Learning from multiple genomic information in cancer for diagnosis and prognosis

Matahi Moarii

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

École doctorale n° 432 :

Sciences des métiers de l'ingénieur

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité " Bio-informatique "

présentée et soutenue publiquement par

Matahi MOARII

le 26 juin 2015

Apprentissage de données génomiques multiples pour le diagnostic et pronostic du cancer

~~~

Learning from multiple genomic information in cancer for diagnosis and prognosis

Directeur de thèse : **Jean-Philippe VERT**
Co-encadrement de la thèse : **Fabien REYAL**

Jury

M. Franck PICARD, Directeur de Recherche, Laboratoire de Biométrie et Biologie Evolutive, UCB Lyon 1	Rapporteur
M. Jörg TOST, Directeur de Recherche,Laboratory for Epigenetics, Centre National de Génotypage Evry	Rapporteur
Mme Sandrine DUDOIT, Professeur, Division of Biostatistics, University of California Berkeley	Président
Mme Véronique STOVEN, Professeur, Center for Computational Biology, Mines Paristech	Examinateur
M. Jean-Philippe VERT, Ingénieur en chef des Mines, Center for Computational Biology, Mines Paristech	Examinateur
M. Fabien REYAL, Directeur de Recherche, Département de Transfert, Institut Curie	Examinateur

MINES ParisTech
Centre de Bio-Informatique
35 rue Saint-Honoré, 77300 Fontainebleau, France

T H È S E
~~~

"There are three kinds of lies: lies, damned lies, and statistics."

Mark Twain

Acknowledgements

En premier lieu, je souhaiterais remercier mes parents. Sans l'éducation qu'ils m'ont apportée et sans leur soutien continu au fil des années, je ne serais sûrement pas où je suis aujourd'hui. Merci infiniment!

Ma reconnaissance va à Jörg Tost et Franck Picard qui ont pris le temps d'être rapporteurs pour cette thèse. Je tiens également à remercier chaleureusement les examinateurs Sandrine Dudoit et Véronique Stoven.

Beaucoup de personnes ont grandement contribué à cette thèse tant d'un point scientifique, personnel et professionnel. Je tiens tout d'abord à remercier très sincèrement mes deux directeurs de thèse Jean-Philippe Vert et Fabien Reyal. Ce fut un réel plaisir de travailler avec vous durant ces années où j'aurai beaucoup appris tant d'un point de vue professionnel que personnel.

Mes amitiés et ma considération vont à mes collègues du CBIO, du RT2Lab et de l'U900 et sans bien sûr oublier mes collègues biologistes de la BDD. J'espère que nous aurons l'occasion de nous re-croiser à l'avenir.

Il serait difficile de ne pas en oublier et donc je tiens à remercier sans les citer tous mes proches: ma famille et mes amis.

Abstract

Learning from multiple genomic information in cancer for diagnosis and prognosis

by Matahi Moarii

De nombreuses initiatives ont été mises en places pour charactériser d'un point de vue moléculaire de grandes cohortes de cancers à partir de diverses sources biologiques dans l'espoir de comprendre les altérations majeures impliquées durant la tumorogénèse. Les données mesurées incluent l'expression des gènes, les mutations et variations de copy-number, ainsi que des signaux épigénétiques tel que la méthylation de l'ADN. De grands consortium tels que "The Cancer Genome Atlas" (TCGA) ont déjà permis de rassembler plusieurs milliers d'échantillons cancéreux mis à la disposition du public. Nous contribuons dans cette thèse à analyser d'un point de vue mathématique les relations existant entre les différentes sources biologiques, valider et/ou généraliser des phénomènes biologiques à grande échelle par une analyse intégrative de données épigénétiques et génétiques.

En effet, nous avons montré dans un premier temps que la méthylation de l'ADN était un marqueur substitutif intéressant pour jauger du caractère clonal entre deux cellules et permettait ainsi de mettre en place un outil clinique des récurrences de cancer du sein plus précis et plus stable que les outils actuels, afin de permettre une meilleure prise en charge des patients.

D'autre part, nous avons dans un second temps permis de quantifier d'un point de vue statistique l'impact de la méthylation sur la transcription. Nous montrons l'importance d'incorporer des hypothèses biologiques afin de pallier au faible nombre d'échantillons par rapport aux nombre de variables.

Enfin, nous montrons l'existence d'un phénomène biologique lié à l'apparition d'un phénotype d'hyperméthylation dans plusieurs cancers. Pour cela, nous adaptons des méthodes de régression en utilisant la similarité entre les différentes tâches de prédictions afin d'obtenir des signatures génétiques communes prédictives du phénotypes plus précises.

En conclusion, nous montrons l'importance d'une collaboration biologique et statistique afin d'établir des méthodes adaptées aux problématiques actuelles en bioinformatique.

Abstract

Learning from multiple genomic information in cancer for diagnosis and prognosis

by Matahi Moarii

Several initiatives have been launched recently to investigate the molecular characterisation of large cohorts of human cancers with various high-throughput technologies in order to understanding the major biological alterations related to tumorogenesis. The information measured include gene expression, mutations, copy-number variations, as well as epigenetic signals such as DNA methylation. Large consortiums such as "The Cancer Genome Atlas" (TCGA) have already gathered publicly thousands of cancerous and non-cancerous samples. We contribute in this thesis in the statistical analysis of the relationship between the different biological sources, the validation and/or large scale generalisation of biological phenomenon using an integrative analysis of genetic and epigenetic data.

Firstly, we show the role of DNA methylation as a surrogate biomarker of clonality between cells which would allow for a powerful clinical tool for to elaborate appropriate treatments for specific patients with breast cancer relapses.

In addition, we developed systematic statistical analyses to assess the significance of DNA methylation variations on gene expression regulation. We highlight the importance of adding prior knowledge to tackle the small number of samples in comparison with the number of variables. In return, we show the potential of bioinformatics to infer new interesting biological hypotheses.

Finally, we tackle the existence of the universal biological phenomenon related to the hypermethylator phenotype. Here, we adapt regression techniques using the similarity between the different prediction tasks to obtain robust genetic predictive signatures common to all cancers and that allow for a better prediction accuracy.

In conclusion, we highlight the importance of a biological and computational collaboration in order to establish appropriate methods to the current issues in bioinformatics that will in turn provide new biological insights.

Contents

Acknowledgements iii

Résumé iv

Abstract v

Contents vi

List of Figures x

List of Tables xii

1 Introduction 1
- 1.1 Préambule . . . 1
- 1.2 Preamble . . . 1
- 1.3 From a macroscopic to a molecular characterization of cancer . . . 2
  - 1.3.1 Histopathology of cancer: the premisces of personalized medicine. A focus on breast cancer. . . . 3
  - 1.3.2 Molecular classifications of cancers: the dawn of personalized medicine 4
  - 1.3.3 The overflow of *-omics* data and the necessity of a statistical framework . . . 4
- 1.4 Epigenetics . . . 5
  - 1.4.1 DNA methylation . . . 6
  - 1.4.2 DNA Methylation in gene regulation . . . 7
  - 1.4.3 An early biomarker in cancer and a source for potential treatments 8
  - 1.4.4 Statistical challenges in DNA methylation analysis . . . 9
- 1.5 Personal contribution and organization of the thesis . . . 9

2 Methods 13
- 2.1 Supervised learning . . . 14
  - 2.1.1 Risk minimization problem . . . 15
  - 2.1.2 The curse of dimensionality . . . 15
  - 2.1.3 Model selection . . . 16
  - 2.1.4 Assessing the performance of a model . . . 18
  - 2.1.5 Interpreting the data . . . 19

Ordinary Least Squares. 19
Ridge Regression. 20
Sparsity-inducing penalties. 20
Feature selection and multiple-testing. 21
Adding prior knowledge. 22
2.2 Unsupervised learning 24
2.2.1 Cluster analysis 24
Choosing a similarity between samples. 24
K-means. 25
Hierachical clustering: an alternative with several advantages. 26
2.2.2 Dimensionality reduction 27
Principal Component Analysis (PCA). 27
Model selection in unsupervised learning. 28
Clinical impact of unsupervised learning. 30
2.3 Conclusion 30

3 Epigenomic alterations in breast carcinoma from primary tumor to locoregional recurrences 31
3.1 Résumé 31
3.2 Abstract 32
3.3 Introduction 33
3.4 Materials and Methods 35
3.4.1 Patients Selection 35
3.4.2 Methylation profiling 36
3.4.3 Clinical Classification. 36
3.4.4 Data analysis 36
3.5 Results 37
3.5.1 Methylation differences between PT and their matched metastasis or recurrence 37
3.5.2 Methylation conservation between PT and their matched metastasis or recurrence 40
3.5.3 Clonality detection based on methylation profiles 42
3.6 Discussion 45

4 Changes in gene expression control by DNA methylation in cancer 51
4.1 Résumé 51
4.2 Abstract 53
4.3 Introduction 53
4.4 Materials and Methods 55
4.4.1 Patients Selection 55
4.4.2 Methylation profiling 55
4.4.3 Gene expression profiling 56
4.4.4 Copy number variations processing 56
4.4.5 Combined CpG island, shores and shelves pattern analysis using dynamic time warping 56
4.4.6 Survival analysis 59
4.4.7 Computing the predictive power of methylation 59

4.5 Results . . . 60
4.5.1 Classification of genes based on their CGI methylation profiles in normal and cancerous tissues . . . 60
4.5.2 Cancer-specific methylation does not repress gene expression but instead targets genes lowly expressed in normal tissues . . . 62
4.5.3 Cancer-specific methylation is an independent predictor of patient survival in breast cancer . . . 67
4.5.4 Methylation of CpG in CGI shores is negatively correlated with gene expression. . . . 71
4.5.5 Regulation of gene expression by DNA methylation is tissue-specific and the process is altered in cancer tissues but overall targets transcription factors. . . . 73
4.5.6 Copy number variations in cancer is an independent factor in gene expression regulation. . . . 75
4.6 Discussion . . . 76

5 Integrative DNA methylation and gene expression profiles to assess the universality of the CpG island methylator phenotype 81
5.1 Résumé . . . 81
5.2 Abstract . . . 82
5.3 Introduction . . . 82
5.4 Material and Methods . . . 83
5.4.1 Patients Selection . . . 83
5.4.2 Methylation profiling . . . 84
5.4.3 Gene expression profiling . . . 84
5.4.4 Mutation profiling . . . 84
5.4.5 CIMP analysis . . . 85
5.4.6 Predicting CIMP status from gene expression profiles . . . 85
5.4.7 Tissue-specific lasso . . . 85
5.4.8 Combined Lasso . . . 85
5.4.9 Group Lasso . . . 85
5.4.10 Survival analysis . . . 87
5.5 Results . . . 87
5.5.1 Are there 2 or 3 CIMP classes? . . . 88
5.5.2 Similar gene expression variations are predictive of CIMP. . . . 93
5.5.3 A genetic signature is associated to CIMP only for colon and gastric cancers . . . 97
5.5.4 Clinical impact of CIMP. . . . 100
5.6 Discussion . . . 100
5.7 Conclusion . . . 103

6 Discussion 105
6.1 "DNA methylation in cancer: too much, but also too little" . . . and more. 105
6.2 "All models are wrong but are some of them actually useful?". . . . 106
6.3 Perspectives in the use of computational tools for epigenetics and biology. 107
Tumor heterogeneity. . . . 107
Alternative splicing. . . . 107
Long Range epigenetic regulation. . . . 107

Bibliography **109**

List of Figures

1.1 Histopathological sections of breast cancer . . . 3
1.2 Methylation . . . 6
1.3 Methylation regulatory mechanism . . . 8

2.1 Curse of dimensionality . . . 17
2.2 Bias-Variance Tradeoff . . . 18
2.3 Joint Regularization of methylome profiles . . . 23
2.4 Similarity . . . 25
2.5 K-means clustering . . . 26
2.6 Hierarchical Clustering . . . 27
2.7 PCA . . . 28
2.8 K-means clustering . . . 29
2.9 Trade-off in dimensionality reduction . . . 29

3.1 Accuracy of the SVM classifier as a function of the number of probes selected . . . 41
3.2 Study of similarity between matched primary tumors and recurrences by hierarchical clustering . . . 43
3.3 Pairwise methylome distance for each samples . . . 44
3.4 Distribution of methylation similarity between samples given the type of pairs . . . 45
3.5 Histogram of the distribution of methylome-similarity score (MS) between unrelated PT/LR pairs . . . 46
3.6 Correlation between methylation and copy-number scores . . . 47
3.7 Kaplan-Meier estimates of the metastasis-free survival between TR and NP for the different classification methods . . . 48

4.1 Standard pairing between two signals . . . 57
4.2 Dynamic time warping pairing between two signals . . . 57
4.3 Distance matrix between signal 1 and signal 2 . . . 58
4.4 CGI+SS patterns in breast tissues . . . 63
4.5 Characteristic profiles for each clusters . . . 64
4.6 Characteristic profiles for each clusters . . . 64
4.7 Characteristic profiles for each clusters . . . 65
4.8 Gene Ontology analysis given the cluster assignment for cancerous breast tissues . . . 65
4.9 Gene Ontology analysis given the cluster assignment for cancerous colon tissues . . . 66

4.10 Gene Ontology analysis given the cluster assignment for cancerous lung tissues . . . 66
4.11 Inter-tissue stability of the cancerous-specific cluster . . . 67
4.12 cluster characteristics analysis in breast tissues . . . 68
4.13 Clinical relevance of the CGI+SS clusters for survival prognosis in breast cancer patients . . . 69
4.14 Hierarchical clustering of breast cancer patients based on the average methylation level of CGI+SS associated with cluster 3down . . . 70
4.15 Impact of DNA methylation in gene expression prediction . . . 73
4.16 Methylation association with gene expression by regions . . . 74
4.17 Tissue-specificity of epigenetic regulation . . . 75
4.18 Shift of epigenetic regulation in cancer . . . 76
4.19 Association between predictive power of methylation and copy-number variations . . . 77

5.1 Methylation profiles hierarchical clustering for each tissue based on the most variant probes . . . 89
5.2 Stability of CIMP clusters given the proportion of variant CGIs . . . 91
5.3 Universal epigenetic signature for CIMP . . . 92
5.4 Stability of CIMP given the number of clusters . . . 94
5.5 Gene expression variations predictive of CIMP . . . 95
5.6 Analysis of a genetic signature associated with CIMP . . . 98
5.7 Comparison of genetic signatures associated with CIMP for colon and gastric cancers . . . 99
5.8 Clinical impact of CIMP on the patient survival . . . 101

List of Tables

3.1 PT/LR clinical and histological features . . . 38
3.2 PT/CL clinical and histological features . . . 39
3.3 PT/AM clinical and histological features . . . 39
3.4 Most significantly differentially methylated genes between PT and AM samples (Top 10) . . . 40
3.5 Comparison of classification methods for clonality between pairs in the PT/LR cohort . . . 49

4.1 Patients dataset . . . 55
4.2 Concordance analysis of CGI+SS patterns clusters between normal tissues. 60
4.3 Concordance analysis of CGI+SS patterns clusters from normal to cancerous tissues . . . 62
4.4 Clinical impact of cancer-specific cluster on patient survival in breast cancer patients . . . 70
4.5 Genes regulation by methylation in different tissues . . . 72

5.1 CIMP Patients datasets . . . 84
5.2 CIMP proportions . . . 85
5.3 Universal epigenetic signature . . . 93
5.4 Matched Meth/GE samples CIMP proportion . . . 94
5.5 Accuracy of CIMP prediction using gene expression profiles . . . 96
5.6 Intersection of the genetic signatures for “Combined-Lasso” and “Group-Lasso” . . . 97
5.7 Clinical impact of CIMP . . . 100

Chapter 1

Introduction

1.1 Préambule

La complétion du "Human Genome Project" a accéléré le développement de nouvelles technologies de mesures liées au génome humain. Ce flux d'information biologique et clinique a considérablement impacté notre manière d'aborder une hypothèse biologique. Nous nous intéressons en particulier dans cette thèse au développement de méthodes statistiques pour l'analyse de données génomiques spécifiques, les données épigénétiques, et leur lien dans le diagnostic et pronostic du cancer. Dans ce chapitre, nous introduisons les principales notions biologiques abordées dans cette thèse. La section 1.3 s'attache à introduire les perspectives actuelles liées à la thérapie du cancer. Le cancer du sein, principale pathologie traitée à l'Institut Curie, sera en particulier abordé. Dans la section 1.4, nous présentons des données épigénétiques et plus spécifiquement de la méthylation de l'ADN.

1.2 Preamble

After the completion of the human genome project more than ten years ago, DNA measurement technologies have witnessed dramatic progress in scope and throughput at constantly decreasing cost. This led to a flood of clinical and biological information routinely collected in hospitals and research laboratories, which impacted our fundamental understanding of biological systems, but which also required new approaches based on statistical analysis to extract biological information from large collections of data.

We focus in this thesis on a particular type of molecular data that can be measured genome-wide, namely, epigenetic data describing the methylation status of particular

bases in DNA, and their relevance for cancer diagnosis and prognosis. In this chapter, we provide a general introduction to the main biological notions used throughout this thesis, and highlight the clinical context motivating the work. In particular, we give in section 1.3 a general introduction to cancer and the current perspectives in cancer therapy. We focus on breast cancer, the main cancer treated at the Institut Curie where I worked during my PhD. In section 1.4, we touch upon epigenetics as "heritable changes not affecting the DNA coding sequence but that affect gene function" [Riggs and Porter, 1996], with a particular focus on DNA methylation. In section 1.5, finally, we summarize the main contributions of this thesis.

1.3 From a macroscopic to a molecular characterization of cancer

Cancer is a major cause of morbidity and mortality worldwide, accounting for 8.2 million deaths in 2012. It occurs when a single cell acquires the ability to reproduce aggressively and to invade other tissues. This phenomenon usually results from successive modifications that alter the function of normal cells and give them specific advantages in favor of uncontrolled proliferation and ability to spread out of the tissue of origin [Knudson, 1971, Hanahan and Weinberg, 2000, Weinberg, 2007, Hanahan and Weinberg, 2011]. Although our understanding of when and where these specific aberrations appear during tumorogenesis has greatly improved over the years, many mechanisms remain elusive.

A difficulty in cancer research is the diversity of diseases it encompasses. Physicians recognize at least 200 types of cancer, with very diverse aspects and clinical implications. Not only does cancer occur in various types of tissues, which greatly affects the patient prognosis (*e.g*, overall survival of breast cancer patients after 5 years is around 90%, but only around 10% for lung cancer patients), but even tumors originating from the same type of tissue can present different characteristics under a microscope and in terms of prognosis and response to treatments.

Understanding and delineating the diversity of cancer is increasingly recognized as a critical issue to improve its treatment. On the one hand, understanding which biological processes are involved during carcinogenesis for a particular subtype of cancer can improve our understanding of the disease at a molecular level, and suggest new drugs targeting new targets. On the other hand, it may also contribute, from a clinical point of view, to give a better characterization of which cancer subtypes can be associated with which specific outcome and respond to which treatment. This should help develop more personalized therapeutic strategies that could be more efficient to a subgroup of patients than the current, still largely "one-size-fits-all"-based approach.

1.3.1 Histopathology of cancer: the premisces of personalized medicine. A focus on breast cancer.

Let us focus more precisely on breast cancer, the most common cancer in women worldwide, which by itself is a very heterogeneous disease. Like most cancers, breast cancer can be divided into different categories based on different criteria, serving different purpose. A very important classification scheme is based on the histopathology of the tumor, that is, how biopsy specimens look like under the microscope. The World Health Organization (WHO) approved in 2003 a histopathological classification of breast cancers into more than 20 major tumor types and subtypes, a few of them being shown in figure 1.1. However, two classes, invasive ductal carcinomas (IDC) and invasive lobular carcinomas (ILC), account for approximately 80% of all breast cancers [Li et al., 2005], suggesting that this classification is not very fine-grained. In fact, the histopathological classification of breast cancer has limited prognostic and predictive value, except for some rare subtypes with clear positive (adenoid-cystic carcinomas [Arpino et al., 2002]) or negative prognosis (metaplastic carcinomas [Colleoni et al., 2012]). Patients within the major subtypes can have very diverse prognostics, while the difference between ILCs and IDCs in terms of positive or negative clinical impact is still subject to debate [Viale et al., 2009].

Overall, the histopathology of cancer therefore fails to grasp the full diversity of breast cancer and has a limited clinical impact in itself, except for a few specific subtypes. Yet it is a first, useful step towards unraveling the complexity and the heterogeneity of cancer.

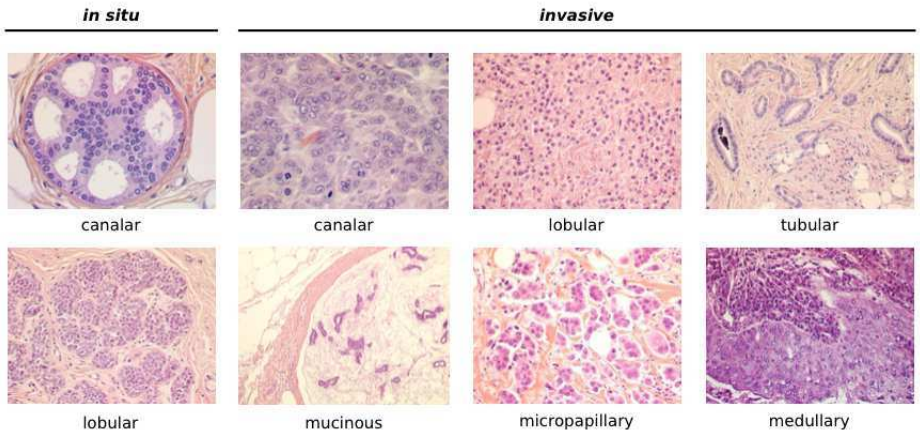


FIGURE 1.1: Histopathology of tumors can distinguish tumors following [Ellis et al., 1992]. Source: Dr. Anne-Vincent Salomon (Institut Curie).

1.3.2 Molecular classifications of cancers: the dawn of personalized medicine

The recent progress in cancer treatments and therapies, such as breast conserving therapies or hormonotherapies, has led to rethink cancer classification in terms of risk assessment (including risk of relapse, progression, metastasis, or survival) and responsiveness to treatments, instead of the classical histopathological classification. The presence or absence of molecular markers such as the presence of estrogen (ER), progesterone (PR) and human epidermal growth factor (HER2) receptors which condition the response to targeted therapies (*e.g* tamoxifen for ER+/PR+ patients and trastuzumab for HER2+ patients) has reorganized the breast cancer groups in terms of treatments. Similarly, the beneficial impact of chemotherapy, which is a very toxic therapy that can often be spared to patient with very good prognosis, can be evaluated from markers of the tumor aggressivity such as the grade, its size and the level of proliferation as measured by molecular markers like Ki67. These markers are routinely assessed by immunohistochemistry on biopsy samples, and the resulting classification of cancers is therefore referred to as the immunohistochemical (IHC) classification. General guidelines and recommendations to specify the IHC classification were assessed and standardized by the American Society of Clinical Oncology (ASCO) in 2007 [Harris et al., 2007].
While of clinical use, several drawbacks to the IHC classification still remain unanswered. First, the clinical impact of these classifications could certainly still be improved with a finer classification. In particular, patients outcome still vary greatly within each IHC tumor groups. Second, the triple-negative breast cancers (ER-/PR-/HER2-) lack all molecular components that could make them benefit from existing targeted therapies for breast cancers. Extending the features used to classify tumors is therefore necessary to assess the effect of new therapies.

1.3.3 The overflow of *-omics* data and the necessity of a statistical framework

In the last 15 years, new technologies to measure thousands or millions of molecular characteristics on each given sample have emerged. Based on microarray or sequencing technologies, they have slowly but steadily transitioned cancer classification from a macroscopic to a molecular level. Being able to measure simultaneously the expression of thousands of genes has paved the way to a better understanding of cancer heterogeneity and to a new molecular classification of breast cancers [Perou et al., 2000, Sørlie et al., 2001, Van't Veer et al., 2002, van de Vijver et al., 2002, Wang et al., 2005]. This classification has already impacted the patients treatments with the recent development of gene

expression profiling platforms (*e.g* MammaPrint, OncotypeDX), which predict the risk of relapse or of treatment response of a patient by combining the expression level of a panel of genes known as a *molecular signature* to aid in the therapeutic strategies [Paik et al., 2004, Parker et al., 2009, Nielsen et al., 2010].

Yet, several statistical and biological issues remain unanswered and question the validity of such methods. Reyal et al. [Reyal et al., 2008] have shown that while of similar performances, different gene-based predictors do not share the same prognostic group assignment. Venet et al. [Venet et al., 2011] demonstrated that most random panels of genes are significantly associated with breast cancer outcome, questioning the biological implications of existing panels. From a statistical point of view, the analysis of such data is usually hindered by the small number of samples available, generally a few hundreds, compared to the thousands of gene expression measurements. This statistical issue commonly referred to as the "small n, big p" issue in the statistical community (where n refers to the sample size and p to the number of features) raises important challenges such as stability and reproducibility of the results [Haury et al., 2011], which we will discuss in more detail in chapter 2.

In summary, the heterogeneity of cancer, while adding a supplementary layer of complexity to the already difficult understanding of the biology of the disease, provides great opportunities for the specific tayloring of therapeutic strategies to the patient, and raises important methodological challenges. While a "perfect" classification from a biological and clinical point of view still remain elusive, we can expect important progress in the coming years in our ability to classify tumors and stratify patients, as we collect more data and improve our methodological approaches to analyze them. In the next chapter, we discuss the clinical and biological interest of specific biological markers, namely DNA methylation markers.

1.4 Epigenetics

The behavior of a cell mostly depends on the proteins it synthesizes, which are themselves governed by the specific regulations of gene expression. While the genome sequence of all somatic cells in an individual is virtually the same, functional variations are therefore controlled by determinants of the gene expression. Transcription factors are proteins that promote or repress gene activity by binding to promoter regions in DNA [Vaquerizas et al., 2009]. Yet, several studies have underlined the insufficiency of a "transcription factors only" model to control gene expression [Itzkovitz et al., 2006, Werner, 2013]. In

particular, recent research has highlighted the crucial role played by *epigenetic* mechanisms in many cellular process including cell differentiation during development [Laurent et al., 2010, Smith and Meissner, 2013] but also in tumorogenesis [Rountree et al., 2001, Ehrlich, 2002, Das and Singal, 2004, Kulis and Esteller, 2010]. The current definition of epigenetics is "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [Riggs and Porter, 1996]. The main epigenetic landmarks in mammals are histone modifications and DNA methylation, the latter being the subject of this thesis. The following subsections will therefore give an introductory description to the biological concepts of DNA methylation, and to its role in gene expression regulation particularly during tumorogenesis. We will also discuss the clinical interest in DNA methylation as an early biomarker in cancer but also as a potential source for treatments. Finally, we will discuss the analysis of DNA methylation data from a statistical point of view.

1.4.1 DNA methylation

DNA methylation refers to the addition of a methyl (CH3) group to a nucleotide in the DNA sequence (figure 1.2). For mammals, this reaction, which is catalyzed by DNA methyltransferases (DNMTs), mostly occurs in the sequence context of a cytosine which is followed by a guanine noted as 5'CG3' or CpG. Three DNMTs have been identified in mammals: DNMT1 guarantees the maintenance of methylation from the methylated parental strand to the unmethylated daughter strand during cell division [Kho et al., 1997], while DNMT3a and DNMT3b trigger *de novo* methylation or demethylation and are specifically important during embryonic development for the establishment of the epigenome [Okano et al., 1999].

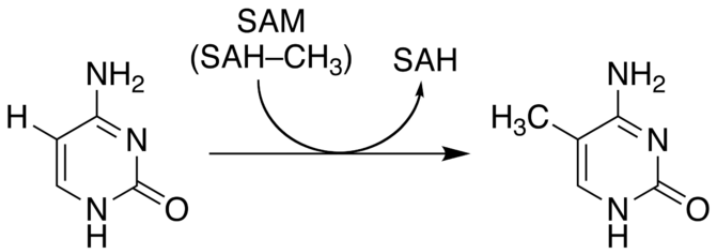


FIGURE 1.2: DNA Methylation of a cytosine by DNMT. SAM: S-adenosylmethionine is a metabolite present in cells and used as a coenzyme in the transfer of the methyl-group. Source: Alice Pinheiro (Institut Curie).

1.4.2 DNA Methylation in gene regulation

CpG sequences are greatly under-represented on average across the genome, but are specifically over-represented in condensed regions of 1 to 5 kb known as CpG Islands (CGIs) [Gardiner-Garden and Frommer, 1987]. The original definition of CGIs based on somehow arbitrary thresholds has been followed to several redefinitions based on either biological or statistical considerations [Takai and Jones, 2002, Wang and Leung, 2004, Wu et al., 2010, Bock et al., 2007]. A particular interest for these regions is due to the fact that 60 to 90% of all genes are associated to CGIs, specifically in their promoter regions, [Saxonov et al., 2006]. The methylation or demethylation of a CGI is known to be related to the initiation of the transcription process of the associated gene, as shown by many studies for specific genes such as housekeeping genes [Deaton and Bird, 2011], imprinted genes [Li et al., 1993], and tissue specific genes [Laurent et al., 2010]. One famous example of gene expression regulation by DNA methylation is the inactivation of one of the two copies of the X-chromosome by DNA methylation [Pollex and Heard, 2012].

Still, the precise mechanism of DNA methylation and its role in gene transcription remains largely unclear. DNA methyltransferases (DNMTs) are responsible for *de novo* methylation and for the maintenance of methylation after cell division [Bird, 2002]. Yet, the signals that govern the pattern of methylation accross the genome is unknown. Moreover, methylation is tightly linked with gene expression but how it clearly regulates expression is still being debated and several hypotheses have been proposed [Klose and Bird, 2006, Bogdanović and Veenstra, 2009, Deaton and Bird, 2011]. A first model is that DNA methylation physically blocks the access of promoter binding sites in particular for specific transcription factors that bind preferably to unmethylated sequences [Rodriguez et al., 2010] 1.3. A second model considers DNA methylation as the initiating mechanism for the establishment of an inactive chromatin state also known as *heterochromatin*. In this case, specific proteins known as methyl-CpG binding domain proteins (MBDs) bind to region of high methylation. These MBDs, in turn, recrute histone deacetylases (HDACs) which compact the chromatin and enforce a inactive state which results in gene silencing (figure 1.3).

While gene promoter methylation has been quite extensively studied, little is known on the role of methylation outside of these regions. Gene body methylation has been positively correlated with gene expression contrary to promoter methylation [Kulis et al., 2013] and has also been associated with alternative splicing [Maunakea et al., 2013]. The role of orphan CGIs, that is CGIs far from any known genes, is yet to be elucidated but could be linked to long range epigenetic regulation [Bert et al., 2013] or be located in actual promoter regions of ancestral genes [Illingworth et al., 2010].

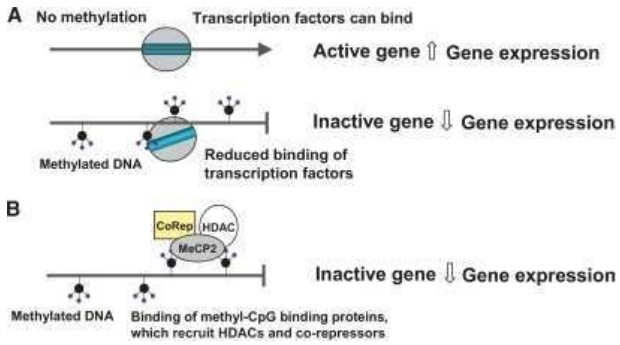


FIGURE 1.3: **Panel A.** DNA methylation physically blocks the access of promoter binding sites and prevent the binding of transcription factors. **Panel B.** Methyl-CpG binding domain proteins (MBDs) bind to region of high methylation which in turn, recruit histone deacetylases (HDACs) which compact the chromatin into an inactive state. Source: [Ling and Groop, 2009] (Copyright: CC BY-NC-ND 3.0)

1.4.3 An early biomarker in cancer and a source for potential treatments

DNA methylation is essential in cell development and differentiation [Smith and Meissner, 2013]. Therefore, abnormalities can lead to several diseases. Disruption of genomic imprinting where one of the paternal or maternal allele is not expressed (*e.g* Prader-Willi Syndrome: deletion of the paternal contribution of seven genes on chromosome 15, Angelman Syndrome: deletion of the same region for the maternal copy, Beckwith-Wiedmann: aberrant parental imprinting on chromosome 11) are strong evidence of the causal link between methylation aberrations and human diseases [Robertson, 2005].

Cancer is the epitome of those methylation-associated diseases. Our current understanding of the role of methylation aberrations in cancer points out to at least two mechanisms. On the one hand, localized hypermethylation of promoter regions of specific genes such as tumor suppressor genes or tissue-specific genes can lead to their inactivation and lead to tumorogenesis [Baylin and Herman, 2000, Esteller et al., 2001]. On the other hand, an overall global hypomethylation can lead to genetic instability and in some cases to the activation of silenced oncogenes [Ehrlich, 2002, Ehrlich, 2009, Hon et al., 2012]. In both cases, abnormal methylation is considered a driver event for abnormal gene expression in cancer, and could therefore potentially be detected before any significant change in gene expression. Thus, understanding the general pattern of a cancer methylomes could pave the way to new schemes for early detection of cancer. This may in turn impact cancer treatment, knowing that time of diagnosis is a crucial factor in prognosis [Richards, 2009].

Several advantages of epigenetic markers over genetic markers place DNA methylation as one of the major interest in cancer research. Current measurement technologies allow for

non-ambiguous mapping of methylation at well defined position on the genome contrary to, *e.g*, mutations of genes. In addition, the fact that DNA can be released from tumor tissues in peripheral tissues and in particular fluids [García-Closas et al., 2013] such as serum, urine or plasma could allow for non-invasive early detection procedures. The reversibility of the methylation marker is also of potential interest for new treatments. Already, demethylant agents such as the 5-aza-2'-deoxycytidine or the 5-azacytidine have already been tested and approved by the FDA to treat myelodysplastic syndromes (MDS) and chronic myelomonocytic leukemia (CML) [Silverman et al., 2002].

1.4.4 Statistical challenges in DNA methylation analysis

Epigenome-wide analyses have become accessible with the development of microarray measurement technologies, and more recently sequencing technologies. We focus in this thesis on bisulphite-based methods such as the illumina HumanMethylation platform, which measures the methylation level of up to 450,000 CpG dinucleotides. Although the methylation of a given CpG in a given cell is a binary attribute, measurements are often issued from a mixture of cell populations with heterogeneous DNA methylation profiles. Therefore, the resulting measurements usually reflect a ratio of methylation for one specific probe as $M/(M+U)$ in which M represents the signal for methylated molecules and U the signal for unmethylation molecules.

This ratio lies in $[0;1]$ and the finite scale of DNA methylation greatly differ from the larger scale of, e.g., gene expression data. Also, DNA methylation measurements are not normally distributed and variance is greatly biased by the mean value of the probe (probes with mean methylation of 0.5 can have variance much larger than probes with mean methylation near 0 or 1). Therefore, the use of standard methods in microarray data analysis such as filtering signal with high standard deviation becomes bias-inducing in DNA methylation analysis. Such observations underly the importance of understanding the data at hand and the underlying technology to build data-driven statistical methods.

1.5 Personal contribution and organization of the thesis

This chapter presented a short overview of some of the current problematics in cancer research. Given the heterogeneity and complexity of the data now available to investigate the diversity of cancers, we chose to focus on the specific role of DNA methylation in tumorogenesis. Each of the following chapters aim at tackling a particular issue in cancer described below.

Chapter 2 introduces the different methods used in this thesis to analyse data. In particular, we discuss the use of supervised and unsupervised learning in a setting where the number n of observations is much smaller than the number p of variables studied p also known as $n \ll p$, the challenges it raises and how to overcome them.

Chapter 3 tackles the issue of relapses in breast cancer. When early breast tumor are detected, an alternative to aggressive treatments such as mastectomy is tumorectomy where only the tumor is removed instead of the whole breast. Although patients survival rates are not significantly different for both therapies, tumorectomy increases the rates of tumor relapse. In case of relapse, being able to characterize the relapse as either a true recurrence or an independent tumor is essential for the treatment of the patient. A true recurrence is often synonym of an aggressive cancer and requires aggressive treatments, while an independent tumor could potentially be treated with less invasive treatments. The monoclonality of cancer [Weinberg, 2007] suggests that being able to characterize the clonality between a primary tumor and its relapse could help tackle this issue. A clinical classification based on the concordance of the histopathological features (stage, grade, ER status, PR status, HER2 status) is used in practice but yields the same drawbacks as the clinical classification of breast cancer subtypes (see section 1.3.1). While a few studies have investigated the use of pangenomic data to tackle breast cancer relapse classification, no method is based on DNA methylation profiles. As methylation is highly conserved in cell division, we hypothesize that it may be a good marker to assess lineage between samples. We therefore investigate the similarity of methylation profiles in different cancer samples, and propose a method based on pairwise analysis of methylation profiles to characterize clonality between samples taken at diagnosis and at relapse.

Chapter 4 tackles the role of genome-wide variations of methylation in gene expression regulations. Aberrant promoter hypermethylation has frequently been observed in cancer but its precise role in tumorogenesis has always been elusive. Epigenetic modifications have been widely studied and have been shown to be associated to gene expression repression in tumour suppressor genes. The high-coverage methylome profiles of hundreds of patients, as well as their matched gene expression and copy number profiles, now available publicly in the cancer genome atlas (TCGA) provides a comprehensive dataset to assess the extent of epigenome-wide regulation of gene expression variations.

Chapter 5 studies the existence of a methylome-based cancer classification. The CpG island methylator phenotype (CIMP), first identified in colorectal cancer, has recently become a major subject of interest and has been observed in several tissues. Yet, these characterizations of CIMP have been tissue-specific and the existence of a biological phenomenon causing the CIMP in cancer is still elusive. In addition, the clinical importance

of CIMP as a predictive factor for prognosis and patients response to treatments is still being validated. We develop a pancancer genome and epigenome-wide CIMP analysis using the large TCGA datasets and demonstrate the existence of a common epigenetic signature of CIMP. Genetic profiling show that CIMP might be linked with a universal genetic signature well-documented in several CIMPs. However, clinical impact of CIMP is still lacking on the TCGA database.

Chapter 2

Methods

Résumé

Ce chapitre présente les méthodes et algorithmes généraux utilisés au cours de cette thèse. Ces éléments sont disponibles dans la littérature mais sont rappelés ici pour rappel et pour permettre une discussion sur les méthodes employées pour l'analyse de données réelles. Ce chapitre est essentiellement composé de 2 sections.

La première section traitera de méthodes d'analyses supervisées particulièrement dans le cadre où le nombre d'échantillons disponibles est beaucoup plus faible que le nombre de variables observées. Nous discuterons particulièrement du problème d'interprétabilité dans un contexte biologique où la compréhension du phénomène biologique est aussi importante que la performance prédictive d'un modèle.

Dans une seconde partie, nous discuterons de méthodes d'analyses non supervisées. Nous décrirons les méthodes de clustering ainsi que de réduction de la dimensionalité. En particulier, nous montrerons la difficulté de sélectionner efficacement et objectivement un modèle (*e.g* nombre de clusters ou nombre de dimensions) dans le cadre de données réelles. Enfin, nous verrons que dans le cadre de données biologiques où le bruit ainsi que le cadre statistique ($n \ll p$) rendent difficiles la détection de clusters robustes, l'impact clinique d'un point de vue pronostic peux jauger de l'importance du modèle.

Abstract

In this chapter, we introduce the statistical methods related to the work present in this manuscript. Most of the methods described here are well discussed in the literature.

Our focus will be on their practical use for biological data analysis. This chapter is essentially composed of 2 sections.

The first section will present a few methods used in this thesis for supervised learning in the context where the number of samples is a lot smaller than the number of features also known as $n \ll p$. We will focus in particular on the interpretability of the data especially in a biological context where the understanding of the underlying biological phenomenon is as important as the prediction performance of a model.

In the second section, we will discuss unsupervised learning methods. We will describe cluster analysis as well as dimensionality reduction techniques. We will show in particular how difficult the task of selecting objectively a model (*e.g*, number of clusters or optimal number of dimensions) can be in the case of real data. Finally, we will see that model selection in a biological setting, where the data are intrinsically noisy and the statistical power is usually poor ($n \ll p$), generally leads to non-robust clustering. In this case, clinical significance of the method in particular for clustering to distinguish classes of different prognosis can be used to measure the importance of the model.

