	4	2	3	0	6†‡§	2	2	3	2	0	4	3	3	3	1	4	4	4
74	P13	3	2	7	3	6	1	5	1	3	18†‡§	4	3	2	2	7	3	3	4	2	2	5	2
35	P14	1	2	7	3	7	3	5	0	3	5	10†‡§	2	1	3	6	3	4	3	2	3	5	4
49	P15	5	2	5	3	4	2	3	0	6‡§	4	1	5†	4	2	3	2	4	3	1	1	2	2
84	P16	2	2	3	2	3	0	2	0	4	2	0	3	23†‡§	1	1	1	3	2	0	3	3	4
53	P18	2	2	9‡	3	3	1	5	1	3	2	3	2	0	2†	7	5	3	2	3	2	3	5
150	P19	9§	4§	18	5	8	2	10§	2§	3	10	5	5	5	7§	42†‡§	13†	11	6	11	10	6	10
93	P20	4	1	6	1	5	0	3	1	2	4	1	2	1	5	7	12†‡	3	4	6	3	3	6
219	P21	2	1	12	3	6	1	5	2§	2	5	3	4	4	6	8	7	63†‡§	6	7	8	3	5
100	P22	5	2	17	5	8	1	10§	1	5	5	5	4	5	3	13	9	10	31†‡§	6	10	5	9
73	P23	7	1	10	3	6	1	7	2§	3	5	5	2	1	5	12	10	6	6	25†‡§	6	3	10
69	P24	6	2	11	5	3	2	6	1	4	5	3	2	3	5	9	5	5	3	7	23†‡§	1	11
42	P25	4	3	9	5	5	2	7	2§	4	5	5	2	2	2	5	4	4	6	1	2	18†‡§	3
88	P26	5	3	11	7	7	1	9	1	6§	5	3	2	4	3	17	5	2	8	5	9	3	43†‡§

* Number of BKPs per tumor. BKP = breakpoint; PT = primary tumor; IBC = ipsilateral breast cancer.

† Numbers correspond to the 22 natural pairs of PTs and their IBCs arising in the same patient; numbers in the other cells correspond to the 462 (22 × 21) artificial pairs of each PT with all other possible IBCs arising in other patients.

‡ Pairs with the most common BKPs per PT.

§ Pairs with the most common BKPs per IBC.

example (Supplementary Fig. 4, available online) the prototype case of PTEN deletion in which the breakpoints are identical between the primary tumor and ipsilateral breast cancer of pair 5 and yet differ in all the other tumors that also harbor a loss of PTEN.

Because clustering is commonly used to determine whether two tumors are clonally related and because it performs better than previously developed similarity scores (18,19), we addressed the issue of whether there was added value in looking at breakpoints rather than at CNAs by comparing clustering by CNAs and by breakpoints to determine the nature of the ipsilateral breast cancer. We concluded from the comparison of clusterings of CNAs and of breakpoints that breakpoint information is more valid than CNA information because when they were discordant, the definition by breakpoints always agreed with the clinical definition, which is routinely used in clinical practice.

A second issue was whether a method could be found to quantify the partial identity between two tumors. We chose to use a partial identity score rather than the results of clustering for a number of reasons. 1) Clustering methods have been designed for exploratory data analysis, so that using a score is more appropriate for a discrimination purpose. 2) A score induces a natural ordering of the pairs from the most dissimilar to the most similar, which is not the case for clustering. 3) The assessment of clonal relatedness by a score can be statistically motivated through the choice of a threshold, as we have demonstrated in the present work. For clustering, clonal relatedness of two tumors depends only on their being clustered apart on the dendrogram, which leads to inconsistent decisions over time. As illustrated by Fig. 3, if pair 2 had not been included in the study, the ipsilateral breast cancer from pair 6 would have been considered as a true recurrence rather than a new primary tumor. Conversely, the assessment of the partial identity score robustness was satisfactory with a narrow range of the cutoff (Supplementary Fig. 2, available online) and with the consistency of the ipsilateral breast cancer status (Supplementary Table 2, available online). Moreover, a score allows one to choose the cutoff that best distinguishes new primary tumors from true recurrences. In this study, we chose a type I error rate at 5% to favor sensitivity for diagnosing true recurrences over the specificity. Further studies will be needed to verify the biologic validity of this choice (Supplementary Fig. 3, available online).

In addition, we chose to weigh the influence of a common breakpoint between the ipsilateral breast cancer and its primary tumor by a factor that takes into account the frequency of this given breakpoint in a population of similar tumors. This weighting changed the determination of three of 22 pairs.

The clinical definition considered an ipsilateral breast cancer as a new primary tumor when the partial identity score did not in three instances. In the first because of a change in location for pairs 12 and 20, in the second because of a lesser degree of differentiation for pair 16, and in the third because of a change in histology for pair 22. The first example illustrates the possibility that a true recurrence can occur at a distance from the first cancer. The second exemplifies the possibility for a true recurrence to have many but not all of the striking alterations present in the primary tumor.

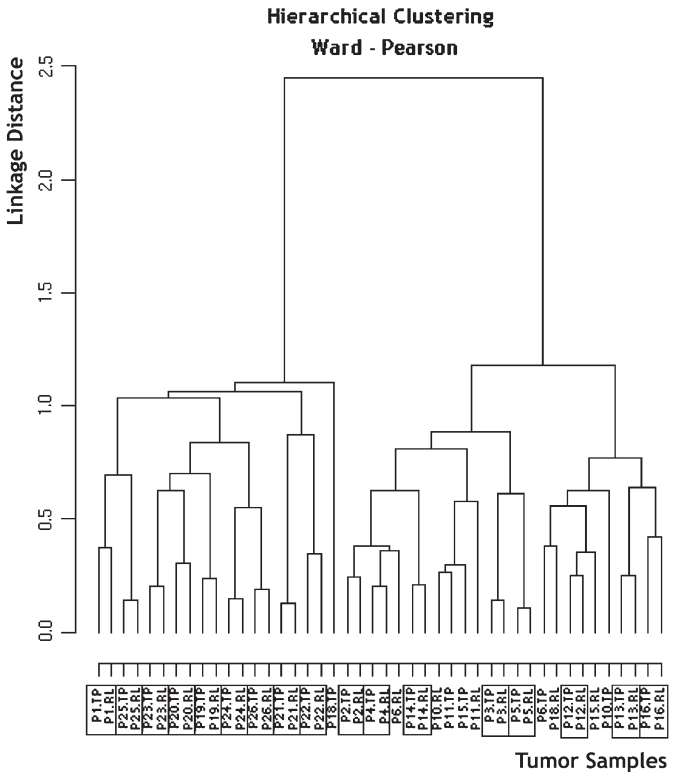


Fig. 2. Dendogram of hierarchical clustering by DNA copy number alterations (Ward–Pearson) of 22 available pairs of primary tumors (TP) and their ipsilateral breast cancer (RL). **Boxes** represent natural pairs with a true recurrence, that is, a pair of tumors from one patient clustered together.

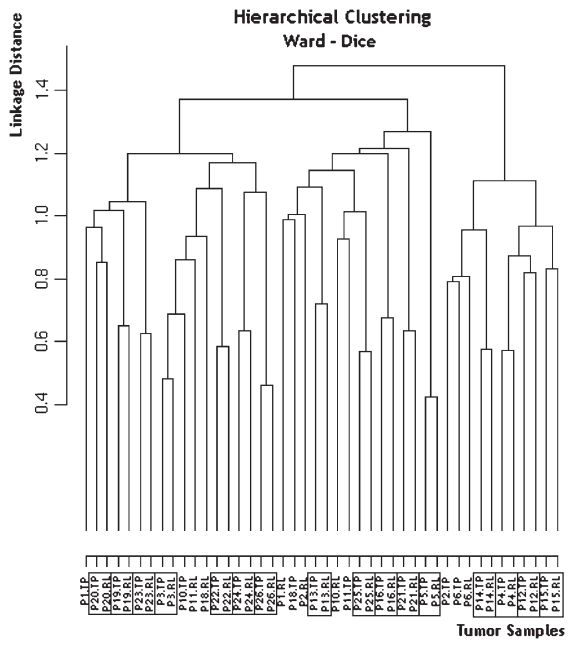


Fig. 3. Dendogram of hierarchical clustering by breakpoints (Ward–Dice) of 22 available pairs of primary tumors (TP) and their ipsilateral breast cancer (RL). **Boxes** represent natural pairs with a true recurrence, that is, a pair of tumors from one patient clustered together.

A criticism that can be made of the clinical definition is that it assumes that a true recurrence is derived from its primary tumor instead of only being related to it. A true recurrence, according to some clinical definitions (5,6,11), cannot be more differentiated than its primary tumor. Usual classifications define differentiation according to histologic grading, DNA ploidy, or the presence of ductal carcinoma in situ. They are based on the assumption that tumors accumulate genetic alterations with time (22,45,46) and that the chronologic order of these alterations reflects the development of a tumor clone. This assumption is, however, challenged by the fact that the ipsilateral breast cancers are neither more aggressive nor more undifferentiated than their primary tumors (47).

The situation with pair 22 illustrates another possible limitation of histologic determination. Here, the clinical status of the ipsilateral breast cancer was of a new primary tumor because its histologic type was a micropapillary carcinoma, whereas the initial tumor was a ductal carcinoma. However, after further histologic analysis, a minor component of micropapillary carcinoma was revealed in the initial carcinoma that otherwise would have been overlooked (Supplementary Fig. 5, available online). This finding implies that, in some instances, the current histologic taxonomy, which is based more on architectural features than on biologic ones, could become obsolete and that some ipsilateral breast cancers could qualify as true recurrences without sharing the same histologic type as their primary tumors.

We observed that patients with true recurrences had lower metastasis-free survival than patients with new primary tumors and that this difference became statistically significant when the partial identity score, instead of clinical definition, was used to define ipsilateral breast cancer types. This observation has been shared by many authors (5,6,10,12). Possible explanations are,

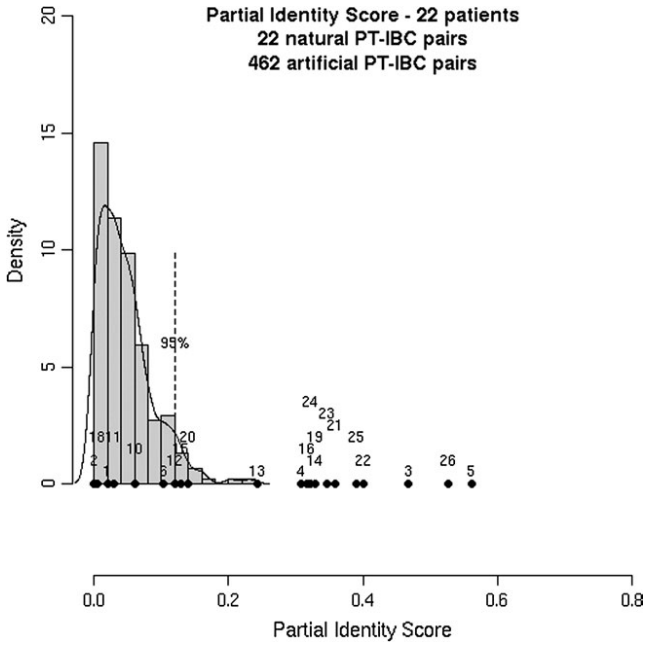


Fig. 4. Partial identity score. Histogram performed on 462 artificial pairs (two different patients) of tumors and representation of the 22 natural (same patient) pairs of primary tumors (PT)/ipsilateral breast cancer (IBC). x-axis: partial identity score (the higher the score, the more likely the IBC is a true recurrence), y-axis: number of artificial pairs in **boxes**. The **vertical dashed bar** represents the upper 5th percentile of the artificial pairs distribution and the threshold above which true recurrences were defined (rejection of the null hypothesis). Each **dot** represents one of the 22 natural pairs (its identifier is written above it).

first, that a true recurrence is the expression of clones that are resistant to adjuvant treatment and therefore could be more difficult to eradicate and, second, that it could be the tip of the iceberg, that is, distant metastases. Conversely, new primary tumors have a prognosis similar to de novo primary cancers but can also reflect a genetic predisposition to develop breast cancer, in the contralateral breast in particular. The clinical implication should therefore be to advocate the use of a systemic treatment in the case of true recurrences and the use of either chemoprevention, such as hormone therapies (48–50) or screening with magnetic resonance imaging (51–53), for patients who are diagnosed with new primary tumors. Here, using breakpoint information led to a better discrimination between new primary tumors and true recurrences in terms of metastasis-free prognosis than the clinical definition.

We also hope that a better distinction among ipsilateral breast cancers of tumors that are genetically related to their primary tumors, that is, true recurrences, will help reveal genetic differences that would provide new information on radioresistance and tumor aggressiveness. To date, little is known about the differential or similarity of the pangenomic expression or the nature of both new primary tumors and ipsilateral breast cancers. Kreike et al. (54) performed a gene expression analysis of 18 000 cDNAs in nine pairs of primary breast cancer with their ipsilateral breast recurrences among women who were younger than 51 years at the time of their initial breast-conserving therapy. Paired data analysis showed no set of genes that had consistently different levels of expression in primary tumors and local recurrences. Another route that has still scarcely been explored is the search for a biologic signature to predict the risk of local recurrence, especially after breast-conserving treatment (54–56). A better distinction between new primary tumors and true recurrences is needed to perform a supervised study based on the occurrence of true recurrences only and not of all ipsilateral breast cancers.

However, our scoring method, which is based on the DNA breakpoint partial identity, has two shortcomings. First, it suffers from the need to conserve unaltered, freshly frozen tissue samples of both the primary tumor and the ipsilateral breast recurrence. This problem should, however, be resolved in time with the possibility of performing the same genomic studies on formalin-fixed paraffin-embedded tissue samples (57–61) or when cryoconservation of either biopsies or fine-needle aspirations (because only 250 ng of DNA is needed, ie, less than 50 000 cells) become standard practice and will make it possible to perform SNP arrays on many more patients. Second, it requires selecting tumors with a cancer cellularity of more than 50%, discarding in the process a number of potentially analyzable tumors. This loss should be diminished in time with both a better selection of frozen tissue material due to the increased experience of the pathologist and the possibility of performing laser capture microdissection.

References

1. Temple WJ, Russell ML, Parsons LL, et al. Conservation surgery for breast cancer as the preferred choice: a prospective analysis. *J Clin Oncol.* 2006;24(21):3367–3373.
2. Clarke M, Collins R, Darby S, et al. Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. *Lancet.* 2005; 366(9503):2087–2106.
3. Engel J, Kerr J, Schlesinger-Raab A, Sauer H, Holzel D. Quality of life following breast-conserving therapy or mastectomy: results of a 5-year prospective study. *Breast J.* 2004;10(3):223–231.
4. Moyer A. Psychosocial outcomes of breast-conserving surgery versus mastectomy: a meta-analytic review. *Health Psychol.* 1997;16(3): 284–298.
5. Haffty BG, Carter D, Flynn SD, et al. Local recurrence versus new primary: clinical analysis of 82 breast relapses and potential applications for genetic fingerprinting. *Int J Radiat Oncol Biol Phys.* 1993;27(3): 575–583.
6. Huang E, Buchholz TA, Meric F, et al. Classifying local disease recurrences after breast conservation therapy based on location and histology: new primary tumors have more favorable outcomes than true local disease recurrences. *Cancer.* 2002;95(10):2059–2067.
7. Gage I, Recht A, Gelman R, et al. Long-term outcome following breast-conserving surgery and radiation therapy. *Int J Radiat Oncol Biol Phys.* 1995;33(2):245–251.
8. Touboul E, Buffat L, Belkacemi Y, et al. Local recurrences and distant metastases after breast-conserving surgery and radiation therapy for early breast cancer. *Int J Radiat Oncol Biol Phys.* 1999;43(1):25–38.
9. Recht A, Silen W, Schnitt SJ, et al. Time-course of local recurrence following conservative surgery and radiotherapy for early stage breast cancer. *Int J Radiat Oncol Biol Phys.* 1988;15(2):255–261.
10. Komoike Y, Akiyama F, Iino Y, et al. Analysis of ipsilateral breast tumor recurrences after breast-conserving treatment based on the classification of true recurrences and new primary tumors. *Breast Cancer.* 2005;12(2): 104–111.
11. Smith TE, Lee D, Turner BC, Carter D, Haffty BG. True recurrence vs. new primary ipsilateral breast tumor relapse: an analysis of clinical and pathologic differences and their implications in natural history, prognoses, and therapeutic management. *Int J Radiat Oncol Biol Phys.* 2000;48(5): 1281–1289.
12. Schlechter BL, Yang Q, Larson PS, et al. Quantitative DNA fingerprinting may distinguish new primary breast cancer from disease recurrence. *J Clin Oncol.* 2004;22(10):1830–1838.
13. Wang ZC, Buraimoh A, Iglehart JD, Richardson AL. Genome-wide analysis for loss of heterozygosity in primary and recurrent phyllodes tumor and fibroadenoma of breast using single nucleotide polymorphism arrays. *Breast Cancer Res Treat.* 2006;97(3):301–309.
14. Vicini FA, Antonucci JV, Goldstein N, et al. The use of molecular assays to establish definitively the clonality of ipsilateral breast tumor recurrences and patterns of in-breast failure in patients with early-stage breast cancer treated with breast-conserving therapy. *Cancer.* 2007;109(7): 1264–1272.
15. van der Sijp JR, van Meerbeeck JP, Maat AP, et al. Determination of the molecular relationship between multiple tumors within one patient is of clinical importance. *J Clin Oncol.* 2002;20(4):1105–1114.
16. Shibata A, Tsai YC, Press MF, Henderson BE, Jones PA, Ross RK. Clonal analysis of bilateral breast cancer. *Clin Cancer Res.* 1996;2(4): 743–748.
17. Kuukasjarvi T, Karhu R, Tanner M, et al. Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. *Cancer Res.* 1997;57(8):1597–1604.
18. Waldman FM, DeVries S, Chew KL, Moore DH 2nd, Kerlikowske K, Ljung BM. Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences. *J Natl Cancer Inst.* 2000;92(4):313–320.
19. Teixeira MR, Ribeiro FR, Torres L, et al. Assessment of clonal relationships in ipsilateral and bilateral multiple breast carcinomas by comparative genomic hybridisation and hierarchical clustering analysis. *Br J Cancer.* 2004;91(4):775–782.
20. Zhao X, Li C, Paez JG, et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* 2004;64(9):3060–3071.
21. Balaton AL, Coindre JM, Collin F, et al. Recommendations for the immunohistochemical evaluation of hormone receptors on paraffin sections of breast cancer. Study Group on Hormone Receptors using