2.1 Supervised learning

Supervised learning refers to a set of statistical methods which try is to make sense out of a series of observations by inferring a relation between an input and an output. The observations can usually be summarized by a set of variables also known as *inputs* represented by X which are usually measured (*e.g* the level of expression of genes), and by an outcome also known as *output* represented by Y which is a feature of interest. Outputs can either be qualitative (*e.g.*, whether or not a relapse will occur with 5 years), or quantitative (*e.g*, toxicity of a drug). We usually refer to a supervised learning problem with qualitative outputs as a *classification* problem compared to a *regression* problem for quantitative outputs. To summarize, the objective is to learn a model that will allow to predict Y given X only. Such a setting can be beneficial for example when measuring Y can be a lot more costly than measuring X (*e.g* annotating each picture on the internet is harder to obtain than the summary of a picture as a set of pixels), or more importantly when Y refers to a future event that we would like to predict in order to adapt our present strategy. Another reason to infer a predictive model is when we are interested in *how* the input influences the output, *e.g* how the genes influence the state of a patient as *healthy* or *cancerous*.

In the following, we will give the general mathematical framework of statistical supervised learning as well as the general set of notations used throughout this thesis. In

particular, we will discuss the issue of learning when the number of observations is small compared to the number of training samples.

2.1.1 Risk minimization problem

Let's suppose that the couple $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ follow a joint probability distribution $\mathbb{P}(X, Y)$. Being able to predict Y given X can be formulated as the problem of finding a function $f : \mathcal{X} \to \mathcal{Y}$ living in a certain space $\mathcal{F}$ such that $\hat{Y} := f(X)$ is a good enough approximation of Y. For that, we first define a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ to quantify the loss $l(\hat{y}, y)$ incurred by a prediction $\hat{y}$ when the true output is y . The risk $R : \mathcal{F} \to \mathbb{R}$ of a function f is defined as the expected loss incurred by the predictions made by function on future samples under the distribution $\mathbb{P}$, i.e;

$$R(f) = \mathbb{E}_{\mathbb{P}} \left[l(f(X), Y) \right]. \tag{2.1}$$

If $\mathbb{P}$ was known, then arguably we should make predictions with the function that will incur the smallest loss, i.e., the one with the smallest risk:

$$f^* = \arg\min_{f \in \mathcal{F}} R(f). \tag{2.2}$$

For example, a common loss function used in regression is the *squared loss* $l(\hat{y}, y) = (\hat{y} - y)^2$, in which case the optimal predictor is

$$\begin{aligned} f^* &= \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)} \left[(Y - f(X))^2 \right] \\ &= \arg\min_{f \in \mathcal{F}} \mathbb{E}_X \mathbb{E}_{Y|X} \left[(Y - f(X))^2 | X \right] \end{aligned} \tag{2.3}$$

which can be solved point wisely by

$$\begin{aligned} \forall x, \quad f^*(x) &= \arg\min_{c \in \mathbb{R}} \mathbb{E}_{Y|X} \left[(Y - c)^2 | X = x \right] \\ &= \mathbb{E} \left[Y | X = x \right]. \end{aligned} \tag{2.4}$$

2.1.2 The curse of dimensionality

Let's describe with an example the different problematics that we might encounter when trying to solve 2.2.

Suppose we are given a training set n of observations $(x_i, y_i)_{i=1,\dots,n}$ where the input variable x_i are p-dimensional vectors in $\mathbb{R}^p$ and the output variable y_i is a real scalar in $\mathbb{R}$. Our objective is to estimate a function $\hat{f}$ that is a good approximation of f^*. A

standard procedure known as the k-nearest neighbours to estimate $f^*(x)$ at a point x consists in averaging the observed outputs $y_i s$ for the k closest $x_i s$ in a neighboorhood of x, that is:

$$\hat{f}(x) = \frac{1}{k} \sum_{i:x_i \in \mathcal{N}_k(x)} y_i, \tag{2.5}$$

where $\mathcal{N}_k(x)$ is the set of the k closest observations from x.

In the low-dimensional setting, when the number n of observations is very large then $\mathcal{N}_k(x)$ tend to be very close to x, making or the k-NN estimator (2.5) a good local estimator of $f^*(x)$ particularly when k is large enough to average out the noise from sufficient local neighbors. When the dimension of the input space p is large, however, observations tend to be far away from each other. To see this, suppose we have n points uniformly distributed in a unit ball in $\mathbb{R}^p$; then the median distance between the origin and the closest point is given by:

$$d(n,p) = \left(1 - \frac{1}{2}^{1/n}\right)^{1/p}, \tag{2.6}$$

which increases to 1 as p increases. Therefore, in large dimension data points tend to be very far from each other (Figure 2.1). Another way to see this phenomenon is to notice that the volume of a ball of radius R is proportional to R^p and therefore the density is proportional to $n^{1/p}$, which means that to obtain the same density of points, the number of observed points needs to grow exponentially with the dimension. This problem, usually referred to as the *curse of dimensionality* [Bellman, 1961], entails that estimators like k-NN can become arbitrarily bad since k (and therefore n) should increase exponentially with p to have a good representation of the neighborhood of x.

Yet in several biological problems, the number of observed points n is generally ~ 100 (*e.g* number of patients in a study) while the dimension of the input space p can be of the order of 10^4 (*e.g* number of genes) to 10^6 (*e.g* SNP data). In the following, we will discuss the different strategies to tackle the problem of learning when the dimension p is much larger than the number of samples n also known as $n \ll p$.

2.1.3 Model selection

Let's suppose that the output Y is related to the input X by

$$Y = f(X) + \epsilon, \tag{2.7}$$

where ϵ is a normally distributed random variable with zero mean and variance σ^2 independent of X. The expected squared prediction error of a given estimate function

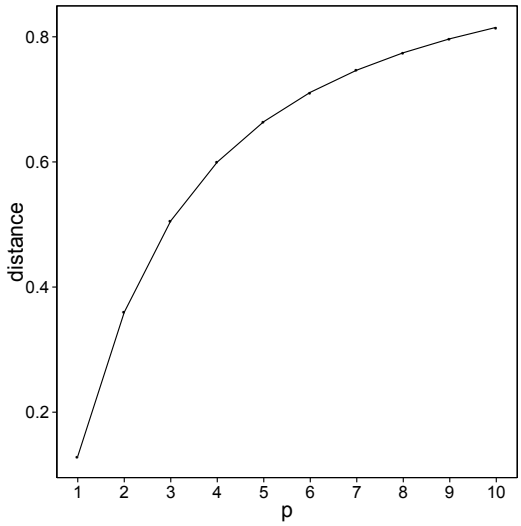


FIGURE 2.1: Median distance between the origin and the closest data point as a function of the dimension of the space ($n = 5$) when samples are uniformly distributed in the unit ball.

$\hat{f}$ at a given x_0 is given by:

$$Err(x_0) = \mathbb{E}\left[(Y - \hat{f}(x_0))^2 | X = x_0\right], \tag{2.8}$$

which can be decomposed as:

$$\begin{aligned} Err(x_0) &= \left(f^*(x) - \mathbb{E}[\hat{f}(x)]\right)^2 + \mathbb{E}\left[\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right]^2 + \sigma^2 \\ &= Bias^2 + Variance + Irreducible\ Error. \end{aligned} \tag{2.9}$$

Amongst the three terms in the right-hand side of (2.9), two can be controlled by the choice of $\hat{f}$, that is, the choice of the modelisation of the relationship between X and Y: the bias and the variance terms. To have a small bias, one typically needs to have a good local estimator, like a k-NN with small k. To have a small variance, one needs to have a procedure that is not too sensitive to individual observations, like k-NN with a large k. As suggested by the k-NN examples, bias and variance generally move in opposite way, and learning in high dimension often boils down to controlling the trade-off between bias and variance.

More generally, if we had an infinite number of observations, then we might be able to reduce both the variance and the bias terms to 0. When the number of sample n is finite, however, we do not have access to the prediction error. An intuitive estimate to estimate the risk of a candidate function f is then to take the *training error*, namely $\frac{1}{n}\sum_i (y_i - f(x_i))^2$. However, simply minimizing the training error over f is not a good idea because it does not account for the complexity of the model f. As the model gets

more and more complex, the training error can decrease and even tend to zero if one finds a function $\hat{f}$ that perfectly reproduces the training examples, i.e., $y_i = \hat{f}(x_i)$ for all $i = 1, \ldots, n$. Yet, a too complicated model might not give the best output for a new observation x, since it may have large bias. On the other hand, a model too simple, with small variance, might be too biased and might also not be able to give a good prediction for a new observation. At the end, a balance between the complexity of the model and its ability fo capture robust information from the training data must be found (figure 2.2).

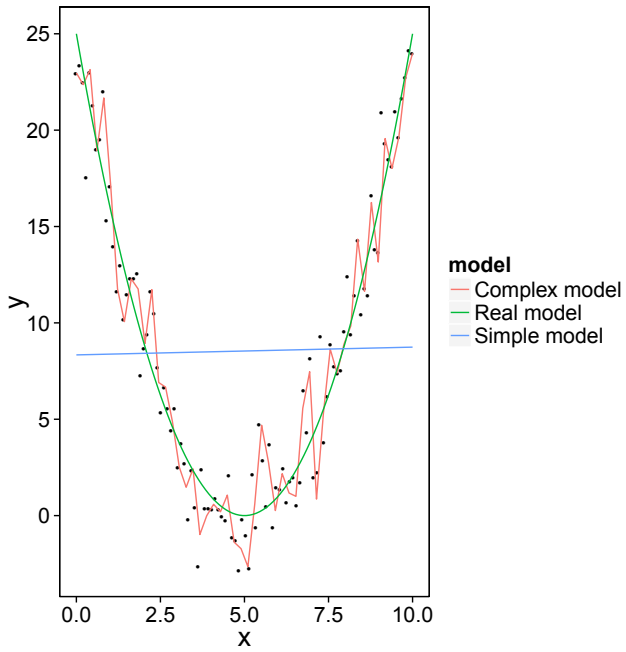


FIGURE 2.2: **Bias-Variance Tradeoff.**. A too simple model might be less sensitive to the noise but is too simplistic to predict Y given X. On the other side, a complicated model, while being a good estimation of Y on the observations, might not be able to give a good prediction on new data points.

2.1.4 Assessing the performance of a model

In the last part, we illustrated the issue of estimating the prediction error by the training error which leads to overfitting. One of the most commonly used method for estimating the prediction error is cross-validation [Stone, 1974, Allen, 1974]. Cross-validation randomly splits the data into K folds of even sizes. A model is trained on $K - 1$ folds and an estimate of the prediction error is then obtained by taking the average error on the K^{th} fold.

For a given fold J, let's define :

$$CV_J(\hat{f}) = \frac{1}{|J|} \sum_{i \in J} L(y_i, \hat{f}^{-J}(x_i)) \tag{2.10}$$

where $|J|$ is the cardinal of fold J and $\hat{f}^{-J}$ is the fitted function computed on the dataset where the observations belonging to the J^{th} fold are removed.

Finally, the cross validation estimate of the prediction error is given by :

$$CV(\hat{f}) = \frac{1}{n} \sum_{J \in [1;K]} |J| CV_J(\hat{f}) \tag{2.11}$$

A set of models f usually involves a tuning parameter α that controls the complexity. A common model selection procedure is given by finding α that minimizes the cross validation estimate:

$$\hat{f}_{CV} = \arg\min_{\alpha} CV(\hat{f}_\alpha) \tag{2.12}$$

An issue that arises in cross-validation is the choice of K. Choosing K small might give a better estimation of the expected error as the training sets in each fold are very different (when $K = 2$ the training sets do not overlap), contrary to choosing K very large where the training sets tend to be very similar (for *leave-one-out* cross-validation, that is $K = n$ the training sets in each fold differ by one observation). However, in a biological context, one also has to take into account the size n of the observations that can be relatively small. In this case, we usually choose K large to reduce the variance of the estimate.

2.1.5 Interpreting the data

In the following, we restrict the set of functions $\mathcal{F}$ to the set of affine functions that is:

$$f(x) = x^\top \omega + \omega_0, \tag{2.13}$$

with $(\omega, \omega_0) \in \mathbb{R}^{p+1}$. To simplify (2.13), we usually "integrate" the constant ω_0 in x by defining artificially the new set of features $x := [1; x]^T \in \mathbb{R}^{p+1}$. Optimizing in f is therefore equivalent to finding the vector of coefficients ω.

Ordinary Least Squares. Given a training set $(x_1, y_1), \cdots, (x_n, y_n)$, a popular estimate of the vector of coefficients ω is given by the *ordinary least squares* methods

that is the vector of coefficients that minimizes the residual sum of squares:

$$RSS(\omega) = \|\mathbf{y} - \mathbf{X}\omega\|^2, \tag{2.14}$$

where $\mathbf{X}$ is the matrix containing all the observations $\mathbf{X} = [x_1; \cdots ; x_n]^T$ and $\mathbf{y}$ is the vector of outputs $\mathbf{y} = (y_1; \cdots ; y_n)^T$.

Minimizing in ω in the case where $\mathbf{X}^T\mathbf{X}$ is non-singular is given by the unique solution:

$$\omega_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.15}$$

Ridge Regression. This estimate is not always an option. In a biological setting, the number of observations p is usually bigger than the number of samples n and thus $\mathbf{X}^T\mathbf{X}$ is singular and the number of solutions is a vector space.

One solution is to modify (2.14) by adding a penalty constraint based on the Euclidean norm of the vector ω, also known as the *ridge regression* [Hoerl and Kennard, 2000]:

$$Ridge(\omega) = \|\mathbf{y} - \mathbf{X}\omega\|^2 + \lambda\|\omega\|^2, \tag{2.16}$$

for $\lambda > 0$. The solution of (2.16) is given by:

$$\omega_{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.17}$$

Ridge regression has thus the interesting property to get rid of the non-singularity issue on $\mathbf{X}^T\mathbf{X}$ by replacing it with $\mathbf{X}^T\mathbf{X} + \lambda I$. In addition, the penalty term $\lambda\|\omega\|^2$ enforces smoothness that is the coefficients to be not too large.

Sparsity-inducing penalties. Another important issue in a biological setting is the interpretability of the results given by the method. Suppose for example that we are interested in predicting the status of a patient (*e.g*, healthy or cancerous) given the expression level of all the genes. A ridge regression estimate on a training set of data containing healthy and cancerous patients might perform well on a new set of data, but may not give particular information about the biological pathways involved in tumorogenesis. The underlying assumption is that amongst the whole set of genes ($\sim$ 25000), not all genes might be involved in defining the status of the patient. In addition, it would take too long to verify biologically all the genes one by one.

An alternative is thus to seek a good model ω with many zero coefficients, entailing that only a subset of features is used in the decision making. This would allow us, for example, to target specifically a small set of genes that can be more easily tested.

Moreover, reducing the number of features used by the model is a way to control the bias-variance trade-off and improve the generalization ability of the model. Such an approach could for example be carried out by penalizing 2.14 by the number of non-zero coefficients instead of the ridge penalty:

$$l_0(\omega) = \|\mathbf{y} - \mathbf{X}\omega\|^2 + \lambda\|\omega\|_0, \tag{2.18}$$

where $\|\omega\|_0 = \#\{i : \omega_i \neq 0\}$. However, solving (2.18) requires an exhaustive search over all possible combinations of p features, a combinatorial problem which becomes quickly intractable for p larger than a few tens. A popular approach to overcome this computational issue is to replace the $l0$ regularization term by the convex $l1$-norm, leading to the Lasso estimate:

$$Lasso(\omega) = \arg\min \|\mathbf{y} - \mathbf{X}\omega\|^2 + \lambda\|\omega\|_1, \tag{2.19}$$

where $\|\omega\|_1 = \sum_{i=1}^{p} |\omega_i|$. (2.19) can be efficiently solved by a variety of algorithms, and leads to sparse models where the number of non-zero entries in ω is controlled by λ.

Feature selection and multiple-testing. Another common method to select a small list of genes that should be retained in a predictive model is to assess the significance of each feature given an outcome. This is commonly done by performing univariate tests that compute a p-value representing the likeliness of a feature j to have the same distribution under different assumptions. These tests can either be *parametric* or *non-parametric*, depending on whether we have some assumptions about the distribution of the features.

The **Student t-test** is a popular *parametric* test where we suppose the data to be distributed according to a Gaussian mixture model for each features, that is, for a given feature j and an output $\epsilon \in \{-1; +1\}$:

$$X_j^{\epsilon} \sim \mathcal{N}(\mu_j^{\epsilon}, \sigma_j^{\epsilon}). \tag{2.20}$$

For each feature j, a t-statistic is calculated as follows:

$$t_j = \frac{\bar{x}_j^{+1} - \bar{x}_j^{-1}}{\sqrt{\frac{(\sigma_j^{+1})^2}{N_{+1}} + \frac{(\sigma_j^{-1})^2}{N_{-1}}}}. \tag{2.21}$$

Under the *null* hypothesis $\mathcal{H}0$ that "the feature j follows the same distribution independently of the output", t_j follows a Student distribution and one can derive the p-value

given by:

$$\begin{aligned} p_{right-tailed} &= Pr(X \geq t_j|\mathcal{H}0) \\ p_{left-tailed} &= Pr(X \leq t_j|\mathcal{H}0) \\ p_{two-tailed} &= 2\min\left\{Pr(X \geq t_j|\mathcal{H}0), Pr(X \leq t_j|\mathcal{H}0)\right\} \end{aligned} \tag{2.22}$$

In the case when the data are not normally distributed, one can instead use a *non-parametric* test such as the **Wilcoxon rank sum test** also known as the **Mann-Whitney U test**. The idea is to compare the ranking of the samples given the output. Define:

$$R_j^{+1} = \sum_{i \in \mathcal{C}^{+1}} r_j^i, \tag{2.23}$$

where $\mathcal{C}^{+1}$ is the subset of samples with positive output and r_j^i is the rank of the observation x_i^j, that is, the value of feature j for patient i. The U_j-statistic is then given by:

$$U_j = R_j^{+1} - \frac{n^{+1}(n^{+1}+1)}{2}, \tag{2.24}$$

where n^{+1} is the size of $\mathcal{C}^{+1}$. Under the *null* hypothesis $\mathcal{H}0$, the distribution of U_j is known [Wilcoxon, 1945, Mann and Whitney, 1947] and a p-value can be computed in a similar fashion as (2.22).

Applying either of the statistical tests returns a list of p-values that naturally gives a ranking of the association between each feature and the output. For a single feature, one generally applies a significance level (usually 5%). However, when the number of features is large, applying such a significance level would lead to several falsely detected features. In this case, one has to correct the p-values for *multiple testing* [Benjamini and Hochberg, 1995, Dudoit and Fridlyand, 2002].

Adding prior knowledge. While sparsity-inducing methods are useful as they allow to reduce the effective number of features, they are not always sufficient to overcome the problem of $n \ll p$. For example, correlated variables can produce unstable signatures using lasso. Similarly, univariate tests do not take into account the joint distribution of features.

In a biological setting, one usually has access to additional information about the data such as the existence of biological pathways that can relate genes working together. One method to incorporate this prior knowledge into a model is to generalize the ridge and

lasso regressions as follow:

$$Penalized(\omega) = \|\mathbf{y} - \mathbf{X}\omega\|^2 + \lambda\Omega(\omega), \tag{2.25}$$

where $\Omega : \mathbb{R}^p \to \mathbb{R}^+$ is a penalty function.

Several studies have investigated the choice of Ω such that optimal solution ω^* has some specific properties:

$$\omega^* = \arg\min \|\mathbf{y} - \mathbf{X}\omega\|^2 + \lambda\Omega(\omega). \tag{2.26}$$

We have already seen that the solution of the lasso ω_{Lasso} has the property of being sparse. As discussed above, this allows to select a subset of genes but sometimes fails to provide a stable signature as genes are often correlated. Knowing biological pathways, one can add this prior information by using a "group-lasso" penalty [Yuan and Lin, 2006, Jacob et al., 2009]:

$$\Omega(\omega) = \sum_{g \in \mathcal{G}} \|\omega_g\|, \tag{2.27}$$

where $\mathcal{G}$ is a subset of $\mathcal{P}(\{1; \cdots ; p\})$. For copy-number or methylation profile analysis, fused-penalties can be used as biological evidence suggest a strong correlation between close features on the genome. In addition, as profiles usually share well-defined biological traits, a joint regularization of signals can improve the detection of breakpoints [Vert and Bleakley, 2010], see an illustration in figure 2.3

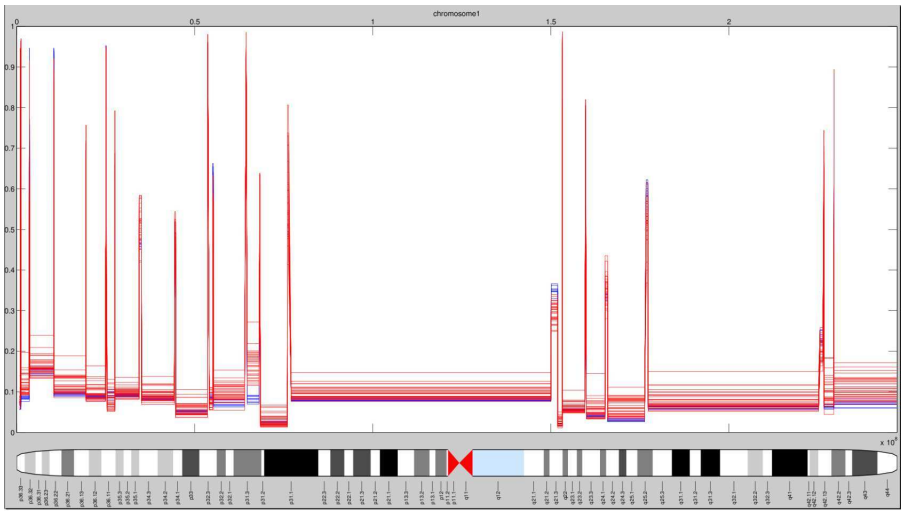


FIGURE 2.3: Joint Regularization of methylation profiles of chromosome 1 using a total-variation penalty show similar breakpoints of hyper and hypo-methylated blocks but different levels for healthy patients (blue) and breast cancer patients (red).

2.2 Unsupervised learning

Unsupervised learning, similarly to its supervised counterpart, also seeks to make sense of a series of observations. However in this case, there is no output Y and the observations are therefore summarized by an input X. The two most common objectives are: *cluster analysis* and *dimensionality reduction*.

Cluster analysis tries to summarize the set of observations by a small number of modes from which the observations are drawn. In other words, the main assumption is that the set of observations is a mixture drawn from a few (generally simple) densities.

Dimensionality reduction can be associated with supervised methods. As seen previously, when the number p of features is high, the number of samples needed n needs to be very high in order to estimate $Pr(X, Y)$. However, while p can sometimes be large, the effective dimension can be much smaller. This is the case for example when most of the data lies in a low-dimensional manifold. In addition, this provides information about the associations between the different features.

The main issue in unsupervised learning is assessing the adequacy of the model. Contrary to supervised learning where one could assess the effectiveness of a method by comparing Y and $\hat{Y}$, the quality of the results in the unsupervised case is often subjective.

In the following, we will discuss a few methods employed in unsupervised analysis for both cluster analysis and dimensionality reduction and illustrate them with specific examples in biology. We will in particular discuss the clinical importance of unsupervised learning despite not having clear model assessment techniques.

2.2.1 Cluster analysis

The main objective behind cluster analysis is to partition the set of observations into K subsets or "clusters". A natural partition is such that observations from a same cluster are more similar than observations from different clusters. These considerations naturally necessitate to introduce a similarity (or dissimilarity) measure over the set of observations.

Choosing a similarity between samples. In general, the set of observations $(x_1, \cdots, x_n)$ lie in a p-dimensional space and one natural dissimilarity measure between two observations is the Euclidean distance between the two vectors in $\mathbb{R}^p$. However other dissimilarity measures exist and can lead to very different clustering results. For

example, one can define a distance using the *Pearson* correlation r as:

$$d_{Pearson}(x_i, x_j) = \frac{1 - r(x_i, x_j)}{2}, \tag{2.28}$$

with

$$r(x_i, x_j) = \frac{\sum_{k=1}^{p}(x_i^k - \bar{x_i})(x_j^k - \bar{x_j})}{\sqrt{\sum_{k=1}^{p}(x_i^k - \bar{x_i})^2 \sum_{k=1}^{p}(x_j^k - \bar{x_j})}} \tag{2.29}$$

can yield significantly different results as illustrated in figure 2.4.

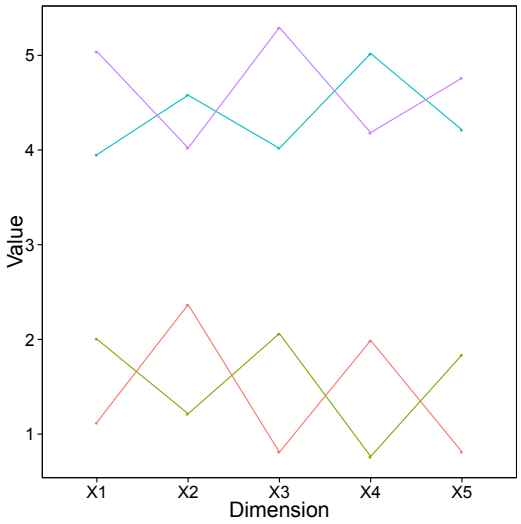


FIGURE 2.4: Similarity between samples is subjective and therefore the clustering procedure might differ. This is a toy example representing 4 samples (red,green, blue,purple) in 5 dimension. A clustering algorithm using the euclidean distance would cluster samples 1 and 2 on one side and 3 and 4 on the other since they are close from a spatial point of view. However using a Pearson distance would cluster samples 1 and 3 on one side and 2 and 4 on the other since their variations are more coordinated (*e.g* the similarity in term of response to a treatment between two proteins can be better explained by the correlation of their abundance over time than by the absolute deviation between abundance over time).

K-means. K-means is a popular clustering methods of the class of partitioning methods which, given a number K, partition the set of samples into K groups or clusters. The goal is therefore to optimize over the set of partitions, a criterion such that the in-cluster similarity is large while the between-cluster similarity is small. In the case of K-means, one seeks a partition $\mathbf{S} = \{S_1, \cdots, S_k\}$ that minimizes:

$$\sum_{i=1}^{K} \sum_{x \in S_i} \|x - \mu_i\|^2, \tag{2.30}$$

where μ_i is the barycenter of the samples in partition S_i. Figure 2.5 illustrate an example of K-means on a toy dataset.

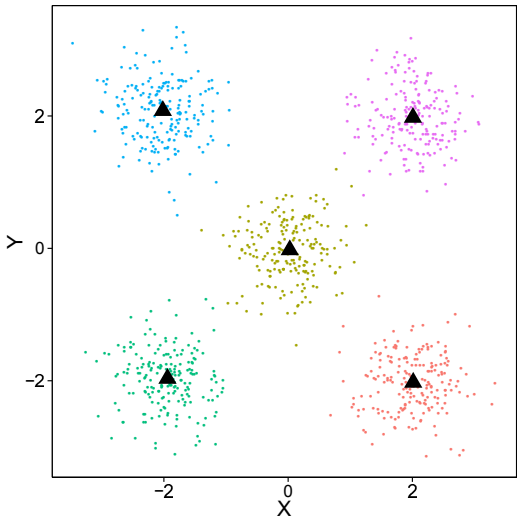


FIGURE 2.5: K-means clustering (k=5) on simulated data sampled following 5 normal distributions. Different colors represent the cluster assignment of each observations while the triangles represents the center of each cluster.

Hierachical clustering: an alternative with several advantages. Partitioning algorithms, such as K-means discussed previously, depend on a choice of the number of clusters K but also have optimization issues (initialization, convergence). At the expense of adding a similarity measurement between groups also known as *linkage*, another type of clustering methods called *hierarchical clustering* does not have the same requirements.

Two main paradigms exists in hierarchical clustering:

- *bottom-up* approaches, where each sample first belongs to its own cluster, and where clusters are merged together iteratively until all samples belong to a unique cluster.
- *top-down* approaches, conversely, start with all samples in a unique cluster and iteratively shatters clusters into two new clusters at each step until all samples are separated.

This particular set of methods yield interesting properties. In particular, visualization can be done using dendrogram, which is a representation of a binary tree where each leaf is a sample and each internal node represents the agglomerative procedure of merging two clusters together. In addition, cutting the dendrogram at a given height yields a clustering of the dataset in K clusters. A particular application in bioinformatics is to perform a clustering on the samples as well as on the features to identify a group of features (*e.g* genes) associated with a group of samples (*e.g* different clusters) (see an example in figure 2.6).

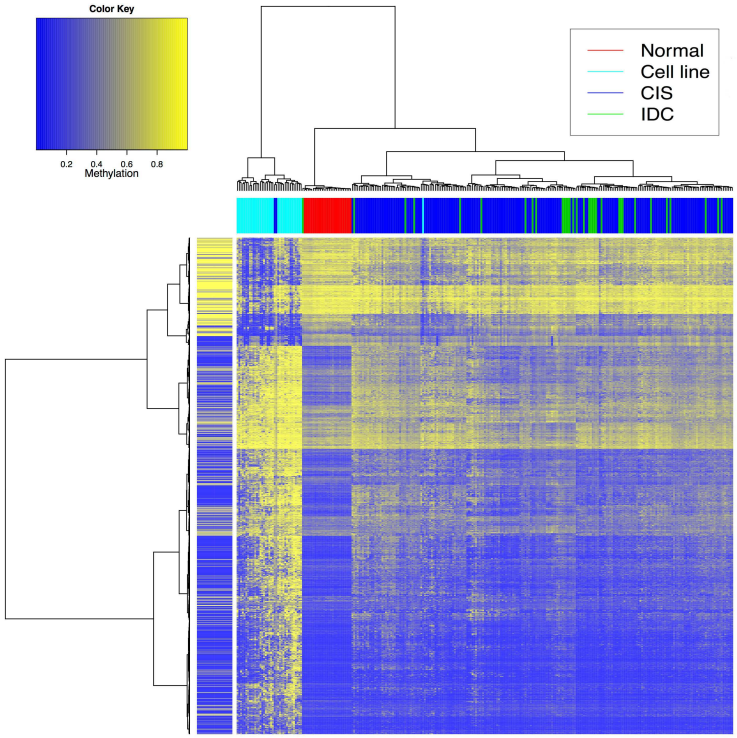


FIGURE 2.6: Hierarchical clustering of breast samples (columns) using CpG methylation as features (rows). The column color panel gives a distinction of the breast samples considered (red=normal tissue, cyan= cell line, dark blue= ductal carcinoma in situ, green= infiltrating ductal carcinoma). The row color panel gives information about the CpGs measured (blue= belonging to a CGI, yellow=outside of a CGI). This bi-clustering is able to distinguish the different types of tissues but also the types of CpGs measured.

2.2.2 Dimensionality reduction

Another important family of unsupervised methods are dimensionality reduction techniques. Previously, we mentioned the *curse of dimensionality* as one important issue that constrains the number of observations to be large enough to be able to learn. In specific cases, although the samples are represented by a d-dimensional vector, the data actually lies in a much smaller space.

Principal Component Analysis (PCA). This is a method to find a sequence of orthogonal vectors that captures the most information about the data by solving :

$$e_k = \arg \max_{\|e_k\|=1, e_k \perp \{e_1,\dots,e_{k-1}\}} e_k^{\mathbf{T}} \hat{X}_k^{\mathbf{T}} \hat{X}_k e_k, \tag{2.31}$$

with

$$\hat{X}_k = X - \sum_{i=1}^{k-1} X e_i e_i^{\mathbf{T}}. \tag{2.32}$$

PCA, like many other dimension reduction methods, is often use to visualize high-dimensional data in 2D, like for example in figure 2.7.

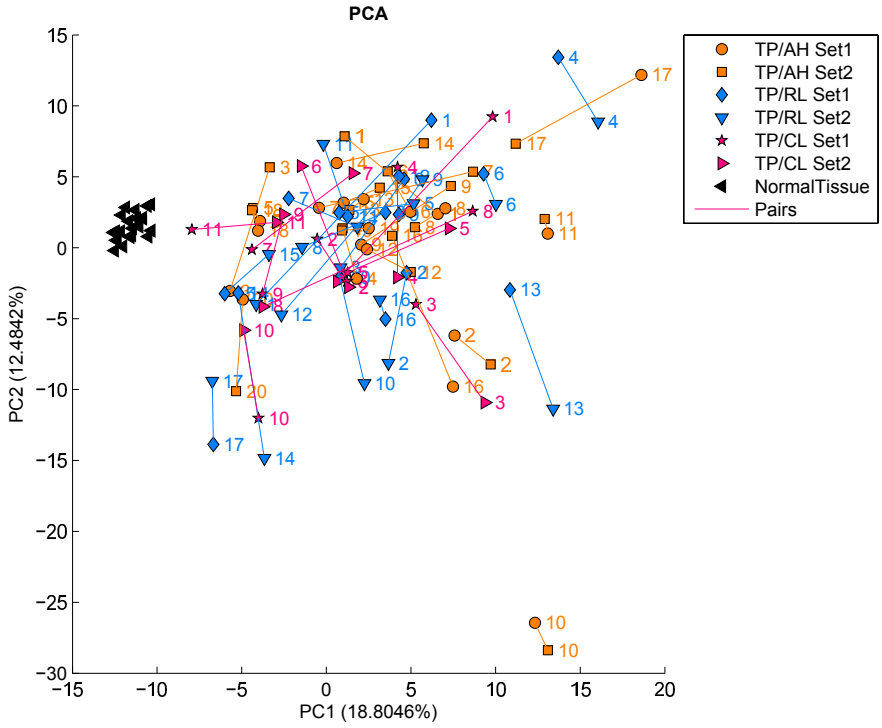


FIGURE 2.7: **Projection of breast methylation profiles on the first two principal components (31% of the total variance explained).** This representation shows that normal tissues are clustered together. Other methylation profiles come from breast primary tumors (Set 1) and are linked with their matched locoregional recurrence (Set2) given the localization (AH= axillary metastasis, RL= ipsilateral relapse, CL= controlateral relapse). See chapter 3 for more information about the data.

Model selection in unsupervised learning. We discussed previously that one particular specificity in partitioning algorithms is the choice of the parameter K. As illustrated in 2.8, an inappropriate choice in K can result in a ill-representation of the data at hand. We showed that hierarchical clustering could circumvent this problem. But even in this case, one generally has to choose a fixed representation (*i.e*, cut the dendrogram at a specific height). In dimensionality reduction, one also has to make a trade-off between the number of components and how well a projection on this subspace still retain enough information about the data (figure 2.9).

Given specific hypotheses about the data (*e.g*, Gaussian distribution), several criteria have been proposed such as the Akaike Information Criterion (AIC) [Akaike, 1973] or the Bayesian Information Criterion (BIC) [Schwarz, 1978] to assess an optimal K. However, these hypotheses rarely apply to real data and various heuristics have therefore

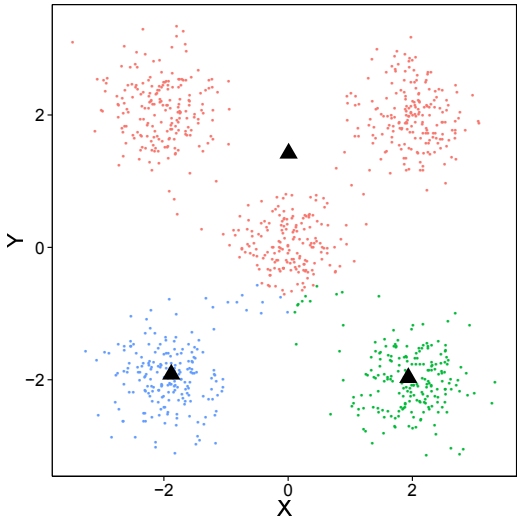


FIGURE 2.8: K-means clustering when K is not well suited for the data (K=3).

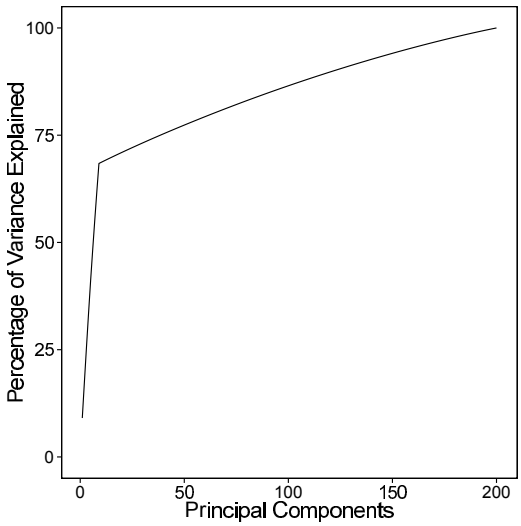


FIGURE 2.9: **Cumulative percentage of variance explained as a function of the number of principal components considered.** A common heuristic to choose a number of principal components is to look at a kink in the cumulative percentage of variance explained. Here on a toy dataset (n=200, p=1000), K=10 allows to explain 70% of the total variance while drastically reducing the number of dimensions.

been proposed too. In particular, looking at the stability of the clustering solutions obtained by perturbing the original dataset [Dudoit and Fridlyand, 2002, Ben-Hur et al., 2002a, Monti et al., 2003] is usually the preferred choice for biological data.

Clinical impact of unsupervised learning. For clinical purposes, assessing statistically the number of clusters (*e.g*, the number of breast cancer subclasses) is often not the primary objective. The clinical importance of the clustering, such as the discovery of subclasses with a significantly worse or better prognosis presenting a particular genomic or epigenomic profile, can help clinicians in proposing the most appropriate therapies to every single patients instead of a generic therapy to all patients.

2.3 Conclusion

In this chapter, we presented a brief overview of relevant statistical methods used throughout the remaining of this thesis in a particular biological context:

- Poor statistical power ($n \ll p$) leans toward simpler (linear) methods incorporating prior biological knowledge.
- A biological interpretation of the results at the cost of accuracy can sometimes be preferred as post-experimental validations can be undertaken to assess definitely the validity of the biological phenomenon.
- The performance of unsupervised methods can be assessed based on their clinical impact instead of mathematical criteria.

Chapter 3

Epigenomic alterations in breast carcinoma from primary tumor to locoregional recurrences

Some content from this chapter has been published as part of a peer-reviewed article in PLoS One [Moarii et al., 2014].

Keywords: Breast cancer, recurrence, metastasis, methylation, clonality, true recurrence.

3.1 Résumé

Les modifications épigénétiques telles que les variations abnormales de la méthylation de l'ADN sont associées à l'apparition de cancers. Comment ce mécanisme influt sur la progression tumorale est cependant encore floue. En comparant le méthylome initial de cancer du sein au méthylome des récidives de cancer chez ces mêmes patients, nous cherchons à déterminer des marqueurs de la progression tumorale dans le cancer du sein.

Pour cela, nous disposons du profils de méthylation de 48 tumeurs primaires ainsi que du profil de méthylation de leur métastase axillaire associée (20 cas), de leur récidive locale *i.e* dans le même sein (17 cas) ou de leur récidive contralatérale *i.e* dans le sein opposé (11 cas). Dans un premier temps, des méthodes d'analyses univariées et multivariées ont permises de determiner des sondes significativement différentiellement méthylées et marqueurs de la progression tumorale. Dans un second temps, nous établissons à partir des profils de méthylation un score de similarité entre deux échantillons, ce qui nous permet d'établir le caractère clonal entre une tumeur primaire et sa récidive locale, primordial dans la stratégie thérapeutique à employer.

Nos résultats montrent qu'un nombre restreint de sondes (49 sondes sur 27000) semblent être caractéristiques de la progression tumorale vers une métastase axillaire. Toutefois, aucune différence consistente n'a observée entre tumeurs primaires et récidive locale ou contralatérale, ce qui témoigne d'un lien moins marqué voire absent entre ces groupes. Dans un second temps, nous observons que les tumeurs primaires sont associées dans la majorités des cas à leur métastase respective (75%) alors que les tumeurs primaires et les récidives contralatérales ne montrent pas plus de similarité que deux tumeurs indépendantes. Ce résultat valide l'utilisation de la méthylation comme marqueur de clonalité entre deux échantillons et nous élaborons un score pour classifier les récidives locales. Cette classification valide la tendance (non-significative) des vraies récidives à être de moins bon pronostic et apporte un intérêt clinique dans l'apport décisionnel pour le traitement des patients.

3.2 Abstract

Epigenetic modifications such as aberrant DNA methylation has long been associated with tumorogenesis. Little is known, however, about how these modifications appear in cancer progres- sion. Comparing the methylome of breast carcinomas and locoregional evolutions could shed light on this process.

We propose to analyze the methylome profiles of 48 primary breast carcinomas (PT) and their matched axillary metastases (PT/AM pairs, 20 cases), local recurrences (PT/LR pairs, 17 cases) or contralateral breast carcinomas (PT/CL pairs, 11 cases) were analyzed. Univariate and multivariate analyzes were performed to determine differentially methylated probes (DMPs), and a similarity score was defined to compare methylation profiles. Correlation with copy-number based score was calculated and metastatic-free survival was compared between methods.

49 DMPs were found for the PT/AM set, but none for the others (FDR $< 5\%$). Hierarchical clustering clustered 75% of the PT/AM, 47% of the PT/LR, and none of the PT/CL pairs together. A methylation-based score (MS) was defined as a clonality measure. The PT/AM set contained a high proportion of clonal pairs while PT/LR pairs were evenly split between high and low MS score, suggesting two groups : true recurrences (TR) and new primary tumors (NP). CL were classified as new tumors. MS score was significantly correlated with copy-number based scores. There was no significant difference between the metastatic-free survival of groups of patients based on different classifications.

Epigenomic alterations are well suited to study clonality and track cancer progression. Methylation-based classification of TR and NP performed as well as clinical and copy-number based methods suggesting that these phenomenons are tightly linked.

3.3 Introduction

Breast conservative therapy, consisting in a partial mastectomy followed by whole breast irradiation, is the standard treatment for patients with early stage breast cancer. Overall survival is not significantly different from more physically and psychologically aggressive treatments such as mastectomy [Van Dongen et al., 2000]. However, patients relapse within 10 years in the same breast as the primary tumor (PT) in approximately 6% of cases [Bartelink et al., 2007], and within 5 years in the contralateral breast in approximately 3.5% of cases [Vichapat et al., 2012] or more in BRCA1/2 mutation carriers [Metcalfe et al., 2004]. Moreover, at the time of diagnosis, early stage breast cancers have already spread to axillary lymph nodes in roughly 30% of cases [Jatoi, 1999].

These different types of locoregional evolutions have different implications in terms of survival and treatments. Axillary metastases (AM) is usually predictive of poor survival [Carter et al., 1989] and is considerably worsen in triple negative breast cancers [Borg et al., 1990]. Local recurrences (LR) have been tightly linked with a greater risk of distant metastasis [Haffty et al., 1996]. Veronesi et al. [Veronesi et al., 1995] distinguished two categories of local recurrences : true recurrences (TR), corresponding to re-growth of resistant cells after initial treatment, and new primary tumors (NP), corresponding to *de novo* cancer. This classification is of potential interest to define adapted treatment scheme, as NP are considered to have an improved survival compared to TR [Smith et al., 1999]. Contralateral breast cancers (CL) are also an heterogeneous entity depending on the synchronism with the primary tumor. Synchronous bilateral breast cancers are developed at the same time, with the same genetic, environmental and hormonal background as the PT. Metachronous CL are usually treated as new cancers [Dawson et al., 1998] although a rare portion are considered as metastases. Overall, CL are still associated with a greater risk of metastasis compared to patients without CL [Healey et al., 1993].

Differences between the PT and either the AM, the LR or the CL have been studied at the genomic, transcriptomic and proteomic levels. Ellsworth et al. [Ellsworth et al., 2005] showed an overall frequency of allelic imbalance greater in PT than in AM. Weigelt et al. [Weigelt et al., 2005] explored the gene expression profile of PT and their

matched AM but were not able to identify a subset of genes to discriminate them, while Feng et al. [Feng et al., 2007] identified a set of 79 genes able to differentiate PT from matched AM. Studies between PT and LR have mainly focused on distinguishing TR and NP. A criterion based on clinical and pathological features was first established but judged insufficiently robust for most clinical applications. Several studies investigated the difference between TR and NP based on pangenomic analyzes of DNA copy number alterations (CNA) [Bollet et al., 2008, Ostrovnaya et al., 2010], intratumoral immune responses [West et al., 2011], loss of heterozigosity [Vicini et al., 2007], to p53 analysis [Van Der Sijp et al., 2002], or X-chromosome inactivation [Shibata et al., 1996]. Finally, studies of PT and CL highlighted the role of synchronism of the CL. Similarity measures based on DNA copy number profiles [Brommesson et al., 2008] or allelic imbalance [Imyanitov et al., 2002] showed a higher level of similarity between PT and synchronous CL compared to PT and metachronous CL.

Epigenetic modifications in cancer has recently been the topic of many studies. In particular the link between hypermethylation and gene silencing is well known [Razin and Riggs, 1980, Tate and Bird, 1993, Bird, 2002]. Several studies have then focused to describe cancer as an epigenetic disease. Baylin et al. [Baylin et al., 2001] have shown that aberrant hypermethylation of specific regions, dominantly CpG islands, are linked with the silencing of tumor suppressor genes and that this phenomenon is present in most cancers. Laird [Laird and Jaenisch, 1994], Ehrlich [Ehrlich, 2002] and Das [Das and Singal, 2004] suggested that a global hypomethylation phenomenon was also linked with tumorogenesis. Jones [Jones and Baylin, 2007] made a complete review of the hallmarks of epigenomics associated with cancer. Moreover, DNA methylation is conserved during cell division [Bird, 2002, Schermelleh et al., 2007] and could serve as a measure for clonality between cells in the classification of LR as either TR or NP.

In this study, epigenetic differences as well as similarities between PTs and either their AMs, LRs or CLs are analyzed. In the first part, univariate and multivariate analyzes are performed between the methylome profiles of primary tumors and their matched recurrences to observe recurrent patterns in cancer progression. Then in the second part, epigenome-wide similarity analyzes on the same samples is performed to observe clonality between tumor cells.

3.4 Materials and Methods

3.4.1 Patients Selection

The patients selected for the study were 49 years old or younger at diagnosis of the initial tumor; all patients were premenopausal; and had no previous history of cancer, except for one nonmelanoma skin cancer. The patients' PT was either ductal or lobular invasive breast carcinoma. However, both types of tumors did not display significantly differentially methylated probes and were thus all included in this study (data not shown).

Specimens from patients with primary breast cancers and breast cancer recurrences were selected from freshly frozen samples of the Institut Curie tissue bank according to the following criteria: 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. Tumor size did not correlate with the overall methylation rate (data not shown).

To ensure that the data would be informative, genomic analyzes were restricted 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. All the therapies were performed posterior to the biopsies of the primary tumors. Therefore, the studied methylation profiles are not modified by any potential effect of the treatments.

The 22 healthy breast tissues are taken from healthy women who underwent cosmetic plastic surgery at the Institut Curie. Part of the PT/AM cohort is identical to the cohort studied by Bollet et al. [Bollet et al., 2008].

All experiments were performed retrospectively and in accordance with the French Bioethics Law 2004-800, the French National Institute of Cancer (INCa) Ethics Charter and after approval by the Institut Curie review board and ethics committee (Comité de Pilotage of the Groupe Sein). In the French legal context, our institutional review board waived the need for written informed consent from the participants. Moreover, women were informed of the research use of their tissues and did not declare any opposition for such researches. Data were analyzed anonymously.

3.4.2 Methylation profiling

For each sample the methylation status at 27,578 positions in the genome was measured with the HumanMethylation27 BeadChip of Infinium technology [Weisenberger et al., 2008] using the standard Illumina protocol. Quality control was assessed using in-built Illumina technology.

3.4.3 Clinical Classification.

Histopathologic characteristics were reviewed by a single pathologist. The histological and biological properties of each sample was 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 [Balaton et al., 1995, Balaton et al., 1996]. The HER2 analysis was performed using the standard ASCO guidelines [Wolff et al., 2013]. In accordance with theories of the clonal evolution of tumor cell populations, LR were clinically defined as TR if they had the same histologic subtype (ductal or lobular) and a similar or increased growth rate, similar estradiol, progesterone and HER2 receptor statuses, and similar or decreased differentiation as the initial tumor [Smith et al., 1999]. TR also had to share with their PT the same breast quadrant. Thus, new PT were clinically defined as such when the LR had occurred in a different location, had a distinct histologic type, or had less aggressiveness features (lower grade, presence of hormonal receptors) than the initial tumor.

3.4.4 Data analysis

A spatial normalization process was applied to all profiles [Sabbah et al., 2011]. Among the 27,578 probes measured on each sample, 5 probes were removed due to missing values for some individuals, and all subsequent analysis was performed on the 27,573 remaining probes.

Differentially methylated probes between PT and their matched AM, LR and CL are obtained using two-sided paired and unpaired Wilcoxon tests, correcting the p-values for multiple testing with the methods of Benjamini and Hochberg [Benjamini and Hochberg, 1995]. Multivariate analysis was performed using a linear support vector machine (SVM) multidimensional classifier on either the complete methylation profile or after

dimensional reduction by considering only the most significant probes based on the Wilcoxon test. A p-value was calculated to assess the significance of the predictor accuracy compared to a predictor that would predict classes randomly. Unsupervised classifications were performed with complete linkage agglomerative clustering using the MATLAB® bioinformatics toolbox, while the support vector machine implemented in LIBSVM [Chang and Lin, 2011] was computed with a linear kernel and nested leave-one-out cross validation for parameter selection for supervised classification.

The similarity between two copy number profiles is assessed with the partial identity score (PIS) as defined by Bollet et al. [Bollet et al., 2008], which is based on the quantity of shared breakpoints between the two profiles and their frequencies. Following [Bollet et al., 2008], a recurrence from a matched PT/LR pair was considered TR based on copy numbers when the PIS between the PT and LR profiles was above the 95% quantile of the empirical PIS distribution between unrelated sample pairs. Similarly, a Methylation-Similarity score (MS) is defined based on the methylation profiles of a PT and its matched LR as the inverse of the Manhattan distance between their methylation profiles considered as 27,573-dimensional vectors. LR are then classified as TR of its matched PT when the MS score is above the 95% quantile of the empirical MS distribution between unrelated pairs. As a baseline, these results were compared to the Manhattan distance between unrelated normal breast tissues.

Metastasis-free survival was estimated by the Kaplan-Meier Method [Kaplan and Meier, 1958] and compared between the group of patients who were diagnosed as TR and the group diagnosed as NP using the log-rank test. The confidence interval of the hazard ratio was obtained using a semi-parametric Cox model [Cox and Oakes, 1984]. Computation was done using MATLAB® packages Logrank [Cardillo, b] and KMPlot [Cardillo, a].

3.5 Results

3.5.1 Methylation differences between PT and their matched metastasis or recurrence

A collection of 17 PT/LR pairs, 11 PT/CL pairs, and 20 PT/AM pairs was analyzed. The methylation data are available in the GEO database record number : GSE44870. Tables 3.1, 3.2 and 3.3 detail the clinico-histopathological properties of each sample. Some of the PT/LR samples match in part the cohort studied by Bollet et al. [Bollet

et al., 2008], and the corresponding sample numbers from both studies are provided in table 3.1.

Within each of the three cohorts, pairs of tumors including a PT and a metastatic or relapse sample can be used to investigate whether particular patterns in methylation profiles can serve as marker for cancer progression.

TABLE 3.1: **PT/LR clinical and histological features.**

			PT					Local Recurrence					
Pair	Cor	Age	Type	Grade	ER	PR	HER2	Type	Grade	ER	PR	HER2	Loc
1	1	23.3	D	3	0	40	0	D	2	90	15	0	1
2	3	42.9	D	3	30	80	0	D	3	60	90	0	1
3	11	49.3	L	3	0	0	0	D	3	0	0	1	1
4	16	48.8	D	2	80	30	0	D	1	20	70	0	1
5	12	49.3	L	2	90	50	0	L	2	90	0	0	0
6	13	45.4	D	2	20	85	0	D	2	95	20	0	1
7	15	46.5	D	2	100	80	0	D	2	70	100	0	1
8	2	42.4	D	2	90	40	0	L	1	90	70	NA	1
9	4	48.6	L	1	90	80	0	L	2	90	80	0	1
10	14	44	L	2	90	60	0	L	2	0	100	0	1
11	18	NA	D	3	0	0	NA	D	2	80	50	NA	1
12	20	47.5	D	3	0	0	1	D	3	0	0	1	0
13	21	46.7	D	2	80	0	NA	D	3	70	0	0	1
14	23	31	D	2	0	0	0	D	3	0	0	0	1
15	24	48.1	D	3	0	0	0	D	3	0	0	0	1
16	25	43.3	D	3	75	70	0	D	3	70	15	0	1
17	26	30.8	D	3	0	0	0	D	3	0	0	0	1

Cor (Correspondence): correspondence number with the Bollet/Servant cohort from [Bollet et al., 2008], **Type** : histological type of the tumor (D= ductal, L= lobular), **Grade** : Aggressiveness of the tumor (1 to 3), **ER** : percentage of estrogen receptor present in the sample, **PR** : percentage of progesterone receptor present in the sample, **Loc** (Location): 1 if the recurrence was located less than 4cm from the PT.

Within each cohort, investigations were made to detect differences at the methylome level between PT and the corresponding matched metastasis (AM) or relapse samples (LR or CL) . Using a paired Wilcoxon test, 49 probes significantly differentially methylated were found between PT and AM samples (at a 5% FDR level). The top 10 probes (ranked by p-value) and the corresponding genes are listed in table 3.4. This suggests that a general signal characteristic of cancer progression from PT to AM might exist. However, no probe was found significantly differentially methylated between PT and LR, and between PT and CL. This may be due to the lack of cancer progression marker at the methylation level between PT and relapse, to the fact that most relapses may

TABLE 3.2: **PT/CL clinical and histological features.**

		PT					Contralateral Recurrence				
Pair	Age	Type	Grade	ER	PR	HER2	Type	Grade	ER	PR	HER2
1	46.6	L	3	80	80	0	NA	NA	90	20	0
2	46.9	D	2	60	100	0	D	2	30	100	0
3	48.4	D	3	70	60	0	D	3	100	10	0
4	42.6	D	2	0	0	0	D-L	2	100	70	0
5	48.5	D	2	70	20	0	D	3	10	20	0
6	44.5	D	2	$\geq$10	$\geq$10	0	Med	2	0	0	0
7	46	D	2	80	30	0	D	1	40	95	0
8	48.9	D	3	90	20	0	Meta	3	0	0	0
9	38.9	D	3	0	0	0	D	3	100	40	0
10	31	D	3	0	0	0	D	3	0	0	0
11	36.3	D	3	10	5	0	D	3	0	0	0

Type : histological type of the tumor (D= ductal, L= lobular, Med=Medullary, Meta=Metaplasic), **Grade** : Aggressiveness of the tumor (1 to 3), **ER** : percentage of estrogen receptor present in the sample, **PR** : percentage of progesterone receptor present in the sample.

TABLE 3.3: **PT/AM clinical and histological features.**

Pair	Age	Type	Grade	ER	PR	HER2
1	45.9	D	3	70	70	0
2	NA	D	3	90	20	0
3	NA	NA	NA	95	30	0
4	48.8	D	1	60	90	0
5	43.6	D	3	0	0	0
6	35.3	D	2	20	70	0
7	45.1	D	3	10	25	0
8	41.9	D	2	70	40	NA
9	43.5	D	1	$\geq$ 10	$\geq$ 10	0
10	43.7	D	3	80	50	NA
11	44.9	D	2	0	0	0
12	43.6	D	1	$\geq$ 10	0	0
13	40.2	D	3	0	0	1
14	32.5	L	3	40	60	1
15	38.5	D	2	0	10	0
16	37.5	D	3	40	50	0
17	39.3	D	3	80	90	0
18	37.6	D	3	0	0	0
19	36.6	D	3	10	50	1
20	35.4	D	3	0	30	0

Age: Age of the patient at diagnosis of the primary tumor in years, **Type** : histological type of the tumor (D= ductal, L= lobular, Meta=Metaplasia), **Grade** : Aggressiveness of the tumor (1 to 3), **ER** : percentage of estrogen receptor present in the sample, **PR** : percentage of progesterone receptor present in the sample.

not be biologically related to the PT, or to the small size of the cohort which limits the power of statistical tests.

On the PT/AM cohort, the SVM model correctly identified the PT and AM in 18 out of 20 held-out pairs (90% success rate, P-value=$2.0 * 10^{-4}$) when considering the whole methylation profile probes. The SVM model obtained after dimensionality reduction by filtering the 22 most significant probes selected according to a Wilcoxon test gave a 100% accuracy. As illustrated in 3.1, good accuracy was still achieved when considering an increasing number of probes (Accuracy $\sim$ 90%). On the PT/LR and PT/CL cohorts, however, the success rate was respectively 58% (10 out of 17 pairs, P-value=0.31) and 27% (3 out of 11 pairs, P-value=0.11) when taking all probes into account. Note that these values are not significantly different from random guess.

TABLE 3.4: **Most significantly differentially methylated genes between PT and AM samples (Top 10).**

CpG	Gene	Pvalue	Methylation Variation
cg04619381	*LOC222171*	0.013	-0.048
cg18140857	*RDHE2*	0.013	0.102
cg23698969	*SLC22A18*	0.013	0.042
cg20161089	*IFI27*	0.013	0.238
cg24959428	*GBP6*	0.02	0.126
cg22630748	*INHBE*	0.02	0.1
cg03623878	*MCF2L*	0.02	-0.05
cg16179125	*CTSZ*	0.02	0.182
cg25115460	*TP73*	0.022	0.109
cg11946165	*CTSK*	0.022	0.098

CpG: CpG probe name. **Gene**: Associated gene. **Pvalue**: FDR corrected p-value. **Methylation Variation**: Mean variation of methylation from the primary tumor to the axillary metastasis.

3.5.2 Methylation conservation between PT and their matched metastasis or recurrence

Instead of searching for differences between PT and their matched metastasis or recurrence, which may characterize markers for cancer progression, the study also focuses on similarities between methylation profiles, which may be useful for example to characterize clonality between a PT and a recurrence. A hierarchical clustering was first performed for all samples within each cohort to characterize the similarities between real matched pairs compared to unrelated samples. The resulting dendrograms are presented in 3.2. Interestingly we see that matched pairs of PT and metastasis/recurrence

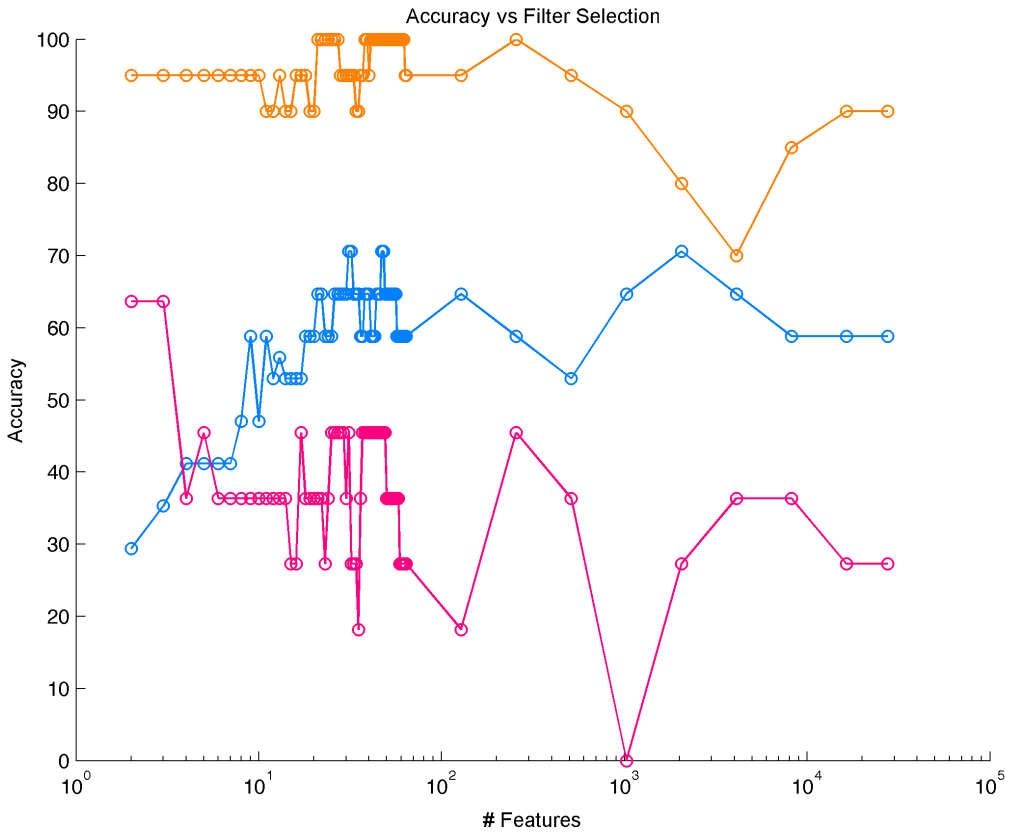


FIGURE 3.1: **Accuracy of the paired-SVM classifier as a function of the number of probes selected obtained through leave-one-out cross-validation for each dataset (orange= PT/AM, blue= PT/LR, pink=PT/CL) .**

samples are usually closer to each other than to any unrelated tissues in the PT/AM cohort (15 out of 20, 75%), less often in the PT/LR cohort (8 out of 17, 41%), and never in the PT/CL cohort. This observation is consistent with decreasing proportions of real clonal pairs from the PT/AM to the PT/CL set.

Another way to see this phenomenon is to assess statistically, within each cohort, how the methylation distances between matched pairs differ from the methylation distances between unmatched pairs. We displays the distributions of methylation distances for different sets of sample pairs in 3.3. We also display in 3.4 the boxplot of methylation distances by groups. Real matched pairs between a PT and its corresponding metastasis or recurrence are significantly closer in terms of global methylation than a random pair of samples taken from two different individuals, both in the PT/AM cohort (P-value=$3.5 * 10^{-7}$) and in the PT/LR cohort (P-value=$1.6 * 10^{-6}$). This is however not true in the PT/CL cohort, where we detect no differences between correctly and randomly matched pairs (P-value=0.44). In addition, we calculated the distribution

of distances between the CL tumors. We performed the same analysis between the PT tumors. We observed that the distribution were not significantly different (data not shown), as expected. This is in agreement with the assumption we made that CL tumors could be considered as new primary tumors. Finally, we also compared the distribution of distances between the healthy breast tissue i and all the other healthy breast tissues from the cohort to assess the heterogeneity between normal breast tissues.

3.5.3 Clonality detection based on methylation profiles

The above results suggests that methylation profiles tend to be conserved during clonal expansion (such as samples in the PT/AM cohort), but strongly differ between unrelated tumors in a given person (such as samples in the PT/CL cohort). Moreover, methylation seems to be a stable mechanism in normal tissues compared to cancerous ones. It is therefore tempting to use methylation distance as a tool to discriminate true recurrences from new tumors in ambiguous cases, that is, for samples in the PT/LR cohort.

As shown in Figure 3, 9 out of 17 PT/LR pairs (52%) have a MS score higher than the threshold given by the 95% percentile of the MS score between unrelated pairs ($MS_{Threshold} = 6.6 * 10^{-4}$); they are therefore considered as clonal pairs from the methylation point of view. The remaining 8 pairs are considered as non-clonal, meaning that the LR may correspond to a new primary tumor.

Comparison between the methylation-based similarity measure MS score with the copy-number-based similarity measure (PIS) developed by [Bollet et al., 2008] show a good correlation overall ($\rho = 0.55$, P-value=$3.7 * 10^{-5}$, see figure 4). Table 5 gives a comparison of the outcomes given by methylation-based, copy-number based and clinical-based classification of LR as TR or NP. The methylation-based classification method agreed with the copy-number based PIS classification method on 14 out of 17 pairs (concordance=82%, P-value=$6.3 * 10^{-3}$) and agreed with the clinical-based classification on 14 out of 17 pairs (concordance=82%, P-value=$6.3 * 10^{-3}$).

Finally, the different classifications of LR as TR or NP were correlated with time-to-recurrence and metastasis-free survivals. The differences in time-to-recurrence for the two groups defined by methylation-based classification or the clinical and histological classification were not statistically significant (P-value=0.83 and P-value=0.12). It was however significant using the partial identity score (P-value=0.03). This is interesting in the sense that one of the main criteria to distinguish TR and NP is the time-to-recurrence. Therefore, methylation-based classification is based on more information than time only.

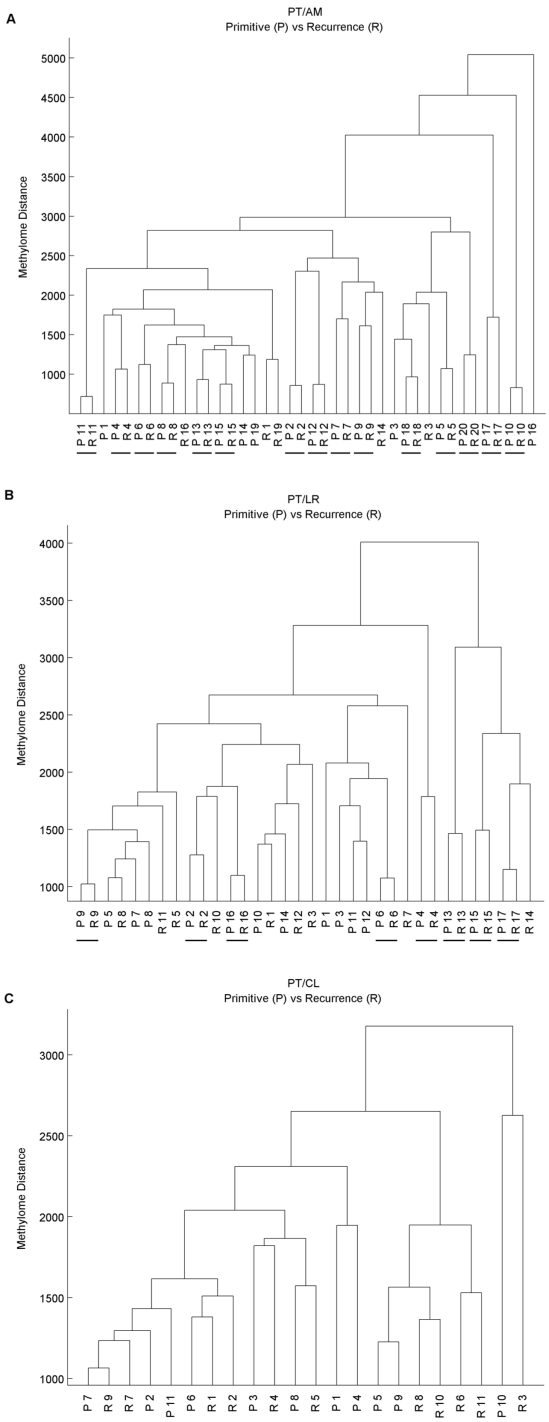


FIGURE 3.2: **Study of similarity between matched primary tumors and recurrences by hierarchical clustering.** Hierarchical clustering based on the manhattan distance between methylome profiles with complete linkage was performed. Real pairs that are closer to each other than to any other samples are underlined. Panel A (resp. B, resp. C) represents the PT/AM (resp. PT/LR, resp. PT/CL) set.

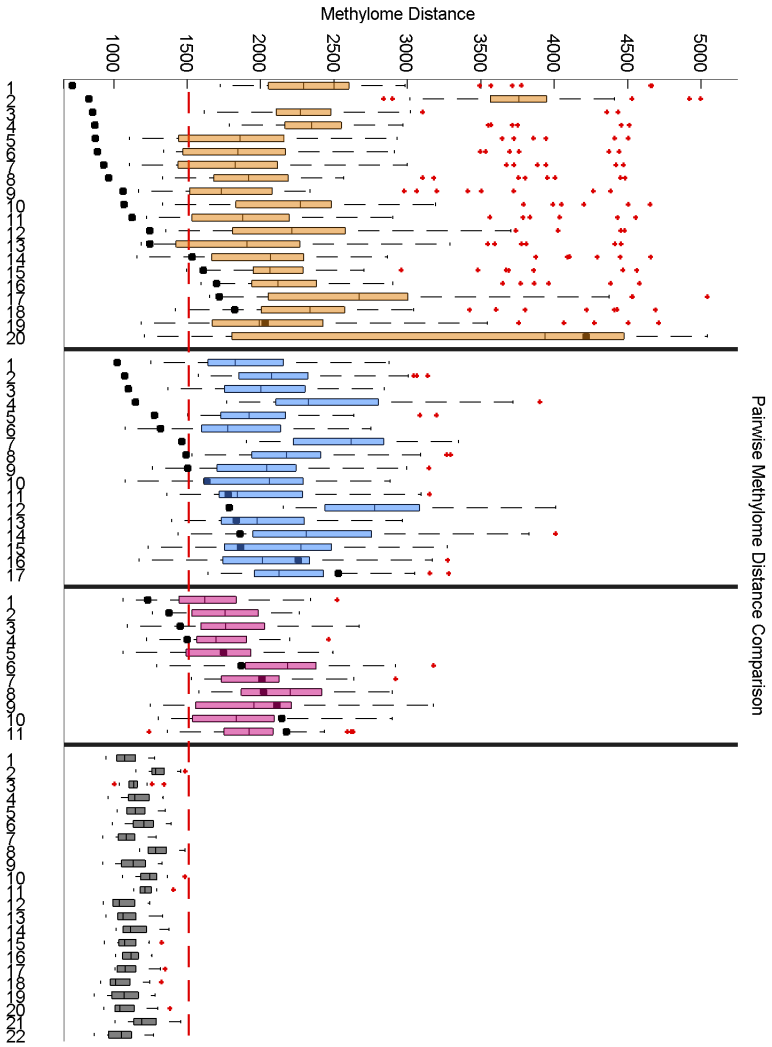


FIGURE 3.3: **Pairwise methylome distance for each samples.** Each boxplot represents the Manhattan distance between primary tumor i and an unrelated locoregional evolution, or the Manhattan distance between locoregional evolution i and an unrelated primary tumor. The black square represent the Manhattan distance between the matched primary tumor and locoregional evolution from sample i. The yellow (resp. blue, resp. pink) panel represents the PT/AM (resp. PT/LR, resp. PT/CL) set. The last panel represents the distribution of distances between the healthy breast tissue i and all the other healthy breast tissues from the cohort.

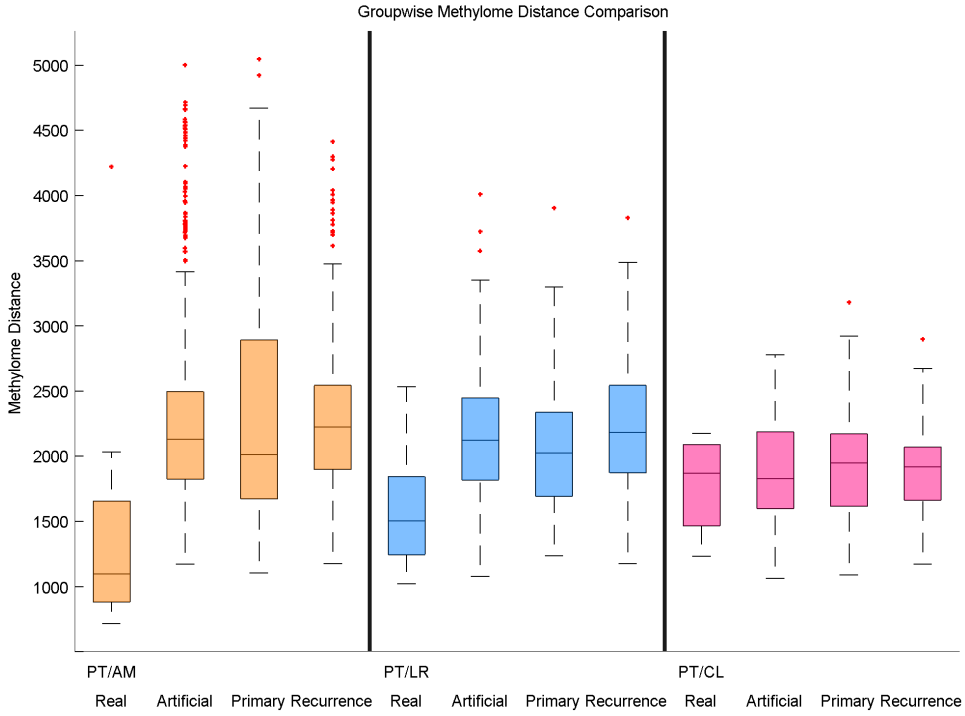


FIGURE 3.4: **Distribution of methylation similiarity between samples given the type of pairs.** Each boxplot represents the distribution of Manhattan distance between matched primary and locoregional evolution ("Real"), between non-matched primary and locoregional evolution ("Artificial"), between two primary tumors ("Primary") or between two locoregional evolution ("Recurrence") for each dataset.

The difference in metastasis-free survival of patients with TR and NP was not significant based on methylation (P-value=0.52, Hazard-Ratio=3.7, 5 year metastasis-free survival=75% for NP), copy-number (P-value=0.15, Hazard-Ratio=16.9, 5 year metastasis-free survival=86% for NP) or clinical features (P-value=0.17, Hazard-Ratio=6.3, 5 year metastasis-free survival=86% for NP) (figure 5).

3.6 Discussion

We studied alterations of methylation profiles from primary breast carcinomas and different types of recurrences, namely, axillary metastases, local recurrences and contralateral breast carcinomas. For this particular dataset, we observed significant methylation differences for 49 CpG probes, which characterizes the progression between a PT and its AM. Consistent with this result, a multivariate analysis with a linear SVM classifier

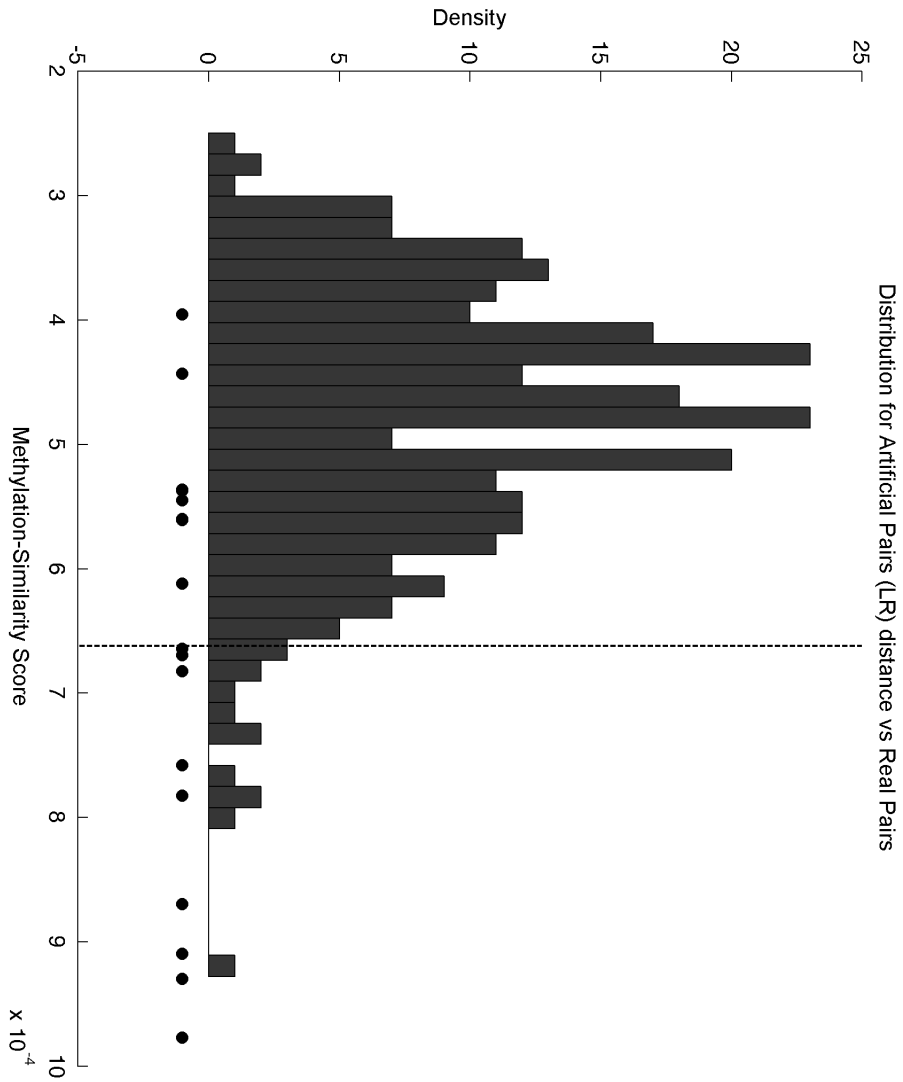


FIGURE 3.5: **Histogram of the distribution of methylome-similarity score (MS) between unrelated PT/LR pairs.** MS score for matched pairs is represented by circles. The vertical dashed line corresponds to the 94% quantile of the distribution of the MS scores for the unrelated pairs, used as a threshold to define clonal pairs ($MS_{Threshold} = 6.6 * 10^{-4}$).

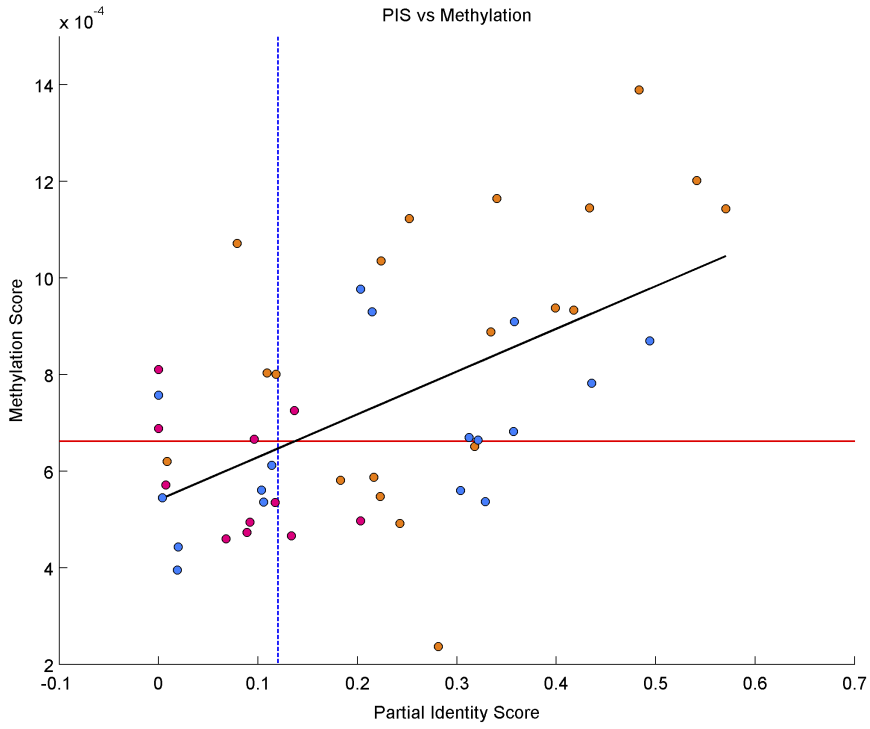


FIGURE 3.6: **Correlation between methylation and copy-number scores.** The horizontal red line (resp. vertical dashed blue line) corresponds to the 95% quantile of the distribution of the methylation-scores (resp. partial identity scores) for the unrelated pairs : $MS_{Threshold} = 6.6 * 10^{-4}$ (resp. $PIS_{Threshold} = 0.12$). PT/AM (resp. PT/LR, resp. PT/CL) pairs are colored in yellow (resp. blue, resp. pink). The black line corresponds to the linear regression between methylation and copy-number scores for all the datasets.

using a small subset of probes perfectly distinguished PTs from AMs with a 100% accuracy. Several significantly differentially methylated probes correspond to genes involved in cancer-related mechanisms such as cell death (*MCF2L*, *RASSF5*, *RASSF6*, *CASZ1*, *SLC22A18*, *IFI27*), tumorogenesis (*CTSZ*, *TP73*, *CTSK*, *PIK3R1*), *KLK11*, cell cycle (*PPM1G*, *RANBP5*, *VAMP8*) and cell differentiation (*SMAF1*, *PAX6*, *PAX8*). On the contrary, for the PT/LR and PT/CL sets, univariate analyzes were not able to find significantly differentially methylated probes. This absence of specific epigenetic alterations between the primary tumors and the local recurrences or the contralateral breast recurrences was confirmed by the poor performances of linear classifiers, unable to separate PT from LR nor PT from CL significantly better than random guesses. Nevertheless, the absence of methylation markers in the PT/LR and the PT/CL groups does not necessarily mean that the primary tumor and the recurrence are independent. We cannot rule out the possibility that the recurrence arises from a specific subclone which does not match the major subclone of the primary tumor. One could for example analyze the methylation profiles of several microdissections samples of the primary tumor to study

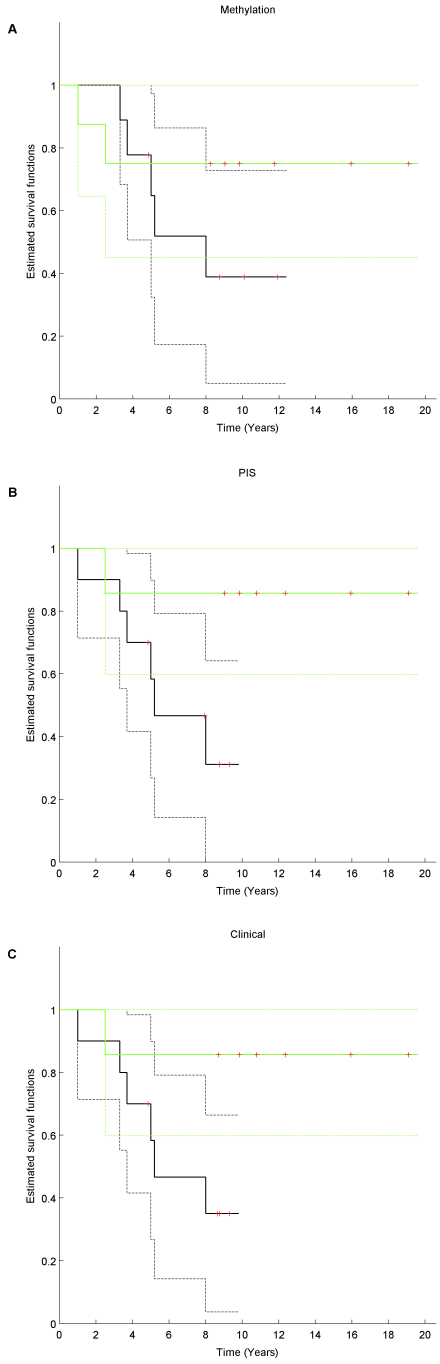


FIGURE 3.7: **Kaplan-Meier estimates of the metastasis-free survival between TR and NP for the different classification methods.** The full black (resp. green) line corresponds to the survival for samples classified as TR (resp. NP) and the corresponding dashed lines correspond to upper and lower 95% CI. The red crosses represent censored data. Panel A (resp. B, resp. C) represent the methylation-based (resp. copy-number based, resp. clinical based) classification.