Immunohistochemistry FNCLCC/AFAQAP. National Federation of Centres to Combat Cancer/French Association for Quality Assurance in Pathology. *Ann Pathol.* 1996;16:144–148.
22. Nowell PC. The clonal evolution of tumor cell populations. *Science.* 1976;194(4260):23–28.
23. Sambrook J, Fritsch EF, Maniatis T. *Molecular Cloning. A Laboratory Manual.* 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989.
24. Hupe P, La Rosa P, Liva S, Lair S, Servant N, Barillot E. ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors. *Oncogene.* 2007;26:6641–6652.
25. Kaplan EL, Meier P. Nonparametric estimation from incomplete observation. *J Am Stat Assoc.* 1958;53:457–481.
26. Cox DR, Oakes D. *Analysis of Survival Data*. London: Chapman & Hall; 1984.
27. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics.* 2004;20(18):3413–3422.
28. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature.* 2006;444(7118):444–454.
29. La Rosa P, Viara E, Hupe P, et al. VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics.* 2006;22(17):2066–2073.
30. Dice LR. Measures of the amount of ecologic association between species. *Ecology.* 1945;26:297–302.
31. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc.* 1963;58:236–244.
32. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer.* 1994; 73(3):643–651.
33. Sobin LH, Wittekind C. TNM classification of malignant tumours. New York: Wiley-Liss; 2002.
34. Vrieling C, Collette L, Fourquet A, et al. Can patient-, treatment- and pathology-related characteristics explain the high local recurrence rate following breast-conserving therapy in young patients?. *Eur J Cancer.* 2003;39(7):932–944.
35. Fourquet A, Campana F, Zafrani B, et al. Prognostic factors of breast recurrence in the conservative management of early breast cancer: a 25-year follow-up. *Int J Radiat Oncol Biol Phys.* 1989;17(4):719–725.
36. Borger J, Kemperman H, Hart A, Peterse H, van Dongen J, Bartelink H. Risk factors in breast-conservation therapy. *J Clin Oncol.* 1994;12(4): 653–660.
37. Elkhuizen PH, van de Vijver MJ, Hermans J, Zonderland HM, van de Velde CJ, Leer JW. Local recurrence after breast-conserving therapy for invasive breast cancer: high incidence in young patients and association with poor survival. *Int J Radiat Oncol Biol Phys.* 1998;40(4): 859–867.
38. Elkhuizen PH, Voogd AC, van den Broek LC, et al. Risk factors for local recurrence after breast-conserving therapy for invasive carcinomas: a case-control study of histological factors and alterations in oncogene expression. *Int J Radiat Oncol Biol Phys.* 1999;45(1):73–83.
39. Oh JL, Bonnen M, Outlaw ED, et al. The impact of young age on locoregional recurrence after doxorubicin-based breast conservation therapy in patients 40 years old or younger: how young is "young"?. *Int J Radiat Oncol Biol Phys.* 2006;65(5):1345–1352.
40. Bollet MA, Sigal-Zafrani B, Mazeau V, et al. Age remains the first prognostic factor for loco-regional breast cancer recurrence in young (<40 years) women treated with breast conserving surgery first. *Radiother Oncol.* 2007;82(3):272–280.
41. Li G, Robinson GW, Lesche R, et al. Conditional loss of PTEN leads to precocious development and neoplasia in the mammary gland. *Development.* 2002;129(17):4159–4170.
42. Sapolsky RJ, Hsie L, Berno A, Ghandour G, Mittmann M, Fan JB. High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays. *Genet Anal.* 1999;14(5–6):187–192.
43. Jonsson G, Staaf J, Olsson E, et al. High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization. *Genes Chromosomes Cancer.* 2007;46(6):543–558.
44. Perez-Tenorio G, Alkhori L, Olsson B, et al. PIK3CA mutations and PTEN loss correlate with similar prognostic factors and are not mutually exclusive in breast cancer. *Clin Cancer Res.* 2007;13(12): 3577–3584.
45. Chen LC, Kurisu W, Ljung BM, Goldman ES, Moore D 2nd, Smith HS. Heterogeneity for allelic loss in human breast cancer. *J Natl Cancer Inst.* 1992;84(7):506–510.
46. Lininger RA, Fujii H, Man YG, Gabrielson E, Tavassoli FA. Comparison of loss heterozygosity in primary and recurrent ductal carcinoma in situ of the breast. *Mod Pathol.* 1998;11(12):1151–1159.
47. Sigal-Zafrani B, Bollet MA, Antoni G, et al. Are ipsilateral breast tumour invasive recurrences in young (40 years) women more aggressive than their primary tumours?. *Br J Cancer.* 2007;97(8):1046–1052.
48. Powles TJ, Ashley S, Tidy A, Smith IE, Dowsett M. Twenty-year follow-up of the Royal Marsden randomized, double-blinded tamoxifen breast cancer prevention trial. *J Natl Cancer Inst.* 2007;99(4):283–290.
49. Cuzick J, Forbes JF, Sestak I, et al. Long-term results of tamoxifen prophylaxis for breast cancer—96-month follow-up of the randomized IBIS-I trial. *J Natl Cancer Inst.* 2007;99(4):272–282.
50. Veronesi U, Maisonneuve P, Rotmensz N, et al. Tamoxifen for the prevention of breast cancer: late results of the Italian Randomized Tamoxifen Prevention Trial among women with hysterectomy. *J Natl Cancer Inst.* 2007;99(9):727–737.
51. Kriege M, Brekelmans CT, Boetes C, et al. Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. *N Engl J Med.* 2004;351(5):427–437.
52. Kuhl CK, Schrading S, Leutner CC, et al. Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer. *J Clin Oncol.* 2005;23(33): 8469–8476.
53. Lehman CD, Gatsonis C, Kuhl CK, et al. MRI evaluation of the contralateral breast in women with recently diagnosed breast cancer. *N Engl J Med.* 2007;356(13):1295–1303.
54. Kreike B, Halfwerk H, Kristel P, et al. Gene expression profiles of primary breast carcinomas from patients at high risk for local recurrence after breast-conserving therapy. *Clin Cancer Res.* 2006;12(19): 5705–5712.
55. Nuyten DS, Kreike B, Hart AA, et al. Predicting a local recurrence after breast-conserving therapy by gene expression profiling. *Breast Cancer Res.* 2006;8(5):R62.
56. Niméus E, Krogh M, Malmström P, Strand C, Fredriksson I, Karlsson P, et al. Gene expression profiling in primary breast cancer distinguishes patients developing local recurrence despite postoperative radiotherapy after breast conserving surgery. In: 29th Annual *San Antonio Breast Cancer Symposium* 2006. San Antonio, TX: 2007;103(1):115–124.
57. Isola J, DeVries S, Chu L, Ghazvini S, Waldman F. Analysis of changes in DNA sequence copy number by comparative genomic hybridization in archival paraffin-embedded tumor samples. *Am J Pathol.* 1994;145(6): 1301–1308.
58. Devries S, Nyante S, Korkola J, et al. Array-based comparative genomic hybridization from formalin-fixed, paraffin-embedded breast tumors. *J Mol Diagn.* 2005;7(1):65–71.
59. Johnson NA, Hamoudi RA, Ichimura K, et al. Application of array CGH on archival formalin-fixed paraffin-embedded tissues including small numbers of microdissected cells. *Lab Invest.* 2006;86(9): 968–978.
60. Oosting J, Lips EH, van Eijk R, et al. High-resolution copy number analysis of paraffin-embedded archival tissue using SNP BeadArrays. *Genome Res.* 2007;17(3):368–376.
61. Schubert EL, Hsu L, Cousens LA, et al. Single nucleotide polymorphism array analysis of flow-sorted epithelial cells from frozen versus fixed tissues for whole genome analysis of allelic loss in breast cancer. *Am J Pathol.* 2002;160(1):73–79.

Funding

Institut Curie, the "Courir pour la vie, Courir pour Curie" association, the "Odyssea" association and the PHRC 2006 (AOM 06 149).

Notes

M. A. Bollet and N. Servant contributed equally to this work. The authors thank the members of the departments of Tumor Biology (Martial Caly, Blandine Massemin, Michèle Galut), Biostatistics (Eléonore Gravier, Chantal Gautier), Translational Research (David Gentien, Cécile Reyes, Audrey Rapinat, Benoît Albaud, Vincent Lepetit), and Bioinformatics (Philippe La Rosa, Séverine Lair) who participated in this study. The authors are also indebted to Anne Vincent-Salomon, Patricia de Crémoux, Dominique Stoppa-Lyonnet, and particularly Olivier Delattre for their very valuable comments on this work. Finally, they thank all the members of the Institut Curie Breast Cancer Group.

The sponsors had no role in the study design, data collection, interpretation of the results, preparation of the manuscript, or the decision to submit the manuscript for publication.

Manuscript received June 4, 2007; revised October 16, 2007; accepted November 13, 2007.

A.2 Genome Alteration Print (GAP)

Method

Open Access

Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays

Tatiana Popova[*†], Elodie Manié[*†], Dominique Stoppa-Lyonnet[*†‡§], Guillem Rigaill[¶¥], Emmanuel Barillot[*#**] and Marc Henri Stern[*†]

Addresses: [*]Centre de Recherche, Institut Curie, 26 rue d'Ulm, Paris, 75248, France. [†]INSERM U830, Institut Curie, 26 rue d'Ulm, Paris, 75248, France. [‡]Department of Tumor Biology, Institut Curie, 26 rue d'Ulm, Paris, 75248, France. [§]University Paris Descartes, 12 rue de l'Ecole de Médecine, Paris, 75270, France. [¶]Translational Research Department, Institut Curie, 1 avenue Claude Vellefaux, Paris, 75475, France. [¥]MIA 518, AgroParisTech/INRA, 16 rue Claude Bernard, Paris, 75231, France. [#]INSERM U900, Institut Curie, 26 rue d'Ulm, Paris, 75248, France. [**]Ecole des Mines ParisTech, 35 rue Saint Honoré, Fontainebleau, 77305, France.

Correspondence: Tatiana Popova. Email: tatiana.popova@curie.fr

Published: 11 November 2009

Genome ***Biology*** 2009, **10:**R128 (doi:10.1186/gb-2009-10-11-r128)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/11/R128

Received: 2 February 2009
Revised: 24 September 2009
Accepted: 11 November 2009

Abstract

We describe a method for automatic detection of absolute segmental copy numbers and genotype status in complex cancer genome profiles measured with single-nucleotide polymorphism (SNP) arrays. The method is based on pattern recognition of segmented and smoothed copy number and allelic imbalance profiles. Assignments were verified by DNA indexes of primary tumors and karyotypes of cell lines. The method performs well even for poor-quality data, low tumor content, and highly rearranged tumor genomes.

Background

Alterations of genomic DNA are hallmarks of cancer [1]. These genetic alterations include point mutations and small insertion/deletion events, translocations, copy-number changes, amplifications, and losses of heterozygosity. Chromosome copy-number alterations and homozygosities (uniparental disomies) acquired during cancer evolution are believed to be selected as the result of the loss of function of tumor-suppressor genes and the gain of function of oncogenes. Recurrent copy-number variations (CNVs) or loss of heterozygosity (LOH) are therefore critical indicators of possible localization of cancer-related genes [1]. Both recurrent regions of alteration and patterns of genomic instability contribute to tumor classification [2]. Single-nucleotide polymorphism (SNP) arrays are presently one of the most efficient technologies for the identification of such alterations [3,4]. SNP arrays simultaneously define copy-number changes and allelic imbalances (including LOH) occurring in a tumor, at high resolution and throughout the whole genome [5].

Genome-wide SNP arrays are available mainly for Affymetrix [6] and Illumina [7] platforms. On both platforms, SNP genotypes are extracted from allele-specific signal intensities after array hybridization. Arbitrarily, the two alleles are designated as A and B, and the ratio of allele-specific signal intensities (A/B, A/(A+B), and so on, depending on the method used) provides an allelic-imbalance value. Chromosomal aberrations are identified by (a) relative copy-number changes and (b) allelic imbalances. Both platforms were originally designed for high-throughput genotyping of *normal* genomes, and they require specific normalization and data-mining strategies to study alterations in *cancer* genomes [8]. Two characteristics of genetic alterations are essential to

extract from SNP data: (a) breakpoints corresponding to the boundaries of the altered regions of genomic DNA, and (b) copy number and genotype status of each such alteration.

Accurate determination of breakpoints has been addressed from many aspects, starting from reduction of nonrelevant variation to optimal breakpoint counts and positioning [9-14]. As compromises between sensitivity and specificity, these methods will perform variably, depending on the specific setting used, the quality of the primary data set, and the complexity of the tumor genomes.

Determination of copy numbers and genotype status of each alteration is more complicated, and no general solution has yet been proposed. Attempts to address this question include a manual interpretation of Affymetrix 500K SNP-array results for glioblastomas presented in [15] and an automatic copy-number recognition method based on allelic imbalances for the Illumina platform, proposed in [16]. Other methods attribute relative gain, loss, or allelic-imbalance status without addressing the determination of absolute copy number and genotype (cnvPartition from Illumina, [17-23]).

Three major sources of problems complicate the estimation of genome-wide copy number in cancer cells with SNP-array technology. The first concerns the determination of the reference point for copy-number variation (the level corresponding to the unaltered status of the tumor genome), which is not trivial for aneuploid cancer genomes with unknown underlying ploidies (diploid, tetraploid, and so on). Eventually, the reference point for a near-diploid cancer genome should correspond to normal genome status: a balanced genotype (AB status) and two copies. In the case of near-tetraploid tumors, a balanced genotype (AABB) and four copies could be proposed as the reference point. Setting the correct reference point thus depends on recognition of the underlying ploidy. This issue is considered in Attiyeh and colleagues [16], in which an aneuploidy correction factor was determined based on intensity-distribution modes in regions with balanced genotypes. Gardina and co-workers [15] directly estimated the chromosome copy-number status by using theoretic allelic ratios indicative of higher ploidy levels and then inferred tumor ploidy.

The second problem arises from the frequent contamination of cancer samples by normal stromal cells. The presence of a significant proportion of normal DNA in a sample diminishes the amplitude of measured signal changes reflecting rearrangements in the tumor DNA. Any fixed threshold-based method of copy-number variation recognition may fail to distinguish the proper regions. A number of publications have addressed this issue [17,18,24]. Staaf and colleagues [17] proposed a strategy for copy-number and LOH recognition based on adjusted thresholds, inferred from their study of dilution series. A model for estimation of normal DNA inclusion on the basis of measured allelic imbalances is considered in [18]. These authors also mentioned that, in addition to negative effects, a small degree of contamination could help in distinguishing somatically acquired homozygosity from germline homozygous regions.

The third problem in mining cancer SNP-array profiles is coming from intratumoral heterogeneity [25]. Although generally arising from a single cell (monoclonal proliferation), cancer progression leads to subpopulations bearing different genomic alterations (subclones) coexisting in most tumor samples. The tumor genomic profile is thus due to (a) genomic alterations shared by all tumor cells and producing few discrete steps of gains and losses, and (b) subclonal events shared by only certain subpopulations of tumor cells and producing a number of intermediate steps in the "main" copy-number profile. CNV and LOH status of an alteration specific for subclones is generally indefinable, as the measured signal reflects the sum of unknown subclonal signals in unknown proportions. An algorithm estimating the proportion of cancer cells harboring the particular alteration event was proposed in [18] and confirmed on known genetic events from a serial dilution of cancer cells with normal matched cells.

In this article, we propose a method for segmental copy-number and genotype detection from SNP arrays that takes advantage of previous findings and addresses the aforementioned issues. This method is based on SNP-array data formalization that we have called the Genome Alteration Print (GAP). The GAP of a tumor sample summarizes segmented CNV and allelic imbalance profiles into a list of segments, characterized by two corresponding averages. GAP visualization reveals the overall genomic ploidy of tumors, pinpoints the possible normal status (reference point for gain and loss), shows the level of contamination, indicates subclones, and generally characterizes the tumor genome. The *model* GAP built on theoretic distribution of CNV and allelic imbalances provides interpretation for a tumor GAP and serves as a basis for automatic recognition of the copy number and genotype of each segment.

Results and discussion

Generation of complex cancer genome data sets

The 300K Illumina SNP-arrays (Human Hap300-Duo) were used to study breast cancer genomes in a series of primary breast carcinomas (40 cases) and two cell lines. This series includes basal-like carcinomas (BLCs) arising in the general population (sporadic BLCs) and in *BRCA1* mutation carriers, who are especially predisposed to BLCs [26]. Both hereditary (in *BRCA1* carriers) and sporadic BLCs are associated with inactivation of *BRCA1* [27], a key protein for DNA repair [28]. Analysis of breast carcinomas by SNP-arrays is complicated by the numerous genomic rearrangements associated with these tumors [29], their high stromal cell content [30], and intratumoral heterogeneity [31].