TABLE 3.5: **Comparison of classification methods for clonality between pairs in the PT/LR cohort.**

Panel A								
		Scores			Classification			
Pair	Cor	PIS	**MS** ($\times 10^{-4}$)	Time (Years)	PIS	**MS**	Clinical	Divergence
1	1	0.019	4.42	6.5	NP	NP	NP	
2	3	0.435	7.82	3.2	TR	TR	TR	
3	11	0.018	3.95	6.4	NP	NP	NP	
4	16	0.303	5.59	3.8	TR	NP	NP	PIS
5	12	0.113	6.11	3.4	NP	NP	NP	
6	13	0.214	9.29	4.6	TR	TR	TR	
7	15	0.105	5.36	3.2	NP	NP	TR	Clinical
8	2	0	7.57	5.2	NP	TR	NP	MS
9	4	0.203	9.76	3.5	TR	TR	TR	
10	14	0.321	6.64	2.4	TR	TR	TR	
11	18	0.003	5.44	2.2	NP	NP	NP	
12	20	0.103	5.60	1.4	NP	NP	NP	
13	21	0.356	6.82	4.2	TR	TR	TR	
14	23	0.328	5.37	0.9	TR	NP	TR	MS
15	24	0.312	6.69	1.4	TR	TR	TR	
16	25	0.357	9.09	2.7	TR	TR	TR	
17	26	0.493	8.69	2.0	TR	TR	TR	

Cor (Correspondence): correspondence number with the Bollet/Servant cohort. **scores**: scores obtained with partial identity (PIS) or methylation (MS). **Time**: time elapsed between diagnosis of the PT and diagnosis of the recurrence. **Classification**: classification of the recurrence based on copy number (PIS), methylation (MS) or clinical features (clinical). **Divergence**: which method deviated from the others.

potential heterogeneity.

The second part of the study focused on observing stability in methylation profiles. It is interesting to note that although PTs and AMs were significantly differentiable using a subset of probes, they also have overall very similar methylation profiles indicating that the tumors might actually be clones with specific alterations characteristic of the lymph node status. The subset of genes determined in the first part, if confirmed, could be associated with bad prognosis. On the other part, although the LRs and the CLs were not significantly different from their primary tumors, they tend to have overall different methylome profiles especially for the CLs. The overall different methylome profiles for the PT/CL set was expected since CLs are usually considered to be independent tumors.

The results above suggested to use global methylation analysis as a measure of clonality to tackle the subclonal populations in the local recurrences as proposed by Veronesi et

al. [Veronesi et al., 1995]. A methylation-based classification was proposed to distinguish LRs as either true recurrences of the first PT or new PT [Smith et al., 1999]. A comparison with both clinical and copy-number based classifications on the same cohorts agreed on 14 out of 17 samples (82% concordance, P-value=6.410^{-3}) for both methods, although comparisons on larger cohorts are needed to assess the performance of methylation-based classification. Moreover, a good correlation between the methylation-based similarity score and the copy-number based similarity score seems to indicate a link between modifications at the genomic and epigenomic levels. Although the role of methylation in gene expression has thoroughly been studied [Bird, 2002,Razin and Riggs, 1980, Tate and Bird, 1993], the relationship between methylation and copy-number still remains unclear. Houseman et al. [Houseman et al., 2009] showed no clear relationship between methylation and copy-number. On the other hand, Lauss et al. [Lauss et al., 2012b] observed associations between the two mechanisms in urothelial carcinoma, while a short report by Kwee et al. [Kwee et al., 2011] tries to build copy number profiles from methylation profiles alone. Our study seems to validate the second hypothesis that methylation and copy-number are well connected mechanisms.

The discordances between the methylation-based classification method and the usual clinical method are discussed here for the samples 7, 8 and 14, although no actual method is a gold standard for classifying TR from NP. Sample 8 filled almost all the requirements for clinical classification as TR (location, receptor status) but failed in aggressiveness and type of tumor (PT was ductal type 2 and LR was lobular type 1). A decrease of aggressiveness of the recurrence could be explained by the use of neoadjuvant therapies. For the change of type, Fisher et al. showed that a mixing of ductal and lobular breast carcinoma was a possibility in 6% of the patients [Fisher et al., 1975] which could explain the change in type. Sample 7 was classified as TR by clinical classification and as NP by both methylation and copy-number based classifications. This suggests some limitations to methods based only on clinical features.

An interesting question for clinical applications would have been to predict whether a primary tumor would relapse (either as AM, LR or CL) or not. However, the patient cohort used in this study does not allow to address this question. Indeed, one would require to compare the methylation profiles of patients who did not display any relapse (AM, LR and CL) to those of the current study.

Chapter 4

Changes in gene expression control by DNA methylation in cancer

Some content from this chapter has been submitted to BMC Genomics.

Keywords: High-density methylation patterns, gene expression epigenetic regulation, epigenetic shift in cancer.

4.1 Résumé

La méthylation de régions à forte densité en CpGs, communément appelées ilôts CpGs, est un mécanisme associé à la régulation du niveau d'expression des gènes dans des cas bien précis. Certaines altérations spécifiques telles que l'hyperméthylation de tels ilôts proches de certains gènes suppresseurs de tumeurs, entrainant leur inactivation, ou encore l'hypométhylation d'ilôts associés à certains oncogènes particuliers, entrainant leur réactivation, sont fréquemment observées dans plusieurs types de cancer. Cependant, le rôle de la méthylation dans la régulation de la transcription de l'ensemble du génôme est encore très peu connue. En particulier, de récentes études ont montré que l'hyperméthylation de certains ilôts CpG n'était pas causal à la répression des gènes mais agissait comme un verrou supplémentaire.

L'analyse des données publiques à grandes échelles disponibles sur "The Cancer Genome Atlas" (TCGA) nous permet aujourd'hui de combiner les données d'expressions de gènes, de méthylation à haute densité, mais également de copy-number pour 672 échantillons sains et cancéreux dans 3 types de cancers différents. A l'aide de diverses méthodes

statistiques, nous analysons le lien entre les variations de méthylation des ilôts CpG et les variations d'expression des gènes associés pour comprendre l'ampleur de la méthylation dans le mécanisme de régulation des gènes.

Nous montrons dans ce chapitre que les profils de méthylation des ilôts CpG chez les patients sains peuvent se résumer par 2 profils caractéristiques : le premier est associé à un ilôt CpG hypométhylé dont les régions voisines ("shores" et "shelves") sont généralement hyperméthylées et le second est associé à une région globalement hyperméthylée. De plus, l'assignation d'un ilôt à un profil caractéristique est globalement conservée entre les différents tissus, ce qui met en évidence la stabilité d'un profil de méthylation associé à un gène donné. Nous observons par ailleurs chez les patients cancéreux l'existence d'un profil caractéristique supplémentaire associé à une région globalement hémi-méthylée. La distribution de l'expression des gènes en fonction de l'appartenance de l'ilôt CpG correspondant à un profil caractéristique montre que de manière générale, le caractère hypo- ou hyperméthylé de l'ilôt CpG n'est pas associé à un niveau plus ou moins élevé de l'expression des gènes. L'expression des gènes associés aux ilôts hémi-méthylés observés uniquement, bien que très fortement réprimés dans les tissus cancéreux, sont également réprimés dans les tissus sains, ce qui remet en question le rôle causal de la méthylation dans la régulation de l'expression. Bien que les profils précédemment décrits, basés sur le niveau moyen de méthylation par sonde à l'échelle d'un sous-groupe de patients n'ai pas montré d'association avec le niveau d'expression, une analyse à l'échelle de chaque individu montre que certaines variations - localisées spécifiquement dans les régions périphériques de l'ilôt CpG (CGI shores) - sont fortement négativement corrélées à la régulation de l'expression du gène associé. Ces gènes, pour lesquels une forte association existe entre la méthylation et l'expression, semblent différer d'un tissu à l'autre mais surtout, entre un tissu sain et un même tissu cancéreux. Une forte association est observée entre ces gènes fortement régulés par la méthylation et les facteurs de transcriptions, ce qui souligne le rôle majeur de la méthylation dans le mécanisme de régulation. Enfin, nous observons un lien complémentaire entre la méthylation et le copy-number dans la prédiction de l'expression des gènes.

Nos résultats suggèrent que durant la tumorogénèse, un mécanisme de reprogrammation épigénétique s'effectue. Ce mécanisme n'a pas un impact direct sur l'expression des gènes associés, mais agit sur la régulation de la transcription en affectant la susceptibilité des gènes aux variations épigénétiques.

4.2 Abstract

Methylation of high-density CpG regions known as CpG Islands (CGIs) has been widely described as a mechanism associated with gene expression regulation. Aberrant promoter methylation is considered a hallmark of cancer involved in silencing of tumor suppressor genes and activation of oncogenes. However, recent studies have also challenged the simple model of gene expression control by promoter methylation in cancer, and the precise mechanism of and role played by changes in DNA methylation in carcinogenesis remains elusive.

Using a large dataset of 672 matched methylomes, gene expression, and copy number profiles accross 3 types of tissues issued from healthy and cancerous patients from The Cancer Genome Atlas (TCGA), we perform a detailed meta-analysis to clarify mechanisms of gene expression control by changes in DNA methylation in normal and cancer tissues. While most genes have their promoter region hypo-methylated in normal samples, we show that a small fraction of genes are hyper-methylated, but not significantly less expressed. This classification is robust across tissues and is also present in some cancer tissues, although in other cancer tissues a significant fraction of genes witness changes in their promoter's methylation. These changes in cancer tissues are not directly accompanied by changes in gene expression levels, since most genes that become hyper-methylated in cancer tissues are already lowly expressed in normal tissues, however large changes in CGI methylation has a prognostic value. A finer analysis of the link between CGI methylation and gene expression in the different types of tissues highlights the presence of many genes whose expression is under control of CGI methylation, particularly through changes in methylation of CpG in the flanking regions of CGIs. These genes are not the same in different tissues, and not the same in normal and cancerous tissues, but are overall enriched in transcription factors.

Our results suggest that epigenetic reprogramming in cancer does not contribute to cancer development via direct gene expression regulation. It may instead modify how some genes are under control of DNA methylation variations, particularly transcription factors, in a cancer-dependent manner.

4.3 Introduction

DNA methylation is one of the main epigenetic mechanisms, alongside histone modifications, that plays a significant role in gene silencing [Newell-Price et al., 2000], tissue differentiation [Laurent et al., 2010], cellular development [Smith and Meissner, 2013], X-chromosome inactivation [Pollex and Heard, 2012], or genetic imprinting [Li et al.,

1993]. Aberrant hyper-methylation of high-density CpG regions known as CpG Islands (CGIs) [Esteller, 2002] and genome-wide hypo-methylation [Ehrlich, 2002] have often been associated with cancer and there has been an increasing effort to understand the specific epigenetic modifications that contribute to carcinogenesis [Laird and Jaenisch, 1994, Das and Singal, 2004, Kulis and Esteller, 2010].

The possibility to measure DNA methylation genome-wide on normal and cancer tissues, with microarray or sequencing technologies, has triggered a lot of data-driven research to clarify the role of methylation in gene regulation and cancer. Several studies have highlighted a correlation between differentially methylated regions near promoter regions and gene expression changes [Meissner et al., 2008, Lister et al., 2009, Zhang et al., 2011, Hansen et al., 2012, Varley et al., 2013]. However, it has also been reported that aberrant over-methylation occurs mostly in normally down-regulated genes, questioning the role of methylation as a causal mechanism for gene repression [Keshet et al., 2006, Sproul et al., 2011, Sproul et al., 2012, Sproul and Meehan, 2013]. More recently, Timp et al. have proposed a model where epigenetic aberrations contribute to carcinogenesis by dysregulating the functions of specific genes that regulate the epigenome itself [Timp and Feinberg, 2013, Timp et al., 2014]. Reddington et al. speculate that epigenetic reprogramming might lead to an altered Polycomb binding landscape which could potential impact genome regulation [Reddington et al., 2014].

To gain further insight into the role of DNA methylation in cancer, we perform a large-scale meta-analysis of methylation profiles of normal and cancerous tissues from The Cancer Genome Atlas (TCGA), focusing for each CGI on (i) how, on average, their methylation level differs between normal and cancer samples and between different tissues, and (ii) how their association with gene expression level, as estimated from inter-individual variability within each sample category, differs. We show in particular that in normal tissues, most CGIs tend to be either hypo- or hyper-methylated, and that the classification is stable across tissues of origin; on cancer samples, on the other hand, a stable subset of the CGIs witness a change in their methylation status in a subset of patients. While this change in methylation has a prognostic value for patient survival in breast cancer, we did not find evidence that it directly impacts gene expression level, as most of the genes concerned are already lowly expressed both in normal and in cancerous tissues. Similar findings were already reported in [Sproul et al., 2011]. A finer analysis of the link between CGI methylation and gene expression in the different types of tissues highlights the presence of many genes whose expression is under control of CGI methylation, particularly through changes in methylation of CpG in the flanking regions of CGIs. These genes are not the same in different tissues, and not the same in normal and cancerous tissues, but are overall enriched in transcription factors. This

suggests that epigenetic reprogramming might contribute to carcinogenesis in part by modifying gene expression susceptibility to changes in DNA methylation.

4.4 Materials and Methods

4.4.1 Patients Selection

All data are issued from TCGA data portal. Cancer types selected are breast, colon and lung adenocarcinomas as consequent matched datasets were available for methylation, gene expression and copy number profiles. The datasets are detailed in 4.1 and the different institutions that released the data are mentioned in the acknowledgement section.

TABLE 4.1: Patients Dataset. Original dataset sizes for methylation (Meth), gene expression (GE) and CNV profiles for normal (N) or cancerous (C) tissues. The "Matched" column represents the final dataset containing samples with matched methylation, gene expression and copy number profiles.

	Meth		GE		CNV		**Matched**	
	N	C	N	C	N	C	N	C
Breast	97	626	100	778	1073	1041	**70**	**474**
Colon	38	291	0	193	0	470	**0**	**33**
Lung	32	452	37	125	568	516	**13**	**82**
Total	167	1370	137	1096	1641	1981	**83**	**589**

4.4.2 Methylation profiling

Methylation profiles were retrieved from level 2 TCGA data obtained the Illumina HumanMethylation450K DNA Analysis BeadChip assay, which is based on genotyping of bisulfite-converted genomic DNA at individual CpG-sites to provide a quantitative measure of DNA methylation [Bibikova et al., 2011]. Following hybridization, the methylation value for a specific probe was calculated as the ratio $M/(M+U)$ where M is the methylated signal intensity and U is the unmethylated signal intensity. 485,577 CpG methylation levels, associated with 27,176 CGIs and 21,231 genes, were measured as such accross the genome.

Following [Irizarry et al., 2009], we considered not only the CGI methylation profile but also included in the analysis proximal regions in the near vicinity (up to 4kb), namely the CGI Shores and Shelves regions in a general CGI+SS methylation profile. As we were interested in the coordinated variations of methylation, we restricted the analysis

to CGI+SS profiles containing at least 20 probes which reduced the analysis to 1827 CGI+SS associated with 2374 genes from the original dataset.

4.4.3 Gene expression profiling

Gene expression profiles were retrieved from level 3 TCGA data, that is obtained from the Illumina HiSeq RNASeq technology and processed following [Mortazavi et al., 2008].

4.4.4 Copy number variations processing

Copy number variations were retrieved from the level 3 TCGA data infered from Affymetrix SNP6.0 data files in GenePattern following [Reich et al., 2006]. For each gene, we then obtained the log ratio copy number score as the segmented log ratio score for the interval containing its transcription start site.

4.4.5 Combined CpG island, shores and shelves pattern analysis using dynamic time warping

CGI+SS patterns were compared using dynamic time warping (DTW) [Rabiner and Juang, 1993] as it is less sensitive to small variations than the Fréchet distance [Efrat et al., 2006] used in [Vanderkraats et al., 2013]. Dynamic time warping was originally applied as a speech signal similarity measure and has been applied with success in several other fields including computer vision [Serra and Berthod, 1994], protein structure matching [Wu et al., 1998] and time series analysis [Keogh and Pazzani, 1999].

A CGI+SS profile i can be represented as a couple of vector $(X^i, Y^i) = ((x_1^i, y_1^i), \ldots, (x_n^i, y_n^i))$ where x_k^i represents the position of the k^{th} CpG associated with the CGI+SS and $y_k^i \in [0; 1]$ represents the mean methylation level for this probe accross a given dataset.

Given two vectors of size m and n respectively. A path w is a vector $(w_1^k, w_2^k)_{(k \in [1:K])}$ in $[1; m] \times [1; n]$ that verifies:

- $w^1 \in \{1\} \times [1; n] \cup [1; m] \times \{1\}$ (partial initialization)
- $\forall i \in \{1; 2\}$, $w_i^{k+1} = w_i^k$ or $w_i^{k+1} = w_i^k + 1$ (monotonicity and continuity)
- $w^K \in \{n\} \times [1; n] \cup [1; m] \times \{n\}$ (partial boundary condition)

For two CGI+SS profiles, we thus compute the DTW distance as:

$$DTW(CGI_1, CGI_2) = \min_{w \in Path} \sum_{k=1}^{length(w)} |y^1_{w^k_1} - y^2_{w^k_2}|^2 \tag{4.1}$$

The main differences between euclidean and dynamic time warping distance that is the pairing system between two signals are illustrated in 4.2. Moreover, the algorithm for DTW is described in 1.

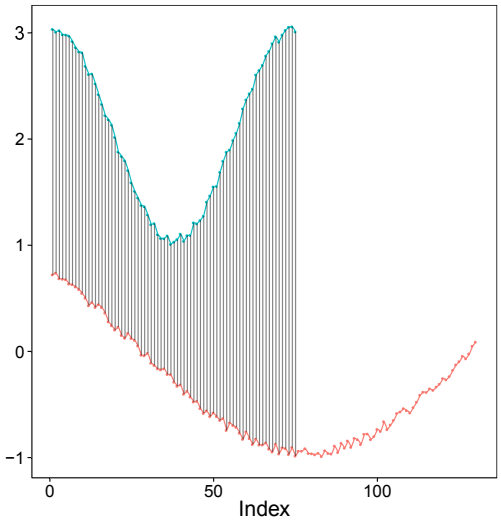


FIGURE 4.1: **Standard pairing between two signals.**

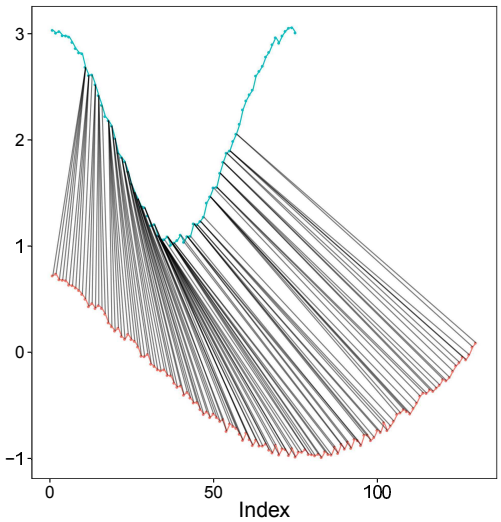


FIGURE 4.2: **Dynamic time warping pairing between two signals.**

Data: $\mathbf{x} = x_1, \ldots, x_n$ and $\mathbf{y} = y_1, \ldots, y_m$

Initialization: $distMat = matrix(0, nrow = n + 1, ncol = m + 1)$;

$distMat(0, 0) = 0$

for *i in 1:n* **do**

  $distMat(i, 0) = \infty$

end

for *j in 1:m* **do**

  $distMat(0, j) = \infty$

end

for *i in 1:n* **do**

  for *j in 1:m* **do**

    $distMat(i + 1, j + 1) =$

    $\|x_{i+1} - y_{i+1}\| + \min(distMat(i, j + 1), distMat(i, j), distMat(i + 1, j))$

  end

end

Result: $DTW(x, y) := distMat(n, m)$

Algorithm 1: DTW algorithm

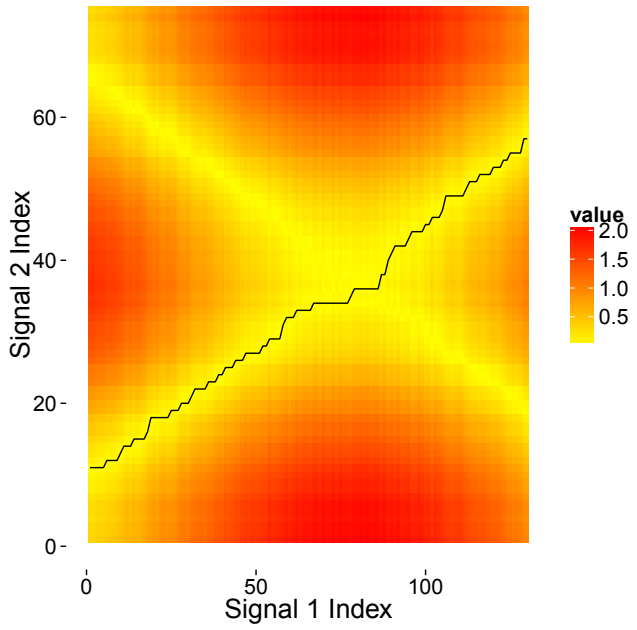


FIGURE 4.3: **Distance matrix between signal 1 and signal 2.** The dark line represents the path that minimizes the total distance between the two signals.

DTW is then applied for each pair of CGI+SS patterns to obtain a dissimilarity matrix that assesses the similarity in shapes between all the CGI+SS methylation profiles. Ward hierarchical clustering is then performed to assess the existence of characteristic patterns amongst the different datasets.

The number of significant clusters is assessed through bootstrapping ($n_{repeats} = 100$) on a random subset of CGI+SS of the initial dataset ($ratio = 80\%$ of the total number of CGI+SS) following Ben-Hur et al [Ben-Hur et al., 2002b].

4.4.6 Survival analysis

Overall survival was estimated using the Kaplan-Meier method [Kaplan and Meier, 1958] to compare the survival between the group of patients with a lower level of methylation in the hemi-methylated CGI+SS compared to the group of patients with a higher level of methylation. A multivariate Cox proportional hazards regression model [Cox and Oakes, 1984] was also fitted to estimate the additional value of this classification as a predictive factor for survival compared to other clinical parameters such as age, tumor size, lymph node status, receptor status and HER2/NEU status.

4.4.7 Computing the predictive power of methylation

We apply ridge [Hoerl et al., 1970] and LASSO [Tibshirani, 1996] multivariate regression methods to predict gene expression using the full CGI+SS methylation profiles as well as univariate least square regression when using only the averaged methylation from the whole CGI+SS profile. Following Acharjee et al. [Acharjee, 2013], we assess the predictive power of the methylation using the predictive goodness of fit R^2 which represents the squared Pearson correlation between observed and fitted values on an independent dataset. The estimation of the predictive power for each gene is obtained through 3-fold cross-validation averaged over 100 repeats. Parameters for both lasso and ridge regression methods were obtained by minimizing the mean squared error function using nested 3-fold cross-validation on the training dataset. The use of the predictive goodness of fit instead of the classic mean squared error as a score allows to compute a comparable score between different predictions. In particular, the mean squared error is highly affected by the absolute level of gene expression while the R^2 is invariant to scaling. It is also important to note that in this case the R^2 computed for least square regression is a prediction R^2 and not just a goodness-of-fit of the given dataset and therefore provides confidence on the generalization of the score on independent datasets.

4.5 Results

4.5.1 Classification of genes based on their CGI methylation profiles in normal and cancerous tissues

We first assess how promoter methylation profiles differ between genes, when for each gene we consider the average methylation profile across normal or cancerous samples. For that purpose, we collected high-density methylation datasets from the cancer genome atlas (TCGA) data portal providing more than 485K CpG methylation levels for 672 normal and cancerous samples from three tissues of origin: breast, colon and lung (4.2). For each CGI, we combine the probes in the CGI and in the shore and shelves of the CGI, defined as the regions up to 4kb outside of the CGI [Irizarry et al., 2009], in a unique CGI, shores and shelves (CGI+SS) methylation profile. We restrict our analysis to the 1827 CGI+SS where at least 20 CpG probes are measured by the technology in order to have high enough coverage to measure the methylation variation within each CGI+SS. For each of the three tissue of origin, and each normal or cancerous set of tissues, we compute the average methylation profile of each CGI+SS by averaging the methylation values of each CpG across the samples. Hence we compute $3 \times 2 = 6$ average profile for each CGI+SS, with we refer to below as *CGI+SS signatures*.

TABLE 4.2: **Concordance analysis of CGI+SS patterns clusters between normal tissues.**

Clusters	**Colon**	
Breast	1	2
1	**1560**	9
2	113	**145**
Clusters	**Lung**	
Colon	1	2
1	**1610**	7
2	63	**147**
Clusters	**Breast**	
Lung	1	2
1	**1549**	20
2	68	**190**

To assess the diversity of CGI+SS signatures across genes, we perform an unsupervised classification of all signatures for each of the 6 types of samples, using Ward hierarchical clustering. Since different CGI+SS may contain a different number of GpG probes, we use a specific distance based on dynamic time warping to compare signatures of different lengths. 4.4 (panel A/C/E) shows the CGI+SS clustering obtained for signatures measured on normal samples from breast (resp. lung and colon) samples. We observe two

stable clusters, which are largely conserved across the 3 tissues of origin (Table 2). To clarify the types of signatures captured by each cluster, we represent on a standardized CGI+SS x-axis the 10 medoid CGI+SS signatures for each cluster and each tissue (4.4). We clearly observe that the large cluster 1, which contains about 90% of all CGI+SS, corresponds to hypo-methylated islands with hemi-methylated CGI shores and hyper-methylated CGI shelves, while the smaller cluster 2 contains about 10% of CGI+SS which are fully hyper-methylated. A closer look at cluster 1 shows that, in some cases, the variation of methylation between islands and shores is unclear, in the sense that some shores are fully hypo-methylated. As CGIs, shores and shelves regions are delimited based on somehow arbitrary criteria, a systematic analysis of these signatures could lead to a refinement of currently accepted boundaries.

Performing the same unsupervised classification independently on signatures obtained from the three types of cancerous tissues leads to different results, with the apparition of a third stable cluster (4.4 panel B/D/F). Comparing the clusters of normal and cancerous tissues shows that, for all types of tissues, the first two clusters found in cancerous tissues are mostly composed of CGI+SS of the corresponding clusters in normal tissues, while the CGI+SS in the third cluster, specifically found in cancerous tissues, tend to come evenly from both clusters in normal tissues (4.3). A look at representative signatures of each cluster (4.5,4.6,4.7) confirms that clusters 1 and 2 contain respectively hypo- and hyper-methylated profiles, just like the respective clusters in normal tissues, while cluster 3 contains CGI+SS signatures which are partly methylated. Separating the CGI+SS in cluster 3 into sub-clusters "3up" and "3down", depending on whether they are in cluster 1 or 2 in normal tissues, we further see that the level of methylation of CGI+SS signatures in the "3up" sub-cluster tends to be lower than the level of methylation of CGI+SS signatures in the "3down" sub-cluster. Interestingly, cluster 3 is mostly conserved between tissues (4.11), suggesting that these epigenetic variations might be associated with a tissue-independent carcinogenesis process.

In summary, this global analysis of methylation signatures suggests the existence of four types of CGI+SS largely conserved across tissues: the majority of them remains hypo-methylated on the CGI and hyper-methylated on the shores and shelve in normal and cancerous tissues (cluster 1); a minority is hyper-methylated in normal and cancerous tissues (cluster 2); finally, a fraction of CGI+SS signatures is hypo-methylated in normal tissues and partly methylated in cancerous tissues (cluster 3up), while another fraction is hyper-methylated in normal tissues and partly methylated in cancerous ones (cluster 3down). To clarify whether these four categories or CGI+SS are associated to particular biological functions, we performed a gene functional enrichment analysis [Yu et al., 2012] of the genes associated to the CGI+SS in each of the four categories, for each tissue. Results are shown in 4.8, 4.9, 4.10. Restricting ourselves to Gene Ontology (GO)

TABLE 4.3: **Concordance analysis of CGI+SS patterns clusters from normal to cancerous tissues.** Each table represents the concordance of clusters between normal and cancerous clustering analysis. Bold numbers in the diagonal shows the stability of clusters between normal and cancerous tissues.

Breast	**Normal**	
Cancerous	1	2
1	**1231**	21
2	9	**109**
3	329	128
Lung	**Normal**	
Cancerous	1	2
1	**1128**	12
2	18	**168**
3	471	30
Colon	**Normal**	
Cancerous	1	2
1	**1112**	11
2	13	**106**
3	548	37

biological processes associated to at least 20 genes, we found that the large cluster 1 is mostly enriched in genes involved in metabolic processes, while the cancer-specific cluster 3up is enriched in genes involved in developmental processes. There was no significant functional enrichment for genes in cluster 2 and 3down.

4.5.2 Cancer-specific methylation does not repress gene expression but instead targets genes lowly expressed in normal tissues

CGI methylation is often associated with gene expression silencing. We therefore assess whether the CGI+SS clusters defined above, corresponding roughly to lowly methylated (clusters 1), highly methylated (cluster 2) or partially methylated in cancer (cluster 3) CGI+SS, are associated with different mean levels of gene expression. In normal breast tissues, we indeed observe that genes near hypo-methylated islands in cluster 1 are slightly but significantly less expressed than genes near an hyper-methylated islands in cluster 2 (4.12, $P_{Breast} = 0.02$). There is however no significant difference between the two clusters in normal lung tissues (4.12, $P_{Lung} = 0.39$), and we could not test the hypothesis on normal colon tissues since we have none with both methylation and expression data (4.1). In cancerous samples, we observe that genes near a CGI+SS in the cancer-specific cluster 3 have a significantly lower expression than other genes (4.12, $P_{Breast}, P_{Lung}, P_{Colon} < 10^{-16}$), particularly for the genes near a CGI+SS in the "3up" cluster. As genes in the "3up" cluster are hypo-methylated in normal tissues, this could suggest that their cancer-specific methylation is a way to repress their expression

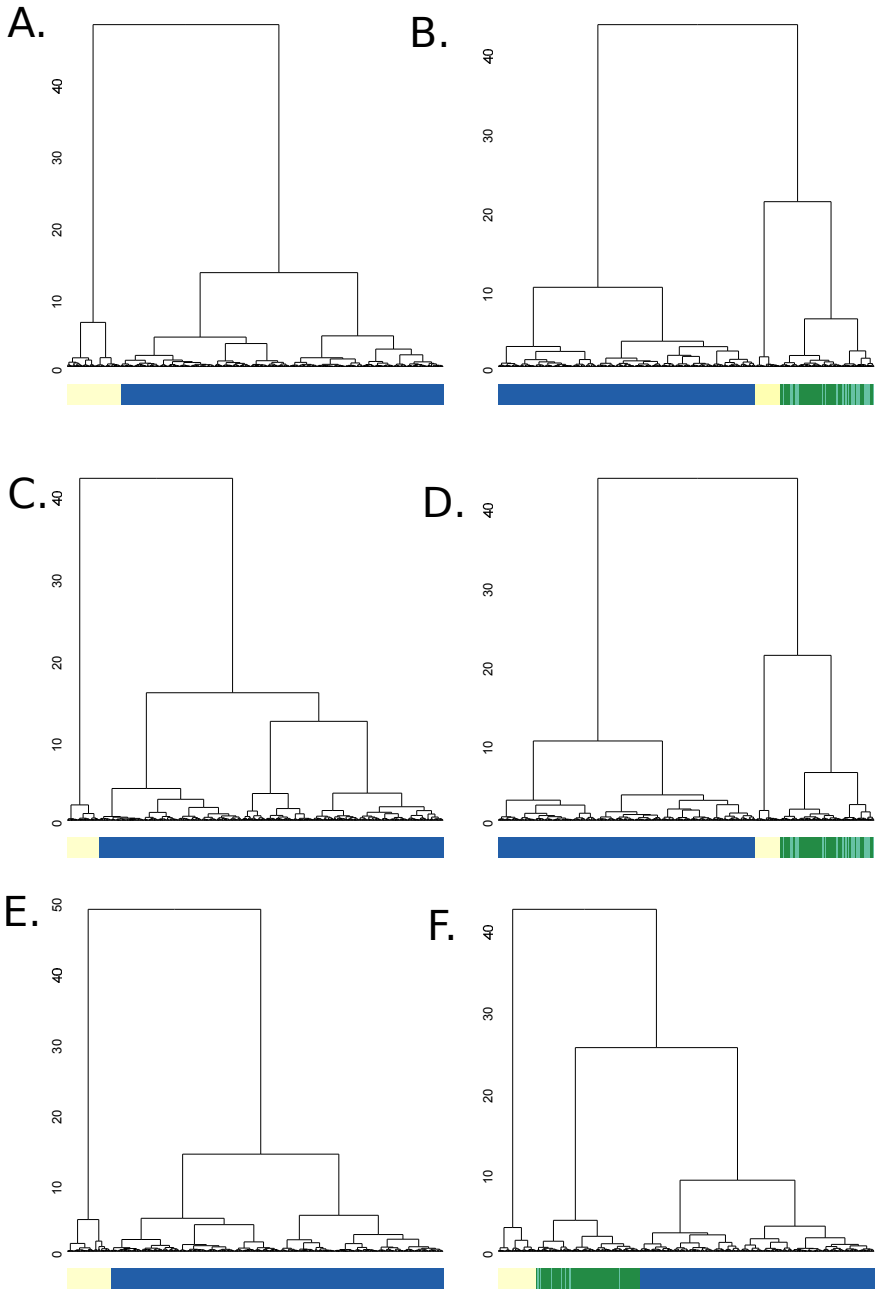


FIGURE 4.4: **CGI+SS patterns in breast tissues.** Hierarchical clustering of CGI+SS DNA methylation patterns for breast normal tissues (panel A) and breast cancerous tissues (panel B) using DTW as a distance metric and a "Ward" linkage. The colorbar represents the clusters association (blue for hypomethylated cluster 1, yellow for cluster 2, dark green for cluster 3down, light green for cluster 3down).

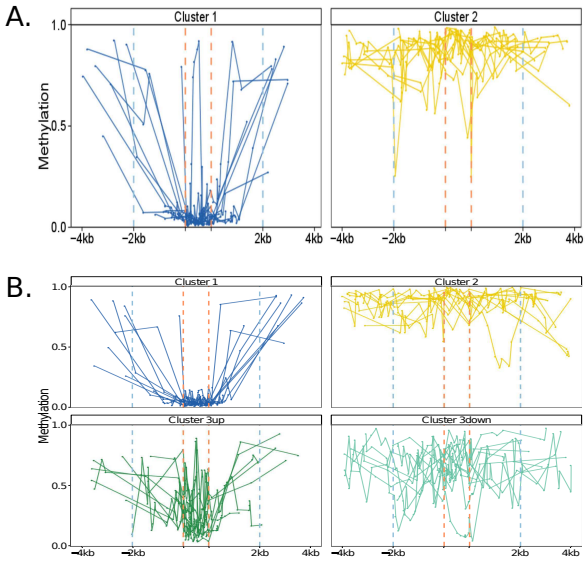


FIGURE 4.5: **Characteristic profiles for each clusters.** Visualization of the CGI+SS DNA methylation signatures as condensed profiles from the 10 medoids profiles for each clusters in breast normal (panel A) or cancerous (panel B) tissues. The two orange dashed lines represent the normalized 1kb long CGI region while the two blue lines represent the 2kb limit between shores and shelves regions.

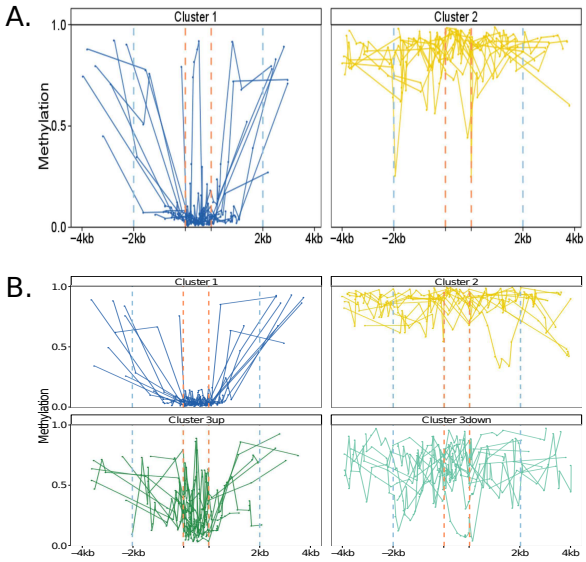


FIGURE 4.6: **Characteristic profiles for each clusters.** Visualization of the CGI+SS DNA methylation signatures as condensed profiles from the 10 medoids profiles for each clusters in colon normal (panel A) or cancerous (panel B) tissues. The two orange dashed lines represent the normalized 1kb long CGI region while the two blue lines represent the 2kb limit between shores and shelves regions.

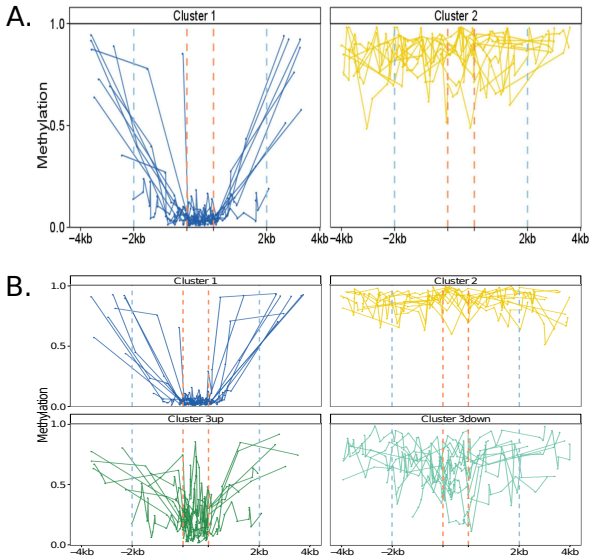


FIGURE 4.7: **Characteristic profiles for each clusters.** Visualization of the CGI+SS DNA methylation signatures as condensed profiles from the 10 medoids profiles for each clusters in lung normal (panel A) or cancerous (panel B) tissues. The two orange dashed lines represent the normalized 1kb long CGI region while the two blue lines represent the 2kb limit between shores and shelves regions.

FIGURE 4.8: **Gene Ontology analysis given the cluster assignment for cancerous breast tissues. Cluster**: Cluster assignment of a gene the CGI+SS methylation pattern . **Description**: Description of the biological processes enriched (top 10 ranked by cluster ratio). **Cluster ratio** (A/B): Ratio between the number of genes (A) associated with the biological process and the total number of genes (B) in a given cluster. **P-val**: Fisher's exact test p-value adjusted for multiple testing.

Cluster	Description	p-value
1	metabolic process	8.85×10^{-168}
1	organic substance metabolic process	5.31×10^{-156}
1	cellular metabolic process	5.56×10^{-151}
1	primary metabolic process	7.47×10^{-150}
1	response to stimulus	5.91×10^{-103}
1	cellular response to stimulus	1.64×10^{-79}
1	cellular component organization or biogenesis	8.00×10^{-79}
1	cellular component organization	6.17×10^{-77}
1	single-organism metabolic process	3.35×10^{-67}
1	localization	1.04×10^{-66}
2	**No significant enrichment**	
3down	**No significant enrichment**	
3up	multicellular organismal process	1.89×10^{-212}
3up	single-multicellular organism process	7.13×10^{-209}
3up	developmental process	2.71×10^{-198}
3up	single-organism developmental process	3.94×10^{-197}
3up	multicellular organismal development	2.32×10^{-191}
3up	anatomical structure development	5.25×10^{-180}
3up	system development	6.83×10^{-169}
3up	organic cyclic compound biosynthetic process	1.13×10^{-131}
3up	aromatic compound biosynthetic process	1.21×10^{-130}
3up	cellular nitrogen compound biosynthetic process	1.21×10^{-130}

FIGURE 4.9: **Gene Ontology analysis given the cluster assignment for cancerous colon tissues. Cluster**: Cluster assignment of a gene the CGI+SS methylation pattern . **Description**: Description of the biological processes enriched (top 10 ranked by cluster ratio). **Cluster ratio** (A/B): Ratio between the number of genes (A) associated with the biological process and the total number of genes (B) in a given cluster. **P-val**: Fisher's exact test p-value adjusted for multiple testing.

Cluster	Description	p-value
1	single-organism metabolic process	1.79×10^{-76}
1	protein metabolic process	7.20×10^{-70}
1	cellular protein metabolic process	6.70×10^{-62}
1	establishment of localization	5.80×10^{-56}
1	transport	2.59×10^{-54}
1	response to stress	1.39×10^{-51}
1	macromolecule modification	2.89×10^{-47}
1	phosphorus metabolic process	4.36×10^{-47}
1	phosphate-containing compound metabolic process	1.50×10^{-46}
1	cellular protein modification process	2.63×10^{-45}
2	**No significant enrichment**	
3down	**No significant enrichment**	
3up	multicellular organismal process	3.51×10^{-241}
3up	single-multicellular organism process	4.35×10^{-235}
3up	developmental process	3.16×10^{-216}
3up	single-organism developmental process	2.64×10^{-213}
3up	multicellular organismal development	1.24×10^{-204}
3up	anatomical structure development	5.60×10^{-198}
3up	system development	6.78×10^{-185}
3up	macromolecule biosynthetic process	3.66×10^{-155}
3up	cellular developmental process	2.77×10^{-154}
3up	cellular differentiation	2.08×10^{-153}

FIGURE 4.10: **Gene Ontology analysis given the cluster assignment for cancerous lung tissues. Cluster**: Cluster assignment of a gene the CGI+SS methylation pattern . **Description**: Description of the biological processes enriched (top 10 ranked by cluster ratio). **Cluster ratio** (A/B): Ratio between the number of genes (A) associated with the biological process and the total number of genes (B) in a given cluster. **P-val**: Fisher's exact test p-value adjusted for multiple testing.

Cluster	Description	p-value
1	metabolic process	8.79×10^{-198}
1	organic substance metabolic process	5.46×10^{-185}
1	cellular metabolic process	5.01×10^{-181}
1	primary metabolic process	8.73×10^{-179}
1	macromolecule metabolic process	1.71×10^{-149}
1	cellular macromolecule metabolic process	1.73×10^{-139}
1	nitrogen compound metabolic process	2.19×10^{-120}
1	cellular nitrogen compound metabolic process	2.49×10^{-115}
1	organic cyclic compound metabolic process	9.72×10^{-114}
1	cellular aromatic compound metabolic process	2.35×10^{-110}
2	**No significant enrichment**	
3down	**No significant enrichment**	
3up	multicellular organismal process	1.83×10^{-246}
3up	single-multicellular organism process	6.68×10^{-242}
3up	single-organism developmental process	2.20×10^{-221}
3up	developmental process	2.20×10^{-221}
3up	multicellular organismal development	5.86×10^{-211}
3up	anatomical structure development	4.94×10^{-199}
3up	system development	2.42×10^{-187}
3up	cellular developmental process	4.81×10^{-145}
3up	cell differentiation	2.83×10^{-144}
3up	organ development	8.20×10^{-130}

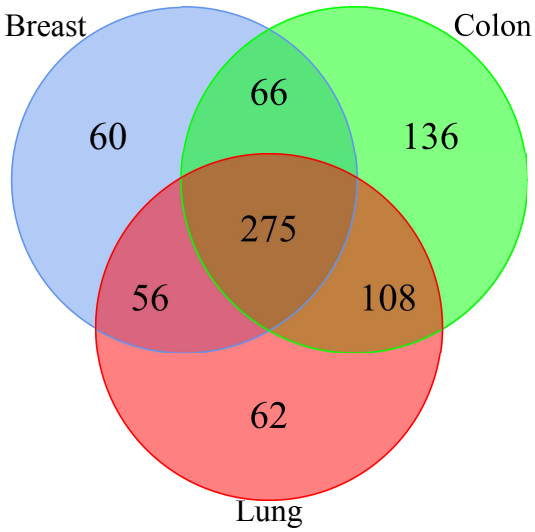


FIGURE 4.11: **Inter-tissue stability of the cancerous-specific cluster.** Venn diagram representing the stability of the cancer-specific CGI+SS cluster 3 between cancerous tissues.

in cancer. However, a closer look at the expression of these genes in normal tissues (4.12) shows that they are already lowly expressed in normal tissues. This suggests that instead of activating CGI methylation to silence to genes, cancer cells instead activates CGI methylation of hypo-methylated genes which are already lowly expressed in normal tissues.

4.5.3 Cancer-specific methylation is an independent predictor of patient survival in breast cancer

Our analysis so far compares CGI+SS in terms of their mean methylation across a set of samples and does not take into account between-sample variations. CGI+SS associated with cluster 1 (resp. 3) are hypo- (resp. hyper-)methylated on average, which indicates that there is little to no variations between samples. However, signatures of CGIs in the cancer-specific cluster 3 are partly methylated, which can either hide the fact that they are hemi-methylated for most cancerous samples, or that they are highly variable between samples. We therefore assess whether the partial methylation of CGI+SS signatures in cluster 3 is related to an overall increase (for cluster 3up) or decrease (for cluster 3down) in methylation for all or most of the patients, or if this it is caused by a subset of patients that become hyper- (resp. hypo-)methylated for these CGI+SS.

For that purpose, we first summarize the methylation of each CGI+SS on each breast cancer sample by a single value, the average methylation of the probes in the CGI+SS.

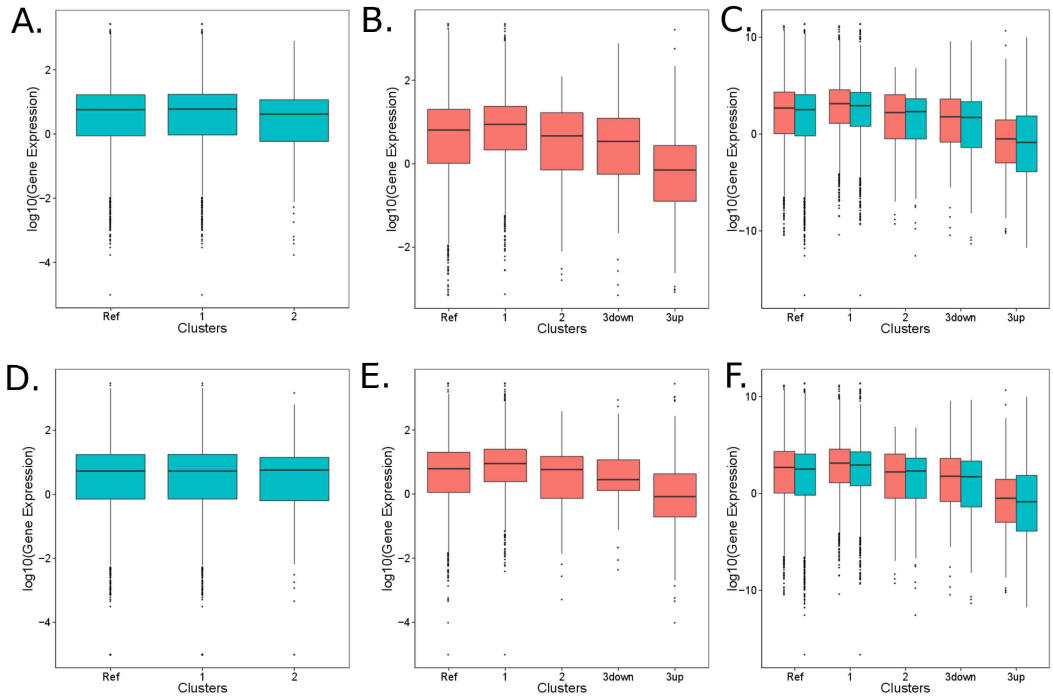


FIGURE 4.12: **cluster characteristics analysis in breast tissues.** Gene expression distribution for genes based on the cluster assignment of their associated CGI+SS. **Panel A/D.** Gene expression distribution in normal tissues shows a slight repression for genes associated with cluster 2 (hypermethylated CGI+SS profiles). "Ref" represents the genome-wide gene expression distribution (Panel A=breast, Panel D=lung) **Panel B/E.** Gene expression profiles in cancerous tissues shows high repression for genes associated with cluster 3 and specifically cluster "3up" (hemi-methylated CGI+SS profiles) (Panel B=breast, Panel E=lung). **Panel C/F.** Gene expression profiles in both normal and cancerous tissues using the cluster assignement in cancerous tissues shows that genes associated with cluster "3up" in cancerous tissues define a cluster of genes already repressed in normal tissues (Panel C=breast, Panel F=lung).

We then represent each sample by the vector of methylation values of the CGI+SS in cluster 3up, and perform a Ward hierarchical clustering of the cancerous samples based on this representation. The resulting clustering is shown in 4.13, where in addition we indicate the ER+, HER2 and survival information for each patient. We observe that the distribution methylation values is very bimodal, and that the hyper-methylation of a given CGI+SS from cluster 3up generally happens in a subset of patients only. Interestingly, we see that the same subset of patients tends to be simultaneously hyper-methylated for all CGI+SS from cluster 3up, suggesting that hyper-methylation of these islands is a characteristics of a subset of the tumors. This allows us to divide the set of breast cancer patients in three clusters given the level of methylation in cluster 3up as either "low", "intermediate", or "high" 4.13. Interestingly, distinguishing patients given the level of methylation from the CGI+SS in cluster 3up is significantly predictive of the patient survival (4.13, log-rank, $p = 0.01$). Surprisingly, the cluster with the lowest survival is the "intermediate" cluster encompassing a portion but not all of the triple

negative breast cancers (65% in cluster 3up "low", 32% in cluster 3up "intermediate" and only 3% in cluster 3up "high"). A multivariate Cox proportional hazards regression model fitted with available clinical parameters (tumor size, lymph node status, hormone receptor status, HER2/NEU status and patient's age) further shows that this stratification of patients based on the methylation level of genes in cluster 3up adds prognostic value independently of other clinical features 4.4. These results support the existence of a CpG island methylator phenotype (CIMP) as introduced by Toyota et al. [Toyota et al., 1999a] that is clinically relevant to assess the survival of patients. More importantly, they suggest that low survival might not be associated with a positive or negative CIMP, but with an intermediate phenotype termed as CIMP-low [Hughes et al., 2013].

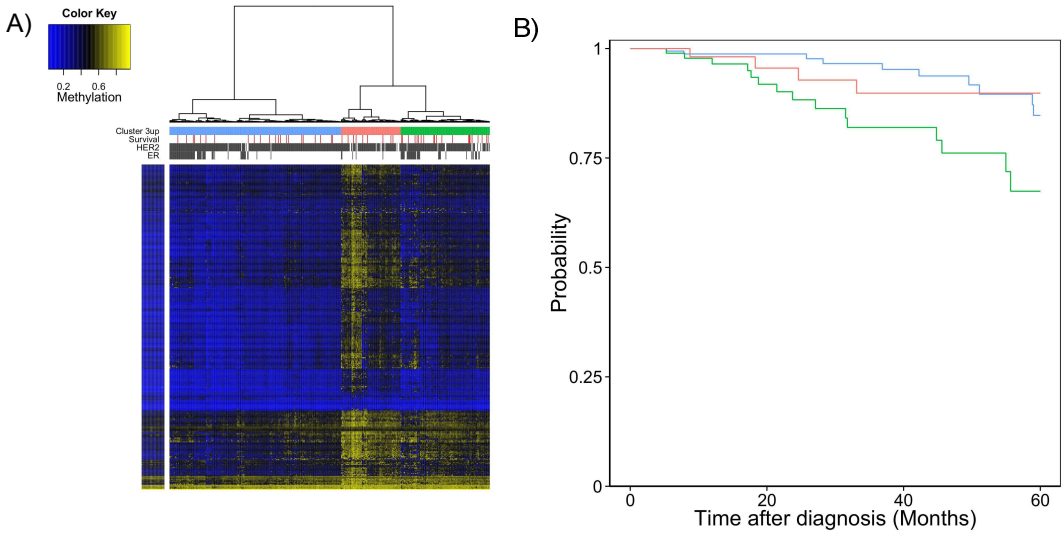


FIGURE 4.13: **Cluster 3up methylation is a predictive factor for survival of patients in breast cancer patients. Panel A.** Hierarchical clustering of breast cancer patients given the average methylation level of all the CGI+SS associated with cluster 3up. The row color bar represents the average methylation level for the same CGI+SS in healthy breast tissues. The column color bar gives clinical information about the patients such as ER and HER2 statuses (grey for negative and white for positive), survival information (white for positive overall survival within 5 years and red for death within 5 years). The top row of the column color bar represents the three classes distinguished by methylation profiles in cluster 3up (blue for cluster 3up "low", green for cluster 3up "intermediate" and pink for cluster 3up "high"). **Panel B.** Kaplan-Meier estimate of breast cancer patient survival given the cluster 3up class (blue for cluster 3up "low", green for cluster 3up "intermediate" and pink for cluster 3up "high") shows that cluster 3up "intermediate" patients have a significantly higher risk of death within 5 years than either cluster 3up "low" or "high" patients (Log-rank, p= 0.01).

A similar analysis on CGI+SS associated with cluster 3down is less conclusive, and does not clearly cluster patients in separate clusters (4.14). A lack of sufficient survival data for colon and lung tissues prevented a similar analysis for these tissues.

TABLE 4.4: Multivariate Cox regression analysis including the level of methylation in the cancer-specific cluster "3up" in addition to significant clinical variables for breast cancer.

Clinical variable (Reference)	HR (95% CI)	p-value
Cluster 3up (Low vs intermediate)	3.44 (1.44-8.23)	0.007
Cluster 3up (Low vs high)	1.92 (0.50-7.34)	0.34
(ER,HER2) (-/- vs +/-)	0.37 (0.15-0.88)	0.026
(ER,HER2) (-/- vs -/+)	1×10^{-8} (0-Inf)	1
(ER,HER2) (-/- vs +/+)	0.53 (0.09-2.94)	0.46
Lymph Node (Negative)	4.51 (1.63-12.44)	0.004

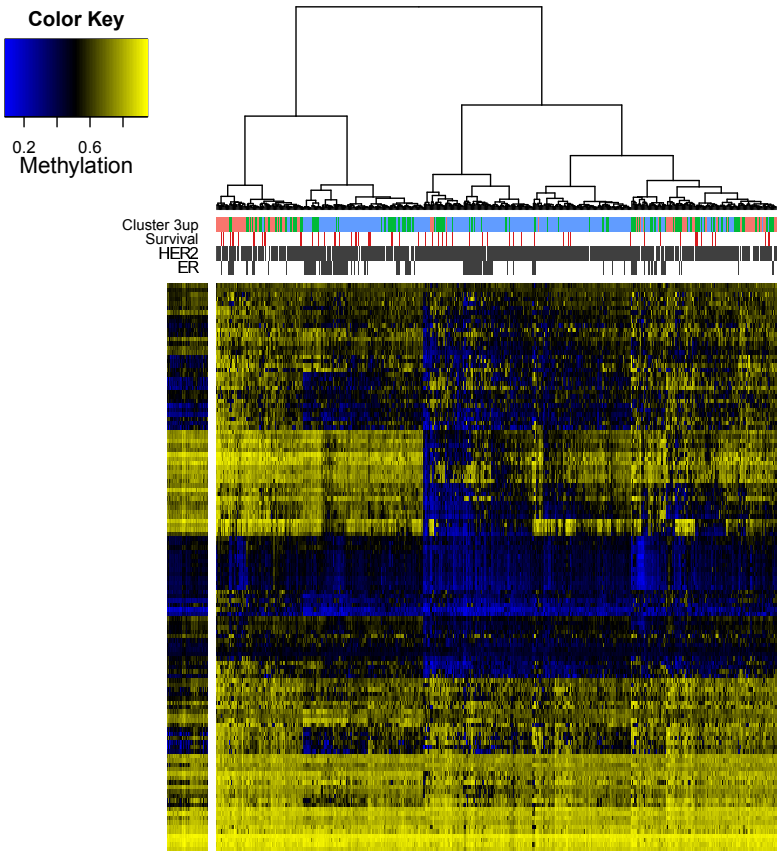


FIGURE 4.14: **Hierarchical clustering of breast cancer patients based on the average methylation level of CGI+SS associated with cluster 3down.**

4.5.4 Methylation of CpG in CGI shores is negatively correlated with gene expression.

Our analysis so far compares CGI+SS to one another, by looking at their average methylation profiles across collections of samples. We found no clear evidence for a correlation between mean methylation level of a CGI and mean expression level of the corresponding genes, but this may be due to the fact that many other factors impact the expression level of a gene, including biological and technical ones. Another way to assess how methylation impacts expression is to look, for each given gene, how variations in expression across samples correlates with variations in methylation of nearby CGIs. For each set of samples (split by tissue of origin and normal/cancerous state), we measure the strength of association between methylation and expression for each gene by computing a predictive goodness of fit R^2 which represents the level of gene expression variation explained by CGI+SS methylation variation. This coefficient is calculated either when the CGI+SS methylation status is summarized by the mean methylation values of all the probes, or by using the full CGI+SS methylation information of each probe.

We observe that the full CGI+SS methylation profile is predictive of gene expression for a subset of genes in each dataset, and that this predictive power is significantly higher than using only the average CGI+SS methylation (4.15, $P_{Breast} < 10^{-16}$, $P_{Lung} = 1.3 \times 10^{-16}$, $P_{Colon} = 3.2 \times 10^{-5}$). We provide in 4.5 the list of the top 50 genes based on their predictive score in cancerous breast, colon and lung tissues. Among the 2374 genes studied, 139 genes are associated with more than one CGI+SS. For these genes, the predictive power is computed using the CGI+SS closest to the TSS. Using all the CGI+SS for these genes do not yield significant improvement over taking only the CGI+SS closest to the TSS except for breast tissues ($P_{Breast} = 0.003$, $P_{Lung} = 0.15$, $P_{Colon} = 0.62$). We also observe no association between the predictive goodness of fit R^2 and the CGI+SS clusters described above ($P_{Breast} = 0.48$, $P_{Lung} = 0.47$, $P_{Colon} = 0.44$).

Since the predictive power of multivariate models based on all CpG probes in a CGI+SS is larger than the predictive power of the mean methylation value only, we now investigate which CpG in a CGI+SS are particularly important predictors of expression. For that purpose, we measure the correlation between the methylation of individual CpG and gene expression for the 50 genes with the largest predictive R^2, and summarize the correlation values based on the position of the probe in the CGI+SS in 4.16. As expected, we observe overall a negative correlation between methylation and gene expression, and notice that this association is stronger in CGI shores than in the CGI itself. This is coherent with results in [Irizarry et al., 2009] stating that variations in the CGI are less critical than variations in proximity regions of the CGI. Performing the

TABLE 4.5: **Genes regulated by methylation in different cancerous tissues.** **Gene**: Top scoring genes ranked by the predictive power of methylation to predict gene expression variation. **Score**: R^2 score associated.

Breast		Colon		Lung	
Gene	Score	Gene	Score	Gene	Score
DQX1	0.699	*C11orf93*	0.786	*PTPRCAP*	0.639
IRS2	0.692	*FAM24B*	0.695	*HOXB2*	0.620
GPSM3	0.669	*SCAND3*	0.679	*LOC254559*	0.606
FOXC1	0.642	*CLIC6*	0.667	*KLC4*	0.598
PSMB9	0.624	*TBX18*	0.639	*SEMA4G*	0.597
HOXC10	0.623	*C11orf92*	0.617	*COL25A1*	0.596
NDRG2	0.623	*FOXD2*	0.601	*HOXC13*	0.591
MAPT	0.607	*ACSF3*	0.586	*SOX9*	0.580
STC2	0.606	*FKBP10*	0.583	*DUSP4*	0.579
ZNF502†	0.585	*TACSTD2*	0.576	*HOXA10*	0.578
PTPRCAP	0.583	*TMEM176B*	0.573	*SIM2*	0.574
SCAND3	0.583	*TMEM176A*	0.568	*FKBP10*	0.568
SLC1A4	0.580	*FAM50B*	0.563	*VAX2*	0.563
TAP1	0.576	*SC65*	0.563	*FAM50B*	0.563
DBNDD2	0.565	*ZIC5*	0.555	*TPD52L1*	0.560
OTX1	0.564	*EFNA3*	0.535	*DQX1*	0.554
TCF7	0.561	*SYS1-DBNDD2*	0.532	*FAM24B*	0.547
LY6G6C	0.561	*DLX6AS*	0.528	*ZNF502*	0.539
FERMT3	0.560	*HOXB6*	0.525	*CSNK1E*	0.531
ZIC4	0.559	*C5orf38*	0.523	*IRX2*	0.528
HLA-B	0.556	*H19*	0.515	*KCTD1*	0.527
GDF9	0.551	*PCDHGA5*	0.512	*ENO3*	0.524
SOX9	0.551	*ME3*	0.502	*ISL2*	0.506
CELSR1	0.550	*CHFR*	0.501	*STMN1*	0.503
SYS1-DBNDD2	0.549	*GPR120*	0.499	*TRIM15*	0.501
HLA-E	0.549	*SLC35C1*	0.497	*HLTF*	0.500
CYP1B1	0.541	*SLC5A6*	0.487	*DMRTA2*	0.497
RUNX3	0.540	*RGL2*	0.481	*ZIC4*	0.497
KIAA1949	0.537	*HOXB2*	0.481	*ALX3*	0.496
RIPK4	0.531	*MGMT*	0.477	*IRS2*	0.494
TPPP2	0.530	*TAP1*	0.474	*SC65*	0.488
HLA-F	0.530	*ETV4*	0.474	*DCLRE1A*	0.485
PPP1R3C	0.529	*PCDHGA12*	0.466	*LIME1*	0.482
HOXB5	0.528	*HOXD9*	0.461	*H2AFY2*	0.469
CELSR3	0.527	*DBNDD2*	0.458	*KIAA1949*	0.468
B3GNT5	0.525	*GPSM3*	0.456	*ZIC5*	0.456
ME3	0.524	*KLC4*	0.454	*BMI1*	0.453
TMC8	0.523	*FARP1*	0.452	*IRX4*	0.448
AIF1	0.522	*FTH1*	0.450	*C11orf93*	0.443
SLC39A6	0.521	*HSPA1L*	0.443	*DNTTIP1*	0.442
HOXC11	0.512	*FSCN1*	0.441	*GATA6*	0.440
ERBB2	0.505	*MUC12*	0.441	*HIST3H2A*	0.434
TBC1D10C	0.503	*WIT1*	0.440	*PIK3R3*	0.433
SIM2	0.503	*SS18L1*	0.439	*PIM3*	0.431
CAMK2N1	0.502	*HOXA1*	0.439	*FAT1*	0.431
RGMA	0.499	*AMH*	0.438	*HOXC9*	0.430
LOC100132215	0.497	*HOXA5*	0.433	*SOCS2*	0.429
PAX6	0.497	*ZNF518B*	0.430	*MGC29506*	0.426
VANGL2	0.496	*EMX1*	0.430	*RDH5*	0.425
DDHD2	0.487	*PDX1*	0.429	*CHST11*	0.424

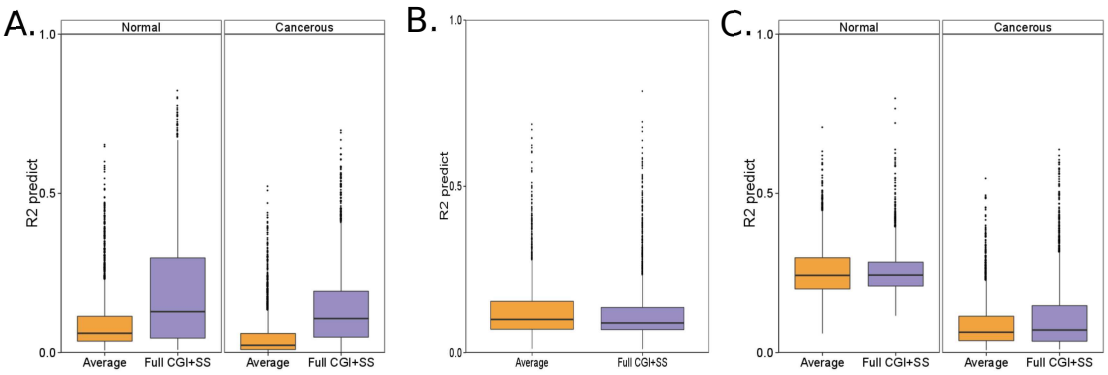


FIGURE 4.15: **Impact of DNA methylation in gene expression prediction.** Predictive power distribution of DNA methylation for gene expression using either the average CGI methylation and least squares (orange) or the full CGI+SS profile and lasso regression (purple) shows that a more complex model allows to better predict gene expression variations in both normal and cancerous tissues (panel A= breast, panel B= colon, panel C=lung).

same analysis by varying the number of genes selected to compute correlations from 20 to 100 gave similar results.

4.5.5 Regulation of gene expression by DNA methylation is tissue-specific and the process is altered in cancer tissues but overall targets transcription factors.

Results in the previous section suggest that for a subset of genes, a regulation of gene expression by methylation of CpG in CGI+SS is likely. To assess whether this regulation is conserved across tissues, we compare the predictive powers of methylation for each genes when it is computed on normal or cancerous samples from different tissues. As shown in 4.17, however, we observe little correlation between the predictive power across tissues in normal and in cancer samples, suggesting that methylation regulates the expression of genes in a tissue-specific manner ($R^{2,Normal}_{Breast/Lung} = 0.04$, $R^{2,Cancerous}_{Breast/Lung} = 0.17$, $R^{2,Cancerous}_{Lung/Colon} = 0.07$, $R^{2,Cancerous}_{Colon/Breast} = 0.06$). We also observe very little correlation between predictive powers in normal and cancerous tissues, which could suggests a shift of the epigenetic regulation mechanism during cancer development (4.18, $R^2_{Breast} = 0.04$, $R^2_{Lung} = 6 \times 10^{-7}$).

Many mechanisms besides DNA methylation are involved in gene expression regulation. In particular, transcription factors (TF) play a critical role in the recruitment of RNA polymerase that allows gene transcription [Struhl, 1999]. We noticed that the list of the 50 genes with the largest predictive R^2 score in each tissue is significantly enriched in TFs as collected from [Zhang et al., 2012], suggesting that methylation plays an

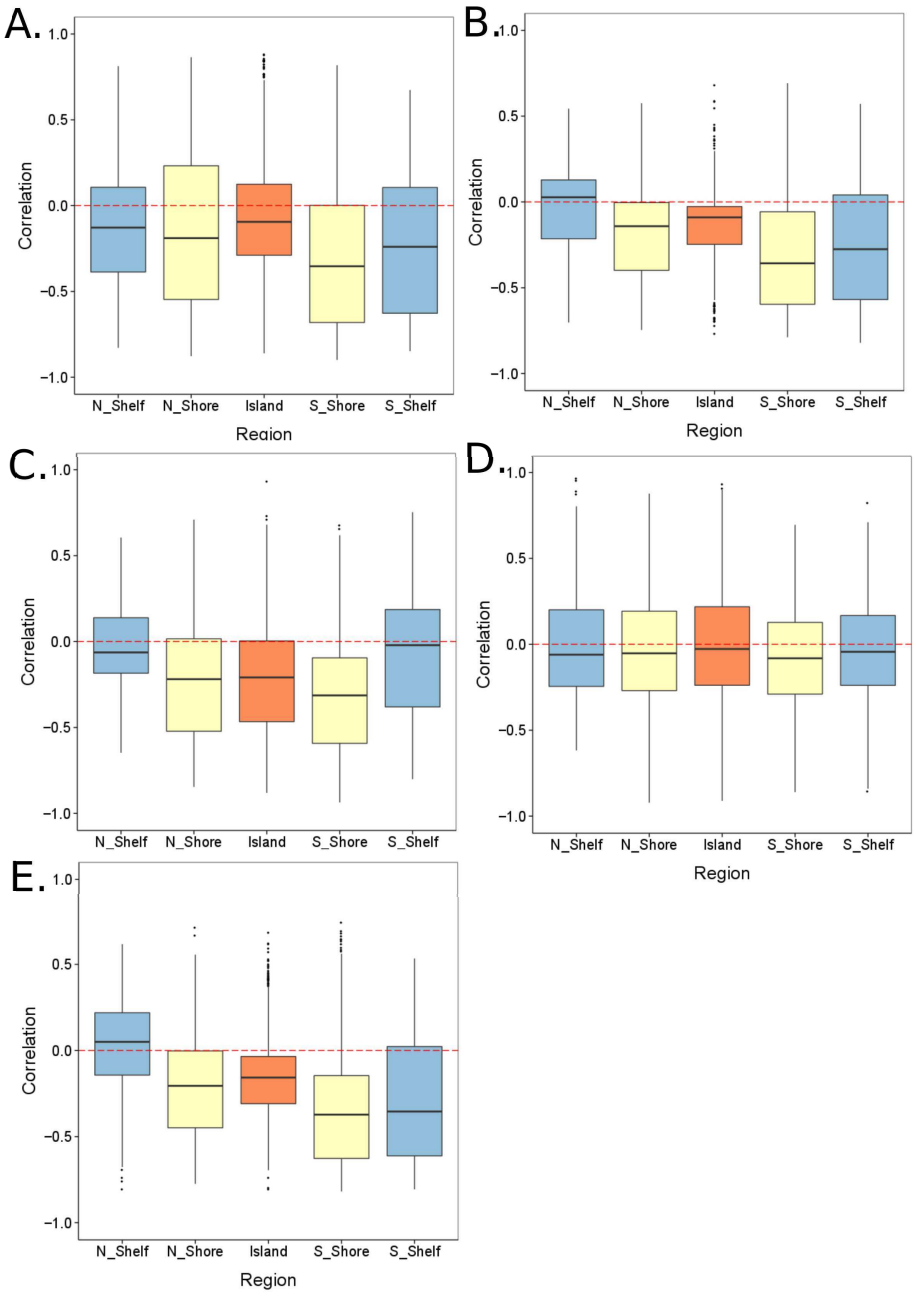


FIGURE 4.16: **Methylation association with gene expression by regions.** Distribution of the correlation between individual probes and gene expression variation for breast top 50 genes ranked by their predictive score by regions related to the CGI exhibits a stronger association for probes located outside of the CGI particularly in shores regions (panel A= normal breast tissues, panel B= cancerous breast tissues, panel C= cancerous colon tissues, panel D= lung normal tissues, panel E= lung cancerous tissues).

important role in the gene regulatory process of transcription factors ($P_{Breast} = 0.03$, $P_{Lung} = 3 \times 10^{-4}$, $P_{Colon} = 0.02$). Using the TF list obtained from [Vaquerizas et al., 2009] yields similar conclusions, as well as varying the number of genes selected from 20 to 100.

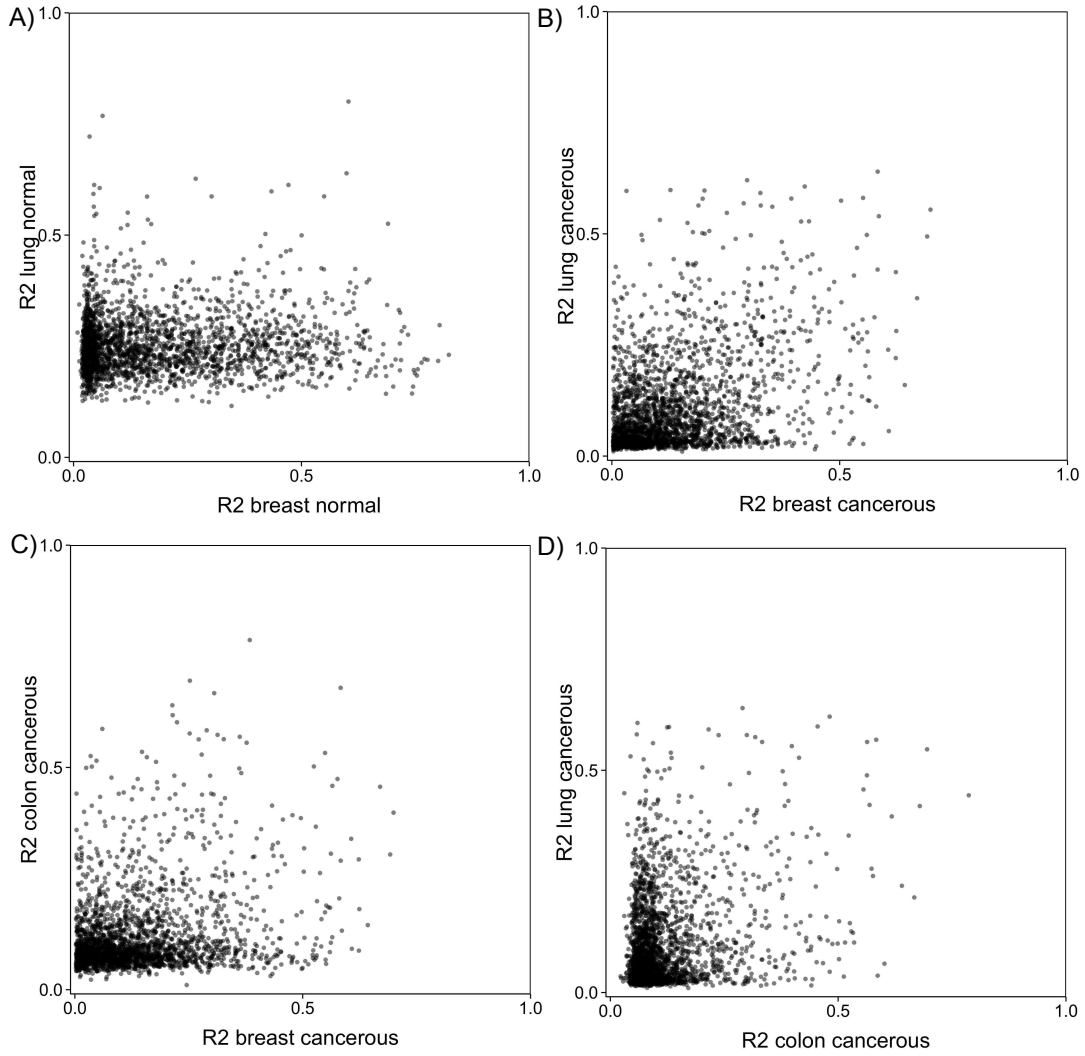


FIGURE 4.17: **Tissue-specificity of epigenetic regulation.** Scatterplot between the predictive power of DNA methylation for gene expression in normal and cancerous between different tissues ($R^{2,Normal}_{Breast/Lung} = 0.04$, $R^{2,Cancerous}_{Breast/Lung} = 0.17$, $R^{2,Cancerous}_{Lung/Colon} = 0.07$, $R^{2,Cancerous}_{Colon/Breast} = 0.06$).

4.5.6 Copy number variations in cancer is an independent factor in gene expression regulation.

In cancer, aberrant DNA copy number variations (CNVs) can have an important impact on gene expression phenotypes [Stranger et al., 2007]. Since genome-wide DNA copy number information is available for all samples analyzed in this study, we now perform

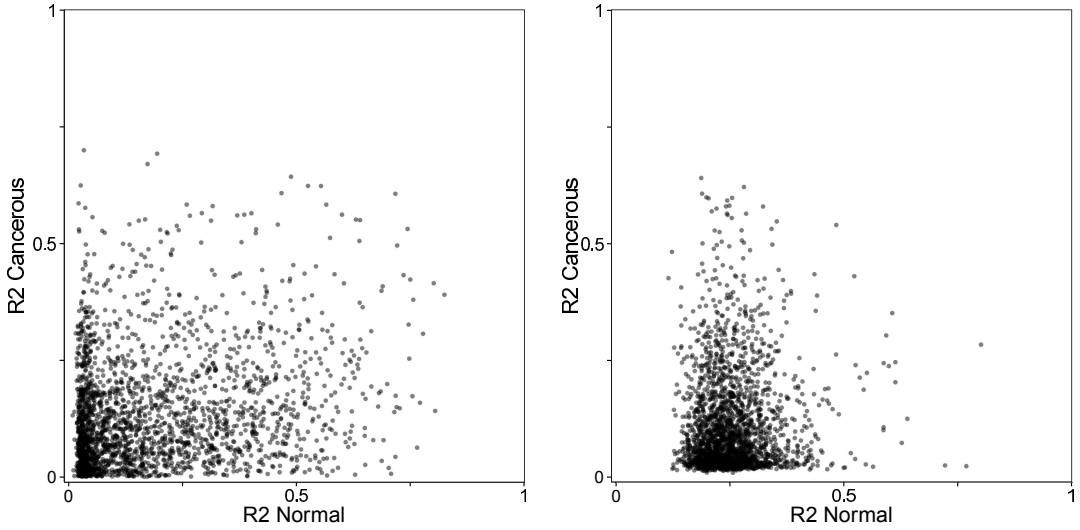


FIGURE 4.18: **Shift of epigenetic regulation in cancer.** Scatterplot between the predictive power of DNA methylation for gene expression in normal and cancerous for breast lung tissues (left: $R^2_{breast} = 0.04$, right: $R^2_{Lung} = 6 \times 10^{-7}$)

an integrated analysis combining methylation, DNA copy number and gene expression. We compute a predictive goodness of fit R^2 to represents the power of DNA copy number information alone to predict gene expression, on the one hand, and a multidimensional regression model combining both the full CGI+SS DNA methylation information and the DNA copy number information, on the other hand. We observe that combining methylation and copy number information leads to significantly better results in predicting gene expression than taking each information separately (4.19, $P_{Breast} < 10^{-16}$, $P_{Lung} < 10^{-9}$,$P_{Colon} < 10^{-8}$). Moreover, correlation analysis between predictive scores using DNA methylation only, on the one hand, and predictive scores using CNVs only, on the other hand, shows very little correlation (4.19, $R^2_{Breast} = 7 \times 10^{-4}$, $R^2_{Lung} = 1 \times 10^{-4}$, $R^2_{Colon} = 1 \times 10^{-3}$). This suggests that both methylation and DNA CNVs are important and non-redundant predictors of gene expression variations.

4.6 Discussion

DNA methylation is a well-described process in normal development and is critical in specific gene expression regulations such as X-chromosome inactivation, genomic imprinting and tissue developpment [Laurent et al., 2010, Smith and Meissner, 2013, Pollex and Heard, 2012, Li et al., 1993]. Since aberrant hyper- and hypo-methylation have also been frequently observed in cancer, it has been often argued that activation of oncogenes or repression of tumor suppressor genes could be caused by these epigenetic variations [Esteller, 2002].