Figure 1a and 1b shows the whole genome profiles of the BLC_B1_T45 sample measured on a 300K Illumina SNP-array. The copy-number variation (CNV) profile is represented by the Log R ratios (LRRs), which are the log-transformed ratios of experimental and normal reference SNP intensities, centered at zero for each sample. Allelic imbalances are represented by the B-allele frequencies (BAFs), which are the normalized proportions of the B alleles in two allele mixtures. Complexity of the profile is characterized by (a) the number of breakpoints in both profiles, and (b) the number of levels in smoothed LRRs and BAFs corresponding to the alteration states in the genomic DNA. The amplitude of both LRR and BAF changes depends on the purity of the tumor sample [17,18,24]. The main challenge is to interpret both segmental LRR and BAF values correctly in terms of *absolute* DNA copy number and LOH status, provided various amplitudes of changes, unknown underlying tumor genome ploidy, and disturbing subclonal intermediates. Specifically, DNA segments including at least 10 SNPs (~40 kb on average) were analyzed, which decreased resolution but minimized the effects of both experimental variations and short CNVs observed in population studies [32,33].

Genome alteration print (GAP)

The method for segmental copy-number and genotypes attribution presented here is based on the structure denoted by GAP. To build the GAP, breakpoints in LRR and BAF profiles are determined separately by the circular binary segmentation (CBS) algorithm (see Materials and methods for details) [12]. Any contiguous region in both LRR and BAF profiles

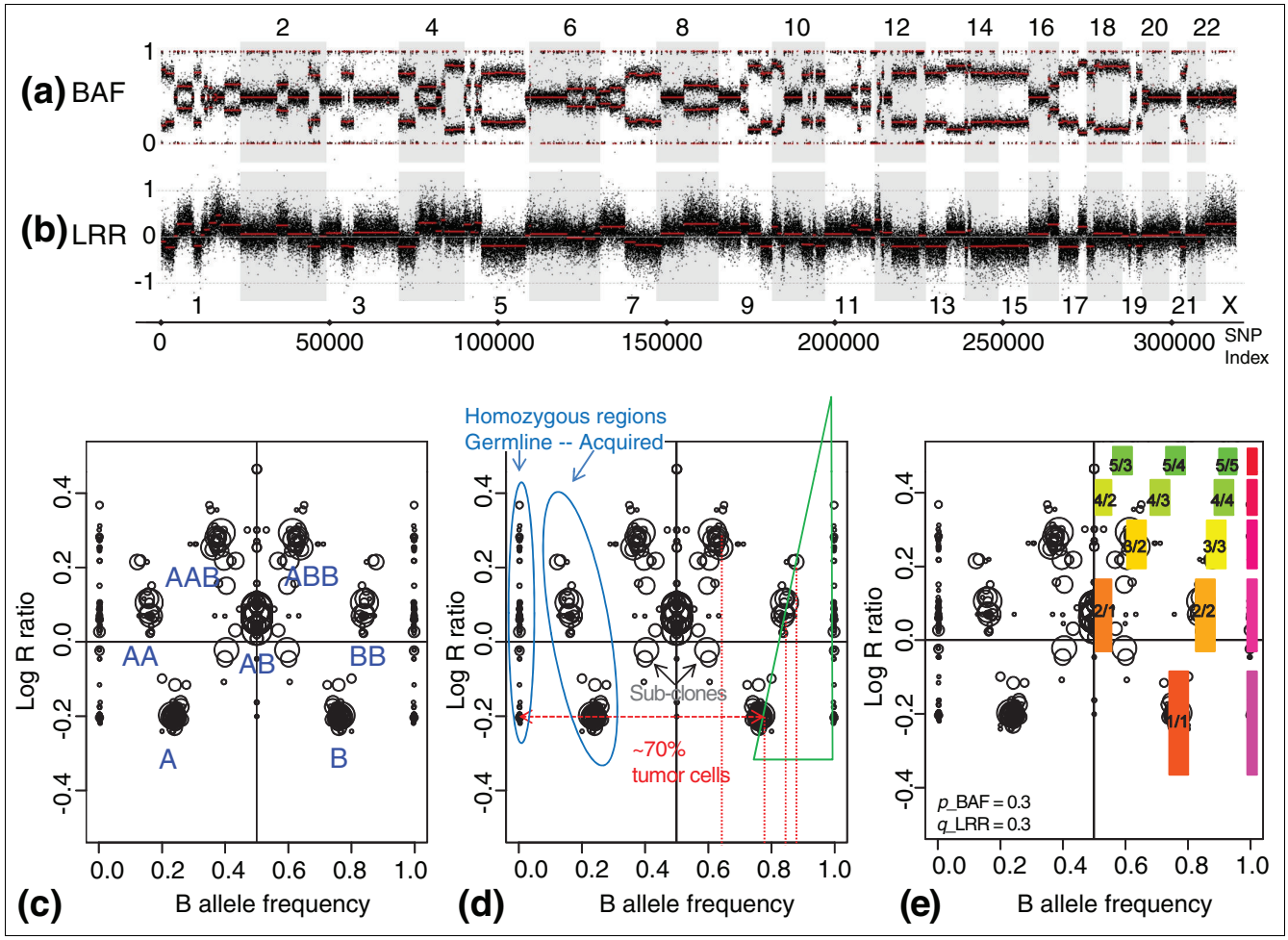


Figure 1

The whole-genome single-nucleotide polymorphism (SNP) array profile and genome alteration print (GAP). The whole-genome profile of genomic rearrangements in the BLC_B1_T45 sample measured by 300K Illumina SNP-array and corresponding GAP. (a) Allelic imbalances are represented by B-allele frequency (BAF). (b) Copy-number variation profile is represented by log R ratio (LRR), centered at zero. (c) The GAP of the sample is a combined sideview projection of segmented LRR and BAF. Each region of the genome is represented by two symmetric circles in the case of allelic imbalance and by one circle centered at BAF = 0.5 in the case of a balanced genotype. Attribution of copy numbers and genotypes corresponds to a near-diploid model of rearrangements. (d) "Reading" GAP pattern: the degree of stromal contamination, acquired and germline homozygosities, and subclones are indicated. (e) The best-fitting model GAP allows interpretation of the cluster structure and estimates contamination by normal DNA and contraction of the pattern on the LRR scale. Clusters are designated by the ratio of copy number to B (or major allele) counts.

(region between two consecutive breakpoints from LRR and BAF breakpoints mixture) is considered to be an alteration unit (possibly unaltered) and characterized by (a) the median of LRR, (b) the modes of BAF distribution, and (c) the length of the corresponding region (in SNP counts). The list and two-dimensional visualization of all alteration units of a measured sample is denoted the GAP.

The GAP of the BLC_B1_T45 sample is shown in Figure 1c. Each alteration unit is represented by a circle, with the center coordinates equal to its BAF (x-axis) and LRR (y-axis) smoothed values. The circle radius is scaled to the relative size of the corresponding chromosome region. In other words, the structure in Figure 1c represents a combined side-view projection of segmented and smoothed profiles of LRR and BAF shown earlier in Figure 1ab. The pattern in Figure 1c has a regular structure: circles corresponding to genomic regions with similar alteration status are assembled in clusters, forming discrete steps in their projection on the LRR scale, and symmetrically disposed on the BAF scale. As "A" and "B" allele names are set arbitrarily, the BAF profile is symmetric relative to 0.5 axis, and one alteration unit is represented by two symmetric circles away from the 0.5 axis on the BAF scale. Clusters centered at BAF = 0.5 present the genome regions with balanced (heterozygous) genotype; that is, an equal representation of both (maternal and paternal) alleles.

According to standard mining of SNP-array results, the GAP pattern shows (a) normal regions, which correspond to the balanced cluster; (b) losses, which are below the level of the balanced cluster; (c) gains, which are above this level; and (d) loss of heterozygosity without copy-number change (uniparental disomy), which are the side clusters of the reference balanced cluster (Figure 1d). The overall pattern of GAP corresponds to rearrangements in a near-diploid tumor.

The balanced cluster representing the normal status is generally not centered at zero on the LRR scale, which is set by normalization. For example, in Figure 1, the functional center (the diploid balanced cluster that represents unaltered regions) is shifted up from the formal center of the LRR profile (zero on LRR scale) because of the prevailing losses versus gains observed in the tumor.

Small germline homozygous regions, detected when more than 50 successive SNPs have a homozygous call (the 50-SNP length was set arbitrarily), form side clusters at the 0 and 1 boundaries of BAF scale. These germline homozygous regions can be easily distinguished from acquired LOH (see Figure 1d). Distances between germline and acquired homozygous clusters reflect the degree of tumor-sample purity [17]. Acquired and germline homozygosities cannot be distinguished in the case of pure tumor sample or (more often) cell line.

It is worth mentioning that (a) allelic imbalance is often treated as LOH, whereas here only single allelic genotypes (A, AA, AAA...) were considered to have an LOH status; (b) although mirrored BAF (see Materials and methods) is used for all computational evaluations, the GAP structure is shown in a complete (symmetric) view for easier association with the initial SNP-array measurement (with symmetric BAF bands).

Influence of tumor dilution and heterogeneity on GAP pattern

Breast carcinomas frequently show a high degree of stromal contamination and heterogeneity seen on the GAP pattern (Figure 1d). The triangle-like figure formed by homozygous clusters has the following interpretation. *P*% of normal DNA adds some proportion of normal (AA, AB, or BB) signal to any measured value. However, (a) this proportion depends on the corresponding copy-number status of a region and, (b) germline homozygous regions would show a pure homozygous signal, whereas cancer homozygous regions (LOH) would show a shift caused by normal heterozygous signal addition. Cancer BAF is modeled depending on the proportion of normal DNA inclusion (p) as the weighted sum of B-allele counts in cancer and normal genotypes related to maximal possible B allele counts at current copy-number level (see Materials and methods). For example, the calculated level of normal stromal DNA in the BLC_B1_T45 sample is approximately 30%. Such BAF dynamics also were illustrated by Nancarrow and colleagues [24] by using computer simulations. The clear linear relation between the measured mirrored BAF (mBAF) and the level of contamination by normal tissue of the tumor sample was demonstrated in [17] in dilution series.

The few isolated circles situated between one- and two-copy levels in Figure 1d could be attributed to losses occurring only in a fraction of the tumor cells (subclones). Following the logic of [18] and using our model of BAF, the proportion of cancer cells harboring this event is approximately 26%. More complicated subclonal mixtures could produce various intermediates in LRR and BAF scales.

The dynamics of change in LRR scale depend on numerous uncontrolled factors and show a high degree of variation from sample to sample. The significant dilution of a cancer sample by normal DNA clearly decreases the contrast (the amplitude of change in LRR corresponding to a copy-number change) [17]. However, universal linear dependence between LRR and contamination, similar to that for BAF, has not yet been described. The observed amplitude of LRR changes is usually smaller than expected by the initial model ($\log_2(CN/2)$), but the proportion between copy-number steps seems to be preserved for well-represented copy-number layers around the mean. LRR is therefore modeled by applying a simple coefficient of contraction q to the standard log ratio, which produces the sequence of LRR values: $-q$, 0, $0.58q$, q, $1.32q$, $2q$,... for corresponding copy-number levels: 1, 2, 3, 4, 5,...

Model GAP that follows theoretic values for BAF and LRR with estimated contamination $p = 0.3$ and coefficient of contraction $q = 0.3$ is superimposed onto the experimental GAP of BLC_B1_T45 sample in Figure 1e.

Diploid and tetraploid GAP patterns

The 40 SNP-array profiles of breast carcinomas and cell lines presented two main types of GAP pattern named "near-diploid" and "near-tetraploid" patterns (Figure 2a and 2b; for more examples, see Additional data file 1). The near-diploid pattern is characterized by a single balanced cluster with one layer of losses (Figures 1 and 2a). The typical near-tetraploid pattern shown in Figure 2b has (a) two balanced modes representing a balanced heterozygous genotype on two- and four-copy levels (AB and AABB); (b) three-copy level (between balanced modes) with the full spectrum of allelic imbalances, including LOH (AAA, AAB, ABB, BBB); (c) few levels higher than four copies accounting for possible five, six, seven... copies.

The near-diploid pattern has genomic DNA mainly presented in one, two, and three copies, whereas the near-tetraploid pattern has the well-represented two-, three-, four-, and five-copy layers. These patterns appeared to be easily distinguished in the case of the high density of alteration events observed in the current series of the breast carcinomas. A unique type of GAP pattern in the series was observed in BLC_T10 (Figure 2c). This pattern is characterized by sparse balanced cluster (due to a single chromosome with balanced genotype) and very strong homozygous clusters on the 3-copy level. This may be interpreted as an almost pure triplication of a haploid genome, possibly similar to the triploid glioblastoma cases described in [15].

DNA index and karyotype were used to verify correspondence between the interpretation of GAP pattern and the actual tumor genomic status. *In silico* DNA indexes inferred from SNP arrays were very close to actual tumor DNA indexes measured with flow cytometry (FCM) analysis for 16 of the 18 breast carcinoma samples tested (Table 1, Additional data file 1). The DNA index provided by FCM characterizes DNA content of tumor genome relative to normal diploid genome, which has a DNA index defined as 1. *In silico* DNA indexes were estimated by averaging segmental copy numbers (divided by 2), inferred from the GAP pattern. For 11 cases, the difference between actual and *in silico* DNA index was less than 0.1; for five cases, it was less than 0.3. With the exception of two outliers, this difference was always less than 0.5, which is the minimal absolute error in the case of wrong assignment of the overall copy-number scale (pattern shift on +1 or -1 copy). For the two outliers (BLC_B1_T22 and BLC_T34), GAP patterns were perfectly near-diploid with a clear contrast, making cluster misattribution unlikely. The discrepancy in DNA index estimation requires further biologic verification (for example, in the case of BLC_B1_T22, there might be a pure and possibly recent duplication of the diploid tumor cells as the *in silico* DNA index was equal to half of the experimental index).

Breast cancer cell lines with known karyotypes were used for another validation of GAP interpretation. The tetraploid breast cancer cell line MDA-MB-175-VII (MDA_175; [34]) has a clear near-tetraploid pattern of GAP (Figure 3a). The

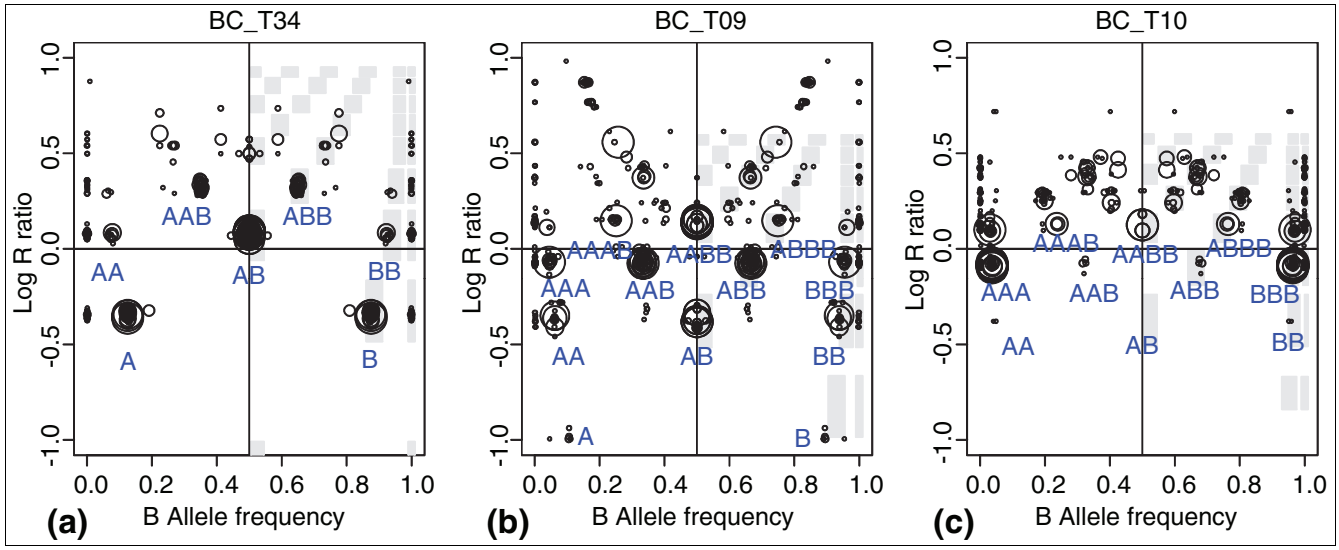


Figure 2
Characteristic genome alteration print (GAP) patterns. Two characteristic **(a, b)** and one unique **(c)** GAP patterns obtained in the analysis of a breast carcinoma series: **(a)** near-diploid pattern, sample BLC_ T34; **(b)** near-tetraploid pattern, sample BLC_T09; and **(c)** possible near-triploid pattern, sample BLC_T10. Attribution of genotypes is based on the type of pattern; best-fitting models are shown.

Table 1

Experimental and *in silico* DNA indexes and parameters of GAP model

Sample ID	DNA index FCM	DNA index GAP	DNA index OverUnder	Tumor content 1-p_{BAF}	Contraction q_{LRR}
BLC_B1_T14	1.14	0.85	0.98	0.85	0.37
BLC_B1_T17	0.84	0.82	0.97	0.77	0.17
BLC_B1_T19	1.6	1.63	2.93	0.4	0.27
BLC_B1_T20	1.41	1.48	3.06	0.4	0.2
BLC_B1_T22[a]	1.98	0.94	1.02	0.87	0.44
BLC_T07	1.68	1.49	3.12	0.44	0.28
BLC_T09	2.02	1.85	1.89	0.92	0.47
BLC_T10	1.88	1.9	1.07	0.95	0.47
BLC_T12	1.51	1.54	2.56	0.65	0.35
BLC_T15	1.11	0.89	0.99	0.74	0.27
BLC_T23	1.32	1.39	2.72	0.41	0.21
BLC_T31	1.91	1.84	1.48	0.84	0.45
BLC_T34[a]	1.55	0.99	1.04	0.87	0.42
BLC_T37	1.51	1.53	1.44	0.89	0.44
L_B1_T24B	1.84	1.64	2.61	0.59	0.29
L_B1_T25A	1.00	1.04	3.03	0.39	0.17
L_B1_T30	1.84	1.83	1.53	0.78	0.42
L_B1_T47	1.00	1.03	1.47	0.45	0.17

[a]Two samples with clear near-diploid pattern of GAP and discordant experimental DNA indexes.

unique balanced cluster must be attributed to a four-copy level because two levels of losses visible below it could not account for 1 and 0 copies, but rather for 3 and 2 copies because of their positions and the absence of a normal contingent in the cell line. Circles on each side of the balanced cluster fit with AAAA and AAAB, and ABBB and BBBB genotypes, respectively, also implying a two-copy level. It is noteworthy that two-copy regions are represented exclusively by homozygous genotypes.