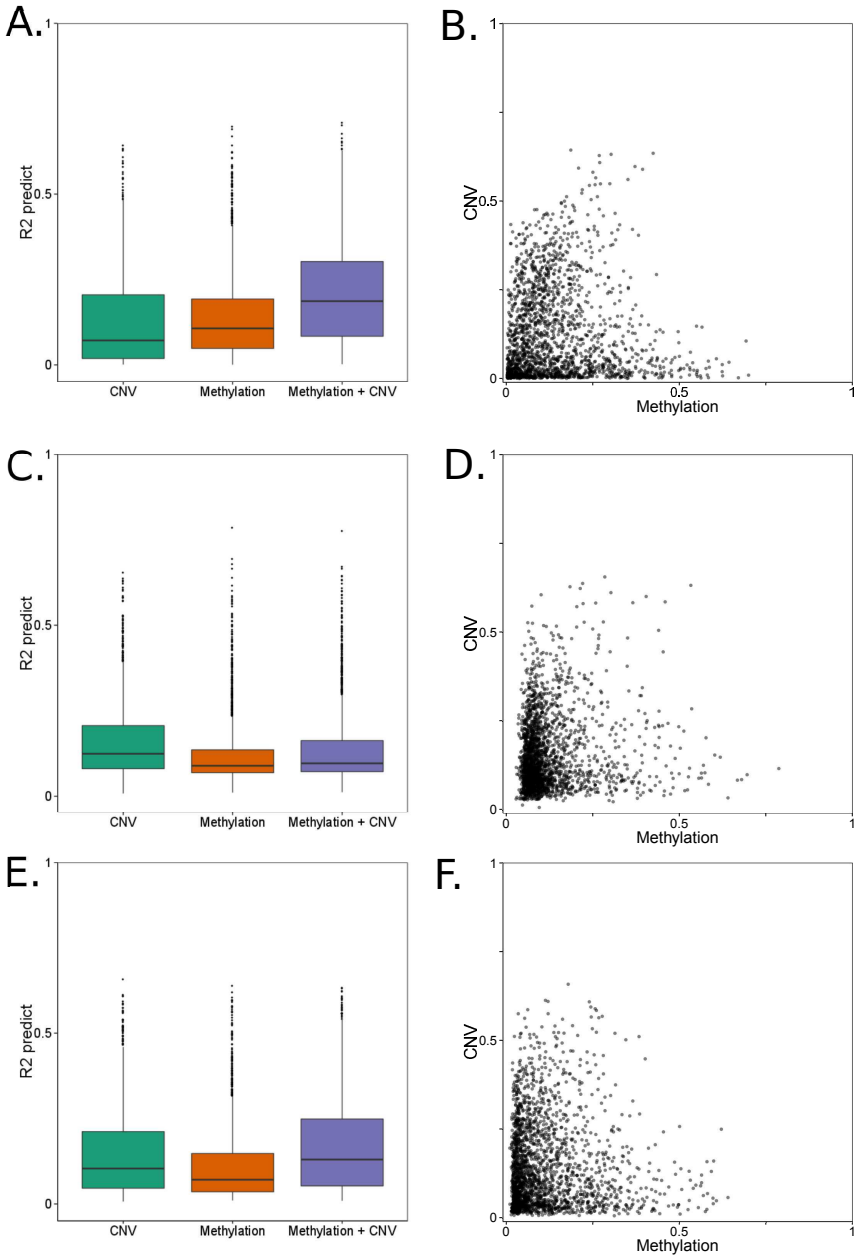


FIGURE 4.19: **Association between predictive power of methylation and copy-number variations. Panels A/C/E.** Predictive power distribution using either CNV data only with least squares, DNA methylation data only with lasso regression or both CNV and DNA methylation data with lasso regression. Combined methylation and CNV information yield significantly higher predictive power (panel A=breast cancerous tissues, panel C= colon cancerous tissues, panel E= lung cancerous tissues). **Panels B/D/F.** Scatterplot of predictive power using DNA methylation only and copy-number information only shows that both regulation mechanisms operate exclusively on genes (panel B= breast cancerous tissues, panel D=colon cancerous tissues, panel E= lung cancerous tissues).

In the present study, we assessed the existence of characteristic CGI+SS DNA methylation signatures in normal tissues and showed a weak association between the hypermethylated signature and gene expression repression. A similar study in cancerous tissues showed the existence of a cancer-specific signature highly associated with repressed genes. However, the corresponding genes are already highly repressed in normal tissues, questioning the causal impact of methylation in gene expression regulation, as already noticed in [Keshet et al., 2006, Sproul et al., 2011, Sproul and Meehan, 2013].

Using regression methods we analyzed whether differences between CGI+SS methylation across samples - independently of signatures - are predictive of gene expression variations. We showed that for certain genes, expression variations across samples can be well predicted from DNA methylation variations and that these genes are not associated with cancer-specific methylation patterns. We also showed that using the full CGI+SS methylation profiles in a multidimensional regression framework yields better predictive power than summarizing the methylation of a CpG island by one mean value, as done in previous studies [Vanderkraats et al., 2013]. Looking at probewise methylation correlation with gene expression for the top scoring genes, we observed that the impact of a CpG methylation on gene expression is largely dependent on its location in or near the island, and that CpGs located outside of CGIs have a bigger impact on gene expression variations than CpG located within the CGI, supporting results from [Irizarry et al., 2009, van Vlodrop et al., 2011]. The impact of CGIs located outside of promoter regions, such as intragenic CGIs is still unclear as it does not seem to contribute significantly to global gene expression regulation. Yet, a few studies point at their potential role in modulating alternative promoters [Maunakea et al., 2013] or in long-range regulation [Kulis et al., 2013].

Reproducing this methodology on different datasets allowed us to compare the variations of gene expression regulation by methylation in normal and cancerous tissues but also between different types of tissues. Our results suggest that genes targeted by methylation are not only very different between different normal tissues, but more importantly that they are very different between normal and cancerous samples of a given tissue suggesting a shift of epigenetic regulation between normal and cancerous tissues. Recently, hydroxymethylation of cytosines (hmC) has been shown to be significantly present in mammalians cells [Kriaucionis and Heintz, 2009] and methylation data generated with Illumina arrays, as done here, are not able to distinguish methylation (mC) from hydroxymethylation [Nestor et al., 2010]. However, hmC are significantly less present in cancer tissues [Haffner et al., 2011, Jin et al., 2011]. It is therefore likely that the epigenetic information measured here is indeed cytosine methylation.

In addition, the association between DNA methylation and other important regulation mechanisms widens our understanding of the role of methylation in the whole gene expression regulation process. While TFs are centric in controlling gene expression, we showed that their activation itself is significantly associated with DNA methylation markers, highlighting the critical role of methylation in the regulatory process. CNVs have been widely analyzed as a source of genetic variation that plays an important role in complex phenotypes such as cancer [Stranger et al., 2007, Henrichsen et al., 2009]. While CNV contribution has been characterized on a genome-wide scale, the link with other regulation mechanisms, particularly DNA methylation, is still unclear [Houseman et al., 2009, Lauss et al., 2012a]. We showed that the impact of both processes in gene expression regulation seems to be non-redundant. The relatively large dataset size gives us confidence in the statistical validity of the results, which are however limited to a fraction of the total genes because of uneven coverage. Methylome sequencing has already been performed and also supports the complexity of methylation patterns but is still limited to very small datasets [Vanderkraats et al., 2013]. Undoubtedly, larger methylome datasets available in the near future will further improve our understanding of the role of DNA methylation in gene expression regulation.

Chapter 5

Integrative DNA methylation and gene expression profiles to assess the universality of the CpG island methylator phenotype

Some content from this chapter has been submitted to Cancer Research.

Keywords: CpG island methylator phenotype, clustering, group-lasso logistic regression, methylation, clinical impact.

5.1 Résumé

Le CpG island methylator phenotype (CIMP) a été introduit par Toyota *et al.* dans le cancer du colon, pour caractériser une sous-population de cancers avec desprofils épigénétiques particuliers marqués par une hyperméthylation coordonnée d'un certains nombres d'ilôts CpG. Depuis, ce phénotype a été étendu à différents profils de tumeurs dont, entre autres, le sein, la vessie, le poumon ou encore l'estomac. Le CIMP a un intérêt clinique majeur car il est associé à un niveau de réponse au traitement différent mais également à un pronostic de survie particulier. Cependant, l'absence d'une base moléculaire au CIMP, commune à tous les cancers pose toujours des questions: est-ce que le CIMP est associé à un phénomène biologique réel ou est-ce qu'il s'agit simplement d'aberrations épigénétiques propres à chaque cancer?

Nous avons analysé de manière systématique les profils de méthylations pangénomique issus d'une technologie unique (Illumina HumanMethylation450K) sur plus de 2000

échantillons tumoraux dans 5 tissus différents et nous montrons l'existence d'une signature épigénétique commune à tous les cancers déterminant du phénotype CIMP. De plus, une analyse intégrative des profils d'expression révèle qu'une signature transcriptomique est également en mesure de prédire ce phénotype avec une très grande précision.

Nos résultats soutiennent l'existence d'un phénomène biologique commun associé au CIMP marqué par la présence d'une signature épigénétique et génétique commune à tous les cancers.

5.2 Abstract

The CpG island methylator phenotype (CIMP) was first characterized in colorectal cancer but since, has been extensively studied in several other tumor types such as breast, bladder, lung, gastric. CIMP is of clinical importance as it has been reported to be associated with prognosis or response to treatment. However, the identification of a universal molecular basis to define CIMP across tumors has remained elusive.

We perform a genome-wide methylation analysis of over 2,000 tumor samples from 5 cancer sites to assess the existence of a CIMP with common molecular basis across cancers. We then show that the CIMP phenotype is associated with specific gene expression variations. However, we do not find a common genetic signature in all tissues associated with CIMP.

Our results suggest the existence of a universal epigenetic and transcriptomic signature that defines the CIMP across several tumor types but does not indicate the existence of a common genetic signature of CIMP.

5.3 Introduction

Epigenetic modifications have been recognized as important players in cancer etiology and development, and constitute promising therapeutic targets for diagnosis or treatment due to their possible reversibility [Jones and Baylin, 2007,Esteller, 2008,Rodriguez–Paredes and Esteller, 2011]. In particular, aberrant methylation of CpG islands (CGIs) located in promoter regions of tumor suppressor and DNA repair genes, leading to their silencing, is now considered a hallmark of cancer playing an important role in neoplasia [Jones, 1986,Baylin and Herman, 2000,Esteller et al., 2001,Esteller, 2008,Jones and Baylin, 2007,Rodriguez-Paredes and Esteller, 2011].

The CpG Island Methylator Phenotype (CIMP) was first defined and observed by [Toyota et al., 1999a] in a subset of colorectal cancers as the joint methylation of several promoter regions, leading to the inactivation of the corresponding genes. The stratification of patients based on CIMP was shown to be clinically relevant, as CIMP positive patients had better prognosis than CIMP negative ones, and could lead to personalized treatments. Since the identification of CIMP in colorectal cancers, many studies have tried to replicate the analysis to find CIMP in different types of cancers including but not limited to colon [Issa et al., 2005, Weisenberger et al., 2006, Estécio et al., 2007, Curtin et al., 2011, Hinoue et al., 2012], breast [Auwera et al., 2010, Fang et al., 2011], lung [Suzuki et al., 2006], stomach [Chen et al., 2012] and glioblastoma [Noushmehr et al., 2010, Baysan et al., 2012, Yilmaz et al., 2012]. While most of these works concluded in the existence of a CIMP in different cancers, other studies did not yield the same conclusions [Bae et al., 2004, Anacleto et al., 2005], and the genes whose promoter CGI methylation are considered to define the CIMP differ between studies. This raises the question of whether the CIMP is tissue specific or is a universal phenomenon with common biological causes affecting common genes across cancers. A recent review of CIMP-related studies across different cancers pointed out the diversity of methods and measurement technologies used to define CIMP, which hinders the establishment of a molecular basis for CIMP in spite of growing evidence linking mutations in specific genes and CIMP in several cancers [Hughes et al., 2013].

In the present study, we investigate the existence and universality of CIMP by performing a systematic genome-wide methylation analyse on several large datasets of different cancer types simultaneously. We propose a simple methodology to assess the existence of a CIMP phenotype in each cancer, and to identify a set of genes whose promoter methylation is a marker for the CIMP. This allows us to compare the different cancer types in search for a cross-cancer CIMP signature, and to analyze the link between CIMP and gene expression in different cancers. Finally, we assess the clinical relevance of CIMP on the overall survival.

5.4 Material and Methods

5.4.1 Patients Selection

All data were retrieved from the TCGA data portal. We selected samples from bladder, breast, colon, lung and gastric adenocarcinomas because large matched datasets were available for methylation, gene expression and mutation profiles. Moreover, all these tissues were previously reported to exhibit a methylator phenotype. The datasets are

detailed in 5.1 and the different institutions that released the data are mentioned in the acknowledgement section.

TABLE 5.1: **Patients Dataset.** Original dataset sizes for methylation (Meth), gene expression (GE) and mutation profiles for cancerous tissues. The "Matched" column represents the number of available samples both methylation and gene expression profiles.

	Meth	GE	**Meth/GE**	**Meth/Mutations**
Bladder	373	56	**43**	**28**
Breast	626	778	**478**	**468**
Colon	291	193	**34**	**219**
Lung	452	125	**82**	**411**
Stomach	338	373	**309**	**199**
Overall	*2090*	*1525*	***941***	***1325***

5.4.2 Methylation profiling

Methylation profiles were retrieved from level 2 TCGA data. They were obtained with the Illumina HumanMethylation450K DNA Analysis BeadChip assay, which is based on genotyping of bisulfite-converted genomic DNA at individual CpG-sites to provide a quantitative measure of DNA methylation [Bibikova et al., 2011]. Following hybridization, the methylation value for a specific probe was calculated as the ratio $M/(M+U)$ where M is the methylated signal intensity and U is the unmethylated signal intensity. 485,577 CpG methylation levels, associated with 27,176 CGIs and 21,231 genes, were measured as such accross the genome.

Following [Irizarry et al., 2009], we considered not only the CGI methylation profile but also included in the analysis proximal regions in the near vicinity (up to 4kb), namely the CGI Shores and Shelves regions in a general CGI+SS methylation profile.

5.4.3 Gene expression profiling

Gene expression profiles were retrieved from level 3 TCGA data. They were obtained from the Illumina HiSeq RNASeq technology and processed following [Mortazavi et al., 2008].

5.4.4 Mutation profiling

Mutations profiles were retrieved from somatic mutations profiles from level 2 TCGA data obtained through whole exome sequencing.

5.4.5 CIMP analysis

To assess the existence of CIMP, we performed Ward hierarchical clustering using euclidean distance. Robustness of the clustering was obtained through consensus clustering [Monti et al., 2003].

TABLE 5.2: **CIMP Proportion.**

	Negative	Positive	Ratio
Bladder	262	111	30%
Breast	509	117	19%
Colon	232	59	20%
Lung	136	316	70%
Stomach	144	194	57%
Overall	*1283*	*797*	*38%*

5.4.6 Predicting CIMP status from gene expression profiles

To predict CIMP using gene expression profiles, we perform logistic regression using a lasso penalty [Tibshirani, 1996] with different settings.

5.4.7 Tissue-specific lasso

We first perform standard logistic regression using lasso to predict CIMP status using a small list of gene expression profiles for each tissue separately. Accuracy is calculated through 3-fold cross-validation averaged over 100 repeats.

5.4.8 Combined Lasso

For the "Combined Lasso", we pool all the samples into a single dataset independently of their tissue of origin. For cross-validation, we separate samples into training and testing by keeping a balanced proportion of samples from each tissues.

5.4.9 Group Lasso

For the "Group Lasso", we predict the CIMP as described below. For sample i belonging to tissue k, we assume that the conditional probability $p_{\beta_\mathbf{k}}(x_i) = \mathbb{P}_\beta(Y = 1|x_i)$ follows 5.1.

$$\eta_{\beta_\mathbf{k}}(x_i) = \beta_k^0 + (x_i^k)^\mathbf{T}\beta_\mathbf{k} \tag{5.1}$$

with

$$\eta_{\beta_k}(x_i) = \log\left(\frac{p_{\beta_\mathbf{k}}(x_i)}{1 - p_{\beta_\mathbf{k}}(x_i)}\right) \tag{5.2}$$

The logistic lasso estimator $\hat{\beta}^k_\lambda$ verifies 5.3:

$$\hat{\beta}^k_\lambda = \arg\min\{-l(\beta) + \lambda\|\beta\|_1\} \tag{5.3}$$

where l is the log-likelihood function:

$$l(\beta) = \sum y_i * \eta_{\beta_k}(x_i) - \log[1 + exp(\eta_{\beta_k}(x_i))] \tag{5.4}$$

To increase the statistical power and given the similarity of the different prediction tasks, we combine the different datasets into a single prediction task as follow:

The vector of output $y \in \mathbb{R}^p$ is given by:

$$y^\mathbf{T} = (y^1_1, \cdots, y^1_{n_1}, \cdots\cdots, y^k_1, \cdots, y^k_{n_k})^\mathbf{T} \tag{5.5}$$

where y^k_i is the CIMP status for patient i in tissue k.

And the combined design matrix X is given by:

$$X = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & & \cdots & \vdots \\ 0 & 0 & \cdots & X_n \end{bmatrix} \tag{5.6}$$

where $X_i \in \mathbb{R}^{n_i \times pK}$ is the gene expression profile matrix of the p genes for the i^{th} tissue dataset of size n_i and K is the number of tissues considered.

We performed group-lasso logistic regression following [Meier et al., 2008] with the groups defined as the set of features corresponding to a given gene for each tissue that is we optimize

$$\hat{\beta}^{group}_\lambda = \arg\min\left\{-l(\beta) + \lambda\sum_{g=1}^{G}\|\beta_g\|_2\right\} \tag{5.7}$$

where

$$G := \left\{ g_i = (e_i, e_{i+p}, \cdots, e_{i+(K-1)p}) \ \forall i \in [1; p] \right\} \tag{5.8}$$

and $(e_i) \in \mathbb{R}^{pK}$ is the vector of zeros except for feature i.

Given the imbalanced proportion of CIMP in each datasets, we defined the "random" predictor as a predictor that always predicts the majority class. The statistical significance of a gene expression based predictor over the "random" predictor was calculated using a Student t-test.

To determine the genetic predictive signature, genes were ranked in their frequency in appearing in the optimal lasso estimator signature averaged over the different folds and repeats [Meinshausen and Bühlmann, 2008]. Genes which frequency was superior to 50% were selected.

5.4.10 Survival analysis

Overall survival was estimated using the Kaplan-Meier method [Kaplan and Meier, 1958] to compare the survival between CIMP positive and CIMP negative tumors. A multivariate Cox proportional hazards regression model [Cox and Oakes, 1984] was also fitted.

5.5 Results

A cross-cancer CIMP signature

We first assess with a common methodology whether a CIMP can be detected on different cancers, and whether CIMP in different cancers share a common signature in terms of which gene promoters are hypermethylated in CIMP positive patients. For that purpose, we collected high-density methylation datasets from the cancer genome atlas (TCGA) data portal providing more than 485,000 CpG methylation levels for more than 2,000 samples from five tissues of origin: bladder, breast, colon, lung and stomach (Table 5.1). For each sample, we aggregate the methylation levels of CpG probes by CGI, including the CGI itself and its shores and shelves, resulting in a single methylation level for each of 21,176 CGIs in each sample.

A CIMP corresponds to the joint hypermethylation of a subset of CGIs in a subset of samples [Toyota et al., 1999a]. To characterize from whole-genome methylation data whether a CIMP exists for a cancer, and which CGIs characterize it, we follow a standard methodology: (i) select the 5% most variant CGIs in the set of samples, which we call the *CIMP signature*, and (ii) check by unsupervised classification whether the samples cluster into two main clusters (CIMP positive and negative clusters) when we restrict them to the methylation values they take on the CGIs in the CIMP signature.

We apply this methodology to each of the five families of tumors, cutting the tree obtained by hierarchical clustering to two clusters in order to enforce a classification of all samples into two subgroups based on the methylation of CGIs in the CIMP signature. Interestingly, in all five cases, one of the two clusters is clearly characterized by an overall hypermethylation of most CGIs in the signature compared to the second cluster, allowing us to characterize it as the CIMP positive cluster, the second one being the CIMP negative cluster (5.1). The proportion of CIMP positive samples according to this definition varies from about 20% for breast and colon cancers to 30% for bladder and about 60% and 70% for stomach and lung cancers respectively (Table 5.2). Proportion of the CIMP-positive group in each tissue is similar to previously reported studies [Hughes et al., 2013]. Varying the size of the CIMP signature from 1% to 10% of all CGIs had a small impact on the clustering stability (5.2).

Comparing the epigenetic signatures that defines CIMP for each tissue, we find a common set of 89 CGIs associated with 51 genes (Figure 5.3, panel B). If the signatures were random subsets of 5% of all CGIs independent from each other, the overlap would contain on average $(5\%)^5 \simeq 3.10^{-5}\%$ of all CGIs, namely 0.006 CGI. This provides a strong evidence that a common set of genes is involved in CIMP in different cancers. We call these 89 CGIs the *cross-cancer CIMP signature* (Table 5.3). A hierarchical clustering on all samples restricted to this cross-cancer CIMP signature is able to cluster CIMP-positive and CIMP-negative patients independently of the tissue of origin (Figure 5.3, panel A), suggesting that CIMP observed in each individual cancer share in common a significant proportion of genes whose promoter CGIs are hypermethylated in all CIMP positive cancers. A functional enrichment analysis of the cross-cancer CIMP signature reveals that it is significantly enriched in genes involved in cell differentiation, neuronal developmental and immune response processes (Figure 5.3, panel C).

5.5.1 Are there 2 or 3 CIMP classes?

Several studies suggest the existence of a third class in CIMP phenotype that corresponds to an intermediate level of methylation [Ogino et al., 2006, Shen et al., 2007, Hinoue

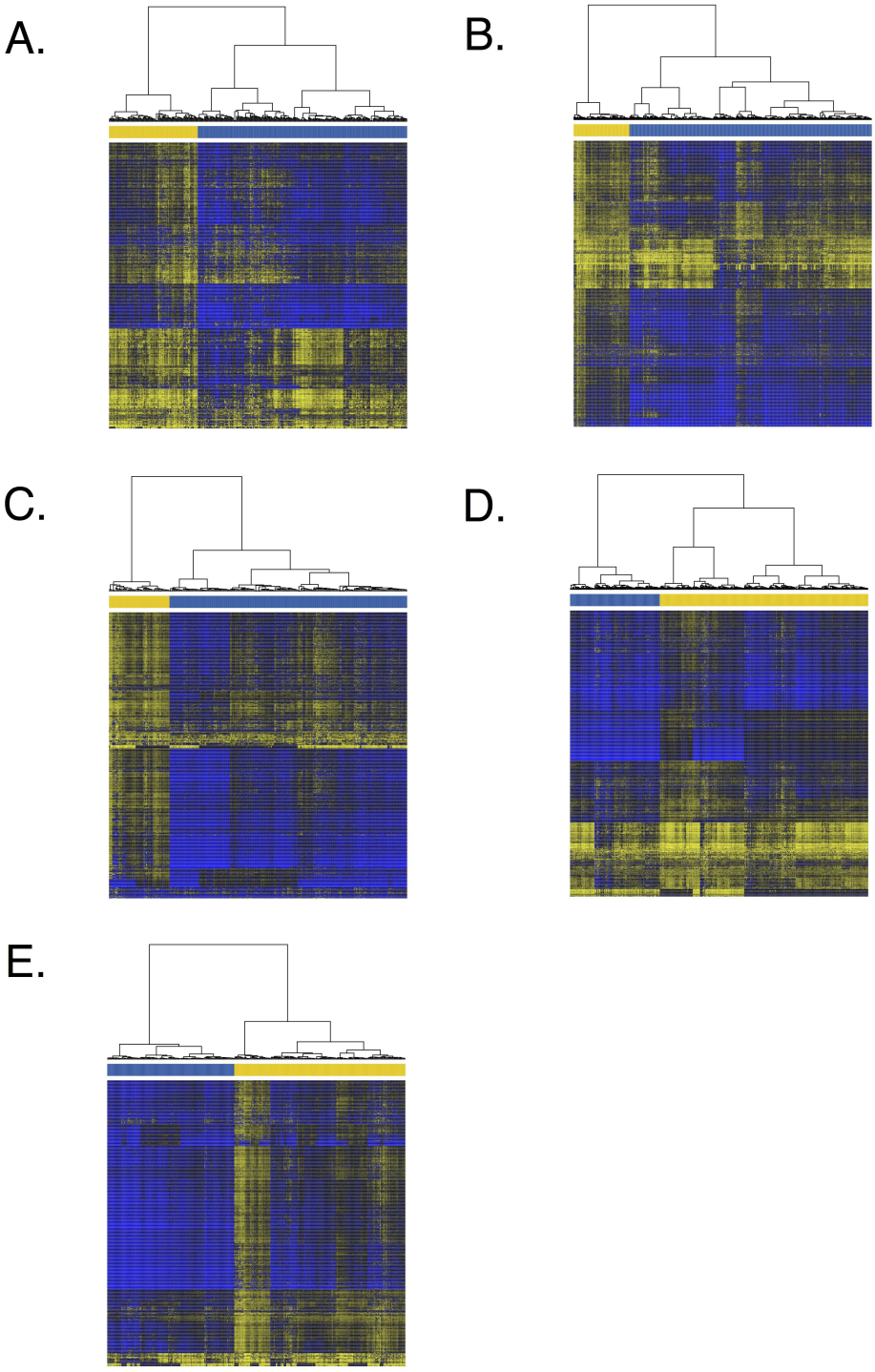
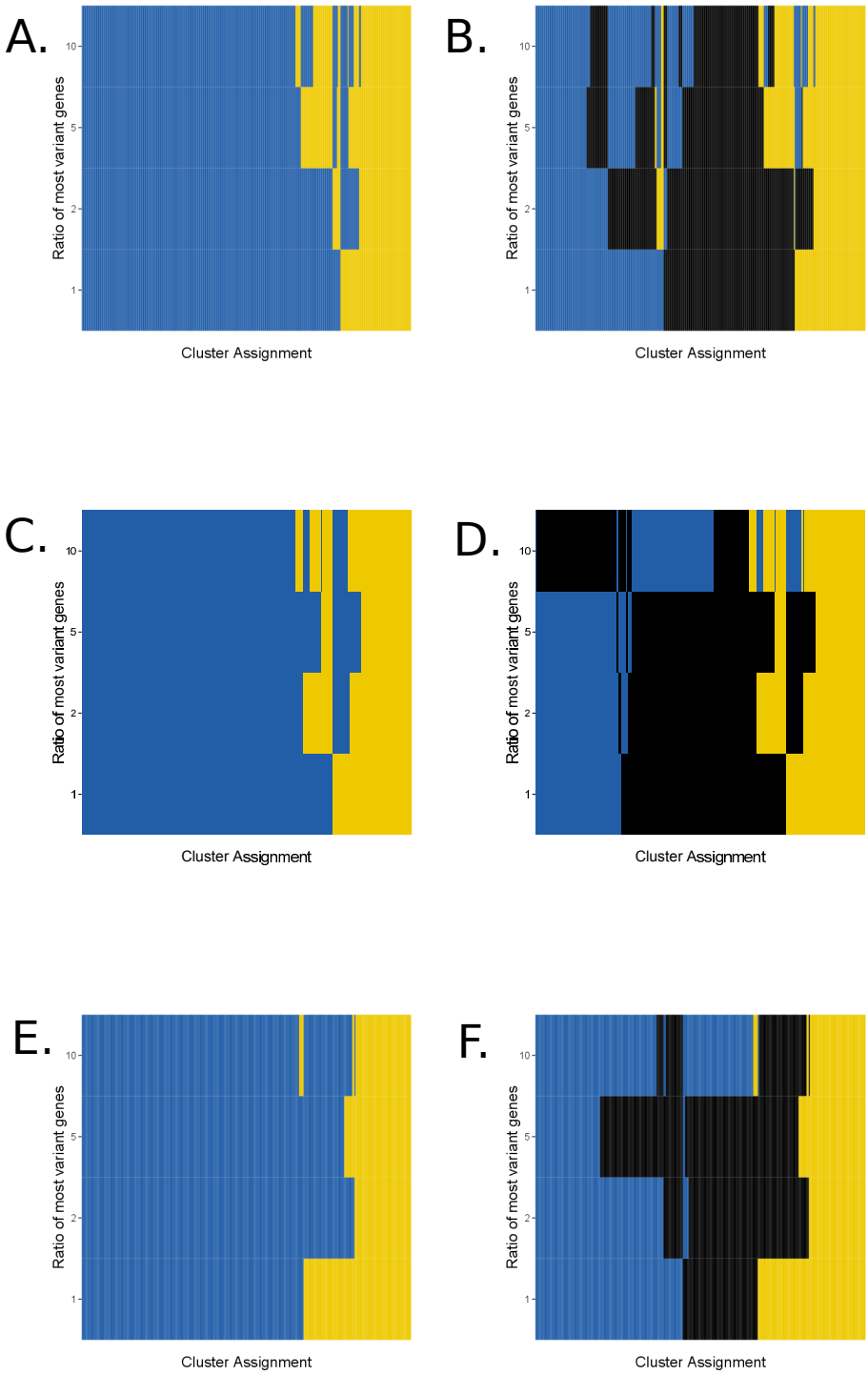
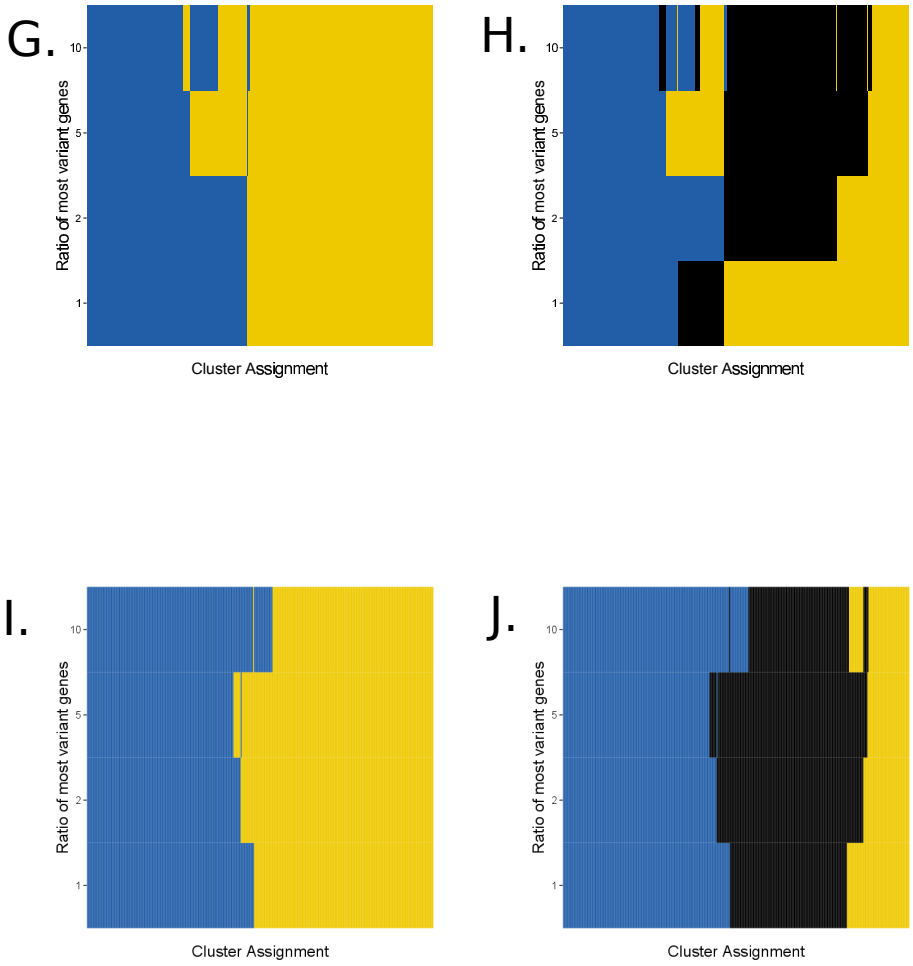


FIGURE 5.1: **Methylation profiles hierarchical clustering for each tissue based on the most variant probes.** Heatmaps range from hypomethylated (blue) to hypermethylated (yellow). The column colorbar represents the CIMP assignment (yellow= CIMP-positive, blue= CIMP-negative). **Panel A.** Bladder **Panel B.** Breast **Panel C.** Colon **Panel D.** Lung **Panel E.** Stomach.


A.
B.
C.
D.
E.
F.
Ratio of most variant genes
10
5
2
1
Cluster Assignment



K.

Tissue	K=2	K=3
Bladder	80%	50%
Breast	80%	26%
Colon	82%	36%
Lung	81%	38%
Stomach	89%	82%

FIGURE 5.2: **Stability of CIMP clusters given the proportion of variant CGIs.** Robustness of cluster assignment for each sample (columns) as a function of the proportion of variant CGIs considered from 1 to 10 % (rows) and given the number of CIMP clusters considered (left panels: K=2, right panels: K=3, yellow=CIMP-positive, blue=CIMP-negative, black=CIMP-low) for bladder (panel A/B), breast (panel C/D), colon (panel E/F), lung (panel G/H), stomach (panel I/J). **Panel K.** Table summarizing the stability of the cluster assignments for each tissue and different number of CIMP clusters considered.

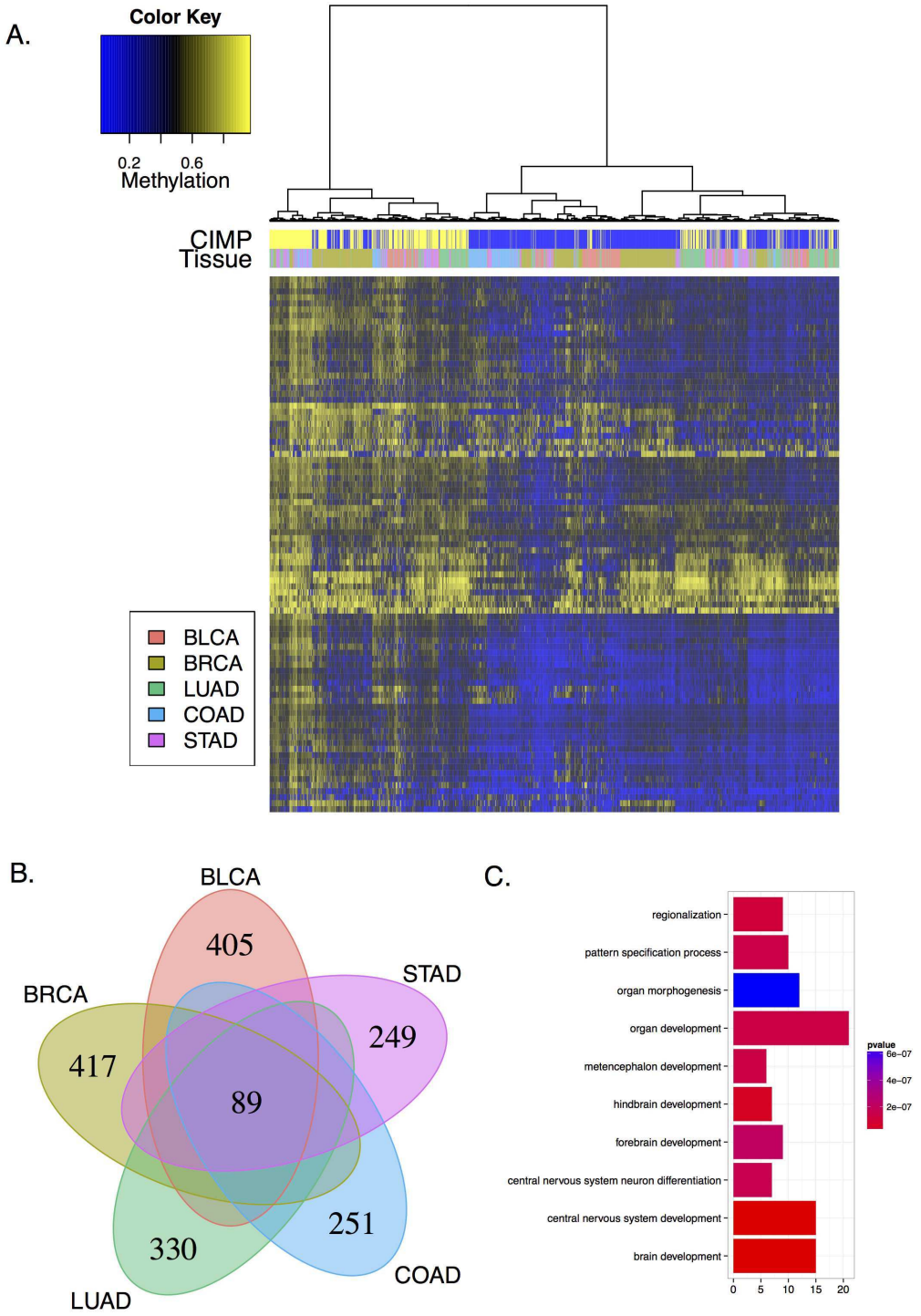


FIGURE 5.3: Universal epigenetic signature for CIMP.

TABLE 5.3: **List of genes associated with the common set of CGIs that define CIMP in each tissue.**

Epigenetic Signature	*LOC339524*, *GSTM1*, *CD1D*, *LMX1A* *CACNA1E* ,*NR5A2*, *WNT3A*, *GNG4* *EMX1*, *CTNNA2* ,*LRRTM1*, *DLX1* *EVX2*, *HOXD13*, *GBX2*, *SYN2* *HAND2*, *NBLA00301*, *EBF1*, *HIST1H2BB* *HIST1H3C*, *HLA-DRB1*, *C6orf186*, *IKZF1* *p16*, *HMX3*, *KNDC1*, *KLHL35* *HOTAIR*, *SLC6A15*, *ALX1*, *RFX4* *CLDN10*, *ADCY4*, *RIPK3*, *NID2* *OTX2*, *OTX2OS1*, *GSC*, *KIF26A* *GREM1*, *SEC14L5*, *HS3ST3B1*, *IGF2BP1* *HOOK2*, *NFIX*, *ZNF577*, *ZNF649* *CPXM1*, *CDH22*, *CHRNA4*

et al., 2012]. While we enforced an analysis with 2 classes to define the CIMP of each sample as positive or negative in the previous section, we now examine whether the data call for a third class. Following [Monti et al., 2003], we assess the existence of an intermediate CIMP phenotype for each tissue by comparing the increase in empirical cumulative distributive distribution $\Delta(K)$ for different values of $K = 2, \ldots, 5$ where K is the number of clusters considered for CIMP.

Figure 5.4 shows how $\Delta(K)$ varies as a function of K for each cancer, suggesting how many clusters exist in each case. We observe that the existence of a third class is not clear-cut. While colon and breast tissues show a significant increase in $\Delta(K)$ for $K = 3$ suggesting a possible third cluster in CIMP, bladder is flat between 2 and 3 clusters, while lung and gastric cancers do not support the presence of 3 classes. In addition, we assess the stability of 3 clusters by varying the number of CGIs that define CIMP and observed that while CIMP clusters are highly robust for $K = 2$, there is some high variability in the cluster definitions for $K = 3$ (5.2). In summary, the presence of 2 clusters is well supported by the data in all cancers, while the third cluster is much more debatable.

5.5.2 Similar gene expression variations are predictive of CIMP.

To shed light on the relationship between methylation and transcription, we now assess to what extent a transcriptomic signature can classify the samples as CIMP positive or negative. For that purpose, we collected for each family of cancer samples with both methylation and gene expression data available, leading to a subset of samples with an overall proportion of CIMP positive samples comparable to that of the original dataset

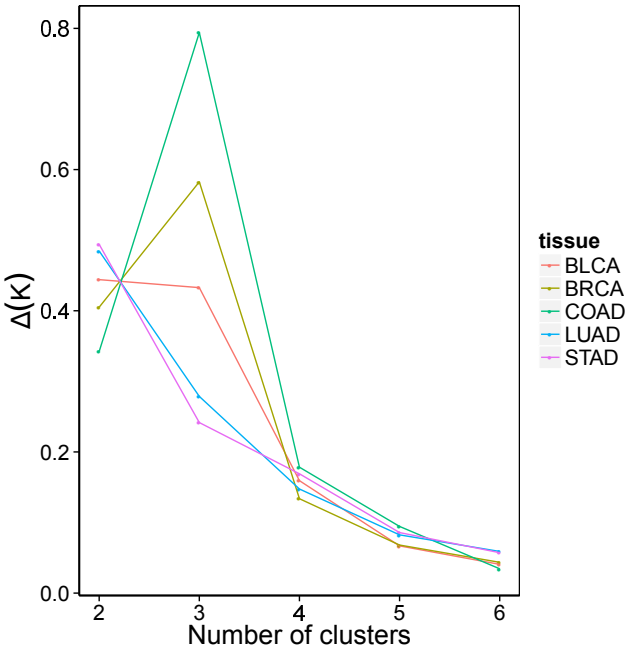


FIGURE 5.4: Stability of CIMP given the number of clusters.

(Table 5.4). We measure by cross-validation how well expression data alone can recover the two CIMP classes.

TABLE 5.4: **Matched Meth/GE samples CIMP Proportion.**

	Negative	Positive	Ratio
Bladder	27	16	37%
Breast	385	93	20%
Colon	27	7	20%
Lung	22	60	75%
Stomach	131	178	58%
Overall	*592*	*354*	*37%*

We first perform a multivariate regression analysis using the lasso technique to assess whether gene expression of a few genes can be predictive of the CIMP status for each tissue separately. The cross-validation accuracies for each family of cancer are shown in Table 5.5. We observe that while a classifier based on gene expression performs significantly better than random to recover CIMP positive samples in breast, lung and stomach cancers, the performance on bladder and colon is not different from a random classifier. Moreover, we compare the lists of genes selected in the transcriptomic signature after bootstrap resampling of the samples in order to assess their robustness and potential biological significance (Figure 5.5, panel C). We observe that very few genes are robustly selected in the signatures, and in particular that no gene is associated with

BLCA-CIMP and COAD-CIMP prediction in more than 15% of the bootstrap resampling. In addition, the transcriptomic signatures of different cancers are very diverse, and no gene is present in all of them (Figure 5.5, panel B). Overall, these results suggest that there is information in the transcriptome related to the CIMP status, but that a robust signature across cancers is difficult to obtain.

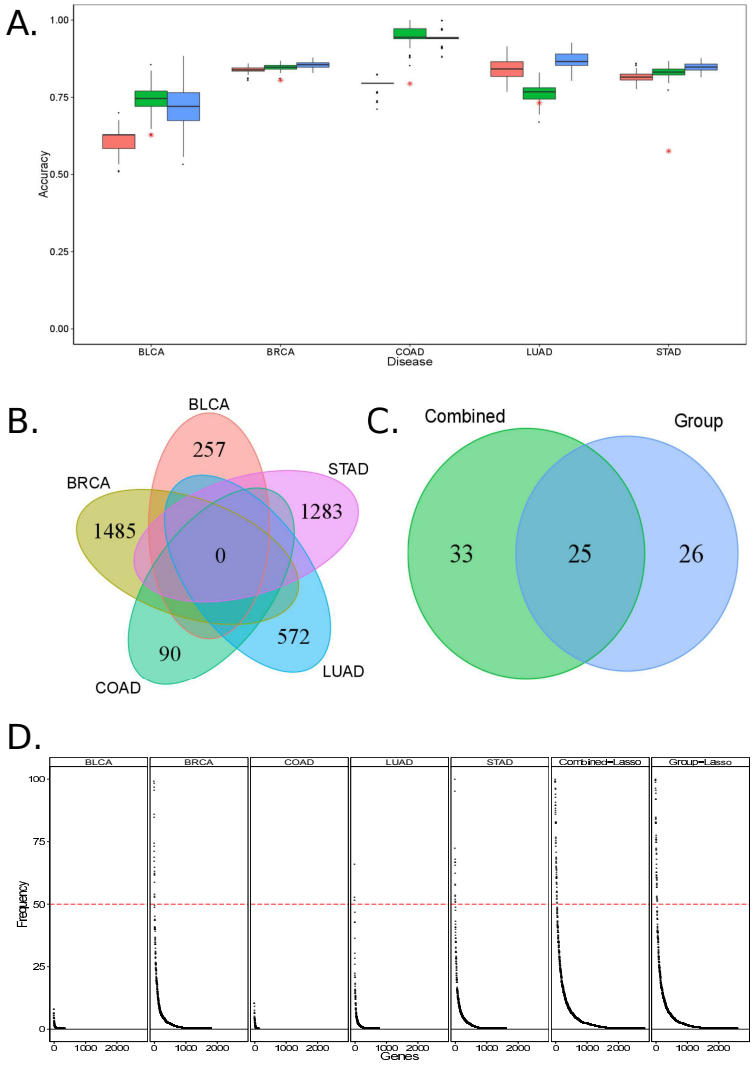


FIGURE 5.5: **Gene expression variations predictive of CIMP. Panel A.** Distribution of the accuracy of the CIMP-phenotype prediction task given the patient gene expression profile using $n = 100$ bootstrap and 3-fold cross-validation for several methods (pink= "tissue-specific" lasso, green= "Combined-Lasso", blue= "Group-Lasso", red star= random prediction). **Panel B.** Venn diagram of the tissue-specific gene signatures using lasso for each tissue separately. **Panel C.** Venn diagram representing the intersection between the "Combined" and "Group" lasso gene signatures. **Panel D.** Stability of each gene signature for each tissue-specific CIMP prediction as well as the "Combined-Lasso" and the "Group-Lasso" CIMP prediction task obtained and ranked by frequency of appearance using bootstrap ($n = 100$ repeats). For bladder and colon CIMP prediction task, the signature was non robust (frequency of the most redundant gene inferior to 10%). The combined prediction task signature outperforms the tissue-specific signatures in robustness.

TABLE 5.5: **Accuracy of CIMP prediction using gene expression profiles.**

	Accuracy		
	Random	Lasso	Group Lasso
Bladder	**62.8**	**62.9** (p=1)	**72.1** ($p \leq 2.10^{-16}$)
Breast	**80.5**	**83.9** ($p \leq 2.10^{-16}$)	**85.5** ($p \leq 2.10^{-16}$)
Colon	**79.4**	**79.5** (p=1)	**94.2** ($p \leq 2.10^{-16}$)
Lung	**73.2**	**84.2** ($p \leq 2.10^{-16}$)	**86.6** ($p \leq 2.10^{-16}$)
Stomach	**57.6**	**81.2** ($p \leq 2.10^{-16}$)	**84.8** ($p \leq 2.10^{-16}$)
Overall	***71.9***	***82.4***	***85***

However, the poor accuracy as well as the non-robustness of genetic signatures to predict CIMP may be due to the small size of some datasets ($n_{BLCA} = 43$, $n_{COAD} = 34$). To overcome the lack of statistical power due to small sample size, we combine in a second analysis the different datasets into a single multivariate regression analysis, based on the assumption that the CIMP signatures of different cancers may share the same genes. We train classifiers to predict CIMP status from gene expression data jointly across cancers using two methods, based on two different assumptions: (i) assuming that all tissues share the same gene signature and coefficients for the prediction task, we run a single Lasso classification on the combined datasets ("Combined-Lasso" prediction) or (ii) assuming that all tissues share the same gene signature but with different coefficients, we jointly train several models with a group Lasso approach to constrain the selected genes to be the same across cancers without imposing their coefficients to coincide ("Group-Lasso" prediction). The rationale for the group lasso approach is that while CIMP may be caused by a common subset of genes, but their impact may vary between tissues. Our results show that both methods significantly outperforms the tissue-specific predictions ($P \leq 2.10^{-16}$, Figure 5.5 panel A, 5.6) in particular for bladder and colon where the size of the initial datasets could not give sufficient statistical power to predict CIMP accurately. There is overall little difference between both methods, with the notable exception of lung cancer where the combined lasso approach is significantly worse than the group lasso (and even the single lasso) model, suggesting that in that case the weights of the genes in the CIMP signature may differ from other cancers. More importantly, each method allows to identify a common genetic signature (51 genes for the "Combined" prediction and 58 genes for the "Group-Lasso" prediction) that distinguishes CIMP-positive and CIMP-negative class for each tumors which is more robust than all the tissue-specific signatures (Figure 5.5 panel C). In addition, these signatures share a large common set of genes (25 common genes). We perform gene ontology analysis on the intersection of the two predictive gene signatures and find specific enrichment only for genetic regulatory processes.

TABLE 5.6: **Intersection of the genetic signatures for "Combined-Lasso" and "Group-Lasso" predictive of CIMP ranked by decreasing level of robustness.**

Over expressed	*ZIC2*, *AMH*, *ZNF300*,*LHX1*, *MLF1*, *ZIC3*, *XKR9*,*TNNT1*, *TNNT1*, *CAMK2N2*,*PCDHB9*, *RAET1K*, *HIST1H2AB*, *C2CD4C*, *FBXL20*, *FBXL20*, *TFCP2L1*, *LDHC*
Under expressed	*MAGEC2*, *ZNF300*,*SLC15A1*,*TSPYL5*, *MLF1*, *ZIC3*, *GATA2*, *MAGEA12*, *LOC441666*, *MAGEA2*, *LOC389493*, *H2AFY2*, *FBXL20*, *TFCP2L1*, *LDHC*, *TFCP2L1*, *LDHC*

5.5.3 A genetic signature is associated to CIMP only for colon and gastric cancers

Several somatic mutations have been found to be tightly associated with epigenetic aberrations in CIMP. Recent studies have pointed out the causal role of IDH1 mutations in Glioblastoma-CIMP [Noushmehr et al., 2010, Yilmaz et al., 2012] and tight associations between IDH2 and TET2 mutations with other CIMPs (leukemia [Figueroa et al., 2010], enchondroma and spindle cell hemangioma [Amary et al., 2011, Pansuriya et al., 2011]). In colon, BRAF and KRAS mutations are associated with microsatellite instability and COAD-CIMP [Weisenberger et al., 2006].

We re-assess the association between mutations in these genes and CIMP in the different types of cancers (Figure 5.6, panel A). We recover a strong association between BRAF mutation and CIMP-positive colon tumors but no specific association with other tumor types. We also find no coordinated association between *IDH1*, *IDH2*, *KRAS*, *BRAF* or *TET2* mutations and CIMP phenotypes for all tissues. In addition we perform genome-wide mutation analysis to assess whether specific gene mutations are associated with CIMP. We find no significant gene mutation association for bladder, breast nor lung CIMPs. For colon and gastric cancer, we find respectively 459 and 1070 gene mutations associated with CIMP with a common intersection of 195 genes (5.7). Gene ontology analysis of this set of genes shows significant enrichment for extracellular matrix organization and cell adhesion but also neuronal developmental processes (5.7).

Finally, we also look at the rate of mutations in each tissue given the CIMP phenotype. We observe a significant association between the number of mutations and the CIMP status for colon and gastric cancer (Figure 5.6 panel B), in accordance with the tight association between CIMP and microsatellite instability for these two tissues [Herman et al., 1998, Weisenberger et al., 2006, Jones et al., 2012, Zang et al., 2012]. However, the same observation could not be made for bladder, breast and lung.

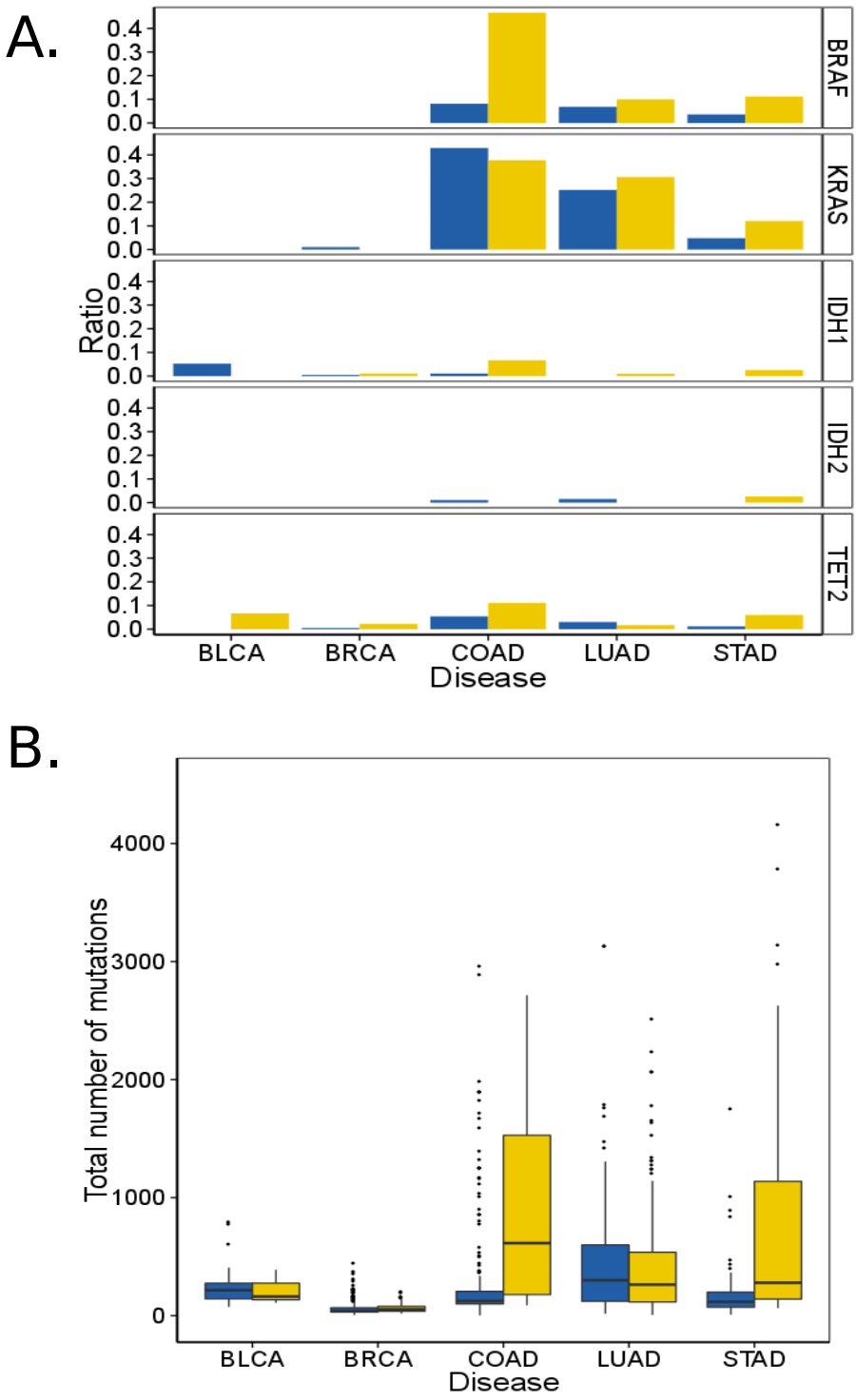


FIGURE 5.6: **Analysis of a genetic signature associated with CIMP. Panel A.** Association between specific mutations (*IDH1*, *IDH2*, *BRAF* and *KRAS*) with the CIMP phenotype for all tissues **Panel B.** Significantly higher mutation rate for CIMP positive tumors is observed for colon and gastric cancers only and is concordant with CIMP association with microsatellite instability for these tissues.

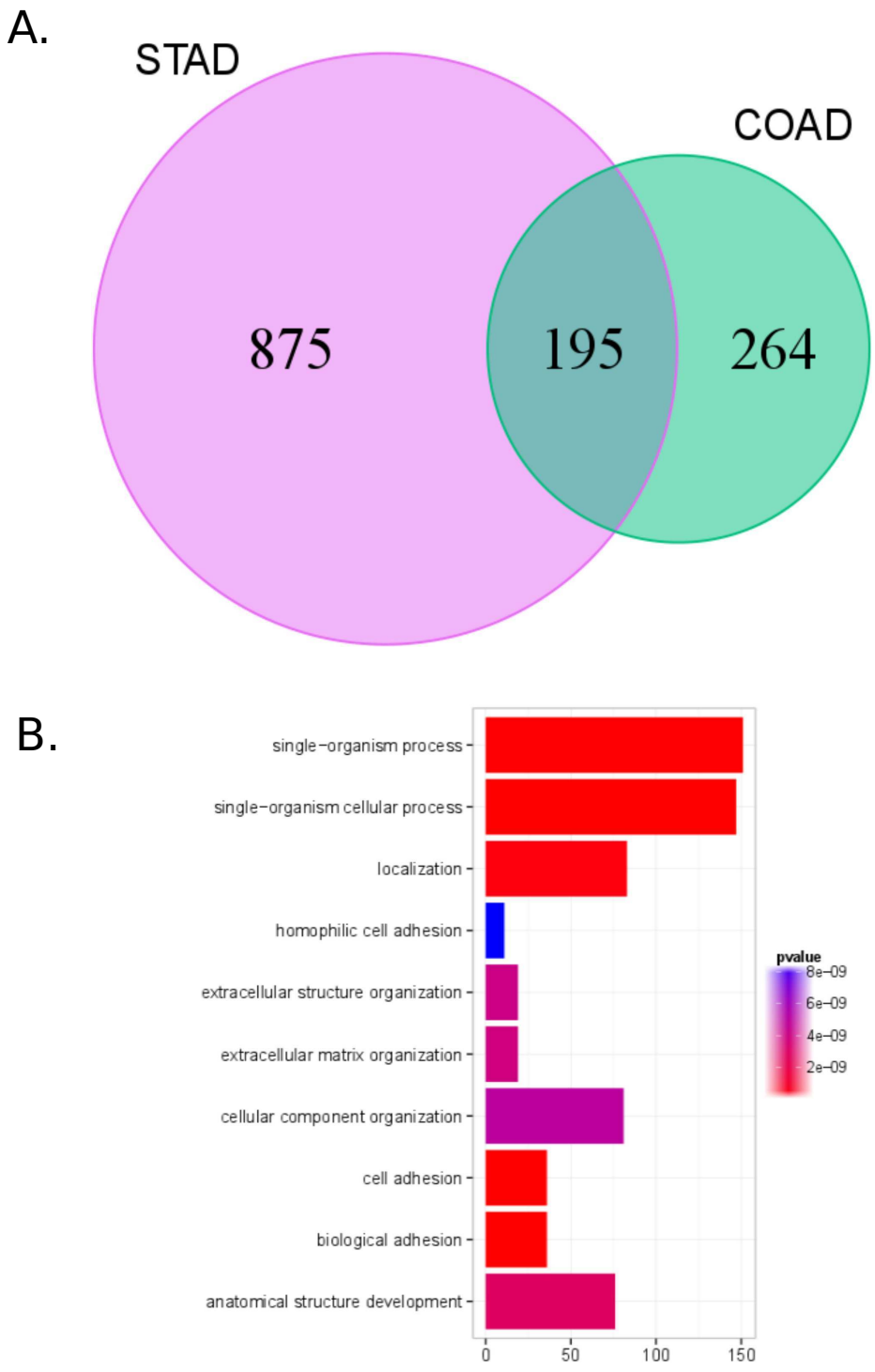


FIGURE 5.7: **Comparison of genetic signatures associated with CIMP for colon and gastric cancers. Panel A.** Venn diagram showing the intersection between the list of genetic mutations significantly associated with CIMP for colon and gastric cancers. **Panel B.** Gene ontology analysis of the common genetic signature associated with CIMP.

5.5.4 Clinical impact of CIMP.

Survival analysis in several CIMP studies has often shown distinct outcome between CIMP positive and negative tumors. However, there is no consensus in the general survival associated with CIMP: while CIMP has been associated with improved survival and lower risk of metastasis in breast [Fang et al., 2011], colorectal [Weisenberger et al., 2006], leukemia [Toyota et al., 2001, Garcia-Manero et al., 2002, Roman-Gomez et al., 2005, Roman-Gomez et al., 2006] or gliomas [Noushmehr et al., 2010], it has also been reportedly associated with poor survival for bladder [Maruyama et al., 2001], lung [Suzuki et al., 2006, Liu et al., 2008] or prostate cancers [Maruyama et al., 2002], and prognosis even remains unclear for gastric cancers [Toyota et al., 1999b, Oue et al., 2003, Kim et al., 2003, Etoh et al., 2004, Kusano et al., 2006].

We perform a systematic survival analysis on the different tissues to assess the clinical impact of CIMP. However, we observe no significant association between CIMP and survival, in any of the tissues (Table 5.7 and 5.8).

TABLE 5.7: **Clinical impact of CIMP.** Overall survival proportion given the CIMP phenotype and the p-value associated with the survival analysis (logrank test).

Tissue	Event		P-value
	CIMP-	CIMP+	
BLCA	47/214	21/96	0.74
BRCA	29/495	9/114	0.20
COAD	28/218	6/54	0.57
LUAD	24/127	67/295	0.49
STAD	26/141	20/193	0.29

5.6 Discussion

CIMP has been thoroughly studied over the past few years in several tissue types but the heterogeneity of the methods and measurement technologies has hindered the assessment of a common epigenetic and genetic signature predictive of CIMP across all cancer sites [Hughes et al., 2013]. In the present study, we analyze a large dataset of over 2,000 tumor methylation profiles measured with a single technology from 5 different tissues types. We observe a universal epigenetic signature that defines CIMP independently from the tissue of origin, which might suggest a common molecular basis to CIMP across tissues. Genes associated with these CGIs are enriched in several biological pathways linked to organ development, and include several interesting genes such as *CDKN2A* coding for p16, a well-characterized tumor suppressor protein [Nobori et al., 1994], which is aberrantly hypermethylated in CIMP positive tumors and might contribute to tumor

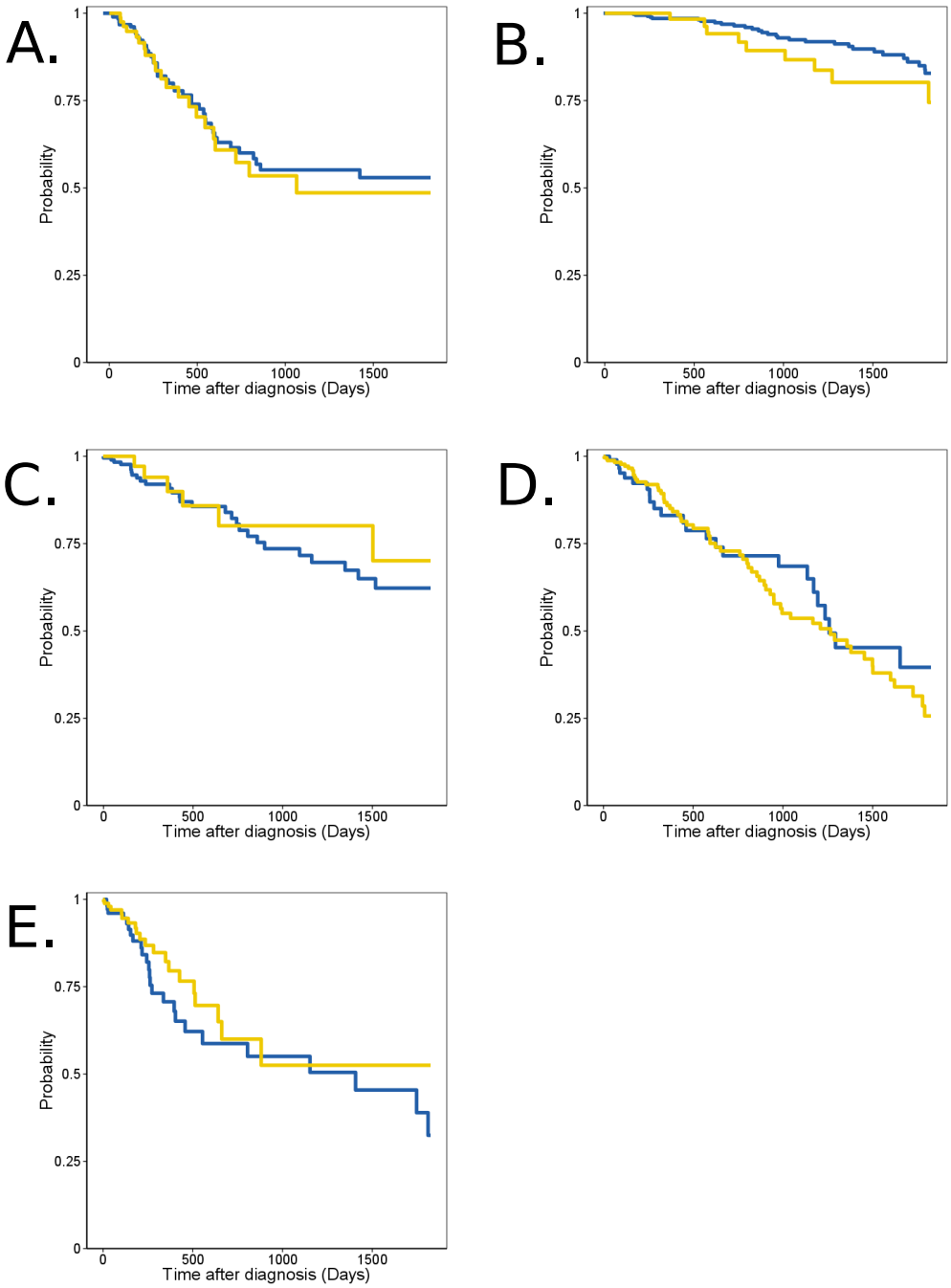


FIGURE 5.8: **Clinical impact of CIMP on the patient survival**. **Panel A.** Bladder **Panel B.** Breast **Panel C.** Colon **Panel D.** Lung **Panel E.** Stomach.

development. Other genes present in the cross-cancer CIMP signature such as *HOTAIR*, which is known to reprogram the chromatin state and is associated with breast cancer metastasis [Gupta et al., 2010], might on the contrary be repressed in CIMP tumors and be linked with a better prognosis for breast cancerous patients. *GREM1* is another gene present in the CIMP signature and is associated with tumor cell proliferation [Sneddon et al., 2006]. Less documented genes present in the CIMP signature could potentially be investigated for a biological validation of their role in tumor development.

Recent studies have pointed out that epigenetic aberrations could be derived from genetic aberrations [Reddington et al., 2014]. By combining the different datasets into a single prediction task, we are able to identify a common set of genes whose expression levels can predict the CIMP status for each tissue. This gene list is enriched mostly in genetic regulatory pathways, suggesting that the epigenetic reprogramming and thus CIMP might be an intermediate step in the regulatory mechanism. Among the genes contained in the signature, *ZIC2*, which is robustly selected in each bootstrap of the CIMP prediction task and is significantly more expressed in CIMP positive tumors for each tissue, has been known to act as a Wnt/β-catenin signalling inhibitor [Pourebrahim et al., 2011] which is usually upregulated in several cancers. Another interesting characteristic of this genetic predictive signature from a clinical point of view is the recurrence of cancer/testis antigens (CTAs) such as *MAGEC2* [von Boehmer et al., 2011, Yang et al., 2014, Reinhard et al., 2014], *MAGEA12* [Heidecker et al., 2000, Mollaoglu et al., 2008], *MAGEA2* [Peche et al., 2012], *LDHC* [Tang and Goldberg, 2009], which are interesting targets for cancer immunotherapy [Scanlan et al., 2002] and are consistently under-expressed in CIMP positive tumors. Recently Gevaert *et al.* [Gevaert, 2015] also showed a strong association between *MAGEA4* hypomethylation and CIMP positive tumors which further supports the link between CTAs and the absence of a methylator phenotype.

Mutation analyses are not very conclusive in defining a set of specific somatic mutations significantly associated with CIMP. In particular, lowly mutated cancer sites such as bladder, breast or even lung do not show any mutations significantly associated with CIMP. For highly mutated cancer sites such as colon or stomach, our results confirm a strong association between *BRAF* mutation and COAD-CIMP [Weisenberger et al., 2006] but do not show any particular associations with *IDH1/2*, which have been reported to be causal in gliomas and leukemia [Yilmaz et al., 2012, Figueroa et al., 2010]. There is a strong association between COAD and STAD-CIMP, and the specific mutations of genes related with extracellular matrix and cell adhesion, both reported to be strongly associated with metastasis [Gilkes et al., 2014, Lu et al., 2012, Bendas and Borsig, 2012, Okegawa et al., 2004]. Interestingly, neuronal developmental processes are highly enriched but affecting different genes from the universal epigenetic signature.

Associations with neuronal development were already mentioned in [Noushmehr et al., 2010].

Studies have often reported a clear distinct clinical prognosis associated with CIMP [Fang et al., 2011, Weisenberger et al., 2006, Toyota et al., 2001, Noushmehr et al., 2010]. This reiterates that a main reason for defining CIMP in each tissue site is its potential use as a prognosis marker. However, CIMP could be associated with a good or bad prognosis depending on the type of tumors. In the current study, we do not observe a significant association with any good nor bad prognosis linked with CIMP.

5.7 Conclusion

This meta-analysis of more than 2,000 samples sheds new light on CIMP across cancers, its link with gene expression, and its clinical relevance. We found strong evidence that a panel of genes, which we call the pan-cancer CIMP signature, is involved simultaneously in the establishment of the CIMP in various cancer sites, which might be an indicator of a universal biological process behind CIMP. We found that differences in the CIMP status of a sample is associated to differences in the transcriptome, and also found a core set of genes whose expression levels differentiates CIMP positive and negative samples, in all cancers studied. Finally, we found little evidence of association between CIMP and mutations, except for the well-known BRAF mutation in colon cancer, and also little association with patient survival.

Chapter 6

Discussion

The main objective of the projects developed in this thesis is the use of computational tools to describe a biological phenomenon. We focused on the role of epigenetics, more precisely *DNA methylation*, and its entanglement with other biological sources such as *gene expression*, *copy-number* or *mutations*. For that, the use of prior information is critical to conform to the current biological knowledge but also to reduce the complexity of the problem. Rigourous approaches thus allow to formalize the extent of the validity of a biological hypothesis but also to generalize to the whole genome a gene-specific observation.

6.1 "DNA methylation in cancer: too much, but also too little" ... and more.

The initial observation from Gardiner-Garden and Frommer [Gardiner-Garden and Frommer, 1987] on the role of DNA methylation in mammals has brought a lot of attention on the study of CpG Islands. The study showed that these small regions with high G+C content and generally located close to the promoter region of genes, could be linked with transcriptional and post-transcriptional repression of gene expression.

Ten years ago, Ehrlich [Ehrlich, 2002] reviewed in "*DNA methylation in cancer: too much, but also too little*" the role of DNA methylation in cancer. She brought the attention in particular on the fact that researchers were at that time too focused on aberrant targeted hypermethylation of CpG Islands in cancer and were probably missing out on the critical role of global hypomethylation in cancer.

More recently, Irizarry et al. [Irizarry et al., 2009] confronted the original assumption that CpG Island methylation was the most important epigenetic feature in gene regulation

and identified neighboring but non-CGI regions for which DNA methylation had an even higher association with transcription.

Finally, Sproul et al. [Sproul et al., 2011] revisited the role of DNA methylation in gene transcriptional regulation and suggested that the initial postulate might not always be true.

The evolution of the scientific community knowledge regarding epigenetics, reviewed here, is characteristic of the subtle trade-off between introducing bias and aiding the computational task.

6.2 "All models are wrong but are some of them actually useful?".

As previously discussed, the technological breakthroughs in biology have accelerated the acquisition of large datasets. Yet, although we can have access to millions of genomic features about a patient, we are still limited by the small number of patients. From a statistical point of view, it is important to make specific assumptions about the data in order to reduce the complexity of the problem. Here, we discuss how relevant it is to make valid biological assumptions about the data and how this can actually affect our results.

In Chapter 3, we used biological properties of DNA methylation, that is the robustness of DNA measurements, in comparison to RNA, and its stability over time, to develop a surrogate marker of the clonality between cells. This straightforward analysis has potentially important implications in the patients therapeutic strategies and illustrate the direct impact of computational tools to the clinic.

However, the other chapters do not share the same straightforwardness in the results. In Chapter 4, we confront the original postulate regarding the causal role of methylation challenged by Sproul et al. [Sproul et al., 2011]. Our results support the original postulate to some extent:

- The relationship between methylation and gene expression is more complex than originally stipulated.
- There is a poor generalization to the whole-genome.

In addition, we can quickly check for new hypotheses:

- How does genome-wide methylation impacts a single gene expression instead of simply its promoter methylation?
- How related is the level of regulation between normal and cancerous tissues?

Chapter 5 is also related to the validation of a biological observation. The hypermethylator phenotype was observed in several cancer but there was no causal biological phenomenon common to tissue specific phenotypes. We showed that adapted regression techniques using similarity between datasets could circumvent the instability of predictive signatures.

While bioinformatics will not replace biological validation, there is an important contribution related to guiding the focus of future experiments. In return, biological knowledge allows to adapt generic models to tackle the $n \ll p$ issue.

6.3 Perspectives in the use of computational tools for epigenetics and biology.

In this last section, we discuss the relevant perspectives to computational analysis in particular for epigenetics. During this thesis, we focused on the validation and generalization of biological phenomena using statistical methods. The recent problematics that arose in biology provide future directions to our results:

Tumor heterogeneity. Clonality between cells as discussed in Chapter 3 do not take into account the existence of several subclones. New methods to combine a deconvolution problem with clonality assessment could allow to discuss the evolution of tumor cells from its diagnosis to its relapse and help characterize the patients response to specific treatments.

Alternative splicing. The existence of orphan CGIs and more generally of non-CGI DNA methylation could be related to more subtle transcriptional regulatory mechanisms than those discussed in Chapter 4. Alternative splicing is an ongoing research subject that could benefit from methylation information.

Long Range epigenetic regulation. Several studies suggest the importance of DNA methylation for long range activation or repression of genes. This could relate

to the 3D structure of the DNA generally not taken into account in longitudinal studies. While the 3D structure reconstruction of the DNA is still an ongoing topic, the integration of such knowledge into further epigenetic analyses could prove useful.

Bibliography

[Acharjee, 2013] Acharjee, A. (2013). Comparison of Regularized Regression Methods for ˜Omics Data. *Journal of Postgenomics Drug & Biomarker Development*, 03(03).

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second international symposium on information theory*, pages 267–281.

[Allen, 1974] Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127.

[Amary et al., 2011] Amary, M., Damato, S., Halai, D., Eskandarpour, M., Berisha, F., and Bonar, F. (2011). Ollier disease and Maffucci syndrome are caused by somatic mosaic mutations of IDH1 and IDH2. *Nature Genetics*, 43.

[Anacleto et al., 2005] Anacleto, C., Leopoldino, A., Rossi, B., Soares, F. A., Lopes, A., Rocha, J. C., Caballero, O., Camargo, A. A., Simpson, A. J., and Pena, S. D. (2005). Colorectal cancer "methylator phenotype": fact or artifact? *Neoplasia*, 7(4):331–5.

[Arpino et al., 2002] Arpino, G., Clark, G. M., Mohsin, S., Bardou, V. J., and Elledge, R. M. (2002). Adenoid cystic carcinoma of the breast: molecular markers, treatment, and clinical outcome. *Cancer*, 94(8):2119–27.

[Auwera et al., 2010] Auwera, I. V. D., Yu, W., Suo, L., Neste, L. V., Dam, P. V., Marck, E. A. V., Pauwels, P., Vermeulen, P. B., Dirix, L. Y., and Laere, S. J. V. (2010). Array-Based DNA Methylation Profiling for Breast Cancer Subtype Discrimination. *PloS one*, 5(9).

[Bae et al., 2004] Bae, Y. K., Brown, A., Garrett, E., Bornman, D., Fackler, M. J., Sukumar, S., Herman, J. G., and Gabrielson, E. (2004). Hypermethylation in histologically distinct classes of breast cancer. *Clinical Cancer research*.

[Balaton et al., 1995] Balaton, A., Baviera, E., Galet, B., Vaury, P., and Vuong, P. (1995). Immunohistochemical evaluation of estrogen and progesterone receptors on

paraffin sections of breast carcinomas. Practical thoughts based on the study of 368 cases. *Arch. Anat. Cytol. Pathol.*, 43(1-2):93–100.

[Balaton et al., 1996] Balaton, A., Coindre, J., Collin, F., Ettore, F., Fiche, M., Jacquemier, J., et al. (1996). Recommandations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/-College of American Pathologists Clinical Practice Guideline Update. *Ann. Pathol.*, 16(2):144–148.

[Bartelink et al., 2007] Bartelink, H., Horiot, J. C., Poortmans, P. M., Van den Bogaert, W., Fourquet, A., Jager, J. J., et al. (2007). impact of a higher radiation dose on local control and survival in breast-conserving therapy of early breast cancer: 10-year results of the randomized boost versus no boost EORTC 22881-10882 trial. 25(22):3259–65.

[Baylin et al., 2001] Baylin, S. B., Esteller, M., Rountree, M. R., Bachman, K. E., Schuebel, K., and Herman, J. G. (2001). Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum. Mol. Genet.*, 10(7):683–692.

[Baylin and Herman, 2000] Baylin, S. B. and Herman, J. G. (2000). DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends in genetics*, 16(4):168–74.

[Baysan et al., 2012] Baysan, M., Bozdag, S., Cam, M. C., Kotliarova, S., Ahn, S., Walling, J., Killian, J. K., Stevenson, H., Meltzer, P., and Fine, H. a. (2012). G-cimp status prediction of glioblastoma samples using mRNA expression data. *PloS one*, 7(11):e47839.

[Bellman, 1961] Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton.

[Ben-Hur et al., 2002a] Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002a). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, 7:6–17.

[Ben-Hur et al., 2002b] Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002b). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 17:6–17.

[Bendas and Borsig, 2012] Bendas, G. and Borsig, L. (2012). Cancer Cell Adhesion and Metastasis: Selectins, Integrins, and the Inhibitory Potential of Heparins. *International Journal of Cell Biology*, (ID 676731).

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, 1:289–300.

[Bert et al., 2013] Bert, S. a., Robinson, M. D., Strbenac, D., Statham, A. L., Song, J. Z., Hulf, T., Sutherland, R. L., Coolen, M. W., Stirzaker, C., and Clark, S. J. (2013). Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer cell*, 23(1):9–22.

[Bibikova et al., 2011] Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., Fan, J.-B., and Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–95.

[Bird, 2002] Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21.

[Bock et al., 2007] Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG island mapping by epigenome prediction. *PLoS computational biology*, 3(6):e110.

[Bogdanović and Veenstra, 2009] Bogdanović, O. and Veenstra, G. J. C. (2009). DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma*, 118(5):549–65.

[Bollet et al., 2008] Bollet, M., Servant, N., Neuvial, P., Decraene, C., Lebigot, I., Meyniel, J. P., et al. (2008). High-Resolution Mapping of DNA Breakpoints to Define True Recurrences among Ipsilateral Breast Cancers. *J. Natl. Cancer Inst.*, 100:48–58.

[Borg et al., 1990] Borg, A., Tandon, A. K., Sigurdsson, H., Clark, G. M., Fernö, M., and Fuqua, S. A. (1990). HER2/neu Amplification predicts poor survival in Node-positive Breast Cancer. *Cancer Res*, 50:4332.

[Brommesson et al., 2008] Brommesson, S., Jönsson, G., Strand, C., Grabau, D., Malmström, P., Ringnér, M., et al. (2008). Tiling array-CGH for the assessment of genomic similarities among synchronous unilateral and bilateral invasive breast cancer tumor pairs. *BMC Clin. Pathol.*, 8:6.

[Cardillo, a] Cardillo, G. Kmplot.

[Cardillo, b] Cardillo, G. Logrank.

[Carter et al., 1989] Carter, C. L., Allen, C., and Henson, D. E. (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer*, 63(1):181–7.

[Chang and Lin, 2011] Chang, C. C. and Lin, C. H. (2011). LIBSVM : a library for support vector machines. *ACM TIST*, 2:27:1–27:27.

[Chen et al., 2012] Chen, H.-Y., Zhu, B.-H., Zhang, C.-H., Yang, D.-J., Peng, J.-J., Chen, J.-H., Liu, F.-K., and He, Y.-L. (2012). High CpG island methylator phenotype is associated with lymph node metastasis and prognosis in gastric cancer. *Cancer science*, 103(1):73–9.

[Colleoni et al., 2012] Colleoni, M., Rotmensz, N., Maisonneuve, P., Mastropasqua, M. G., Luini, a., Veronesi, P., Intra, M., Montagna, E., Cancello, G., Cardillo, a., Mazza, M., Perri, G., Iorfida, M., Pruneri, G., Goldhirsch, a., and Viale, G. (2012). Outcome of special types of luminal breast cancer. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, 23(6):1428–36.

[Cox and Oakes, 1984] Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

[Curtin et al., 2011] Curtin, K., Slattery, M. L., and Samowitz, W. S. (2011). CpG island methylation in colorectal cancer: past, present and future. *Pathology research international*, 2011:902674.

[Das and Singal, 2004] Das, P. M. and Singal, R. (2004). DNA methylation and cancer. *Journal of Clinical Oncology*, 22(22):4632–42.

[Dawson et al., 1998] Dawson, L., Chow, E., and Goss, P. E. (1998). Evolving perspectives in contralateral breast cancer. *Eur J Cancer*, 34(13):2000–2009.

[Deaton and Bird, 2011] Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & development*, 25(10):1010–22.

[Dudoit and Fridlyand, 2002] Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7).

[Efrat et al., 2006] Efrat, A., Fan, Q., and Venkatasubramanian, S. (2006). Curve Matching, Time Warping, and Light Fields: New Algorithms for Computing Similarity between Curves. *Journal of Mathematical Imaging and Vision*, 27(3):203–216.

[Ehrlich, 2002] Ehrlich, M. (2002). DNA methylation in cancer : too much , but also too little. *Oncogene*, 21:5400–5413.

[Ehrlich, 2009] Ehrlich, M. (2009). DNA hypomethylation in cancer cells. *Epigenomics*, 1(2):239–259.

[Ellis et al., 1992] Ellis, I. O., Galea, M., Broughton, N., Locker, A., Blamey, R. W., and Elston, C. W. (1992). Pathological prognostic factors in breast cancer. ii. histological type. relationship with survival in a large study with long-term follow-up. *Histopathology*.