A near-tetraploid genome implies the number of chromosomes to be close to 92 (88 autosomes = two sets of diploid genomes). Copy-number summary for centromeric regions was considered a surrogate measure of chromosome number. As no SNP measurements can be performed at centromeres because of their highly repetitive DNA structure, pericentric regions were used to estimate the copy-number status of the chromosomes. The status of 39 pericentric regions (two for each of the 17 metacentric autosomes and one for each of the five acrocentric autosomes) was determined according to GAP. The number of autosomes in MDA_175 was estimated to be 86.5, which is close to the description in [34] (model number was 84 chromosomes; range, 82 to 89; verified on the cell line used for the SNP-array). Table 2 shows the frequency of occurrence of the inferred copy number of pericentric regions (also for other tumor samples considered in this study, with more-detailed information presented in Additional data file 2). A similar analysis was performed with the MDA-MB-468 (MDA_468) cell line; this hypotetraploid breast cancer cell line (modal number, 64; range, 60 to 67) [34] showed a typical tetraploid GAP pattern (Figure 3b). Estimated autosome number (71.5) matched the description, and the slight overestimation was likely due to segmental amplification in one pericentric region (Table 2 and Additional data file 2). Taken together, these results indicate correct local assignment with our approach.

It should be noted that determination of the reference point for gain and loss attribution for complex highly rearranged cancer genomes is not always obvious, even with known patterns of rearrangements and absolute copy numbers. Samples displaying a near-diploid GAP pattern (as in Figure 2a) represent a simple situation, as their unique balanced cluster corresponding to 2-copy indicates the reference point. Near-tetraploid patterns with a unique balanced cluster at four copies (such as that of the cell line in Figure 3a) and inferred autosome numbers close to 88 indicate underlying tetraploidy, and it is logical to set the reference point to four copies in these cases. Underlying ploidy is less clear for intermediate DNA index or autosome number (between one and two, or 44 and 88, respectively), and the GAP shows a tetraploid pattern with two balanced clusters (as for BLC_B1_T19 and BLC_B1_T20 samples). Correct interpretation of gains and losses in such cases requires further biologic validation.

Automatic recognition of segmental copy numbers and genotypes

The GAP pattern can be easily mined by automatic procedures. This procedure includes (a) recognition of a GAP pattern and (b) assignment of segmental copy numbers and

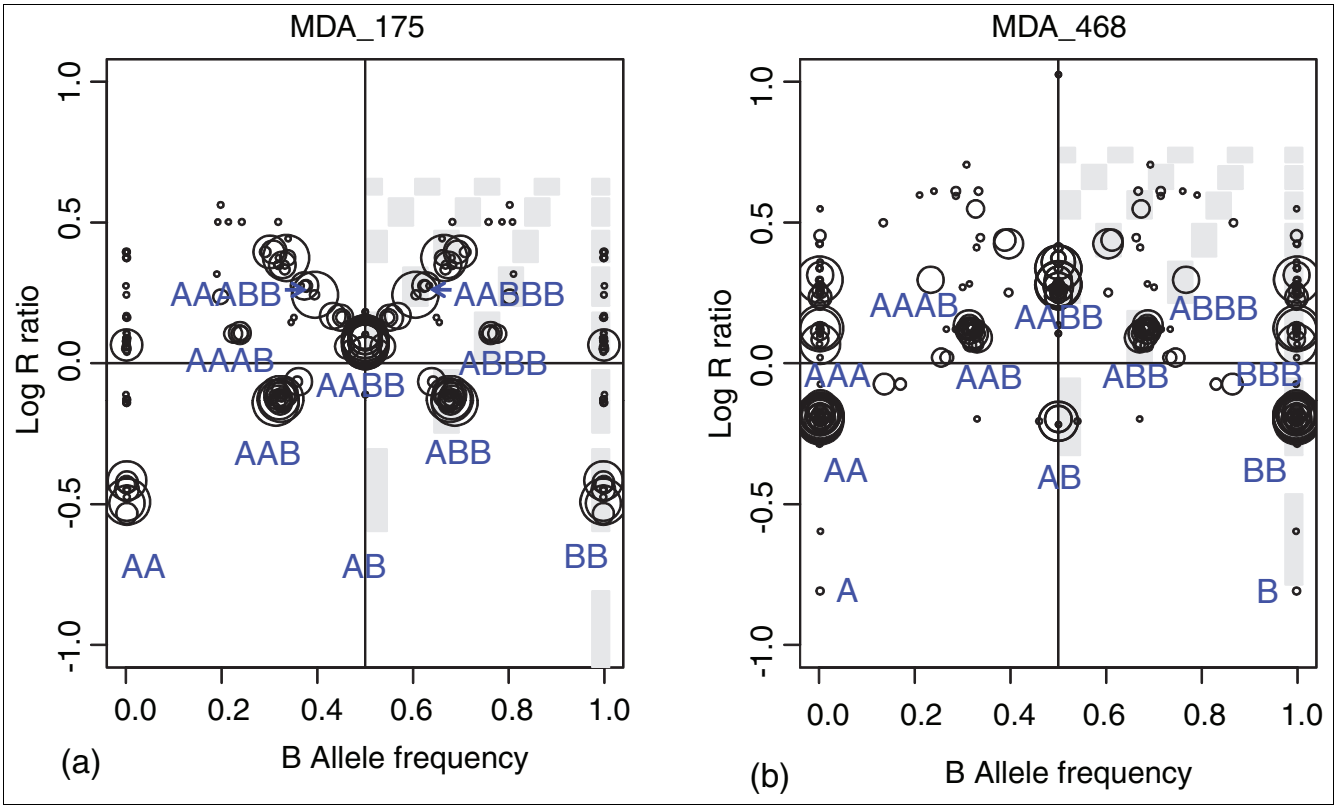


Figure 3
Genome alteration prints (GAPs) for breast cancer cell lines. GAPs for breast cancer cell lines: **(a)** MDA_175; and **(b)** MDA_468. Both GAPs show a near-tetraploid pattern, and genotypes were assigned accordingly.

genotypes to a corresponding tumor genome based on this pattern. As described earlier, the GAP is characterized by two parameters: *p*, which is the proportion of tumor contamination by normal DNA affecting BAF values, and *q*, which is a coefficient of contraction of LRR values. The automatic recognition procedure searches for parameters and position of a model GAP that best fits to the experimental GAP. Quality of fitness is assessed by genome coverage in terms of number of SNPs that are explained by the model (see Material and methods for details). In other words, the model GAP template that most closely corresponds to the experimental GAP is selected. In the second round, the model GAP is used as the basis for interpretation of the experimental GAP, and segmental copy numbers and genotypes are assigned accordingly.

The quality of pattern recognition was tested on 42 in-house samples, including the samples validated by DNA index. The procedure performed 41 correct and one erroneous recognitions, as compared with manual assessment. The problematic sample presented a high variance and low contrast, and the correct solution had a high but not the highest score. In general, the method tolerates contamination of tumor samples by normal DNA and experimental variations, as shown by correct recognition of our validated series with up to 60% of normal contamination and up to 0.17 contraction of LRR scale (see Table 1).

We considered subclones as segments located essentially between designated clusters. They could be artefacts from incorrect segmentation, or true tumor heterogeneity. An interesting case is represented by sample BLC_T31. Its first interpretation was that of a near-tetraploid pattern, but its second interpretation with a very similar score was that of a near-diploid pattern because of poor representation of the three-copy level interpreted as subclones in the latter case. The DNA index determined by FCM indicated near-tetraploidy, supporting the first interpretation (see Additional data file 1).

It should be stressed that (a) correct recognition requires good contrast between clusters and multiplicity of genetic events (for example, patterns consisting of AB and A∅ genotypes versus AABB and AA cannot be distinguished when no other evidence of a four-copy pattern exists); (b) the robustness of the quality criterion used in our method is not always satisfactory: the correct solution often differs from incorrect solutions by less than 1%; (c) the linear models used in the method diverge from experimental data in both the LRR and BAF scales when copy numbers were higher than 6-copy.

Table 2

Frequency of inferred copy numbers at pericentric regions and deduced autosome numbers

	Copy number[a]									
Sample ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Autosome number**	**Pattern[b]**
MDA_175		5	9	16	4	4	1		86.5	2
MDA_468[c]		17	8	8	3	2		1	71.5	2
BLC_B1_T14	12	25	2						38	1
BLC_B1_T17	14	21	3	1					38.5	1
BLC_B1_T19		7	10	16	4	2			76.5	2
BLC_B1_T20	1	16	9	8	3	1	1		66.5	2
BLC_B1_T22	12	24	3						37.5	1
BLC_T07		15	13	8		1	1	1	70	2
BLC_T09		3	16	13		3	2	2	85.5	2
BLC_T10			21	9	2	3	3	1	87	1.5
BLC_T12		11	10	12	4	1		1	73.5	2
BLC_T15	10	27	2						40	1
BLC_T23	3	13	11	6	2			4	68.5	2
BLC_T31		5	4	25	1	2	1	1	84.5	2
BLC_T34	3	36							42	1
BLC_T37	1	16	9	6	1	4		2	70.5	2
L_B1_T24B	1	11	12	7	6	2			72	2
L_B1_T25A		37		2					46	1
L_B1_T30		3	11	23	1	1			79	2
L_B1_T47	1	34	2	1	1				47	1

[a]Frequency of inferred copy numbers (1 to 8 are indicated) at pericentric regions. [b]1, 2, 1.5 indicates a near-diploid, near-tetraploid, and near-triploid patterns of Genome Alteration Print (GAP), respectively (Figure 2, Additional data file 1). [c]Estimated high chromosome copy number (= 8) is likely to result from a segmental amplification in one pericentric region, leading to overestimation of the autosome number in MDA_468.

However, no universal rule to correct this effect was identified on the basis of the 41 tumors examined.

Comparative testing of GAP recognition

When characterizing rearrangements in tumor genome measured by SNP array, it is essential to extract from data (a) the degree of genomic instability displayed by the number and distribution of breakpoints, and (b) the type of each alteration. The GAP method is based on both LRR and BAF breakpoints and is therefore not directly suitable for breakpoint counting. To minimize double counting of a single breakpoint, LRR and BAF breakpoints separated by a small region (arbitrary defined as 10 SNPs) were simply merged. More complicated pooling of LRR and BAF breakpoints could allow more accurate breakpoint counting, and this would not be expected to influence the performance of the GAP method. Another way to address breakpoint detection in highly rearranged cancer genomes with possible low tumor content and noisy profiles is to use the GAP pattern as a source for secondary optimization.

The GAP method is elaborated for determination of alteration events in *complex*, highly rearranged cancer genomes (in contrast, it would be of little help for interpretation of a stable genome with few amplifications). The methods specifically developed for analysis of cancer genomes include SOMATICS [18] and BAFsegmentation [17], which reveal segments with allelic imbalances based on various models but do not produce copy numbers and genotypes. The OverUnder algorithm presented by Attiyeh and associates [16] estimates ploidy, as well as copy numbers and genotypes, and has been shown to outperform PennCNV, IlluminaCN Estimate, and CBS for the analysis of cancer genomes. We compared our automatic GAP fitting method with the OverUnder algorithm in terms of quality and consistency of recognition.

The OverUnder algorithm (available as Illumina Beadstudio plug-in) was initially applied to our validated series of breast carcinomas to estimate the DNA indexes (Table 1). OverUnder results for seven samples clearly deviate from experimental data (Figure 4). These samples are characterized by high levels of normal DNA contamination, as estimated by the GAP model. The GAP method tolerated normal contamination, demonstrating better overall performance.

The self-consistency of the methods was tested on the basis of dilution series available in the GEO database (GEO:GSE11976) [17]. The HCC1395/CRL2324 cell line [34] measured in this series is genetically complex and poorly defined. However, estimated copy numbers and LOH regions

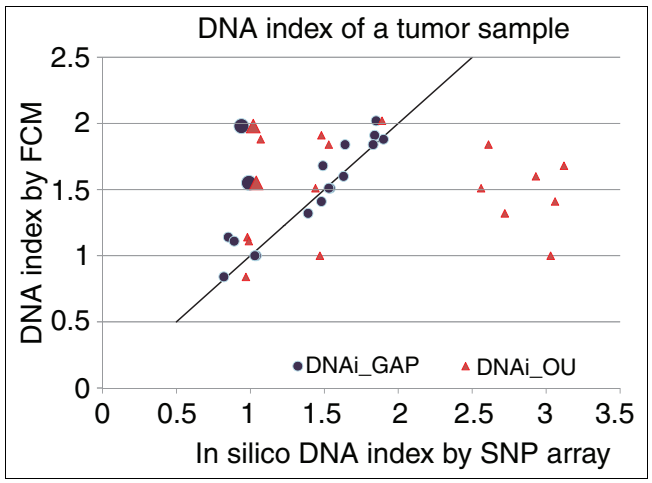


Figure 4
Comparison of genome alteration print (GAP) and OverUnder-based *in silico* DNA indexes with experimental DNA indexes. GAP indexes (blue circles) show excellent correspondence with experimental DNA indexes. OverUnder indexes (red triangles) show more outliers with overestimation of the DNA index. Both methods show consistent results, but not corresponding to the experimental DNA indexes (1.98 and 1.5) for two samples, designated by enlarged markers.

must be consistent for all CRL2324 samples with various proportions of tumor DNA. The results of the self-consistency test are presented in Table 3 (more details in Additional data file 3). The better self-consistency of the GAP method is obvious in terms of copy numbers and LOH. Structural reproducibility of tumor GAP pattern with various proportions of normal DNA is illustrated in Additional data file 4.

GAP for Affymetrix SNP platform

Affymetrix GeneChip SNP 6.0 array was used to generate SNP profiles of the BLC_B1_T45 sample. The GAP was obtained according to the same strategy as for Illumina SNP data but by using the profile-recognition method described in [14].

Comparison of the data generated on these two platforms is shown in Figure 5. Affymetrix SNP measurements are represented by Log Copy Number Ratio and Allelic Differences as compared with Illumina LRR and BAF, respectively. Germline homozygous SNPs were omitted if fewer than 50 in a row, and are therefore represented by small clusters along two parallel lines at 0 and 1 limits of the BAF scale in an Illumina plot. Homozygous SNPs were always included in Affymetrix GAP and therefore formed large clusters represented along divergent diagonal lines (as allelic differences are dependent on copy-number levels) in the Affymetrix plot. Genome regions localized and attributed to a specific copy number in an Illu-

Table 3

Self-consistency of copy numbers and LOH in dilution series by using GAP and OverUnder analyses

GAP	CN	LOH			DNA index	Tumor DNA	1-*p* BAF	*q* LRR
CRL2324	1	1			1.45	1	1	0.45
CRL2324_79	0.93	0.98			1.46	0.79	0.8	0.35
CRL2324_50	0.9	0.97			1.44	0.5	0.42	0.22
CRL2324_47	0.78	0.96			1.49	0.47	0.42	0.24
CRL2324_45	0.81	0.96			1.5	0.45	0.35	0.18
CRL2324_34	0.69	0.93			1.53	0.34	0.27	0.16
CRL2324_30	0.73	0.93			1.52	0.3	0.25	0.12
CRL2324_23	0.72	0.93			1.59	0.23	0.26	0.14
CRL2324_21	0.7	0.92			1.64	0.21	0.14	0.12
OverUnder	CN	LOH	CN ± 1	CN CBS	DNA index	Tumor DNA		
CRL2324	1	1	1	1	2.48	1		
CRL2324_79	0.36	0.94	0.74	0.46	2.16	0.79		
CRL2324_50	0.21	0.45	0.65	0.1	2.64	0.5		
CRL2324_47	0.22	0.45	0.68	0.1	2.5	0.47		
CRL2324_45	0.34	0.45	0.71	0.11	2.85	0.45		
CRL2324_34	0.24	0.44	0.56	0.19	2.57	0.34		
CRL2324_30	0.14	0.45	0.48	0.23	2.54	0.3		
CRL2324_23	0.31	0.45	0.68	0.23	2.51	0.23		
CRL2324_21	0.07	0.47	0.12	0.05	1.11	0.21		

CN = copy number; LOH = loss of heterozygosity; tumor DNA = proportion of tumor DNA in the dilution; DNA index = *in silico* DNA index with each algorithm; 1-*p* BAF and *q* LRR are parameters of the model GAP; CN ± 1, copy numbers are considered to be consistent when the difference is less than or equal to 1; CN CBS, consistency is calculated on averaged (by median) and rounded copy-number assignments in CBS determined segments.

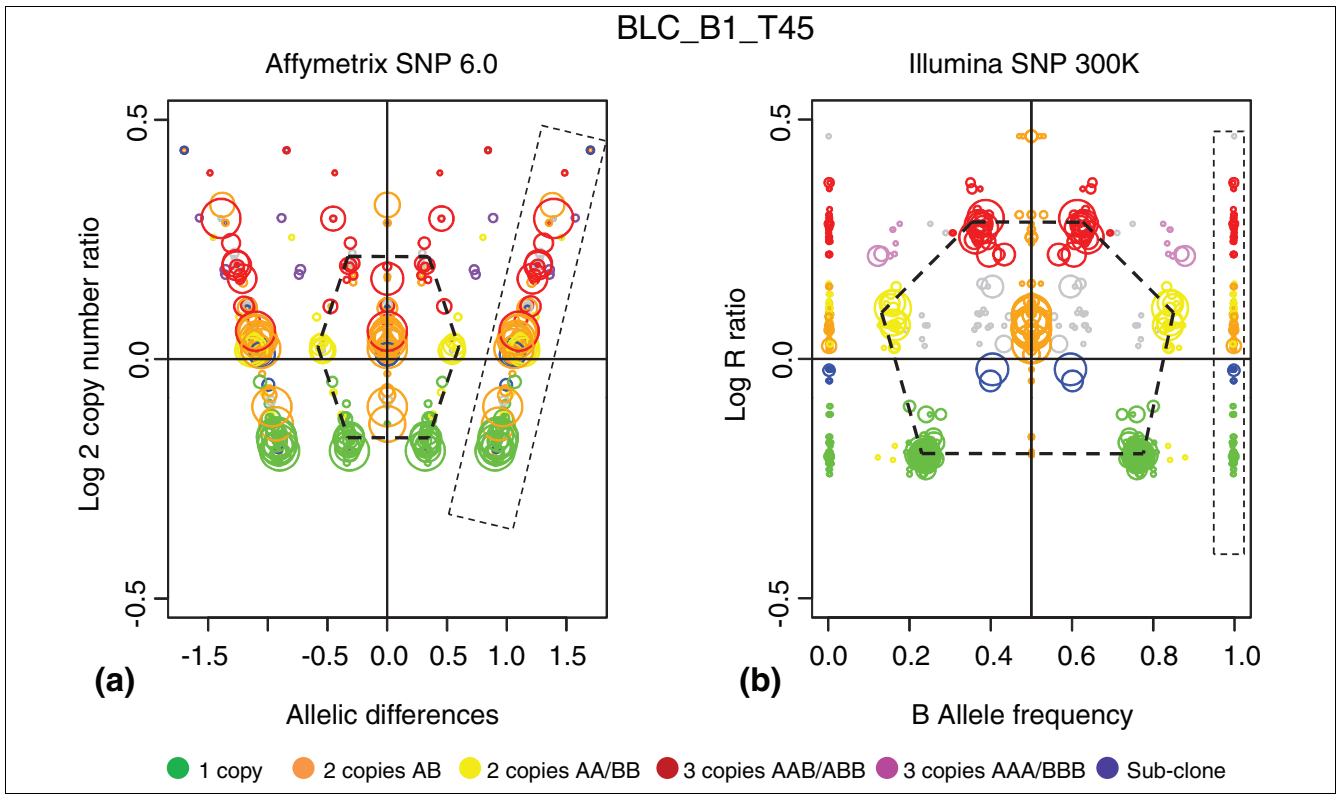


Figure 5
Genome alteration print (GAP) for Affymetrix single-nucleotide polymorphism (SNP) GeneChip SNP 6.0 array. BLC_B1_T45 tumor sample measured on two SNP-array platforms, analyzed by using GAP, and superimposed by color code: **(a)** GAP for Affymetrix; and **(b)** GAP for Illumina. Copy numbers obtained from the Illumina GAP were coded by colors indicated at the bottom of the Figure. Concordance between Affymetrix and Illumina patterns is illustrated by the relevant Illumina-derived color gradation on Affymetrix GAP. Germline homozygous regions are boxed. The main cluster patterns are indicated by hexagonal frames. The differences in relative cluster sizes are due to different distributions of SNPs measured along the genome in Illumina and Affymetrix chips.

mina-profiled genome were used to color code the regions in the Affymetrix SNP profile. Excellent concordance was observed between Affymetrix and Illumina patterns, as shown by relevant Illumina-derived color gradation on Affymetrix GAP. Visible differences in relative cluster sizes are due to different distributions of measured SNPs along the genome in Illumina and Affymetrix chips. The main conclusions from this comparison are (a) excellent correspondence between the two technologies in terms of copy-number variation; and (b) GAP can be used for analysis of complex cancer genomes on Affymetrix platforms.

Conclusions

We present a method to mine complex genome alteration profiles measured with SNP-arrays. We introduce genome alteration print (GAP), a combined side-view projection of LRR and BAF segmented and smoothed profiles. The method, based on GAP pattern recognition, is fully automatic and provides segmental copy numbers and genotypes. It also estimates tumor-sample contamination by normal DNA. The method performs well, even for poor-quality data, low tumor content, and highly rearranged tumor genomes. Visualization of the GAP recognition pattern characterizes overall rearrangements in a tumor sample and can be used to verify the results. The GAP method is designed for Illumina SNP-array, but can be easily applied to Affymetrix SNP-arrays. This method could be a valuable tool to identify recurrent alterations in complex tumor-genome profiles.

Materials and methods

Illumina arrays

A series of 40 breast carcinomas, including cases described in [35], was analyzed, as well as the breast cancer cell lines MDA-MB-175-VII (MDA_175) and MDA-MB-468 (MDA_468) [34]. DNA was extracted from samples, and genomic profiling of the tumor samples was performed at Integragen [36] on 300K Illumina SNP-arrays (Human Hap300-Duo). SNP-array data are available through Gene Expression Omnibus [37] [GEO:GSE18799].

Data processing

Normalization of raw data was performed with Illumina Beadstudio software version 3.3 by using standard settings (all supporting files are provided by Illumina [7]). The normalization procedure tQN proposed in [9] also was used to make BAF symmetric.

LRR and BAF segmentation and construction of the GAP

The circular binary segmentation (CBS) algorithm (DNAcopy package, Bioconductor) [12,38] was applied to LRR and filtered BAF data separately to define breakpoints (the minimal level of significance was defined as 10^{-2} for LRR and 10^{-3} for BAF profiles). Smoothing of outliers was performed in both cases. LRR was smoothed by the median between breakpoints. To obtain one banded BAF profile, (a) non-informative homozygous SNPs were filtered out, based on the threshold (mBAF > 0.97), as suggested in [17]; (b) tQN normalized and reflected relative to the 0.5 axis version of BAF, named mirrored BAF (mBAF) [17], was segmented. In addition, the boundaries of germline homozygous regions, detected when more than 50 successive SNPs had a homozygous call (the number of SNPs was set arbitrarily), were included into the set of breakpoints. The mode estimation (dip-test package, [38]) was used for smoothing of the mBAF profile to maintain the contrast between balanced and slightly shifted imbalances.

Any region between two consecutive breakpoints from the LRR and BAF breakpoint mixture was considered to be an alteration unit (possibly unaltered) and characterized by (a) the averaged LRR, (b) the mode of mBAF distribution, and (c) the length of the corresponding region (in SNP counts). The list and the two-dimensional visualization of all alteration units of a measured sample were denoted the genome alteration print (GAP). For GAP visualization, each alteration unit was represented by a circle centered on BAF (x-axis) and LRR (y-axis) smoothed values, and the radius was scaled to the relative size of the corresponding chromosome region.

Comments on stability of GAP

The CBS algorithm was used to favor sensitivity over specificity in the breakpoint-detection process, as "false" breakpoints do not significantly change the overall GAP pattern. False alteration units often appeared as artefacts at joining LRR and BAF breakpoints, but were not visible, provided the true alteration units were significantly longer. More problems were observed when the robust profile estimators were applied to poor-quality data: the absence of true breakpoints could significantly alter the GAP pattern.

Model GAP

The model GAP was determined by the independent combination of BAF and LRR models. The BAF model was used to determine the position of clusters on the horizontal scale, and the LRR model was used to determine the *relative* position of clusters on the vertical scale.

The cancer BAF was modeled as the weighted sum of B-allele counts in cancer and normal genotypes, as a ratio of the maximal possible B allele counts at the current copy-number level:

$$BAF^M = \frac{(1-p)\cdot n_B^c + p n_B^n}{(1-p)\cdot(n_B^c + n_A^c) + 2p}$$

where p is normal DNA proportion (and hence $(1 - p)$ is tumor DNA proportion); n_B^c and n_A^c are the B and A allele counts in the tumor genotype; $(n_B^c + n_A^c)$ is considered to be the copy-number level; and n_B^n is the B allele count in normal genome (n_B^n = 0, 1, 2). A similar model was described in [17,24]; a model proposed in [18] could also be used for GAP-method settings.

To estimate normal DNA contamination in a measured tumor sample, at least one cluster annotation (copy number and genotype) and its position on the BAF scale must be known. For example, projections of cluster centers in the experimental GAP pattern (Figure 1e) were assessed to be as follows: BAF^M = 0.765 for the B cluster, BAF^M = 0.845 for the BB cluster, BAF^M = 0.885 for the BBB cluster, and BAF^M = 0.628 for the ABB cluster. Substitution of BAF^M, B allele counts, and copy numbers in the model provides an estimation of the contamination coefficient p = 0.307, 0.31, 0.309, and 0.312, respectively. As expected, inferred coefficients were very close to each other and estimated the normal DNA contamination around 30% for this sample.

The same method was used to estimate the proportion of tumor cells bearing a given rearrangement (subclone); in the case shown in Figure 1d: BAF^M = 0.575, $n_B^c = 1$, $n_B^c + n_A^c = 1$, $n_B^n = 1$ gave the normal content estimation $p \approx 0.74$ and hence the tumor content was $1 - p \approx 0.26$.

LRR was modeled by applying a simple coefficient of contraction q to the standard log ratio: $\mathrm{LRR_n} = q \cdot \log_2(\frac{n}{2})$; n is the copy number, which produces the sequence of LRR values: $-q$, 0, $0.58q$, q, $1.32q$, $2q$,... for corresponding copy number levels: 1, 2, 3, 4, 5... The LRR of zero copy (homozygous deletion) was arbitrarily set at $-3q$ ($\log_2 0 = -\infty$, variation in real LRR is usually very large and not followed by the model).

Fitting model GAP and copy number and genotype recognition

Automatic recognition of the tumor GAP pattern consisted of an exhaustive search for (a) the best centering of the model GAP on the LRR scale for each pair of contamination proportion (p) and coefficient of contraction (q), and (b) the best (p, q) couple satisfying a few necessary conditions. The genome coverage in terms of the number of SNPs explained by the model was used as the quality criterion. The necessary conditions were used to filter unusual interpretations.

GAP pattern-recognition algorithm: 1) Initiation of a grid with 0.005 cell dimension on the BAF × LRR plane and definition of the densities of alteration units in SNP counts on the grid; 2) Smoothing of the densities by averaging adjacent cells and filtering of low densities to enhance the contrast (densities were set to 0 in 95 to 98% of cells in the grid); 3) Choosing model parameters (p, $p \in \{0, 0.02, 0.04, \ldots, 0.86\}$; q, $q \in \left\{\frac{2}{32}, \frac{3}{32}, \ldots 1\right\}$) and setting of the GAP template with one to five copies by determining the centers and sizes of model clusters on the grid; 4) For a given pair (p, q), searching for the best centering of the GAP template on the grid in terms of maximal density falling into designated clusters; 5) Checking all possible combinations of p and q and ranking templates; 6) Filtering of templates according to necessary conditions. This removes from further consideration redundant interpretations with many empty clusters; 7) Choosing the best interpretation, superimposing the model GAP to the experimental one, and ascribing copy number and genotype to each alteration unit after the annotation of its closest cluster on the template.

In the case of low contrast between clusters, additional adjustments of recognition are necessary to attribute correctly the alteration units located between designated clusters. A confidence score is attributed to all alteration units (depending on the distance to the nearest model cluster(s)), and the *linear* copy number and genotype profiles are adjusted by keeping confident assignments and correcting less-confident assignments.

Experimental estimation of ploidy and karyotyping

The DNA content of tumor samples was obtained with flow cytometry (FCM) analysis after propidium iodine staining, as described in [39]. The DNA index is equal to 1 for normal diploid cells. A karyotype of MDA_175 was obtained by a routine procedure [40].

Estimation of DNA content and chromosome number based on SNP data

The inferred copy-number profile was averaged along the genome, providing the DNA content of the corresponding cancer sample.

Chromosome copy numbers were characterized by the status of pericentric regions, defined as the alteration units directly before or after the centromeric part of the chromosome (which has no SNP measurement *per se*). The definition of pericentric region therefore depends on the SNP chip used for genotyping. Regions less than 10 SNPs were ignored. If the pericentric alteration unit is a small region (less than 100 SNPs), setting the chromosome copy number on the basis of this alteration unit could be erroneous and could therefore interfere with karyotype assessment.

Dilution series

The dilution series described in [17], measured by Illumina 370 K array and available in the GEO database [37] [GEO:GSE11976], were processed in a similar way to in-house tumor samples: (a) normalization by the method proposed in [9]; (b) segmentation of LRR and BAF profiles by CBS, in the same way as described in subsection 3 of the Materials and Methods; (c) construction of GAP, GAP pattern recognition, and copy-number assignment.

Results of the OverUnder [16] algorithm

The OverUnder plug-in was applied to the data normalized in BeadStudio 3.3, with window length equal to 51. OverUnder produces continuous copy-number values, which were rounded to discrete values and then summarized in comparison tables. As rounding can introduce artificial discrepancies, the procedure was slightly modified so that copy numbers were considered to be equal when they differed by no more than 1 unit (column CN ± 1, Table 3). CBS sections also were used to average (by median) and to round copy-number assignments (column CN CBS, Table 3).

Affymetrix SNP data

One BLC sample also was analyzed on the Genome-Wide Human SNP-array 6.0, according to the manufacturer's instructions (Affymetrix Inc., Santa Clara, CA). Normalization was performed by using the Genotyping Console™ (Affymetrix), and profile recognition was performed by using the method described in [14].

Availability

An implementation of the proposed GAP pattern recognition and detection of copy numbers and genotypes based on segmented profiles is available, together with the supporting data [41]. SNP array data for the 19 primary tumors and the two cell lines shown here are available through Gene Expression Omnibus [37] [GEO:GSE18799].

Abbreviations

BAF: B allele frequency; BLC: basal-like breast carcinoma; CBS: circular binary segmentation; CN: copy number; CNV: copy-number variation; FCM: flow cytometry; LOH: loss of heterozygosity; LRR: Log R ratios; mBAF: mirrored BAF;

MDA_175: MDA-MB-175-VII breast cancer cell line; MDA_468: MDA-MB-468 breast cancer cell line; SNP: single-nucleotide polymorphism.

Additional data files

The following additional data are included with the online version of this article.

A table of images of GAP patterns and copy-number recognition templates for a series of breast carcinomas with available DNA indexes (Additional data file 1), a table listing copy-number status of pericentric regions inferred on the basis of GAP pattern for a series of breast carcinomas and cell lines (Additional data file 2), two tables indicating self-consistency in copy-number attribution in dilution series calculated for two methods of recognition: GAP method and OverUnder algorithm (Additional data file 3), and GAP patterns and copy-number recognition templates for the dilution series of cell line CRL2324 (Additional data file 4).

Authors' contributions

TP performed research, analyzed data, and wrote the paper; EM performed experiments and analyzed data; DSL designed the general project; GR provided the Affymetrix data and analyzed data; EB supervised bioinformatics analyses; and MHS designed the research and wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We thank X. Sastre-Garau, N. Gruel, and I. Lebigot for providing breast cancer DNAs, T. Dubois for giving us access to data, J. Couturier for karyotype data, A. Vincent-Salomon for tumor characterization, A. Zinovyev for helpful discussion and support, and C. Lucchesi and S. Pook for critical reading. This work is part of the "Cancéropôle Ile-de-France-Région Ile-de-France-Hereditary Breast Cancer" program coordinated by DSL and MHS. This work also was supported by the INSERM, the Institut Curie, and its Translational Research Department. TP is supported by a grant from the Cancéropôle Ile-de-France-Région Ile-de-France. GR is supported by a grant from the Institut National du Cancer (INCa).

References

1. Albertson DG, Collins C, McCormick F, Gray JW: **Chromosome aberrations in solid tumors.** *Nat Genet* 2003, **34:**369-376.
2. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM, Esserman L, Albertson DG, Waldman FM, Gray JW: **Genomic and transcriptional aberrations linked to breast cancer pathophysiologies.** *Cancer Cell* 2006, **10:**529-541.
3. Engle LJ, Simpson CL, Landers JE: **Using high-throughput SNP technologies to study cancer.** *Oncogene* 2006, **25:**1594-1601.
4. Bacolod MD, Schemmann GS, Giardina SF, Paty P, Notterman DA, Barany F: **Emerging paradigms in cancer genetics: some important findings from high-density single nucleotide polymorphism array studies.** *Cancer Res* 2009, **69:**723-727.
5. Dutt A, Beroukhim R: **Single nucleotide polymorphism array analysis of cancer.** *Curr Opin Oncol* 2007, **19:**43-49.
6. **Affymetrix** [http://www.affymetrix.com]
7. **Illumina** [http://www.illumina.com]
8. Oosting J, Lips EH, van Eijk R, Eilers PH, Szuhai K, Wijmenga C, Morreau H, van Wezel T: **High-resolution copy number analysis of paraffin-embedded archival tissue using SNP BeadArrays.** *Genome Res* 2007, **17:**368-376.
9. Staaf J, Vallon-Christersson J, Lindgren D, Juliusson G, Rosenquist R, Hoglund M, Borg A, Ringner M: **Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios.** *BMC Bioinformatics* 2008, **9:**409.
10. Rigaill G, Hupe P, Almeida A, La Rosa P, Meyniel JP, Decraene C, Barillot E: **ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays.** *Bioinformatics* 2008, **24:**768-774.
11. Wiel MA van de, Brosens R, Eilers PH, Kumps C, Meijer GA, Menten B, Sistermans E, Speleman F, Timmerman ME, Ylstra B: **Smoothing waves in array CGH tumor profiles.** *Bioinformatics* 2009, **25:**1099-1104.
12. Venkatraman ES, Olshen AB: **A faster circular binary segmentation algorithm for the analysis of array CGH data.** *Bioinformatics* 2007, **23:**657-663.
13. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E: **Analysis of array CGH data: from signal ratio to gain and loss of DNA regions.** *Bioinformatics* 2004, **20:**3413-3422.
14. Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ: **A statistical approach for array CGH data analysis.** *BMC Bioinformatics* 2005, **6:**27.
15. Gardina PJ, Lo KC, Lee W, Cowell JK, Turpaz Y: **Ploidy status and copy number aberrations in primary glioblastomas defined by integrated analysis of allelic ratios, signal ratios and loss of heterozygosity using 500K SNP Mapping Arrays.** *BMC Genomics* 2008, **9:**489.
16. Attiyeh EF, Diskin SJ, Attiyeh MA, Mosse YP, Hou C, Jackson EM, Kim C, Glessner J, Hakonarson H, Biegel JA, Maris JM: **Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy.** *Genome Res* 2009, **19:**276-283.
17. Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Goransson H, Juliusson G, Rosenquist R, Hoglund M, Borg A, Ringner M: **Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays.** *Genome Biol* 2008, **9:**R136.
18. Assie G, LaFramboise T, Platzer P, Bertherat J, Stratakis CA, Eng C: **SNP arrays in heterogeneous tissue: highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples.** *Am J Hum Genet* 2008, **82:**903-915.
19. Huang J, Wei W, Chen J, Zhang J, Liu G, Di X, Mei R, Ishikawa S, Aburatani H, Jones KW, Shapero MH: **CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays.** *BMC Bioinformatics* 2006, **7:**83.
20. Lamy P, Andersen CL, Dyrskjot L, Torring N, Wiuf C: **A hidden Markov model to estimate population mixture and allelic copy-numbers in cancers using Affymetrix SNP arrays.** *BMC Bioinformatics* 2007, **8:**434.
21. Li C, Beroukhim R, Weir BA, Winckler W, Garraway LA, Sellers WR, Meyerson M: **Major copy proportion analysis of tumor samples using SNP arrays.** *BMC Bioinformatics* 2008, **9:**204.
22. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M: **PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.** *Genome Res* 2007, **17:**1665-1674.
23. Bengtsson H, Irizarry R, Carvalho B, Speed TP: **Estimation and assessment of raw copy numbers at the single locus level.** *Bioinformatics* 2008, **24:**759-767.
24. Nancarrow DJ, Handoko HY, Stark MS, Whiteman DC, Hayward NK: **SiDCoN: a tool to aid scoring of DNA copy number changes in SNP chip data.** *PLoS ONE* 2007, **2:**e1093.
25. Shipitsin M, Campbell LL, Argani P, Weremowicz S, Bloushtain-Qimron N, Yao J, Nikolskaya T, Serebryiskaya T, Beroukhim R, Hu M, Halushka MK, Sukumar S, Parker LM, Anderson KS, Harris LN, Garber JE, Richardson AL, Schnitt SJ, Nikolsky Y, Gelman RS, Polyak K: **Molecular definition of breast tumor heterogeneity.** *Cancer Cell* 2007, **11:**259-273.
26. Foulkes WD, Stefansson IM, Chappuis PO, Begin LR, Goffin JR, Wong N, Trudel M, Akslen LA: **Germline BRCA1 mutations and a basal epithelial phenotype in breast cancer.** *J Natl Cancer Inst* 2003, **95:**1482-1485.

27. Turner NC, Reis-Filho JS, Russell AM, Springall RJ, Ryder K, Steele D, Savage K, Gillett CE, Schmitt FC, Ashworth A, Tutt AN: **BRCA1 dysfunction in sporadic basal-like breast cancer.** *Oncogene* 2007, **26:**2126-2132.
28. Gudmundsdottir K, Ashworth A: **The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability.** *Oncogene* 2006, **25:**5864-5874.
29. Vincent-Salomon A, Gruel N, Lucchesi C, MacGrogan G, Dendale R, Sigal-Zafrani B, Longy M, Raynal V, Pierron G, de Mascarel I, Taris C, Stoppa-Lyonnet D, Pierga JY, Salmon R, Sastre-Garau X, Fourquet A, Delattre O, de Cremoux P, Aurias A: **Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity.** *Breast Cancer Res* 2007, **9:**R24.
30. Kreike B, van Kouwenhove M, Horlings H, Weigelt B, Peterse H, Bartelink H, Vijver MJ van de: **Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas.** *Breast Cancer Res* 2007, **9:**R65.
31. Vincent-Salomon A, Ganem-Elbaz C, Manie E, Raynal V, Sastre-Garau X, Stoppa-Lyonnet D, Stern MH, Heard E: **X inactive-specific transcript RNA coating and genetic instability of the X chromosome in BRCA1 breast tumors.** *Cancer Res* 2007, **67:**5134-5140.
32. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA: **Systematic assessment of copy number variant detection via genome-wide SNP genotyping.** *Nat Genet* 2008, **40:**1199-1203.
33. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM: **Common deletion polymorphisms in the human genome.** *Nat Genet* 2006, **38:**86-92.
34. **American Type Culture Collection (ATCC)** [http://www.lgcstandards-atcc.org/]
35. Manie E, Vincent-Salomon A, Lehmann-Che J, Pierron G, Turpin E, Warcoin M, Gruel N, Lebigot I, Sastre-Garau X, Lidereau R, Remenieras A, Feunteun J, Delattre O, de The H, Stoppa-Lyonnet D, Stern MH: **High frequency of TP53 mutation in BRCA1 and sporadic basal-like carcinomas but not in BRCA1 luminal breast tumors.** *Cancer Res* 2009, **69:**663-671.
36. **Integragen** [http://www.integragen.com]
37. **Gene Expression Omnibus (GEO)** [http://www.ncbi.nlm.nih.gov/geo/]
38. R Development Core Team: ***R*: A language and environment for statistical computing.** R Foundation for Statistical Computing. Vienna, Austria; 2009.
39. Flagiello D, Gerbault-Seureau M, Sastre-Garau X, Padoy E, Vielh P, Dutrillaux B: **Highly recurrent der(1;16)(q10;p10) and other 16q arm alterations in lobular breast cancer.** *Genes Chromosomes Cancer* 1998, **23:**300-306.
40. Fletcher J: **Metaphase harvest and cytogenetic analysis of solid tumor cultures.** *Current Protocol in Human Genetics* 2007, **Chapter 10:**. Unit 10.3
41. **Institut Curie Bioinformatics: GAP download Site** [http://bioinfo.curie.fr/projects/snp_gap]

Bibliography

Adélaïde, J., Finetti, P., Bekhouche, I., Repellini, L., Geneix, J., Sircoulomb, F., Charafe-Jauffret, E., Cervera, N., Desplans, J., Parzy, D., Schoenmakers, E., Viens, P., Jacquemier, J., Birnbaum, D., Bertucci, F., and Chaffanet, M. (2007), "Integrated Profiling of Basal and Luminal Breast Cancers," *Cancer Research*, 67, 11565–11575.

Aguilera, A. and Gomez-Gonzalez, B. (2008), "Genome instability: a mechanistic view of its causes and consequences," *Nat Rev Genet*, 9, 204–217.

Ai, L., Tao, Q., Zhong, S., Fields, C. R., Kim, W., Lee, M. W., Cui, Y., Brown, K. D., and Robertson, K. D. (2006), "Inactivation of Wnt inhibitory factor-1 (WIF1) expression by epigenetic silencing is a common event in breast cancer," *Carcinogenesis*, 27, 1341–1348, PMID: 16501252.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002), *Molecular Biology of the Cell, Fourth Edition*, Garland Science, 4th ed.

Albertson, D. G., Collins, C., McCormick, F., and Gray, J. W. (2003), "Chromosome aberrations in solid tumors," *Nature Genetics*, 34, 369–376, PMID: 12923544.

Andre, F., Job, B., Dessen, P., Tordai, A., Michiels, S., Liedtke, C., Richon, C., Yan, K., Wang, B., Vassal, G., Delaloge, S., Hortobagyi, G. N., Symmans, W. F., Lazar, V., and Pusztai, L. (2009), "Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array," *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 15, 441–451, PMID: 19147748.

Aparicio, S. A., Caldas, C., and Ponder, B. (2000), "Does massively parallel transcriptome analysis signify the end of cancer histopathology as we know it?" *Genome Biology*, 1, REVIEWS1021, PMID: 11178245.

Argos, M., Kibriya, M. G., Jasmine, F., Olopade, O. I., Su, T., Hibshoosh, H., and Ahsan, H. (2008), "Genomewide scan for loss of heterozygosity and chromosomal amplification in breast carcinoma using single-nucleotide polymorphism arrays," *Cancer Genetics and Cytogenetics*, 182, 69–74, PMID: 18406867.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000), "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, 25, 25–29, PMID: 10802651.

Auger, I. and Lawrence, C. (1989), "Algorithms for the optimal identification of segment neighborhoods," *Bulletin of Mathematical Biology*, 51, 39–54.

Ballester, B., Johnson, N., Proctor, G., and Flicek, P. (2010), "Consistent annotation of gene expression arrays," *BMC Genomics*, 11, 294, PMID: 20459806.

Bärlund, M., Tirkkonen, M., Forozan, F., Tanner, M. M., Kallioniemi, O., and Kallioniemi, A. (1997), "Increased copy number at 17q22-q24 by CGH in breast cancer is due to high-level amplification of two separate regions," *Genes, Chromosomes & Cancer*, 20, 372–376, PMID: 9408753.

Beissbarth, T. and Speed, T. P. (2004), "GOstat: find statistically overrepresented Gene Ontologies within a group of genes," *Bioinformatics*, 20, 1464–1465.

Bellman, R. (1961), "On the approximation of curves by line segments using dynamic programming," *Commun. ACM*, 4, 284.

Ben-Yaacov, E. and Eldar, Y. C. (2008), "A fast and flexible method for the segmentation of aCGH data," *Bioinformatics*, 24, i139–145.

Bergamaschi, A., Kim, Y. H., Wang, P., Sørlie, T., Hernandez-Boussard, T., Lonning, P. E., Tibshirani, R., Børresen-Dale, A., and Pollack, J. R. (2006), "Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer," *Genes, Chromosomes and Cancer*, 45, 1033–1040.

Biernacki, C., Celeux, G., and Govaert, G. (2000), "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.

Binder, H., Fasold, M., and Glomb, T. (2009), "Mismatch and G-stack modulated probe signals on SNP microarrays," *PloS One*, 4, e7862, PMID: 19924253.

Binder, H., Krohn, K., and Preibisch, S. (2008), ""Hook"-calibration of GeneChip-microarrays: chip characteristics and expression measures," *Algorithms for Molecular Biology: AMB*, 3, 11, PMID: 18759984.

Binder, H., Preibisch, S., and Berger, H. (2010), "Calibration of microarray gene-expression data," *Methods in Molecular Biology (Clifton, N.J.)*, 576, 375–407, PMID: 19882273.

Blamey, R. W., Ellis, I. O., Pinder, S. E., Lee, A. H. S., Macmillan, R. D., Morgan, D. A. L., Robertson, J. F. R., Mitchell, M. J., Ball, G. R., Haybittle, J. L., and Elston, C. W. (2007), "Survival of invasive breast cancer according to the Nottingham Prognostic Index in cases diagnosed in 1990-1999," *European Journal of Cancer (Oxford, England: 1990)*, 43, 1548–1555, PMID: 17321736.

Bollet, M. A., Servant, N., Neuvial, P., Decraene, C., Lebigot, I., Meyniel, J., Rycke, Y. D., Savignoni, A., Rigaill, G., Hupé, P., Fourquet, A., Sigal-Zafrani, B., Barillot, E., and Thiery, J. (2008), "High-resolution mapping of DNA breakpoints to define true recurrences among ipsilateral breast cancers," *Journal of the National Cancer Institute*, 100, 48–58, PMID: 18159071.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003), "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics (Oxford, England)*, 19, 185–193, PMID: 12538238.

Boras-Granic, K. and Wysolmerski, J. J. (2008), "Wnt signaling in breast organogenesis," *Organogenesis*, 4, 116–122, PMID: 19279723 PMCID: 2634257.

Broet, P. and Richardson, S. (2006), "Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model," *Bioinformatics*, 22, 911–918.

Bui, T. D., Tortora, G., Ciardiello, F., and Harris, A. L. (1997), "Expression of Wnt5a is downregulated by extracellular matrix and mutated c-Ha-ras in the human mammary epithelial cell line MCF-10A," *Biochemical and Biophysical Research Communications*, 239, 911–917, PMID: 9367869.

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., Stebbings, L. A., Leroy, C., Edkins, S., Hardy, C., Teague, J. W., Menzies, A., Goodhead, I., Turner, D. J., Clee, C. M., Quail, M. A., Cox, A., Brown, C., Durbin, R., Hurles, M. E., Edwards, P. A. W., Bignell, G. R., Stratton, M. R., and Futreal, P. A. (2008), "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing," *Nature Genetics*, 40, 722–729, PMID: 18438408.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W., Lapuk, A., Neve, R. M., Qian, Z., and Ryder, T. (2006), "Genomic and transcriptional aberrations linked to breast cancer pathophysiologies," *Cancer Cell*, 10, 529–541.

Chin, S. F., Teschendorff, A. E., Marioni, J. C., Wang, Y., Barbosa-Morais, N. L., Thorne, N. P., Costa, J. L., Pinder, S. E., van de Wiel, M. A., Green, A. R., Ellis, I. O., Porter, P. L., Tavaré, S., Brenton, J. D., Ylstra, B., and Caldas, C. (2007), "High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer," *Genome Biology*, 8, R215, PMID: 17925008.

Clevers, H. (2006), "Wnt/beta-catenin signaling in development and disease," *Cell*, 127, 469–480, PMID: 17081971.

Cochran, W. G. and Cox, G. M. (1992), *Experimental Designs, 2nd Edition*, Wiley, 2nd ed.

Collaborative Group on Hormonal Factors in Breast Cancer (2001), "Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease," *Lancet*, 358, 1389–1399, PMID: 11705483.

Collu, G., Meurette, O., and Brennan, K. (2009), "Is there more to Wnt signalling in breast cancer than stabilisation of beta-catenin?" *Breast Cancer Research*, 11, 105.

Courjal, F., Cuny, M., Simony-Lafontaine, J., Louason, G., Speiser, P., Zeillinger, R., Rodriguez, C., and Theillet, C. (1997), "Mapping of DNA amplifications at 15 chromosomal localizations in 1875 breast tumors: definition of phenotypic groups," *Cancer Research*, 57, 4360–4367, PMID: 9331099.

Cowling, V. H., D'Cruz, C. M., Chodosh, L. A., and Cole, M. D. (2007), "c-Myc transforms human mammary epithelial cells through repression of the Wnt inhibitors DKK1 and SFRP1," *Molecular and Cellular Biology*, 27, 5135–5146, PMID: 17485441.

Cox, D. R. (1992), *Planning of Experiments*, Wiley-Interscience.

Cui, X., Kerr, M. K., and Churchill, G. A. (2003), "Transformations for cDNA microarray data," *Statistical Applications in Genetics and Molecular Biology*, 2, Article4, PMID: 16646782.

Delmar, P., Robin, S., and Daudin, J. J. (2005), "VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data," *Bioinformatics*, 21, 502–508.

DiMeo, T. A., Anderson, K., Phadke, P., Feng, C., Perou, C. M., Naber, S., and Kuperwasser, C. (2009), "A Novel Lung Metastasis Signature Links Wnt Signaling with Cancer Cell Self-Renewal and Epithelial-Mesenchymal Transition in Basal-like Breast Cancer," *Cancer Research*, 69, 5364–5373.

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M., and Wang, K. (2008), "Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms," *Nucl. Acids Res.*, 36, e126.

Do, J. H. and Choi, D. (2006), "Normalization of microarray data: single-labeled and dual-labeled arrays," *Molecules and Cells*, 22, 254–261, PMID: 17202852.

Doane, A. S., Danso, M., Lal, P., Donaton, M., Zhang, L., Hudis, C., and Gerald, W. L. (2006), "An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen," *Oncogene*, 25, 3994–4008.

Efron, B. and Tibshirani, R. (2007), "On testing the significance of sets of genes," *The Annals of Applied Statistics*, 1, 107–129.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005), "Outcome signature genes in breast cancer: is there a unique set?" *Bioinformatics (Oxford, England)*, 21, 171–178, PMID: 15308542.

Ellsworth, D. L., Ellsworth, R. E., Liebman, M. N., Hooke, J. A., and Shriver, C. D. (2004), "Genomic instability in histologically normal breast tissues: implications for carcinogenesis," *The Lancet Oncology*, 5, 753–758, PMID: 15581548.

Engler, D. A., Mohapatra, G., Louis, D. N., and Betensky, R. A. (2006), "A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations," *Biostat*, 7, 399–421.

Evers, B., Helleday, T., and Jonkers, J. (2010), "Targeting homologous recombination repair defects in cancer," *Trends in Pharmacological Sciences*, 31, 372–380.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., Macgrogan, G., Bergh, J., Cameron, D., Goldstein, D., Duss, S., Nicoulaz, A., Brisken, C., Fiche, M., Delorenzi, M., and Iggo, R. (2005), "Identification of molecular apocrine breast tumours by microarray analysis," *Oncogene*, 24, 4660–4671, PMID: 15897907.

Fearnhead, P. (2005), "Exact Bayesian curve fitting and signal segmentation," *IEEE Transactions on Signal Processing*, 53, 2160–2166.

Fong, P. C., Boss, D. S., Yap, T. A., Tutt, A., Wu, P., Mergui-Roelvink, M., Mortimer, P., Swaisland, H., Lau, A., O'Connor, M. J., Ashworth, A., Carmichael, J., Kaye, S. B., Schellens, J. H., and de Bono, J. S. (2009), "Inhibition of Poly(ADP-Ribose) Polymerase in Tumors from BRCA Mutation Carriers," *New England Journal of Medicine*, 361, 123–134.

Foulkes, W. D., Stefansson, I. M., Chappuis, P. O., Bégin, L. R., Goffin, J. R., Wong, N., Trudel, M., and Akslen, L. A. (2003), "Germline BRCA1 mutations and a basal epithelial phenotype in breast cancer," *Journal of the National Cancer Institute*, 95, 1482–1485, PMID: 14519755.

Fournier, H. and Vigneron, A. (2008), "Fitting a Step Function to a Point Set," in *Proceedings of the 16th annual European symposium on Algorithms*, Karlsruhe, Germany: Springer-Verlag, pp. 442–453.

Fridlyand, J., Snijders, A., Ylstra, B., Li, H., Olshen, A., Segraves, R., Dairkee, S., Tokuyasu, T., Ljung, B., Jain, A., McLennan, J., Ziegler, J., Chin, K., Devries, S., Feiler, H., Gray, J., Waldman, F., Pinkel, D., and Albertson, D. (2006), "Breast tumor copy number aberration phenotypes and genomic instability," *BMC Cancer*, 6, 96.

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004), "Hidden Markov models approach to the analysis of array CGH data," *J. Multivar. Anal.*, 90, 132–153.

Galea, M. H., Blamey, R. W., Elston, C. E., and Ellis, I. O. (1992), "The Nottingham prognostic index in primary breast cancer," *Breast Cancer Research and Treatment*, 22, 207–219.

Galfalvy, H. C., Erraji-Benchekroun, L., Smyrniotopoulos, P., Pavlidis, P., Ellis, S. P., Mann, J. J., Sibille, E., and Arango, V. (2003), "Sex genes for genomic analysis in human brain: internal controls for comparison of probe level data extraction," *BMC Bioinformatics*, 4, 37, PMID: 12962547.

Gey, S. and Lebarbier, E. (2008), "Using CART to Detect Multiple Change Points in the Mean for Large Sample," http://hal.archives-ouvertes.fr/hal-00327146_v1/.

Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2004), "A global test for groups of genes: testing association with a clinical outcome," *Bioinformatics (Oxford, England)*, 20, 93–99, PMID: 14693814.

Gray, J. W., Collins, C., Henderson, I. C., Isola, J., Kallioniemi, A., Kallioniemi, O. P., Nakamura, H., Pinkel, D., Stokke, T., and Tanner, M. (1994), "Molecular cytogenetics of human breast cancer," *Cold Spring Harbor Symposia on Quantitative Biology*, 59, 645–652, PMID: 7587125.

Guédon, Y. (2008), "Exploring the segmentation space for the assessment of multiple change-point models," Tech. Rep. 6619, INRIA.

Guha, S., Li, Y., and Neuberg, D. (2008), "Bayesian Hidden Markov Modeling of Array CGH Data," *Journal of the American Statistical Association*, 103, 485–497.

Gusterson, B. (2009), "Do 'basal-like' breast cancers really exist?" *Nat Rev Cancer*, 9, 128–134.

Gusterson, B. A., Ross, D. T., Heath, V. J., and Stein, T. (2005), "Basal cytokeratins and their relationship to the cellular origin and functional classification of breast cancer," *Breast Cancer Research: BCR*, 7, 143–148, PMID: 15987465.

Han, W., Jung, E., Cho, J., Lee, J. W., Hwang, K., Yang, S., Kang, J. J., Bae, J., Jeon, Y. K., Park, I., Nicolau, M., Jeffrey, S. S., and Noh, D. (2008), "DNA copy number alterations and expression of relevant genes in triple-negative breast cancer," *Genes, Chromosomes & Cancer*, 47, 490–499, PMID: 18314908.

Harchaoui, Z. and Levy-Leduc, C. (2008), "Catching Change-points with Lasso," in *Advances in Neural Information Processing Systems 20*, eds. Platt, J., Koller, D., Singer, Y., and Roweis, S., Cambridge, MA: MIT Press, pp. 617–624.

Harr, B. and Schlötterer, C. (2006), "Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons," *Nucleic Acids Research*, 34, e8, PMID: 16432259.

Haverty, P. M., Fridlyand, J., Li, L., Getz, G., Beroukhim, R., Lohr, S., Wu, T. D., Cavet, G., Zhang, Z., and Chant, J. (2008), "High-resolution genomic and expression analyses of copy number alterations in breast tumors," *Genes, Chromosomes & Cancer*, 47, 530–542, PMID: 18335499.

Hennessy, B. T., Gonzalez-Angulo, A., Stemke-Hale, K., Gilcrease, M. Z., Krishnamurthy, S., Lee, J., Fridlyand, J., Sahin, A., Agarwal, R., Joy, C., Liu, W., Stivers, D., Baggerly, K., Carey, M., Lluch, A., Monteagudo, C., He, X., Weigman, V., Fan, C., Palazzo, J., Hortobagyi, G. N., Nolden, L. K., Wang, N. J., Valero, V., Gray, J. W., Perou, C. M., and Mills, G. B. (2009), "Characterization of a Naturally Occurring Breast Cancer Subset Enriched in Epithelial-to-Mesenchymal Transition and Stem Cell Characteristics," *Cancer Research*, 69, 4116–4124.

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K. E., Jones, L. P., Assefnia, S., Chandrasekharan, S., Backlund, M. G., Yin, Y., Khramtsov, A. I., Bastein, R., Quackenbush, J., Glazer, R. I., Brown, P. H., Green, J. E., Kopelovich, L., Furth, P. A., Palazzo, J. P., Olopade, O. I., Bernard, P. S., Churchill, G. A., Dyke, T. V., and Perou, C. M. (2007), "Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors," *Genome Biology*, 8, R76, PMID: 17493263.

Hicks, J., Krasnitz, A., Lakshmi, B., Navin, N. E., Riggs, M., Leibu, E., Esposito, D., Alexander, J., Troge, J., Grubor, V., Yoon, S., Wigler, M., Ye, K., Børresen-Dale, A., Naume, B., Schlicting, E., Norton, L., Hägerström, T., Skoog, L., Auer, G., Månér, S., Lundin, P., and Zetterberg, A. (2006), "Novel patterns of genome rearrangement and their association with survival in breast cancer," *Genome Research*, 16, 1465–1479, PMID: 17142309.

Hopkins, A. L. and Groom, C. R. (2002), "The druggable genome," *Nature Reviews. Drug Discovery*, 1, 727–730, PMID: 12209152.

Howe, L. R. and Brown, A. M. C. (2004), "Wnt signaling and breast cancer," *Cancer Biology & Therapy*, 3, 36–41, PMID: 14739782.

Hu, Z., Fan, C., Oh, D., Marron, J., He, X., Qaqish, B., Livasy, C., Carey, L., Reynolds, E., Dressler, L., Nobel, A., Parker, J., Ewend, M., Sawyer, L., Wu, J., Liu, Y., Nanda, R., Tretiakova, M., Orrico, A., Dreher, D., Palazzo, J., Perreard, L., Nelson, E., Mone, M., Hansen, H., Mullins, M., Quackenbush, J., Ellis, M., Olopade, O., Bernard, P., and Perou, C. (2006), "The molecular portraits of breast tumors are conserved across microarray platforms," *BMC Genomics*, 7, 96.

Huang, J., Gusnanto, A., O'Sullivan, K., Staaf, J., Borg, A., and Pawitan, Y. (2007), "Robust smooth segmentation approach for array CGH data analysis," *Bioinformatics*, 23, 2463–2469.

Huguet, E. L., McMahon, J. A., McMahon, A. P., Bicknell, R., and Harris, A. L. (1994), "Differential Expression of Human Wnt Genes 2, 3, 4, and 7B in Human Breast Cell Lines and Normal and Disease States of Human Breast Tissue," *Cancer Research*, 54, 2615–2621.

Hupé, P., Stransky, N., Thiery, J., Radvanyi, F., and Barillot, E. (2004), "Analysis of array CGH data: from signal ratio to gain and loss of DNA regions," *Bioinformatics*, 20, 3413–3422.

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E., and van Noort, V. (2009), "Repeatability of published microarray gene expression analyses," *Nature Genetics*, 41, 149–155, PMID: 19174838.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a), "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Research*, 31, e15, PMID: 12582260.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b), "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics (Oxford, England)*, 4, 249–264, PMID: 12925520.

Jiang, N., Leach, L., Hu, X., Potokina, E., Jia, T., Druka, A., Waugh, R., Kearsey, M., and Luo, Z. (2008), "Methods for evaluating gene expression from Affymetrix microarray datasets," *BMC Bioinformatics*, 9, 284.

Johannsdottir, H. K., Jonsson, G., Johannesdottir, G., Agnarsson, B. A., Eerola, H., Arason, A., Heikkila, P., Egilsson, V., Olsson, H., Johannsson, O. T., Nevanlinna, H., Borg, A., and Barkardottir, R. B. (2006), "Chromosome 5 imbalance mapping in breast tumors from BRCA1 and BRCA2 mutation carriers and sporadic breast tumors," *International Journal of Cancer. Journal International Du Cancer*, 119, 1052–1060, PMID: 16570289.

Jong, K., Marchiori, E., van der Vaart, A., Ylstra, B., Weiss, M., and Meijer, G. (2003), "Chromosomal Breakpoint Detection in Human Cancer," in *Applications of Evolutionary Computing*, Springer-Verlag, pp. 107–116.

Jönsson, G., Staaf, J., Vallon-Christersson, J., Ringnér, M., Holm, K., Hegardt, C., Gunnarsson, H., Fagerholm, R., Strand, C., Agnarsson, B. A., Kilpivaara, O., Luts, L., Heikkilä, P., Aittomäki, K., Blomqvist, C., Loman, N., Malmström, P., Olsson, H., Johannsson, O. T., Arason, A., Nevanlinna, H., Barkardottir, R. B., and Borg, A. (2010), "Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics," *Breast Cancer Research: BCR*, 12, R42, PMID: 20576095.

Kallioniemi, A. (2008), "CGH microarrays and cancer," *Current Opinion in Biotechnology*, 19, 36–40, PMID: 18162393.

Kallioniemi, A., Kallioniemi, O., Sudar, D., Rutovitz, D., Gray, J., Waldman, F., and Pinkel, D. (1992), "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors," *Science*, 258, 818–821.

Kanehisa, M. and Goto, S. (2000), "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, 28, 27–30, PMID: 10592173.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000), "Analysis of variance for gene expression microarray data," *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 7, 819–837, PMID: 11382364.

Khramtsov, A. I., Khramtsova, G. F., Tretiakova, M., Huo, D., Olopade, O. I., and Goss, K. H. (2010), "Wnt/beta-catenin pathway activation is enriched in basal-like breast cancers and predicts poor outcome," *The American Journal of Pathology*, 176, 2911–2920, PMID: 20395444.

Kirikoshi, H., Sekihara, H., and Katoh, M. (2001), "Expression of WNT14 and WNT14B mRNAs in human cancer, up-regulation of WNT14 by IFNgamma and up-regulation of WNT14B by beta-estradiol," *International Journal of Oncology*, 19, 1221–1225, PMID: 11713592.

Klarmann, G. J., Decker, A., and Farrar, W. L. (2008), "Epigenetic gene silencing in the Wnt pathway in breast cancer," *Epigenetics: Official Journal of the DNA Methylation Society*, 3, 59–63, PMID: 18398311.

Kronenwett, U., Huwendiek, S., Östring, C., Portwood, N., Roblick, U. J., Pawitan, Y., Alaiya, A., Sennerstam, R., Zetterberg, A., and Auer, G. (2004), "Improved Grading of Breast Adenocarcinomas Based on Genomic Instability," *Cancer Research*, 64, 904–909.

Kwei, K. A., Kung, Y., Salari, K., Holcomb, I. N., and Pollack, J. R. (2010), "Genomic instability in breast cancer: pathogenesis and clinical implications," *Molecular Oncology*, 4, 255–266, PMID: 20434415.

Lai, T. L., Xing, H., and Zhang, N. (2008), "Stochastic segmentation models for array-based comparative genomic hybridization data analysis," *Biostat*, 9, 290–307.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005), "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data," *Bioinformatics (Oxford, England)*, 21, 3763–3770, PMID: 16081473.

Lavielle, M. (2005), "Using penalized contrasts for the change-point problem," *Signal Processing*, 85, 1501–1510.

Lebarbier, E. (2005), "Detecting multiple change-points in the mean of Gaussian process by model selection," *Signal Processing*, 85, 717–736.

Lejeune, S., Huguet, E. L., Hamby, A., Poulsom, R., and Harris, A. L. (1995), "Wnt5a cloning, expression, and up-regulation in human primary breast cancers." *Clinical Cancer Research*, 1, 215–222.

Li, Y., Lu, W., He, X., Schwartz, A. L., and Bu, G. (2004), "LRP6 expression promotes cancer cell proliferation and tumorigenesis by altering [beta]-catenin subcellular distribution," *Oncogene*, 23, 9129–9135.

Lin, S., Xia, W., Wang, J. C., Kwong, K. Y., Spohn, B., Wen, Y., Pestell, R. G., and Hung, M. (2000), "β-Catenin, a novel prognostic marker for breast cancer: Its roles in cyclin D1 expression and cancer progression," *Proceedings of the National Academy of Sciences of the United States of America*, 97, 4262 –4266.

Liu, C., Prior, J., Piwnica-Worms, D., and Bu, G. (2010), "LRP6 overexpression defines a class of breast cancer subtype and is a target for therapy," *Proceedings of the National Academy of Sciences*, 107, 5136–5141.

Liu, H., Zeeberg, B. R., Qu, G., Koru, A. G., Ferrucci, A., Kahn, A., Ryan, M. C., Nuhanovic, A., Munson, P. J., Reinhold, W. C., Kane, D. W., and Weinstein, J. N. (2007), "AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets," *Bioinformatics*, 23, 2385–2390.

Lizárraga, F., Poincloux, R., Romao, M., Montagnac, G., Dez, G. L., Bonne, I., Rigaill, G., Raposo, G., and Chavrier, P. (2009), "Diaphanous-Related Formins Are Required for Invadopodia Formation and Invasion of Breast Tumor Cells," *Cancer Research*, 69, 2792–2800.

Loo, L. W. M., Grove, D. I., Williams, E. M., Neal, C. L., Cousens, L. A., Schubert, E. L., Holcomb, I. N., Massa, H. F., Glogovac, J., Li, C. I., Malone, K. E., Daling, J. R., Delrow, J. J., Trask, B. J., Hsu, L., and Porter, P. L. (2004), "Array Comparative Genomic Hybridization Analysis of Genomic Alterations in Breast Cancer Subtypes," *Cancer Research*, 64, 8541–8549.

Loo, L. W. M., Ton, C., Wang, Y., Grove, D. I., Bouzek, H., Vartanian, N., Lin, M., Yuan, X., Lawton, T. L., Daling, J. R., Malone, K. E., Li, C. I., Hsu, L., and Porter, P. L. (2008), "Differential patterns of allelic loss in estrogen receptor-positive infiltrating lobular and ductal breast cancer," *Genes, Chromosomes & Cancer*, 47, 1049–1066, PMID: 18720524.

Manié, E., Popova, T., Vincent-Salomon, A., Rigaill, G., Dubois, T., Stoppa-Lyonnet, D., and Stern, M. (2009), "A genomic portrait of BRCA1 and sporadic Basal-Like Carcinomas." Advances in Breast Cancer Research 2009.

Manié, E., Vincent-Salomon, A., Lehmann-Che, J., Pierron, G., Turpin, E., Warcoin, M., Gruel, N., Lebigot, I., Sastre-Garau, X., Lidereau, R., Remenieras, A., Feunteun, J., Delattre, O., de Thé, H., Stoppa-Lyonnet, D., and Stern, M. (2009), "High frequency of TP53 mutation in BRCA1 and sporadic basal-like carcinomas but not in BRCA1 luminal breast tumors," *Cancer Research*, 69, 663–671, PMID: 19147582.

Marioni, J. C., Thorne, N. P., Valsesia, A., Fitzgerald, T., Redon, R., Fiegler, H., Andrews, T. D., Stranger, B. E., Lynch, A. G., Dermitzakis, E. T., Carter, N. P., Tavaré, S., and Hurles, M. E. (2007), "Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization," *Genome Biology*, 8, R228.

Marty, B., Maire, V., Gravier, E., Rigaill, G., Vincent-Salomon, A., Kappler, M., Lebigot, I., Djelti, F., Tourdés, A., Gestraud, P., Hupé, P., Barillot, E., Cruzalegui, F., Tucker, G. C., Stern, M., Thiery, J., Hickman, J. A., and Dubois, T. (2008), "Frequent PTEN genomic alterations and activated phosphatidylinositol 3-kinase pathway in basal-like breast cancer cells," *Breast Cancer Research: BCR*, 10, R101, PMID: 19055754.

Matsuda, Y., Schlange, T., Oakeley, E., Boulay, A., and Hynes, N. (2009), "WNT signaling enhances breast cancer cell motility and blockade of the WNT pathway by sFRP1 suppresses MDA-MB-231 xenograft growth," *Breast Cancer Research*, 11, R32.

Mavaddat, N., Antoniou, A. C., Easton, D. F., and Garcia-Closas, M. (2010), "Genetic susceptibility to breast cancer," *Molecular Oncology*, 4, 174–191, PMID: 20542480.

McCabe, N., Turner, N. C., Lord, C. J., Kluzek, K., Białkowska, A., Swift, S., Giavara, S., O'Connor, M. J., Tutt, A. N., Zdzienicka, M. Z., Smith, G. C., and Ashworth, A. (2006), "Deficiency in the Repair of DNA Damage by Homologous Recombination and Sensitivity to Poly(ADP-Ribose) Polymerase Inhibition," *Cancer Research*, 66, 8109–8115.

Mendes-Pereira, A. M., Martin, S. A., Brough, R., McCarthy, A., Taylor, J. R., Kim, J., Waldman, T., Lord, C. J., and Ashworth, A. (2009), "Synthetic lethal targeting of PTEN mutant cells with PARP inhibitors," *EMBO Molecular Medicine*, 1, 315–322, PMID: 20049735.

Michiels, S., Koscielny, S., and Hill, C. (2005), "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *Lancet*, 365, 488–492, PMID: 15705458.

Moinfar, F. (2008), "Is 'basal-like' carcinoma of the breast a distinct clinicopathological entity? A critical review with cautionary notes," *Pathobiology: Journal of Immunopathology, Molecular and Cellular Biology*, 75, 119–131, PMID: 18544967.

Nagahata, T., Shimada, T., Harada, A., Nagai, H., Onda, M., Yokoyama, S., Shiba, T., Jin, E., Kawanami, O., and Emi, M. (2003), "Amplification, up-regulation and over-expression of DVL-1, the human counterpart of the Drosophila disheveled gene, in primary breast cancers," *Cancer Science*, 94, 515–518, PMID: 12824876.

Naylor, T. L., Greshock, J., Wang, Y., Colligon, T., Yu, Q. C., Clemmer, V., Zaks, T. Z., and Weber, B. L. (2005), "High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization," *Breast Cancer Research: BCR*, 7, R1186–1198, PMID: 16457699.

Neuvial, P., Hupé, P., Brito, I., Liva, S., Manié, E., Brennetot, C., Radvanyi, F., Aurias, A., and Barillot, E. (2006), "Spatial normalization of array-CGH data," *BMC Bioinformatics*, 7, 264, PMID: 16716215.

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J., and Tong, F. (2006), "A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes," *Cancer Cell*, 10, 515–527.

Nusse, R. and Varmus, H. E. (1982), "Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome," *Cell*, 31, 99–109, PMID: 6297757.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004), "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics (Oxford, England)*, 5, 557–572, PMID: 15475419.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009), "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes," *J Clin Oncol*, 27, 1160–1167.

Parkin, D. M. (2004), "International variation," *Oncogene*, 23, 6329–6340, PMID: 15322508.

Peppercorn, J., Perou, C. M., and Carey, L. A. (2008), "Molecular subtypes in breast cancer evaluation and management: divide and conquer," *Cancer Investigation*, 26, 1–10, PMID: 18181038.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000), "Molecular portraits of human breast tumours," *Nature*, 406, 747–752, PMID: 10963602.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J. (2005), "A statistical approach for array CGH data analysis," *BMC Bioinformatics*, 6, 27, PMID: 15705208.

Pinkel, D. and Albertson, D. G. (2005), "Array comparative genomic hybridization and its applications in cancer," *Nature Genetics*, 37 Suppl, S11–17, PMID: 15920524.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W., Chen, C., Zhai, Y., Dairkee, S. H., marie Ljung, B., Gray, J. W., and Albertson, D. G. (1998), "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays," *Nat Genet*, 20, 207–211.

Ploner, A., Miller, L. D., Hall, P., Bergh, J., and Pawitan, Y. (2005), "Correlation test to assess low-level processing of high-density oligonucleotide microarray data," *BMC Bioinformatics*, 6, 80, PMID: 15799785.

Podo, F., Buydens, L. M. C., Degani, H., Hilhorst, R., Klipp, E., Gribbestad, I. S., Huffel, S. V., van Laarhoven, H. W. M., Luts, J., Monleon, D., Postma, G. J., Schneiderhan-Marra, N., Santoro, F., Wouters, H., Russnes, H. G., Sørlie, T., Tagliabue, E., and Børresen-Dale, A. (2010), "Triple-negative breast cancer: present challenges and new perspectives," *Molecular Oncology*, 4, 209–229, PMID: 20537966.

Polakis, P. (2007), "The many ways of Wnt in cancer," *Current Opinion in Genetics & Development*, 17, 45–51, PMID: 17208432.

Polyak, K. (2007), "Breast cancer: origins and evolution," *The Journal of Clinical Investigation*, 117, 3155–3163, PMID: 17975657.

Popova, T., Manié, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E., and Stern, M. (2009), "Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays," *Genome Biology*, 10, R128.

Pusztai, L., Mazouni, C., Anderson, K., Wu, Y., and Symmans, W. F. (2006), "Molecular classification of breast cancer: limitations and potential," *The Oncologist*, 11, 868–877, PMID: 16951390.

Rakha, E. A., Reis-Filho, J. S., and Ellis, I. O. (2008), "Basal-Like Breast Cancer: A Critical Review," *J Clin Oncol*, 26, 2568–2581.

Ralhan, R., Kaur, J., Kreienberg, R., and WiesmÃ$\frac{1}{4}$ller, L. (2007), "Links between DNA double strand break repair and breast cancer: accumulating evidence from both familial and nonfamilial cases," *Cancer Letters*, 248, 1–17, PMID: 16854521.

Reis-Filho, J. S. and Tutt, A. N. J. (2008), "Triple negative tumours: a critical review," *Histopathology*, 52, 108–118, PMID: 18171422.

Richardson, A. L., Wang, Z. C., Nicolo, A. D., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J. D., Livingston, D. M., and Ganesan, S. (2006), "X chromosomal abnormalities in basal-like human breast cancer," *Cancer Cell*, 9, 121–132, PMID: 16473279.

Rigaill, G. (2010a), "Exact and fast segmentation of large SNP/CGH profiles." SMPGD2010 conference: http://iml.univ-mrs.fr/sta/SMPGD2010/.

— (2010b), "Pruned dynamic programming for optimal multiple change-point detection," *Arxiv:1004.0887*.

Rigaill, G., Aurélie Dumont, B. M., Maire, V., Richardson, M., and Dubois, T. (2010a), "Transcriptomic pattern of the Wnt pathway in TNBC." In preparation.

Rigaill, G., Dumont, A., Marty, B., Maire, V., Richardson, M., Dubois, T., and Vincent-Salomon, A. (2010b), "Correspondence between the IHC and transcriptomic classification of TNBC." In preparation.

Rigaill, G., Hupe, P., Almeida, A., Rosa, P. L., Meyniel, J., Decraene, C., and Barillot, E. (2008), "ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays," *Bioinformatics*, 24, 768–774.

Rigaill, G., Lebarbier, E., and Robin, S. (2010c), "Exact posterior distributions over the segmentation space and model selection for multiple change-point detection problems," *Arxiv:1004.4347*.

— (2010d), "Exact posterior distributions over the segmentation space and model selection for multiple change-point detection problems," Accepted in the proceeding of COMPSTAT 2010.

Rigaill, G., Philippe Hupé, J. M., Decraene, C., and Barillot, E. (2007), "GINCOS: a method to normalize Affymetrix GeneChip Human Mapping 50K Set." SMPGD2007 conference:http://www.lsp.ups-tlse.fr/Biopuces/SMPGD07.

Rouzier, R., Perou, C. M., Symmans, W. F., Ibrahim, N., Cristofanilli, M., Anderson, K., Hess, K. R., Stec, J., Ayers, M., Wagner, P., Morandi, P., Fan, C., Rabiul, I., Ross, J. S., Hortobagyi, G. N., and Pusztai, L. (2005), "Breast Cancer Molecular Subtypes Respond Differently to Preoperative Chemotherapy," *Clinical Cancer Research*, 11, 5678–5685.

Ryo, A., Nakamura, M., Wulf, G., Liou, Y. C., and Lu, K. P. (2001), "Pin1 regulates turnover and subcellular localization of beta-catenin by inhibiting its interaction with APC," *Nature Cell Biology*, 3, 793–801, PMID: 11533658.

Saal, L. H., Gruvberger-Saal, S. K., Persson, C., Lövgren, K., Jumppanen, M., Staaf, J., Jönsson, G., Pires, M. M., Maurer, M., Holm, K., Koujak, S., Subramaniyam, S., Vallon-Christersson, J., Olsson, H., Su, T., Memeo, L., Ludwig, T., Ethier, S. P., Krogh, M., Szabolcs, M., Murty, V. V. V. S., Isola, J., Hibshoosh, H., Parsons, R., and Borg, A. (2008), "Recurrent gross mutations of the PTEN tumor suppressor gene in breast cancers with deficient DSB repair," *Nature Genetics*, 40, 102–107, PMID: 18066063.

SantaLucia, J. (1998), "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proceedings of the National Academy of Sciences of the United States of America*, 95, 1460–1465, PMID: 9465037.

Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A. M., Yu, J., Klijn, J. G., Foekens, J. A., and Martens, J. W. (2008), "Subtypes of Breast Cancer Show Preferential Site of Relapse," *Cancer Research*, 68, 3108–3114.

Smyth, G. K. (2004), "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, 3, Article3, PMID: 16646809.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Döhner, H., Cremer, T., and Lichter, P. (1997), "Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances," *Genes, Chromosomes & Cancer*, 20, 399–407, PMID: 9408757.

Sørlie, T. (2003), "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proceedings of the National Academy of Sciences*, 100, 8418–8423.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., and Børresen-Dale, A. (2001a), "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences of the United States of America*, 98, 10869 –10874.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., and Børresen-Dale, A. L. (2001b), "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences of the United States of America*, 98, 10869–10874, PMID: 11553815.

Staaf, J., Jonsson, G., Ringner, M., and Vallon-Christersson, J. (2007), "Normalization of array-CGH data: influence of copy number imbalances," *BMC Genomics*, 8, 382.

Staaf, J., Jönsson, G., Ringnér, M., Vallon-Christersson, J., Grabau, D., Arason, A., Gunnarsson, H., Agnarsson, B. A., Malmström, P., Johannsson, O. T., Loman, N., Barkardottir, R. B., and Borg, A. (2010), "High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer," *Breast Cancer Research: BCR*, 12, R25, PMID: 20459607.

Stange, D. E., Radlwimmer, B., Schubert, F., Traub, F., Pich, A., Toedt, G., Mendrzyk, F., Lehmann, U., Eils, R., Kreipe, H., and Lichter, P. (2006), "High-resolution genomic profiling reveals association of chromosomal aberrations on 1q and 16p with histologic and genetic subgroups of invasive breast cancer," *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 12, 345–352, PMID: 16428471.

Stemke-Hale, K., Gonzalez-Angulo, A. M., Lluch, A., Neve, R. M., Kuo, W., Davies, M., Carey, M., Hu, Z., Guan, Y., Sahin, A., Symmans, W. F., Pusztai, L., Nolden, L. K., Horlings, H., Berns, K., Hung, M., van de Vijver, M. J., Valero, V., Gray, J. W., Bernards, R., Mills, G. B., and Hennessy, B. T. (2008), "An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer," *Cancer Research*, 68, 6084–6091, PMID: 18676830.

Stephens, P. J., McBride, D. J., Lin, M., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., Greenman, C. D., Jia, M., Latimer, C., Teague, J. W., Lau, K. W., Burton, J., Quail, M. A., Swerdlow, H., Churcher, C., Natrajan, R., Sieuwerts, A. M., Martens, J. W. M., Silver, D. P., Langerød, A., Russnes, H. E. G., Foekens, J. A., Reis-Filho, J. S., van 't Veer, L., Richardson, A. L., Børresen-Dale, A., Campbell, P. J., Futreal, P. A., and Stratton, M. R. (2009), "Complex landscapes of somatic rearrangement in human breast cancer genomes," *Nature*, 462, 1005–1010, PMID: 20033038.

Stingl, J. (2009), "Detection and analysis of mammary gland stem cells," *The Journal of Pathology*, 217, 229–241, PMID: 19009588.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005), "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 15545–15550, PMID: 16199517.

Suzuki, H., Toyota, M., Carraway, H., Caraway, H., Gabrielson, E., Ohmura, T., Fujikane, T., Nishikawa, N., Sogabe, Y., Nojima, M., Sonoda, T., Mori, M., Hirata, K., Imai, K., Shinomura, Y., Baylin, S. B., and Tokino, T. (2008), "Frequent epigenetic inactivation of Wnt antagonist genes in breast cancer," *British Journal of Cancer*, 98, 1147–1156, PMID: 18283316.

Tavassoli, F. A., Devilee, P., and for Research on Cancer, I. A. (2003), *Pathology and genetics of tumours of the breast and female genital organs*, IARC.

Tibshirani, R. and Wang, P. (2008), "Spatial smoothing and hot spot detection for CGH data using the fused lasso," *Biostatistics (Oxford, England)*, 9, 18–29, PMID: 17513312.

Toullec, A., Gerald, D., Despouy, G., Bourachot, B., Cardon, M., Lefort, S., Richardson, M., Rigaill, G., Parrini, M., Lucchesi, C., Bellanger, D., Stern, M., Dubois, T., Sastre-Garau, X., Delattre, O., Vincent-Salomon, A., and Mechta-Grigoriou, F. (2010), "Oxidative stress promotes myofibroblast differentiation and tumour spreading," *EMBO Molecular Medicine*, 2, 211–230.

Tucker, C. L. and Fields, S. (2003), "Lethal combinations," *Nature Genetics*, 35, 204–205, PMID: 14593402.

Turashvili, G., Bouchal, J., Burkadze, G., and Kolar, Z. (2006), "Wnt Signaling Pathway in Mammary Gland Development and Carcinogenesis," *Pathobiology*, 73, 213–223.

Turner, N. C., Reis-Filho, J. S., Russell, A. M., Springall, R. J., Ryder, K., Steele, D., Savage, K., Gillett, C. E., Schmitt, F. C., Ashworth, A., and Tutt, A. N. (2006), "BRCA1 dysfunction in sporadic basal-like breast cancer," *Oncogene*, 26, 2126–2132.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001), "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, 98, 5116–5121, PMID: 11309499.

van de Wiel, M. A., Brosens, R., Eilers, P. H. C., Kumps, C., Meijer, G. A., Menten, B., Sistermans, E., Speleman, F., Timmerman, M. E., and Ylstra, B. (2009), "Smoothing waves in array CGH tumor profiles," *Bioinformatics*, 25, 1099–1104.

van de Wiel, M. A., Picard, F., van Wieringen, W. N., and Ylstra, B. (2010), "Preprocessing and downstream analysis of microarray DNA copy number profiles," *Briefings in Bioinformatics*, PMID: 20172948.

Vargo-Gogola, T. and Rosen, J. M. (2007), "Modelling breast cancer: one size does not fit all," *Nature Reviews. Cancer*, 7, 659–672, PMID: 17721431.

Veeck, J., Noetzel, E., Bektas, N., Jost, E., Hartmann, A., KnÃ$\frac{1}{4}$chel, R., and Dahl, E. (2008), "Promoter hypermethylation of the SFRP2 gene is a high-frequent alteration and tumor-specific epigenetic marker in human breast cancer," *Molecular Cancer*, 7, 83, PMID: 18990230.

Veeck, J., Ropero, S., Setien, F., Gonzalez-Suarez, E., Osorio, A., Benitez, J., Herman, J. G., and Esteller, M. (2010), "BRCA1 CpG Island Hypermethylation Predicts Sensitivity to Poly(Adenosine Diphosphate)-Ribose Polymerase Inhibitors," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, PMID: 20679605.

Venkatraman, E. S. and Olshen, A. B. (2007), "A faster circular binary segmentation algorithm for the analysis of array CGH data," *Bioinformatics*, 23, 657–663.

Vincent-Salomon, A. (2008), "Hétérogénéité biologique des carcinomes mammaires canalaires infiltrants de type basal-like et in situ." Ph.D. thesis, Ecole Doctorale de Cancérologie.

Vincent-Salomon, A., Gruel, N., Lucchesi, C., MacGrogan, G., Dendale, R., Sigal-Zafrani, B., Longy, M., Raynal, V., Pierron, G., de Mascarel, I., Taris, C., Stoppa-Lyonnet, D., Pierga, J., Salmon, R., Sastre-Garau, X., Fourquet, A., Delattre, O., de Cremoux, P., and Aurias, A. (2007), "Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity," *Breast Cancer Research: BCR*, 9, R24, PMID: 17417968.

Volik, S., Zhao, S., Chin, K., Brebner, J. H., Herndon, D. R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W., Magrane, G., Jong, P. D., Gray, J. W., and Collins, C. (2003), "End-sequence profiling: sequence-based analysis of aberrant genomes," *Proceedings of the National Academy of Sciences of the United States of America*, 100, 7696–7701, PMID: 12788976.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., and Tibshirani, R. (2005), "A method for calling gains and losses in array CGH data," *Biostat*, 6, 45–58.

Wang, Z. C., Lin, M., Wei, L., Li, C., Miron, A., Lodeiro, G., Harris, L., Ramaswamy, S., Tanenbaum, D. M., Meyerson, M., Iglehart, J. D., and Richardson, A. (2004), "Loss of Heterozygosity and Its Correlation with Expression Profiles in Subclasses of Invasive Breast Cancers," *Cancer Research*, 64, 64–71.

Weigelt, B., Baehner, F. L., and Reis-Filho, J. S. (2010a), "The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade," *The Journal of Pathology*, 220, 263–280, PMID: 19927298.

Weigelt, B., Geyer, F. C., and Reis-Filho, J. S. (2010b), "Histological types of breast cancer: How special are they?" *Molecular Oncology*, 4, 192–208.

Willenbrock, H. and Fridlyand, J. (2005), "A comparison study: applying segmentation to array CGH data for downstream analyses," *Bioinformatics (Oxford, England)*, 21, 4084–4091, PMID: 16159913.

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004), "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays," *Journal of the American Statistical Association*, 99, 909–917.

Yager, J. D. and Davidson, N. E. (2006), "Estrogen carcinogenesis in breast cancer," *The New England Journal of Medicine*, 354, 270–282, PMID: 16421368.

Yao, Y. (1984), "Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches," *The Annals of Statistics*, 12, 1434–1447.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001), "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, 17, 977–987.

Zardawi, S. J., O'Toole, S. A., Sutherland, R. L., and Musgrove, E. A. (2009), "Dysregulation of Hedgehog, Wnt and Notch signalling pathways in breast cancer," *Histology and Histopathology*, 24, 385–398, PMID: 19130408.

Zhang, L., Miles, M. F., and Aldape, K. D. (2003), "A model of molecular interactions on short oligonucleotide microarrays," *Nature Biotechnology*, 21, 818–821, PMID: 12794640.

Zhang, N. R. and Siegmund, D. O. (2007), "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data," *Biometrics*, 63, 22–32, PMID: 17447926.