[Ellsworth et al., 2005] Ellsworth, R. E., Ellsworth, D. L., Neatrour, D. M., Denyarmin, B., Lubert, S. M., Sarachine, M. J., et al. (2005). Allelic Imbalance in Primary Breast Carcinomas and Metastatic Tumors of the Axillary Lymph Nodes. *Mole. Can. Res.*, 3:71–77.

[Estécio et al., 2007] Estécio, M. R. H., Yan, P. S., Ibrahim, A. E. K., Tellez, C. S., Shen, L., Huang, T. H.-M., and Issa, J.-P. J. (2007). High-throughput methylation profiling by MCA coupled to CpG island microarray. *Genome research*, 17(10):1529–36.

[Esteller, 2002] Esteller, M. (2002). CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene*, 21(35):5427–40.

[Esteller, 2008] Esteller, M. (2008). Epigenetics in cancer. *The New England journal of medicine*, 358(11):1148–59.

[Esteller et al., 2001] Esteller, M., Corn, P. G., Baylin, S. B., and Herman, J. G. (2001). A gene hypermethylation profile of human cancer. *Cancer Research*, 61(8):3225–3229.

[Etoh et al., 2004] Etoh, T., Kanai, Y., Ushijima, S., Nakagawa, T., Nakanishi, Y., Sasako, M., et al. (2004). Increased DNA methyltransferase 1 (DNMT1) protein expression correlates significantly with poorer tumor differentiation and frequent DNA hypermethylation of multiple CpG islands in gastric cancers. *Am J Pathol*, (164):689–99.

[Fang et al., 2011] Fang, F., Turcan, S., Rimner, A., Kaufman, A., Giri, D., Morris, L. G. T., Shen, R., Seshan, V., Mo, Q., Heguy, A., Baylin, S. B., Ahuja, N., Viale, A., Massague, J., Norton, L., Vahdat, L. T., Moynahan, M. E., and Chan, T. a. (2011). Breast cancer methylomes establish an epigenomic foundation for metastasis. *Science translational medicine*, 3(75):75ra25.

[Feng et al., 2007] Feng, Y., Sun, B., Li, X., Zhang, L., Niu, Y., Xiao, C., et al. (2007). Differentially expressed genes between primary cancer and paired lymph node metastases predict clinical outcome of node-positive breast cancer patients. *Breast Cancer Res. Treat.*, 103:319–329.

[Figueroa et al., 2010] Figueroa, M., Abdel-Wahab, O., Lu, C., Ward, P., Patel, J., Shih, A., et al. (2010). Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell*, (18):553–67.

[Fisher et al., 1975] Fisher, E. R., Gregorio, R. M., Fisher, B., Redmond, C., Vellios, F., and Sommers, S. C. (1975). The pathology of invasive breast cancer. A syllabus derived from findings of the National Surgical Adjuvant Breast Project (protocol no. 4). *Cancer*, 36:1–85.

[García-Closas et al., 2013] García-Closas, M., Gail, M. H., Kelsey, K. T., and Ziegler, R. G. (2013). Searching for blood DNA methylation markers of breast cancer risk and early detection. *Journal of the National Cancer Institute*, 105(10):678–80.

[Garcia-Manero et al., 2002] Garcia-Manero, G., Daniel, J., Smith, T., Kornblau, S., Lee, M., Kantarjian, H., et al. (2002). DNA methylation of multiple promoter-associated CpG islands in adult acute lymphocytic leukemia. *Clinical Cancer Research*, (8):2217–24.

[Gardiner-Garden and Frommer, 1987] Gardiner-Garden, M. and Frommer, M. (1987). CpG Islands in Vertebrate Genomes. *Journal of Molecular Biology*, 2(196):261–282.

[Gevaert, 2015] Gevaert, O. (2015). Pancancer analysis of DNA methylation-driven genes using MethylMix. *Genome Biology*, 16(17).

[Gilkes et al., 2014] Gilkes, D. M., Semenza, G. L., and Wirtz, D. (2014). Hypoxia and the extracellular matrix: drivers of tumour metastasis. *Nature Reviews Cancer*, (14):430–439.

[Gupta et al., 2010] Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., et al. (2010). Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464:1071–1076.

[Haffner et al., 2011] Haffner, M., Chaux, A., Meeker, A., Esopi, D., Gerber, J., Pellakuru, L., Toubaji, A., Argani, P., Iacobuzio-Donahue, C., Nelson, W., Netto, G., De Marzo, A., and Yegnasubramian, S. (2011). Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. *Oncotarget*, 2:627–637.

[Haffty et al., 1996] Haffty, B. F., Reiss, M., Beinfield, M., Fischer, D., Ward, B., and McKhann, C. (1996). Ipsilateral breast tumor recurrence as a predictor of distant disease: implications for systemic therapy at the time of local relapse. 14(1):52–57.

[Hanahan and Weinberg, 2000] Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 1(100):57–70.

[Hanahan and Weinberg, 2011] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 5(144):646–74.

[Hansen et al., 2012] Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, 13(10):R83.

[Harris et al., 2007] Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., Somerfield, M. R., Hayes, D. F., and Bast, R. C. (2007). American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of Clinical Oncology*, 25(33):5287–312.

[Haury et al., 2011] Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, 6(12):e28210.

[Healey et al., 1993] Healey, E. A., Cook, E. F., Orav, E. J., Schnitt, S. J., Connolly, J. L., and Harris, J. R. (1993). Contralateral breast cancer: clinical characteristics and impact on prognosis. 11(8):1545–1552.

[Heidecker et al., 2000] Heidecker, L., Brasseur, F., Probst-Kepper, M., Guéguen, M., Boon, T., and Van den Eynde, B. J. (2000). Cytolytic T lymphocytes raised against a human bladder carcinoma recognize an antigen encoded by gene MAGE-A12. *Journal of immunology (Baltimore, Md. : 1950)*, 164(11):6041–6045.

[Henrichsen et al., 2009] Henrichsen, C. N., Chaignat, E., and Reymond, A. (2009). Copy number variants, diseases and gene expression. *Human molecular genetics*, 18(R1):R1–8.

[Herman et al., 1998] Herman, J., Umar, A., Polyak, K., Graff, J., Ahuja, N., Issa, J., et al. (1998). Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci U S A*, (95):6870–5.

[Hinoue et al., 2012] Hinoue, T., Weinsenberger, D., Lange, C., Shen, H., Byun, H., Van Den Berg, D., et al. (2012). Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Research*, (22):271–82.

[Hoerl and Kennard, 2000] Hoerl, A. E. and Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 42(1):80–86.

[Hoerl et al., 1970] Hoerl, A. E., Kennard, R. W., and Kennard, W. (1970). Ridge Regression : Applications to Nonorthogonal Problems. *Technometrics*, 12(1):69–82.

[Hon et al., 2012] Hon, G. C., Hawkins, R. D., Caballero, O. L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L. E., Camargo, A. A., Stevenson, B. J., Ecker, J. R., Bafna, V., Strausberg, R. L., Simpson, A. J., and Ren, B. (2012).

Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome research*, 22(2):246–58.

[Houseman et al., 2009] Houseman, E. A., Christensen, B. C., Karagas, M. R., Wrensch, M. R., Nelson, H. H., Wiemels, J. L., Zheng, S., Wiencke, J. K., Kelsey, K. T., and Marsit, C. J. (2009). Copy number variation has little impact on bead-array-based measures of DNA methylation. *Bioinformatics (Oxford, England)*, 25(16):1999–2005.

[Hughes et al., 2013] Hughes, L. A. E., Melotte, V., de Schrijver, J., de Maat, M., Smit, V. T. H. B. M., Bovee, J. V. M. G., French, P. J., van den Brandt, P. A., Schouten, L. J., de Meyer, T., Van Criekinge, W., Ahuja, N., Herman, J. G., Weijenberg, M. P., and van Engeland, M. (2013). The CpG island methylator phenotype: what's in a name? *Cancer research*.

[Illingworth et al., 2010] Illingworth, R. S., Gruenewald-Schneider, U., Webb, S., Kerr, A. R. W., James, K. D., et al. (2010). Orphan cpg islands identify numberous conserved promoters in the mammalian genome. *PLoS Genetics*, (10.1371/journal.pgen.1001134).

[Imyanitov et al., 2002] Imyanitov, E. N., Suspitsin, E. N., Grigoriev, M. Y., Togo, A. V., Belogubova, E. V., Pozharisski, K. M., et al. (2002). Concordance of Allelic Imbalance Profiles in Synchronous and Metachronous Bilateral Breast Carcinomas. *Int. J. Cancer*, 100:557–564.

[Irizarry et al., 2009] Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J. B., Sabunciyan, S., and Feinberg, A. P. (2009). Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*, 41(2):178–186.

[Issa et al., 2005] Issa, J.-P. J., Shen, L., and Toyota, M. (2005). CIMP, at last. *Gastroenterology*, 129(3):1121–4.

[Itzkovitz et al., 2006] Itzkovitz, S., Tlusty, T., and Alon, U. (2006). Coding limits on the number of transcription factors. *BMC genomics*, 7:239.

[Jacob et al., 2009] Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. *Proceedings of the 26th International Conference on Machine Learning*.

[Jatoi, 1999] Jatoi, I. (1999). Management of the axilla in primary breast cancer. *Surg. Clin. North. Am*, 79(5):1061–1073.

[Jin et al., 2011] Jin, S., Jiang, Y., Qiu, R., Rauch, T., Wang, Y., Schackert, G., Krex, D., and Lu, Q. (2011). 5-hydroxymethylcytosine is strongly depleted in human cancers, but its levels do not correlate with IDH1 mutations. *Cancer Research*, 71:7360–7365.

[Jones, 1986] Jones, P. (1986). DNA methylation and cancer. *Cancer Research*, 46(2):461–466.

[Jones and Baylin, 2007] Jones, P. a. and Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, 128(4):683–92.

[Jones et al., 2012] Jones, S., Li, M., Parsons, D., Zhang, X., Wesseling, J., Kristel, P., et al. (2012). Somatic mutations in the chromatin remodeling gene ARID1A occur in several tumor types. *Hum Mutat*, (33):100–3.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, D. (1958). Nonparametric estimation from incomplete observation. *J. Am. Statist.*, 58:457–481.

[Keogh and Pazzani, 1999] Keogh, E. J. and Pazzani, M. J. (1999). Scaling up Dynamic Time Warping to Massive Datasets. *Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (KDD)*, 1704:1–11.

[Keshet et al., 2006] Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R. a., Niveleau, A., Cedar, H., and Simon, I. (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nature genetics*, 38(2):149–53.

[Kho et al., 1997] Kho, M. R., Baker, D. J., Laayoun, A., and Smith, S. S. (1997). Stalling of human dna (cytosine-5) methyltransferase at single-strand conformers from a site of dynamic mutation. *Journal of Molecular Biology*, 275:67–79.

[Kim et al., 2003] Kim, H., Kim, Y., Kim, S., Kim, N., and Noh, S. (2003). Concerted promoter hypermethylation of hMLH1, p16INK4A, and E-cadherin in gastric carcinomas with microsatellite instability. *J Pathol*, (200):23–31.

[Klose and Bird, 2006] Klose, R. J. and Bird, A. P. (2006). Genomic DNA methylation: the mark and its mediators. *Trends in biochemical sciences*, 31(2):89–97.

[Knudson, 1971] Knudson, A. (1971). Mutation and Cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. USA*, 68(4):820–823.

[Kriaucionis and Heintz, 2009] Kriaucionis, S. and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, 324:929–930.

[Kulis and Esteller, 2010] Kulis, M. and Esteller, M. (2010). DNA methylation and cancer. *Advances in genetics*, 70(10):27–56.

[Kulis et al., 2013] Kulis, M., Queirós, A. C., Beekman, R., and Martín-Subero, J. I. (2013). Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochimica et biophysica acta*, 1829(11):1161–74.

[Kusano et al., 2006] Kusano, M., Toyota, M., Suzuki, H., Akino, K., Aoki, F., Fujita, M., et al. (2006). Genetic, epigenetic, and clinicopathologic features of gastric carcinomas with the CpG island methylator phenotype and an association with Epstein-Barr virus. *Cancer*, (106):1467–79.

[Kwee et al., 2011] Kwee, I., Rinaldi, A., Rancoita, P., Rossi, D., Capello, D., Forconi, F., et al. (2011). Integrated DNA copy number and methylation profiling of lymphoid neoplasms using a single array. *Br. J. Haematol*, 156(3):354–357.

[Laird and Jaenisch, 1994] Laird, P. W. and Jaenisch, R. (1994). DNA methylation and cancer. *Human molecular genetics*, 3:1487–95.

[Laurent et al., 2010] Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Kin Sung, K. W., Rigoutsos, I., Loring, J., and Wei, C.-L. (2010). Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–31.

[Lauss et al., 2012a] Lauss, M., Aine, M., Sjödahl, G., Veerla, S., Patschan, O., Gudjonsson, S., Chebil, G., Lövgren, K., Fernö, M. r., Må nsson, W., Liedberg, F., Ringnér, M., Lindgren, D., and Höglund, M. (2012a). DNA methylation analyses of urothelial carcinoma reveal distinct epigenetic subtypes and an association between gene copy number and methylation status. *Epigenetics*, 7(8):858–867.

[Lauss et al., 2012b] Lauss, M., Aine, M., Sjödahl, G., Veerla, S., Patschan, O., Gudjonsson, S., et al. (2012b). DNA methylation analyses of urothelial carcinoma reveal distinct epigenetic subtypes and an association between gene copy number and methylation status. *Epigenetics*, 7(8):858–867.

[Li et al., 2005] Li, C. I., Uribe, D. J., and Daling, J. R. (2005). Clinical characteristics of different histologic types of breast cancer. *British journal of cancer*, 93(9):1046–52.

[Li et al., 1993] Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, 366:362–365.

[Ling and Groop, 2009] Ling, C. and Groop, L. (2009). Epigenetics: a molecular link between environmental factors and type 2 diabetes. *Diabetes*, 12(58):2718–2725.

[Lister et al., 2009] Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, a. H., Thomson, J. a., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–22.

[Liu et al., 2008] Liu, Z., Zhao, J., Chen, X., Li, W., Liu, R., Lei, Z., et al. (2008). CpG island methylator phenotype involving tumor suppressor genes located o chromosome 3p in non-small cell lung cancer. *Lung Cancer*, (62):15–22.

[Lu et al., 2012] Lu, P., Weaver, V. M., and Werb, Z. (2012). The extracellular matrix: A dynamic niche in cancer progression. *Journal of Cell Biology*, 196(4):395–406.

[Mann and Whitney, 1947] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

[Maruyama et al., 2001] Maruyama, R., Toyooka, S., Toyooka, K., Harada, K., Virmani, A., Zochbauer-Muller, S., et al. (2001). Aberrant promoter methylation profile of bladder cancer and its relationship to clinicopathological features. *Cancer Research*, (61):8659–63.

[Maruyama et al., 2002] Maruyama, R., Toyooka, S., Toyooka, K., Virmani, A., Zochbauer-Muller, S., Farinas, A., et al. (2002). Aberrant promoter methylation profile of prostate cancers and its relationship to clinicopathological features. *Clinical Cancer Research*, (8):514–9.

[Maunakea et al., 2013] Maunakea, A. K., Chepelev, I., Cui, K., and Zhao, K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell research*, 23(11):1256–69.

[Meier et al., 2008] Meier, L., Geer, S. V. D., Bühlmann, P., and Zürich, E. T. H. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*.

[Meinshausen and Bühlmann, 2008] Meinshausen, N. and Bühlmann, P. (2008). Stability selection. *arXiv*, (0809.2932v1).

[Meissner et al., 2008] Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., and Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770.

[Metcalfe et al., 2004] Metcalfe, K., Lynch, H. T., Ghadirian, P., Tung, N., Olivotto, I., Warner, E., et al. (2004). Contralateral breast cancer in BRCA1 and BRCA2 mutation carriers. 22(12):2328–2335.

[Moarii et al., 2014] Moarii, M., Pinheiro, A., Sigal-Zafrani, B., Fourquet, A., Caly, M., Servant, N., Stoven, V., Vert, J., and Reyal, F. (2014). Epigenomic Alterations in Breast Carcinoma from Primary Tumor to Locoregional Recurrences. *PLoS ONE*, (9):e103986.

[Mollaoglu et al., 2008] Mollaoglu, N., Vairaktaris, E., Nkenke, E., Neukam, F. W., and Ries, J. (2008). Expression of MAGE-A12 in oral squamous cell carcinoma. *Disease markers*, 24(1):27–32.

[Monti et al., 2003] Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus Clustering : A Resampling-Based Method for Class Discovery and Visualization of Gene. *Machine Learning*, 52:91–118.

[Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):1–8.

[Nestor et al., 2010] Nestor, C., Ruzov, A., Meehan, R., and Dunican, D. (2010). Enzymatic approaches and bisulfite sequencing cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine in DNA. *Biotechniques*, 48:317–319.

[Newell-Price et al., 2000] Newell-Price, J., Clark, A. J., and King, P. (2000). DNA Methylation and Silencing of Gene Expression. *Trends in Endocrinology & Metabolism*, 11(4):142–148.

[Nielsen et al., 2010] Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S. R., Snider, J., Stijleman, I. J., Reed, J., Cheang, M. C. U., Mardis, E. R., Perou, C. M., Bernard, P. S., and Ellis, M. J. (2010). A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 16(21):5222–32.

[Nobori et al., 1994] Nobori, T., Miura, K., Wu, D. J., Lois, A., Takabayashi, K., and Carson, D. A. (1994). Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers. *Nature*, 368:753–756.

[Noushmehr et al., 2010] Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., Pan, F., Pelloski, C. E., Sulman, E. P., Bhat, K. P., Verhaak, R. G. W., Hoadley, K. a., Hayes, D. N., Perou, C. M., Schmidt, H. K., Ding,

L., Wilson, R. K., Van Den Berg, D., Shen, H., Bengtsson, H., Neuvial, P., Cope, L. M., Buckley, J., Herman, J. G., Baylin, S. B., Laird, P. W., and Aldape, K. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell*, 17(5):510–22.

[Ogino et al., 2006] Ogino, S., Kawasaki, T., Kirkner, G. J., Loda, M., and Fuchs, C. S. (2006). CpG island methylator phenotype-low (CIMP-low) in colorectal cancer: possible associations with male sex and KRAS mutations. *J Mol Diagn*, 8:582–8.

[Okano et al., 1999] Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99:247–257.

[Okegawa et al., 2004] Okegawa, T., Pong, R., and Hsieh, J. (2004). The role of cell adhesion molecule in cancer progression and its application in cancer therapy. *Acta. Biochim. Pol.*, 51(2):445–57.

[Ostrovnaya et al., 2010] Ostrovnaya, I., Olshen, A. B., Seshan, V. E., Orlow, I., Albertson, D. G., and Begg, C. B. (2010). A Metastasis or a Second Independent Cancer? Evaluating the clonal origin of tumors using array copy number data. *Stat. Med.*, 29(15):1608–1621.

[Oue et al., 2003] Oue, N., Oshimo, Y., Nakayama, H., Ito, R., Yoshida, K., Matsusaki, K., et al. (2003). DNA methylation of multiple genes in gastric carcinoma: association with histological type and CpG island methylator phenotype. *Cancer Science*, (94):901–5.

[Paik et al., 2004] Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J., and Wolmark, N. (2004). A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *New England Journal of Medicine*, pages 2817–2826.

[Pansuriya et al., 2011] Pansuriya, T., van Eijk, R., d'Adamo, P., van Ruler, M., Kuijjer, M., Oosting, J., et al. (2011). Somatic mosaic IDH1 and IDH2 mutations are associated with enchondroma and spindle cell hemangioma in Ollier disease and Maffucci syndrome. *Nature Genetics*, 43:1256–61.

[Parker et al., 2009] Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., 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. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 27(8):1160–7.

[Peche et al., 2012] Peche, L. Y., Scolz, M., Ladelfa, M. F., Monte, M., and Schneider, C. (2012). MageA2 restrains cellular senescence by targeting the function of PMLIV/p53 axis at the PML-NBs. *Cell Death and Differentiation*, 19(6):926–936.

[Perou et al., 2000] Perou, C. M., Sørlie, T., Eisen, M. B., Rijn, M. V. D., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, I., Pergamenschikov, A., Williams, C., Zhu, S. X., Lùnning, P. E., Brown, P. O., Botstein, D., and Grant, S. (2000). Molecular portraits of human breast tumours. *Nature*, 533(May):747–752.

[Pollex and Heard, 2012] Pollex, T. and Heard, E. (2012). Recent advances in X-chromosome inactivation research. *Current opinion in cell biology*, 24(6):825–32.

[Pourebrahim et al., 2011] Pourebrahim, R., Houtmeyers, R., Ghogomu, S., Janssens, S., Thelie, A., Tran, H., et al. (2011). Transcription factor Zic2 inhibits Wnt/beta-catenin protein signaling. *J Biol Chem*, 286(43):37732–40.

[Rabiner and Juang, 1993] Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Razin and Riggs, 1980] Razin, A. and Riggs, A. D. (1980). DNA Methylation and Gene Function. *Science*, 210:604–10.

[Reddington et al., 2014] Reddington, J. P., Sproul, D., and Meehan, R. R. (2014). DNA methylation reprogramming in cancer: does it act by re-configuring the binding landscape of Polycomb repressive complexes? *Bioessays*, 36(2).

[Reich et al., 2006] Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J. P. (2006). GenePattern 2.0. *Nature genetics*, 38(5):500–1.

[Reinhard et al., 2014] Reinhard, H., Yousef, S., Luetkens, T., Fehse, B., Berdien, B., Kröger, N., and Atanackovic, D. (2014). Cancer-testis antigen MAGE-C2/CT10 induces spontaneous CD4+ and CD8+ T-cell responses in multiple myeloma patients. *Blood cancer journal*, 4:e212.

[Reyal et al., 2008] Reyal, F., van Vliet, M. H., Armstrong, N. J., Horlings, H. M., de Visser, K. E., Kok, M., Teschendorff, A. E., Mook, S., van 't Veer, L., Caldas, C., Salmon, R. J., van de Vijver, M. J., and Wessels, L. F. a. (2008). A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast cancer research*, 10(6):R93.

[Richards, 2009] Richards, M. a. (2009). The size of the prize for earlier diagnosis of cancer in England. *British journal of cancer*, 101 Suppl 2(S2):S125–9.

[Riggs and Porter, 1996] Riggs, A. D. and Porter, T. N. (1996). Cold Spring Harbor Laboratory Press.

[Robertson, 2005] Robertson, K. D. (2005). Dna methylation and human disease. *Nat Rev Genet*, 8(6):597–610.

[Rodriguez et al., 2010] Rodriguez, C., Borgel, J., Court, F., Cathala, G., Forne, T., and Piette, J. (2010). CTCF is a DNA methylation-sensitive positive regulator of the INK/ARF locus. *Biochem. Biophys. Res. Commun.*, 392:129–134.

[Rodriguez-Paredes and Esteller, 2011] Rodriguez-Paredes, M. and Esteller, M. (2011). Cancer epigenetics reaches mainstream oncology. *Nature Medicine*, pages 330–339.

[Roman-Gomez et al., 2006] Roman-Gomez, J., Jimenez-Velasco, A., Agirre, X., Castillejo, J., Navarro, G., Calasanz, M., et al. (2006). CpG island methylator phenotype redefines the prognostic effect of t(12;21) in childhood acute lymphoblastic leukemia. *Clinical Cancer Research*, (12):4845–50.

[Roman-Gomez et al., 2005] Roman-Gomez, J., Jimenez-Velasco, A., Agirre, X., Prosper, F., Heiniger, A., and Torres, A. (2005). Lack of CpG island methylator phenotype defines a clinical subtype of T-cell acute lymphoblastic leukemia associated with good prognosis. *Journal of Clinical Oncology*, (23):7043–9.

[Rountree et al., 2001] Rountree, M. R., Bachman, K. E., Herman, J. G., and Baylin, S. B. (2001). DNA methylation, chromatin inheritance, and cancer. *Oncogene*, 20(24):3156–65.

[Sabbah et al., 2011] Sabbah, C., Mazo, G., Paccard, C., Reyal, F., and Hupe, P. (2011). SMETHILLIUM: Spatial normalisation METHod for ILLumina InfinIUM HumanMethylation BeadChip. *Bioinformatics*, 27(12):1693–5.

[Saxonov et al., 2006] Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci.*, 103(5):1412–1417.

[Scanlan et al., 2002] Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J., and Chen, Y.-T. (2002). Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunological Reviews*, 188:22–32.

[Schermelleh et al., 2007] Schermelleh, L., Haemmer, A., Spada, F., Rösing, N., Meilinger, D., Rothbauer, U., et al. (2007). Dynamics of Dnmt1 interaction with

the replication machinery and its role in postreplicative maintenance of DNA methylation. *Nucleic Acids Res.*, 35:4301–4312.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Serra and Berthod, 1994] Serra, B. and Berthod, M. (1994). Subpixel contour matching using continuous dynamic programming. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 202–207.

[Shen et al., 2007] Shen, L., Toyota, M., Kondo, Y., Lin, E., Zhang, L., Guo, Y., et al. (2007). Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci USA*, (104):18654–9.

[Shibata et al., 1996] Shibata, A., Tsai, Y. C., Press, M. F., Henderson, B. E., Jones, P. A., and Ross, R. K. (1996). Clonal Analysis of bilateral breast cancer. *Cancer Res*, 2:743–748.

[Silverman et al., 2002] Silverman, L. R., Demakos, E. P., Peterson, B. L., Kornblith, A. B., Holland, J. C., Odchimar-Reissig, R., Stone, R. M., Nelson, D., Power, B. L., DeCastro, C. M., Ellerton, J., Larson, R. A., Schiffer, C. A., and Holland, J. F. (2002). Randomized Controlled Trial of Azacitidine in Patients With the Myelodysplastic Syndrome: A Study of the Cancer and Leukemia Group B. *Journal of Clinical Oncology*, 20(10):2429–2440.

[Smith et al., 1999] Smith, T. E., Lee, D., Turner, B. C., Carter, D., and Haffty, B. G. (1999). 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. Col.*, 48:1281–1289.

[Smith and Meissner, 2013] Smith, Z. D. and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature reviews. Genetics*, 14(3):204–20.

[Sneddon et al., 2006] Sneddon, J., Zhen, H., Montgomery, K., van de Rijn, M., Tward, A., West, R., et al. (2006). Bone morphogenetic protein antagonist gremlin 1 is widely expressed by cancer-associated stromal cells and can promote tumor cell proliferation. *Proc Natl Acad Sci USA*, 103(40):14842–7.

[Sørlie et al., 2001] Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Rijn, M. V. D., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein, P., and Børresen-dale, A.-l. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.*, 98(19).

[Sproul et al., 2012] Sproul, D., Kitchen, R. R., Nestor, C. E., Dixon, J. M., Sims, A. H., Harrison, D. J., Ramsahoye, B. H., and Meehan, R. R. (2012). Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biology*, 13(R84).

[Sproul and Meehan, 2013] Sproul, D. and Meehan, R. R. (2013). Genomic insights into cancer-associated aberrant CpG island hypermethylation. *Briefings in functional genomics*, 12(3):174–90.

[Sproul et al., 2011] Sproul, D., Nestor, C., Culley, J., Dickson, J. H., Dixon, J. M., Harrison, D. J., Meehan, R. R., Sims, A. H., and Ramsahoye, B. H. (2011). Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 108(11):4364–9.

[Stone, 1974] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147.

[Stranger et al., 2007] Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E., and Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y.)*, 315(5813):848–53.

[Struhl, 1999] Struhl, K. (1999). Fundamentally Different Logic of Gene Regulation in Eukaryotes and Prokaryotes. *Cell*, 98:1–4.

[Suzuki et al., 2006] Suzuki, M., Shigematsu, H., Lizasa, T., Hiroshima, K., Nakatani, Y., Minna, J., et al. (2006). Exclusive mutation in epidermal growth factor receptor gene, HER-2, and KRAS, and synchronous methylation of nonsmall cell lung cancer. *Cancer*, (106):2200–7.

[Takai and Jones, 2002] Takai, D. and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *PNAS*, 99(6):3740–3745.

[Tang and Goldberg, 2009] Tang, H. and Goldberg, E. (2009). Homo sapiens lactate dehydrogenase c (Ldhc) gene expression in cancer cells is regulated by transcription factor Sp1, CREB, and CpG island methylation. *Journal of andrology*, 30(2):157–167.

[Tate and Bird, 1993] Tate, P. H. and Bird, A. P. (1993). Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr. Opin. Genet. Dev.*, 3:226–31.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

[Timp et al., 2014] Timp, W., Bravo, H. C., McDonald, O. G., Goggins, M., Umbricht, C., Zeiger, M., Feinberg, A. P., and Irizarry, R. A. (2014). Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Medicine*, 6(61).

[Timp and Feinberg, 2013] Timp, W. and Feinberg, A. P. (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Epigenetics and Genetics*, 13(July).

[Toyota et al., 1999a] Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B., and Issa, J.-P. J. (1999a). CpG island methylator phenotype in colorectal cancer. *Proc. Natl. Acad. Sci.*, 96(July):8681–8686.

[Toyota et al., 1999b] Toyota, M., Ahuja, N., Suzuki, H., Itoh, F., Ohe-Toyota, M., Imai, K., et al. (1999b). Aberrant methylation in gastric cancer associated with the CpG island methylator phenotype. *Cancer Research*, 59:5438–42.

[Toyota et al., 2001] Toyota, M., Kopecky, K., Toyota, M., Jair, K., Willman, C., and Issa, J. (2001). Methylation profiling in acute myeloid leukemia. *Blood*, (97):2823–9.

[van de Vijver et al., 2002] van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., Van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A Gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009.

[Van Der Sijp et al., 2002] Van Der Sijp, J. R., van Meerbeeck, J. P., Maat, A. P., Zondervan, P. E., Sleddens, H. F., and van Geel, A. N. (2002). Determination of the molecular relationship between multiple tumors within one patient is of clinical importance. 20:1105–1114.

[Van Dongen et al., 2000] Van Dongen, J. A., Voogd, A. C., Fentiman, I. S., Legrand, C., Sylvester, R., Tong, D., et al. (2000). Long-Term Results of a Randomized Trial Comparing Breast-Conserving Therapy with Mastectomy : European Organization for Research and Treatment of Cancer 10801 Trial. *J. Natl. Cancer Inst.*, 92:1143–50.

[van Vlodrop et al., 2011] van Vlodrop, I. J. H., Niessen, H. E. C., Derks, S., Baldewijns, M. M. L. L., van Criekinge, W., Herman, J. G., and van Engeland, M. (2011). Analysis of promoter CpG island hypermethylation in cancer: location, location, location! *Clinical Cancer Research*, 17(13):4225–31.

[Vanderkraats et al., 2013] Vanderkraats, N. D., Hiken, J. F., Decker, K. F., and Edwards, J. R. (2013). Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic acids research*, 41(14):6816–27.

[Van't Veer et al., 2002] Van't Veer, L. J., Dai, H., Van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., Van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(345):530–536.

[Vaquerizas et al., 2009] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. a., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–63.

[Varley et al., 2013] Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F., Cross, M. K., Williams, B. a., Stamatoyannopoulos, J. a., Crawford, G. E., Absher, D. M., Wold, B. J., and Myers, R. M. (2013). Dynamic DNA methylation across diverse human cell lines and tissues. *Genome research*, 23(3):555–67.

[Venet et al., 2011] Venet, D., Dumont, J. E., and Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS computational biology*, 7(10):e1002240.

[Veronesi et al., 1995] Veronesi, U., Marubini, E., Del Vecchio, M., Manzari, A., Greco, M., Luini, A., et al. (1995). Local Recurrences and Distant Metastases After Conservative Breast Cancer Treatments: Partly Independent Events. *J. Natl. Cancer Inst.*, 87:19–27.

[Vert and Bleakley, 2010] Vert, J.-P. and Bleakley, K. (2010). Fast detection of multiple change-points shared by many signals using group lars. *Advances in Neural Information Processing Systems 2013*, pages 2343–2351.

[Viale et al., 2009] Viale, G., Rotmensz, N., Maisonneuve, P., Orvieto, E., Maiorano, E., Galimberti, V., Luini, A., Colleoni, M., Goldhirsch, A., and Coates, A. S. (2009). Lack of prognostic significance of "classic" lobular breast carcinoma: a matched, single institution series. *Breast cancer research and treatment*, 117(1):211–4.

[Vichapat et al., 2012] Vichapat, V., Garmo, H., Holmqvist, M., Liljegren, G., Wärnberg, F., Lambe, M., et al. (2012). Tumor Stage Affects Risk and Prognosis of Contralateral Breast Cancer: Results From a Large Swedish Population Based Study. 30:3478–3485.

[Vicini et al., 2007] Vicini, F. A., Antonucci, J. V., Goldstein, N., Wallace, M., Kestin, L., Krauss, D., et al. (2007). 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*, 109:1264–72.

[von Boehmer et al., 2011] von Boehmer, L., Keller, L., Mortezavi, A., Provenzano, M., Sais, G., Hermanns, T., Sulser, T., Jungbluth, A. a., Old, L. J., Kristiansen, G., van den Broek, M., Moch, H., Knuth, A., and Wild, P. J. (2011). MAGE-C2/CT10 protein expression is an independent predictor of recurrence in prostate cancer. *PLoS ONE*, 6(7):1–7.

[Wang et al., 2005] Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Gelder, M. E. M.-v., Yu, J., Jatkoe, T., Berns, E. M. J. J., Atkins, D., and Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679.

[Wang and Leung, 2004] Wang, Y. and Leung, F. C. C. (2004). An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics*, 20(7):1170–1177.

[Weigelt et al., 2005] Weigelt, B., Wessels, L. F., Bosma, A. J., Glas, A. M., Nuyten, D. S., He, Y. D., et al. (2005). No common denominator for breast cancer lymph node metastasis. *Br. J. Cancer*, 93:924–932.

[Weinberg, 2007] Weinberg, R. A. (2007). Garland Science.

[Weisenberger et al., 2006] Weisenberger, D. J., Siegmund, K. D., Campan, M., Young, J., Long, T. I., Faasse, M. a., Kang, G. H., Widschwendter, M., Weener, D., Buchanan, D., Koh, H., Simms, L., Barker, M., Leggett, B., Levine, J., Kim, M., French, A. J., Thibodeau, S. N., Jass, J., Haile, R., and Laird, P. W. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nature genetics*, 38(7):787–93.

[Weisenberger et al., 2008] Weisenberger, D. J., Van Den Berg, D., Pan, F., Berman, B. P., and Laird, P. (2008). Comprehensive DNA methylation analysis on the illumina infinium assay platform. Technical report.

[Werner, 2013] Werner, E. (2013). What Transcription Factors Can't Do: On the Combinatorial Limits of Gene Regulatory Networks. *arXiv*, page 12.

[West et al., 2011] West, N. R., Panet-Raymond, V., Truong, P. T., Alexander, C., Babinsky, S., Milne, K., et al. (2011). Intratumoral Immune Responses can Distinguish

New Primary and True Recurrence Types of Ipsilateral Breast Tumor Recurrences (IBTR). *Breast Cancer (Auckl)*, 5:105–115.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparison by ranking methods. *Biometrics Bulletin*, 1(6).

[Wolff et al., 2013] Wolff, A. C., Hammond, E. H., Hicks, D. G., Dowsett, M., McShane, L. M., H, A. K., et al. (2013). Recommandations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Update. 31(31):3997–4013.

[Wu et al., 2010] Wu, H., Caffo, B., Jaffee, H. a., Irizarry, R. a., and Feinberg, A. P. (2010). Redefining CpG islands using hidden Markov models. *Biostatistics (Oxford, England)*, 11(3):499–514.

[Wu et al., 1998] Wu, T. D., Schmidler, S. C., Hastie, T., and Brutlag, D. L. (1998). Regression analysis of multiple protein structures. *Journal of computational biology*, 5(3):585–595.

[Yang et al., 2014] Yang, F., Zhou, X., Miao, X., Zhang, T., Hang, X., Tie, R., Liu, N., Tian, F., Wang, F., and Yuan, J. (2014). MAGEC2, an epithelial-mesenchymal transition inducer, is associated with breast cancer metastasis. *Breast Cancer Research and Treatment*, 145(1):23–32.

[Yilmaz et al., 2012] Yilmaz, E., Campos, C., Fabius, A. W. M., Lu, C., Ward, P. S., Viale, A., Morris, L. G. T., Huse, J. T., and Mellinghoff, I. K. (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*, 483(7390):479–483.

[Yu et al., 2012] Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, 16(5):284–7.

[Yuan and Lin, 2006] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, (68):49–67.

[Zang et al., 2012] Zang, Z., Cutcutache, I., Poon, S., Zhang, S., McPherson, J., Tao, J., et al. (2012). Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nature Genetics*, (44):570–4.

[Zhang et al., 2012] Zhang, H.-M., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H., and Guo, A.-Y. (2012). AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic acids research*, 40(Database issue):D144–9.

[Zhang et al., 2011] Zhang, Y., Liu, H., Lv, J., Xiao, X., Zhu, J., Liu, X., Su, J., Li, X., Wu, Q., Wang, F., and Cui, Y. (2011). QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic acids research*, 39(9):e58.

Apprentissage de données génomiques multiples pour le diagnostic et pronostic du cancer

RÉSUMÉ : De nombreuses initiatives ont été mises en places pour charactériser d'un point de vue moléculaire de grandes cohortes de cancers à partir de diverses sources biologiques dans l'espoir de comprendre les altérations majeures impliquées durant la tumorogénèse. Les données mesurées incluent l'expression des gènes, les mutations et variations de copy-number, ainsi que des signaux épigénétiques tel que la méthylation de l'ADN. De grands consortium tels que "The Cancer Genome Atlas" (TCGA) ont déjà permis de rassembler plusieurs milliers d'échantillons cancéreux mis à la disposition du public. Nous contribuons dans cette thèse à analyser d'un point de vue mathématique les relations existant entre les différentes sources biologiques, valider et/ou généraliser des phénomènes biologiques à grande échelle par une analyse intégrative de données épigénétiques et génétiques.

En effet, nous avons montré dans un premier temps que la méthylation de l'ADN était un marqueur substitutif intéressant pour jauger du caractère clonal entre deux cellules et permettait ainsi de mettre en place un outil clinique des récurrences de cancer du sein plus précis et plus stable que les outils actuels, afin de permettre une meilleure prise en charge des patients.

D'autre part, nous avons dans un second temps permis de quantifier d'un point de vue statistique l'impact de la méthylation sur la transcription. Nous montrons l'importance d'incorporer des hypothèses biologiques afin de pallier au faible nombre d'échantillons par rapport aux nombre de variables.

Enfin, nous montrons l'existence d'un phénomène biologique lié à l'apparition d'un phénotype d'hyperméthylation dans plusieurs cancers. Pour cela, nous adaptons des méthodes de régression en utilisant la similarité entre les différentes tâches de prédictions afin d'obtenir des signatures génétiques communes prédictives du phénotypes plus précises.

En conclusion, nous montrons l'importance d'une collaboration biologique et statistique afin d'établir des méthodes adaptées aux problématiques actuelles en bioinformatique.

Mots clés : Apprentissage statistique, connaissances a priori, cancer, méthylation, diagnostic et pronostic, médecine personnalisée.

Learning from multiple genomic information in cancer for diagnosis and prognosis

ABSTRACT : Several initiatives have been launched recently to investigate the molecular characterisation of large cohorts of human cancers with various high-throughput technologies in order to understanding the major biological alterations related to tumorogenesis. The information measured include gene expression, mutations, copy-number variations, as well as epigenetic signals such as DNA methylation. Large consortiums such as "The Cancer Genome Atlas" (TCGA) have already gathered publicly thousands of cancerous and non-cancerous samples. We contribute in this thesis in the statistical analysis of the relationship between the different biological sources, the validation and/or large scale generalisation of biological phenomenon using an integrative analysis of genetic and epigenetic data.

Firstly, we show the role of DNA methylation as a surrogate biomarker of clonality between cells which would allow for a powerful clinical tool for to elaborate appropriate treatments for specific patients with breast cancer relapses.

In addition, we developed systematic statistical analyses to assess the significance of DNA methylation variations on gene expression regulation. We highlight the importance of adding prior knowledge to tackle the small number of samples in comparison with the number of variables. In return, we show the potential of bioinformatics to infer new interesting biological hypotheses.

Finally, we tackle the existence of the universal biological phenomenon related to the hypermethylator phenotype. Here, we adapt regression techniques using the similarity between the different prediction tasks to obtain robust genetic predictive signatures common to all cancers and that allow for a better prediction accuracy.

In conclusion, we highlight the importance of a biological and computational collaboration in order to establish appropriate methods to the current issues in bioinformatics that will in turn provide new biological insights.

Keywords : Machine learning, prior knowledge, cancer, methylation, diagnosis and prognosis, personalized medicine.

MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY