HAL
open science

Statistical methods for deciphering intra-tumor hereterogeneity : challenges and opportunities for cancer clinical management

Judith Abecassis

To cite this version:

Judith Abecassis. Statistical methods for deciphering intra-tumor hereterogeneity : challenges and opportunities for cancer clinical management. Bioinformatics [q-bio.QM]. Université Paris sciences et lettres, 2020. English. NNT : 2020UPSLM065 . tel-03228677

HAL Id: tel-03228677

https://pastel.hal.science/tel-03228677

Submitted on 18 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à MINES ParisTech

Statistical methods for deciphering intra-tumor hereterogeneity: challenges and opportunities for cancer clinical management

Méthodes statistiques pour la détection de l'hétérogénéité intra-tumorale : défis et pertinence pour la prise en charge du cancer

Soutenue par

Judith Abécassis

Le 9 Mars 2020

École doctorale n°621

Ingénierie des Systèmes, Matériaux, Mécanique, Énergétique

Spécialité

Bioinformatique

Composition du jury :

Stefan MICHIELS Group Leader, Institut Gustave Roussy	*President*
Florian MARKOWETZ Group Leader, Cancer Research UK Cambridge Institute	*Rapporteur*
Alessandra CARBONE Professeur, Sorbonne Université	*Rapporteur*
Donate WEGHORN Group Leader, Centre for Genomic Regulation	*Examinateur*
Fabien REYAL Group Leader, Institut Curie	*Directeur de thèse*
Jean-Philippe VERT Professeur, Mines ParisTech & Google Brain	*Directeur de thèse*

MINES ParisTech | PSL

Acknowledgments

Je pense que la réalisation de cette thèse a été un vrai défi pour moi, pour plusieurs raisons, certaines personnelles puisqu'entre le début de mon M2 et aujourd'hui, quelques changements radicaux ont eu lieu, et pour d'autres raisons plus professionnelles, comme le tiraillement perpétuel entre les mathématiques et la biologie, l'acceptation de l'incertitude inhérente à la recherche, mon adaptation à mes deux directeurs de thèse, et une certaine solitude face à l'ampleur de la tâche. Mais au milieu de cette longue nuit, j'ai eu la chance d'être soutenue et de découvrir de nombreuses personnes incroyables que je tiens à remercier ici.

First to the jury

Let me first thank Alessandra Carbone and Florian Markowetz who accepted to review my thesis, and of course to Donate Weghorn and Stefan Michiels who are part of this jury, sometimes with a long journey. I am very humbled to have this opportunity to discuss my PhD work with such prominent scientists, whose work has accompanied and inspired me during the past years.

To my great welcoming labs

Je tiens à remercier infiniment Jean-Philippe Vert et Fabien Reyal d'avoir accepté de m'encader pour la réalisation de cette thèse, chacun avec ses spécialités et ses méthodes de management. J'ai été particulièrement impressionnée par la bienveillance et la disponibilité de Jean-Philippe, toujours là au moment clé, capable de répondre avant la fin de la phrase où j'explique mon problème à la question que je n'ai pas encore posée ! Fabien m'a permis d'entrevoir l'autre côté de la recherche, celui des patients, où finalement seul le résultat compte.

Au-delà de ces deux figures marquantes, j'ai eu la chance d'être aussi soutenue et conseillée par toute une famille d'autres chercheurs à la fois au CBIO et au RT2Lab. Tout d'abord Véronique, pleine d'humour face aux hauts et aux bas des montagnes russes doctorales (en plus de toutes ses autres aventures), toujours une oreille attentive pour m'aider à mieux m'organiser et reprendre le dessus pour gravir la montagne. Thomas, toujours enthousiaste et disponible pour discuter de nouvelles idées. Chloé, qui scientifiquement et personnellement a été un exemple à suivre, et m'a encouragé sur de nombreux projets, parfois indirectement, mais en ouvrant des voies où il était alors facile de lui emboîter le pas. Anne-Sophie enfin, ma référence sur de très nombreux sujets, un savant mélange de pragmatisme efficace, de vision pour aller plus loin, et surtout une énergie bouillonnante qui redonne un petit élan juste en tournant la tête au bureau pour voir sa détermination.

All this very rich and diverse environment is of course completed by outstanding lab members, to support each others during harder times, have fun, and share great

tips and knowledge all the time. Chronologically, I was first very impressed when I joined the lab by all the PhD already there, with publications, relevant comments in group meetings, and strong specialties in the lab, with Nelle, the Python and open source software advocate, Alice and her inspiring passion for science, Elsa, always radiant and spreading encouraging words, Erwan, both earnest and joking around, Matahi, so eager to enjoy what life has to offer, Yunlong, organizer of all CBIO beers, and Nino of course, with a lot of hidden talents, and who also guided me through my first attempts of projects, with kindness and a lot of patience until I finally took her advice. And then was a younger generation, as brilliant of course, but maybe a little less intimidating for the newcomer I was, with Marine, awesome and funny office neighbor, Peter, maybe the one in CBIO that shared my viewpoint best, I hope we will finish our little common side-project soon (on it!), Benoit, always here for some philosophical considerations, Hector, a model of quiet strength, calmly working, making awesome science, while bringing fun and joy to the CBIO, along with Joe of course, always very relaxed, and Lotfi, always having some theory in his sleeves. And finally, a little renewal came, with Romain, half-provocative, half-on another planet, taking this whole PhD thing with some welcome perspective, Asma and Maguette, for more women in the team, and Antoine and Viven, the return of the postdocs, and Arthur, always up for new challenges, Tristan, Matthieu, looking forward to know more now that I will be out of the tunnel. The RT2Lab also had its share of great personalities, with so much fun at the beginning with Hélène, Cécile, Matahi and Anne-Sophie, and now some new awesome dynamic with the arrival of Eric, Elise, Beatriz and Nadir! And of course some great U932 colleagues, Ares, Wilfrid.

And of course, a huge thank to Dominique, Isabelle and Caroline, Pamela and Victoria, administrative superheroines, that have saved me from chaos so many times over the years.

Bien sûr ma famille

Tout d'abord, un immense merci à Louis, qui m'accompagne quotidiennement, supportant tous les aléas de la thèse, avec qui nous formons une super équipe pour continuer à garder du temps pour nos aventures néo-rurales, et plein d'autres à venir !! Mais y prenant aussi une part intégrante avec quelques aides techniques en Python, en terminal, en architecture efficace pour gagner du temps, en join et select pour remplacer mes affreuses boucles, et tous ses conseils que je n'ai pas assez suivis sur les tests.

Merci à ma mère, qui m'a transmis (je pense) sa force, son souci d'aller au fond des choses, sa rigueur, qui permettent de s'attaquer à tous les problèmes. Elle trouverait sûrement que j'aurais pu faire mieux, mais je finis par comprendre que c'est sa façon d'exprimer l'estime qu'elle a pour moi.

A mon père, lui aussi dédié à son travail, mais fin équilibriste pour être auprès de sa famille le plus souvent possible, en prenant sur lui beaucoup des difficultés rencontrées pour nous protéger.

A ma sœur Sarah, fidèle voisine d'école PSL, pour notre déjeuner presque hebdomadaire, nos discussions sincères, tous les efforts que tu as faits pour notre relation et que nous puissions compter l'une sur l'autre, ta patience pour écouter ce que je faisais, faire des figures pour moi, relire mon manuscrit, aider voire gérer la logistique de tout ce que j'essaie d'organiser, toujours partante pour me suivre dans des projets farfelus. Et Jonathan, que j'ai connu plus récemment, qui partage sa joie de vivre. A ma sœur Raphaëlle, ma première "challengeuse", qui m'a donné une habitude de

la remise en cause permanante de mes croyances, en même temps que des exemples inspirants d'autres façons de penser, d'autres valeurs, d'autres arts que le scolaire.

Merci à mes oncles et tantes, Nadine, Olivier, Frédéric, Laura, Alain, Martine, et mes oncles et tantes d'adoption, Agnès, Thierry, Michelle, Martine , à mes cousins Delphine, Amaury, Clémentine, Camille, Marianne, Daniel, Charlotte, Gautier, Emmanuel, Manon, Justine, Thomas, Clémentine pour leur soutien, et tous les moments loin de la thèse passés ensemble, des cousinades aux macaronnades, les petits clins d'œil RPZ qui me permet de toujours vous avoir à mes côtés. Et un merci particulier à Frédéric, Laura, Agnès, Thierry, Michelle, Emmanuel, Charlotte, et Justine, un peu plus académiciens, pour tous leurs conseils avisés sur le fonds, la forme, et le management.

A ma nouvelle famille d'adoption, Isabelle, Patrice, qui prennent presque autant soin de moi que de Louis, Clémence, Aurélien, pour leur soutien et nos échanges de bons procédés brico-jardin vs SQL, et enfin, Jean-Christophe et Séverine, qui m'ont accueillie à San Francisco sans m'avoir jamais rencontrée, qui réussissent à fabriquer du temps malgré leurs voyages et nombreux cours ; j'ai la chance d'avoir trouvé en Séverine un nouvel exemple de formidable chercheuse dont la vocation m'influence petit à petit.

Des supers amis

Je tiens tout d'abord à remercier Laure, une des premières à m'avoir encouragée à accepter et revendiquer ma personnalité "geek", Aurélia, qui a réussi à insuffler beaucoup de rires au lycée. Le groupe 13, Aurore et France, sans qui ça aurait été dur de traverser les colles de Bio, Sophie, inspiration de curiosité scientifique, Adèle, un exemple d'engagement complet, associé au courage de changer d'avis, et Vero, pétillante, Loulette, et son énergie magnétique. Et enfin, l'ENS, presque une maison, de nombreuses nouveautés avec le COF, les Pompoms, et le club œno, qui ne seraient rien sans les personnes avec qui j'ai pu faire tout ça, Aurore, meilleure coloc (ex-aequo avec Dylan), meilleure voisine du vert, co-panière, Noémie, mot d'ordre lentilles, légumes et sport, Veronica, pour son énergie entrainante et dansante, toujours là pour un peu de maths ou un peu de R, mais surtout beaucoup de blagues et de discussions, Clément, préfigurateur de mon amour de pandas, des soirées séries, Julien (et Claire), pour la suite des soirées séries, et toutes nos aventures berlinoises, Samuel, toujours là pour les moments graves, sérieux et rigolos, Constance, élégance et finesse, Arnaud et Léa, toujours dans le partage, de connaissances, bons repas et bons vins, Lauriane, Marylou et Danijela, pour la team femmes scientifiques, Jessica, Timothée, Arthur, Alexandre, pour leur amitié et leur soutien sans qui je n'aurai jamais fait le MVA, Célian, Antoine, Claire, Raphaël, Erwan, Victor, Nathanaël, Léo, Marc. Et depuis, Olivier et Itsuko, pour nos super discussions, toujours drôles, et toujours bien nourries. Et bien sûr, le club des sages d'Alice et Charlotte, pour un soutien avisé et collégial.

Après un apprentissage long

Merci à mes anciens professeurs qui m'ont menée jusque là, en particulier M. Korn, M. Palaquet, M. Limon, évidemment M. Boucekkine, qui a vraiment cru à mon projet maths-bio et a tout fait pour m'encourager, des feuilles d'exercices récréatives, au groupe de travail, et enfin m'a recommandé à David Bessis pour mon année de césure-stage chez tinyclues qui m'a permis de découvrir le monde du machine learning, guidée par David et Artem, qui ont fait le pari de tout m'apprendre alors que je ne partais de rien, à Mme Gazeau, qui m'a appris 95% de mon savoir biologique, je

n'ai fait que compléter quelques concepts par la suite, mais surtout une rigueur et une précision sans laquelle on ne peut pas faire de science, et qui me guide encore aujourd'hui dans mes travaux, M. Dejean de la Batie, qui m'a un peu réconciliée avec la physique, jamais avec le sens physique, mais qui m'a surtout transmis sa passion pour l'informatique et sa curiosité pour les nouvelles choses, et enfin Mme Tobailem, pour ses vérifications.

Open science

My PhD work would simply not have existed without bioRχiv, as the main articles my work relies on were published in journals in 2020, this faster circulation of ideas is full of promises. Thanks to GitHub for hosting all my code and documents, either to share them with the community or, with the academic plan, to just save them smartly and safely to structure my reflexion. And finally, I'd like to thank sci-hub, for its contribution to shareable science, and faster than any other access way.

Contents

Introduction 1
Preamble . . . 1
Organization and contributions of the thesis . . . 2

1 Elements of cancer genomics 3
1.1 Interpretation of genomic features . . . 5
1.1.1 Driver alterations . . . 5
1.1.2 ITH and cancer evolution . . . 6
1.1.2.1 Origin of ITH . . . 6
1.1.2.2 A few generalities on ITH inference . . . 6
1.1.2.3 Clinical implications of ITH . . . 7
1.1.3 Mutational signatures . . . 7
1.1.3.1 Relation with mutational processes . . . 7
1.1.3.2 Approaches for signature deconvolution in cancer genomes 10
1.1.3.3 Future challenges . . . 10
1.2 Specificities of sequencing for cancer research . . . 10
1.2.1 Overview of sequencing techniques . . . 11
1.2.2 Extraction of relevant features . . . 12
1.2.2.1 Variant calling . . . 13
1.2.2.2 Copy number and Structural variants . . . 14

2 Computational methods to unravel tumor evolution from genomic data 15
2.1 Overview of existing methods . . . 17
2.1.1 Selection of ITH methods . . . 17
2.1.2 ITH method features, and attribution strategies . . . 17
2.1.2.1 Input description . . . 18
2.1.2.2 Output description . . . 19
2.1.2.3 Preliminary algorithmic characterization . . . 21
2.1.3 Main classes of methods . . . 22
2.2 Challenges for method evaluation . . . 25
2.2.1 Different inputs, different outputs, different problems . . . 25
2.2.2 Choice of a benchmarking dataset . . . 25
2.2.2.1 Simulated data . . . 26
2.2.2.2 Real data . . . 27
2.2.3 Metrics . . . 27
2.2.4 Previous comparisons of ITH methods . . . 28
2.3 Open questions for ITH inference . . . 30
2.3.1 Directions for future developments . . . 30
2.3.2 Method evaluation . . . 31

3 Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data 33
3.1 Introduction . . . 36
3.2 Materials and methods . . . 37
3.2.1 Data . . . 37
3.2.2 Variant calling filtering . . . 37
3.2.3 ITH methods . . . 38
3.2.3.1 Published methods . . . 38
3.2.3.2 Consensus (CSR) . . . 38
3.2.4 Clinical variables . . . 38
3.2.5 Survival regression . . . 39
3.2.5.1 Model . . . 39
3.2.5.2 Evaluation procedure . . . 39
3.2.6 Immune signatures . . . 39
3.2.7 Correlations . . . 40
3.2.7.1 Comparison metrics . . . 40
3.2.8 WES and single cell paired dataset . . . 40
3.2.8.1 Data availability and preprocessing . . . 40
3.2.8.2 Evaluation metrics . . . 41
3.3 Results . . . 41
3.3.1 Assessing ITH on TCGA samples . . . 41
3.3.2 Methods quantifying ITH exhibit inconsistent results . . . 43
3.3.3 ITH is a weak and non robust prognosis factor . . . 47
3.3.4 ITH prognosis signal is redundant with other known factors . . . 48
3.4 Discussion . . . 48
3.4.1 Comparison to similar studies . . . 48
3.4.2 Can we truly measure ITH? . . . 49
3.4.3 Association with survival, link with other variables . . . 50
3.4.4 Can we build a gold standard dataset for benchmark? . . . 50

4 CloneSig: Joint Inference of intra-tumor heterogeneity and signature deconvolution in tumor bulk sequencing data 53
4.1 Introduction . . . 56
4.2 Results . . . 57
4.2.1 Joint estimation of ITH and mutational processes with CloneSig 57
4.2.2 Performance for subclonal reconstruction . . . 58
4.2.3 Performance for signature deconvolution . . . 61
4.2.4 Pan-cancer overview of signature changes . . . 62
4.2.5 Clinical relevance of ITH and signature changes . . . 65
4.3 Discussion . . . 66
4.3.1 Improved ITH and signature detection in WES . . . 67
4.3.2 Clinical relevance of signature variations . . . 68
4.3.3 Importance of input signatures and challenges . . . 69
4.4 Materials and methods . . . 69
4.4.1 CloneSig model . . . 69
4.4.2 Parameter estimation . . . 70
4.4.3 Test of mutational signature changes . . . 70
4.4.4 Simulations . . . 70
4.4.4.1 Default simulations . . . 71
4.4.4.2 Simulations for comparison with other ITH and signature methods . . . 71

4.4.4.3 Simulations without signature change between clones 71
4.4.4.4 Simulations to assess the separating power of CloneSig 72
4.4.4.5 Simulations to assess the sensitivity of the statistical test . . . 72
4.4.5 Evaluation metrics . . . 72
4.4.5.1 Metrics evaluating the subclonal decomposition . . . 72
4.4.5.2 Metrics evaluating the identification of mutational signatures . . . 73
4.4.6 Implementation . . . 73
4.4.7 Data . . . 74
4.4.8 Copy number calling and purity estimation . . . 74
4.4.9 Variant calling filtering . . . 74
4.4.10 Construction of a curated list of signatures associated with each cancer type . . . 74
4.4.11 Survival analysis . . . 75

5 Closing remarks **77**
5.1 Conclusion . . . 77
5.2 Perspectives . . . 78
5.2.1 How relevant is the number of clones to quantify tumor evolution? 78
5.2.2 The necessity to go beyond the TCGA . . . 79

References **81**

Appendix A Supplementary materials for the ITH methods comparison **101**

Appendix B Supplementary materials for CloneSig **129**
B.1 Supplementary methods . . . 129
B.1.1 EM algorithm for parameter estimation . . . 129
B.1.2 Selecting the number of clones . . . 132
B.1.3 Statistical test for signature change . . . 137
B.1.4 Several "modes" to run CloneSig . . . 140
B.2 Full benchmarking results . . . 141
B.3 Complete overview of TCGA results . . . 157

Appendix C Participation to the Dream challenge for ensemble variant calling **193**
C.1 Description of available data . . . 193
C.2 Materials and Methods . . . 194
C.2.1 Selection of pipelines . . . 194
C.2.2 Feature engineering . . . 194
C.2.3 Implemented algorithms . . . 194
C.3 Results . . . 195
C.4 Discussion . . . 196

Appendix D Supplementary for intra-tumor heterogeneity methods review **197**

List of Figures

1.1 Acquisition of ITH in the course of tumor growth and evolution. . . . 6
1.2 The 96 mutation types. The 16 possible mutation types of the substitution class C>A are shown as an example. source `https://commons.wikimedia.org/wiki/File:MutationTypes_v3.jpg` 8
1.3 The 67 SBS signatures of COSMIC 9

2.1 Schematic representation of the ITH reconstruction problem 18
2.2 Influence of the cancer cell population CNV at SNV locus, and tumor purity on CCF derivation 20
2.3 Typology of ITH methods 23
2.4 Evaluation of a new published method and comparison with existing ones. 26
2.5 Number of evaluations per method 27
2.6 Number of ITH methods published per year 30

3.1 Intersection of successful runs among the 4 considered ITH methods . 42
3.2 Runtime of the different ITH methods as a function of the number of mutations in each sample 43
3.3 Distribution of number of clones called by ITH methods with public and protected mutations sets as inputs 44
3.4 Correlation between various measures of ITH, and other potential confounding variables measured using WES and trancriptomics data . 45
3.5 Prognostic power of ITH measured using different ITH method and input mutation set combination 47
3.6 Prognostic power of ITH-derived features compared to other prognostic factors 48

4.1 CloneSig analysis of 246 SNVs obtained by WES of a sarcoma sample (patient TCGA-3B-A9HI) 58
4.2 Probabilistic graphical model for CloneSig 59
4.3 Comparison of CloneSig, TrackSig, Palimpsest, PyClone, SciClone and Ccube for subclonal reconstruction 60
4.4 Accuracy of CloneSig as a function of the difference in the CCF between the two clones, and of the cosine distance between their mutational profiles 62
4.5 Comparison of CloneSig, TrackSig, Palimpsest, and deconstructSigs for signature deconvolution 63
4.6 Mutational signature changes in the TCGA cohort 64
4.7 Kaplan-Meier curves for all TCGA patients stratified using CloneSig's results 66

4.8 Stratification of the TCGA patients by CloneSig in each cancer type 67

A.1 Prognostic power of diverse combination of ITH-derived features . . . 102
A.2 Pairwise computation of score1B for the different ITH methods and inputs . . . 105
A.3 Pairwise computation of score1C for the different ITH methods and inputs . . . 106
A.4 Pairwise computation of score2A for the different ITH methods and inputs . . . 107

B.1 Evolution of the loglikelihood and BIC criterion for 2 simulated samples, with the same parameters and 200 mutations and 2000 mutations 133
B.2 Variation of the degree of freedom of a subset of cancer type-specific signatures (35 distinct types) or for all available signatures depending on the number of signatures . . . 134
B.3 Test accuracy of various model selection criteria . . . 135
B.4 Number of clones found with different model selection criteria on the test set . . . 136
B.5 Empirical distribution of $-2\log(\lambda)$ . . . 137
B.6 Correlation of $-2\log(\lambda)$, with $\lambda = \frac{\ell_{sigCst}}{\ell_{sigChange}}$ with potentially relevant covariates . . . 138
B.7 Empirical distribution of the p-values of the calibrated test of significance of signature change for negative simulated samples . . . 139
B.8 Percentage of significant tests depending on the max distance between 2 clones . . . 139
B.9 Percentage of significant tests depending several variables . . . 139
B.10 CloneSig's performance for 3 different input signature strategies . . . 140
B.11 Score_1B for ITH methods on simulated data . . . 142
B.12 Score_2A for ITH methods on simulated data . . . 143
B.13 Score_2C (area under the curve) for ITH methods on simulated data 144
B.14 Score_2C (sensitivity) for ITH methods on simulated data . . . 145
B.15 Score_2C (specificity) for ITH methods on simulated data . . . 146
B.16 Score_sig_1B for signature deconvolution methods on simulated data 147
B.17 Score_sig_1C (accuracy) for signature deconvolution methods on simulated data . . . 148
B.18 Score_sig_1C (sensitivity) for signature deconvolution methods on simulated data . . . 149
B.19 Score_sig_1C (specificity) for signature deconvolution methods on simulated data . . . 150
B.20 Score_sig_1D for signature deconvolution methods on simulated data 151
B.21 Maximal cosine distance between the true and estimated mutation type profile for signature deconvolution methods on simulated data . 152
B.22 Median cosine distance between the true and estimated mutation type profile for signature deconvolution methods on simulated data . . . 153
B.23 Proportion of SNVs with cosine distance between the true and estimated mutation type profile under 0.05 for signature deconvolution methods on simulated data . . . 154
B.24 Proportion of SNVs with cosine distance between the true and estimated mutation type profile under 0.10 for signature deconvolution methods on simulated data . . . 155

B.25 Runtime for ITH reconstruction and signature deconvolution methods on simulated data . . . 156
B.26 Clonal and subclonal signature activities estimated by CloneSig in ACC 158
B.27 Clonal and subclonal signature activities estimated by CloneSig in BLCA . . . 159
B.28 Clonal and subclonal signature activities estimated by CloneSig in BRCA . . . 160
B.29 Clonal and subclonal signature activities estimated by CloneSig in CESC . . . 161
B.30 Clonal and subclonal signature activities estimated by CloneSig in CHOL . . . 162
B.31 Clonal and subclonal signature activities estimated by CloneSig in COADREAD . . . 163
B.32 Clonal and subclonal signature activities estimated by CloneSig in DLBC . . . 164
B.33 Clonal and subclonal signature activities estimated by CloneSig in ESCA . . . 165
B.34 Clonal and subclonal signature activities estimated by CloneSig in GBM . . . 166
B.35 Clonal and subclonal signature activities estimated by CloneSig in HNSC . . . 167
B.36 Clonal and subclonal signature activities estimated by CloneSig in KICH . . . 168
B.37 Clonal and subclonal signature activities estimated by CloneSig in KIRC . . . 169
B.38 Clonal and subclonal signature activities estimated by CloneSig in KIRP . . . 170
B.39 Clonal and subclonal signature activities estimated by CloneSig in LGG 171
B.40 Clonal and subclonal signature activities estimated by CloneSig in LIHC . . . 172
B.41 Clonal and subclonal signature activities estimated by CloneSig in LUAD . . . 173
B.42 Clonal and subclonal signature activities estimated by CloneSig in LUSC . . . 174
B.43 Clonal and subclonal signature activities estimated by CloneSig in MESO . . . 175
B.44 Clonal and subclonal signature activities estimated by CloneSig in OV 176
B.45 Clonal and subclonal signature activities estimated by CloneSig in PAAD . . . 177
B.46 Clonal and subclonal signature activities estimated by CloneSig in PCPG . . . 178
B.47 Clonal and subclonal signature activities estimated by CloneSig in PRAD . . . 179
B.48 Clonal and subclonal signature activities estimated by CloneSig in SARC . . . 180
B.49 Clonal and subclonal signature activities estimated by CloneSig in SKCM . . . 181
B.50 Clonal and subclonal signature activities estimated by CloneSig in STAD . . . 182

B.51 Clonal and subclonal signature activities estimated by CloneSig in TGCT . . . 183
B.52 Clonal and subclonal signature activities estimated by CloneSig in THCA . . . 184
B.53 Clonal and subclonal signature activities estimated by CloneSig in THYM . . . 185
B.54 Clonal and subclonal signature activities estimated by CloneSig in UCEC . . . 186
B.55 Clonal and subclonal signature activities estimated by CloneSig in UCS 187
B.56 Clonal and subclonal signature activities estimated by CloneSig in UVM . . . 188
B.57 An example of empirical distribution of the total copy number . . . 189
B.58 CloneSig's ability to distinguish 2 clones depending on the CCF distance between the two clones . . . 190
B.59 Kaplan-Meier curves for all TCGA samples (8625) distinguishing tumors only along the number of clones or along the number of clones and the presence of a significant change in signatures along tumor evolution using the public input mutation sets . . . 190

C.1 Precision and recall of different methods on synthetic datasets . . . 195
C.2 Results from the challenge leaderboard . . . 196

List of Tables

2.1 Sub-Challenges and participation to the ITH Dream Challenge 30

3.1 Main characteristics of ITH methods tested 42
3.2 Results on the single cell-WES dataset 47

A.2 Clinical variables significance for single-variable cox model for BLCA 104
A.1 Summary statistics of the number of protected and public mutations per sample 108
A.3 Clinical variables significance for single-variable cox model for BRCA 110
A.4 Clinical variables significance for single-variable cox model for HNSC 112
A.5 Signatures adapted from Bindea et al. [2013] 128

B.1 Values for the coefficients α for different penalty shapes and training subset 135
B.2 Values for the coefficient α for different penalty shapes and training subset 138
B.3 Characteristics of the TCGA cohort used in this study 191
B.4 Table of presence of signatures in the different cancer types 192

C.1 Characteristics of synthetic datasets from the DREAM Challenge . . 194

D.1 Characteristics of ITH methods 198

List of abbreviations

ACC Adrenocortical Carcinoma

AUC Area Under the Curve

BAF B Allele Frequency

BIC Bayesian Inference Criterion

BLCA Bladder Urothelial Carcinoma

bp base pair

BRCA Breast Invasive Carcinoma

CCF Cancer Cell Fraction

CESC Cervical Squamous cell Carcinoma and Endocervical Adenocarcinoma

CGH Comparative Genomic Hybridization

ChIP Chromatin Immunoprecipitation

CHOL Cholangiocarcinoma

CI Concordance Index

CNA/CNV Copy Number Alteration or Variation

COAD Colon Adenocarcinoma

COSMIC Catalogue of Somatic Mutations in Cancer

CRC Colorectal Cancer

DBS Doublet Base Substitution

ddNTP di-deoxyNucleotideTriPhosphates

DLBC Lymphoid Neoplasm Diffuse Large B-cell Lymphoma

DNA DeoxyriboNucleic Acid

dNTP deoxyNucleotideTriPhosphates

EM Expectation Maximization

ESCA Esophageal Carcinoma

FDR False Discovery Rate

GBM Glioblastoma multiforme

GDC Genomic Data Commons

HNSC Head and Neck Squamous cell Carcinoma

HR Hazard Ratio

HRD Homologous recombination Repair Defect

ICGC International Cancer Genome Consortium

ID Insertion and Deletion

IGV Integrative Cenomics Viewer

indel insertion or deletion

ITH Intra-Tumor Heterogeneity

KICH Kidney Chromophobe

KIRC Kidney Renal clear cell Carcinoma

KIRP Kidney Renal Papillary cell carcinoma

LDA Latent Dirichlet Allocation

LGG Brain Lower Grade Glioma

LIHC Liver Hepatocellular Carcinoma

LOH Loss of Heterozygosity

LUAD Lung Adenocarcinoma

LUSC Lung Squamous cell Carcinoma

MCMC Markov Chain Monte Carlo

MESO Mesothelioma

MMR Mismatch Repair

NGS Next Generation Sequencing

NMF Non negative Matrix Factorization

OV Ovarian serous cystadenocarcinoma

PAAD Pancreatic Adenocarcinoma

PCAWG Pan-Cancer Analysis of Whole Genomes

PCPG Pheochromocytoma and Paraganglioma

PCR Polymerase Chain Reaction

PRAD Prostate Adenocarcinoma

READ Rectum Adenocarcinoma

RNA RiboNucleic Acid

ROC Receiver Operating Characteristic

SARC Sarcoma

SBS Single Base Substitution

SKCM Skin Cutaneous Melanoma

SNP Single Nucleotide Polymorphism

SNV Single Nucleotide Variant

STAD Stomach Adenocarcinoma

SV Structural Variant

SVM Support Vector Machine

TCGA The Cancer Genome Atlas

TGCT Testicular Germ Cell Tumors

THCA Thyroid Carcinoma

THYM Thymoma

TNBC Triple Negative Breast Cancer

TSSB Tree-Structured Stick-Breaking

UCEC Uterine Corpus Endometrial Carcinoma

UCS Uterine Carcinosarcoma

UV UltraViolet

UVM Uveal Melanoma

VAF Variant Allele Frequency

VST Variance Stabilizing Transformation

WES Whole Exome Sequencing

WGS Whole Genome Sequencing

Introduction

Preamble

Cancer is a versatile disease, that denotes a large variety of situations, body localizations, degree of pathogenicity, treatment sensitivity, and outcomes. Until two decades ago, tumor characterization and treatment strategy only depended on clinical features such as tumor location, size, histology and grade, but progressively have included some molecular phenotypes in the framework of precision medicine.

This characterization is further complexified by intra-tumor heterogeneity (ITH). Acquisition of ITH is concomitant with tumor progression, as all cells in the human body acquire genomic alterations at each division, and tumor cells at an even faster rate. Most of those mutations do not impact cellular functions, but some may provide an advantage to their carrier, and lead to their progressive outgrowing of other tumor cells. This new subpopulation can entirely replace existing ones, or coexist with them, resulting in a composite structure. Recently, much work was dedicated to unravel the underlying population composition of tumors from the sequencing of one or several heterogeneous tumor samples. The accuracy of such reconstruction is critical for further clinical application.

The actual impact of the recently obtained catalog of cancer genome alterations on clinical practice is controversial [Kaiser, 2018]. Several hypotheses have been formulated to explain that setback, such as the existence of other involved mechanisms, namely gene expression regulation, epigenetic alterations, interaction with the tumor micro-environment, and ITH. Several processes could rely on this latter characteristic, such as treatment resistance, metastatic ability or immune system escapement. Indeed, this evolutionary framework allows researchers to consider and model not only tumor characteristics at the time of diagnosis, but also consider the full history of tumorigenesis, and its potential implications for future evolution, that can lead to better-suited therapeutic strategies.

The contributions of this thesis lie in the fields of computational methods conceived to estimate and describe genomic ITH from high throughput sequencing data. We propose a broad overview of existing methods developed to solve this problem, that highlights the associated computational challenges. A second aspect of this work focuses on the problem of evaluating existing (and to be developed) methods, which is a difficult question, as the truth is hidden and experimentally challenging to measure. Finally, with the development of a new approach, CloneSig, we illustrate the opportunity to integrate several aspects of tumor evolution in the inference for improved performance, and exploration of potentially informative patterns.

Organization and contributions of the thesis

We describe here the organization of the thesis, and our contributions to the field of intra-tumor heterogeneity deconvolution, and the offered perspectives for clinical application.

Chapter 1 introduces the intra-tumor heterogeneity inference problem, and technical knowledge and concepts regarding cancer genomics and sequencing methods, necessary to a good understanding of the thesis contributions.

Chapter 2 proposes a very broad overview of existing methods for ITH deconvolution. We review to the best of our knowledge the main classes of existing methods, the associated algorithms and explore their limitations, and other challenges of the field. These constitute important motivations for the work presented in the subsequent chapters.

Chapter 3 and Appendix A presents an attempt at evaluating existing ITH methods on real data, and their potential for clinical applications. We highlight the high dependency of results on preceding preprocessing steps, and their lack of stability and consistency in situations of noisy experimental settings.

Chapter 4 and Appendix B describes a new approach for ITH inference, to unravel the evolutionary history of genome alterations acquisition, and additionally reconstitute the past dynamic of mutational processes associated with those events. This constitutes a step towards a more functional description of tumor evolution.

Appendix C briefly outlines results obtained for the DREAM Somatic Mutation Calling Meta-pipeline Challenge, focusing on consensus approaches for variant calling, a crucial pre-processing step for ITH deconvolution.

Chapter 1

Elements of cancer genomics

Abstract

This chapter introduces key concepts of cancer genomics that are useful for the reader's understanding of the work presented in this thesis and its motivations. Cancer is characterized by abnormal cells dividing without proper control, and eventually forming masses throughout the body. A large majority of cancers is associated with genetic alterations, and a better understanding of their exact nature and role in disease development holds promising applications for treatment in the framework of precision medicine. We first present current approaches developed to describe a given tumor, by delineating the role of specific alterations, and deciphering the history of a tumor using its genome. We then outline more technical details on the sequencing approaches that allow researchers and physicians to obtain genomic data, as the methods we will describe in this thesis aim at exploiting the data's underlying structure to recapitulate as accurately as possible the unobserved past events of the tumor development.

Résumé

Ce chapitre rappelle le contexte de cette thèse en fournissant au lecteur les connaissances nécessaires en génomique oncologique pour appréhender le travail présenté dans cette thèse et ses enjeux. Le cancer est une maladie caractérisée par la présence de cellules anormales proliférant sans contrôle adéquat, et formant finalement des masses à différents endroits de l'organisme. La plupart des cancers ont pour origine des anomalies génétiques, et une meilleure compréhension de leur nature exacte et de leur rôle dans le développement de la maladie offre des perspectives prometteuses pour la mise au point de traitements personnalisés plus efficaces. Nous présenterons dans un premier temps les approches existantes pour décrire une tumeur, en déterminant l'influence spécifique de chaque mutation, et en retraçant l'histoire de la tumeur à partir de son génome. Nous nous attacherons ensuite à décrire plus en détails les techniques de séquençage qui permettent aux chercheurs et aux médecins d'obtenir ces données génomiques, dans la mesure où les méthodes que nous étudierons dans cette thèse sont conçues pour exploiter au mieux la structure sous-jacente de ces données pour accéder aux étapes passées du développement de la tumeur, qui ne sont plus observables au moment du diagnostique.

Contents

1.1 Interpretation of genomic features **5**
1.1.1 Driver alterations 5
1.1.2 ITH and cancer evolution 6
1.1.2.1 Origin of ITH 6
1.1.2.2 A few generalities on ITH inference 6
1.1.2.3 Clinical implications of ITH 7
1.1.3 Mutational signatures 7
1.1.3.1 Relation with mutational processes 7
1.1.3.2 Approaches for signature deconvolution in cancer genomes 10
1.1.3.3 Future challenges 10
1.2 Specificities of sequencing for cancer research **10**
1.2.1 Overview of sequencing techniques 11
1.2.2 Extraction of relevant features 12
1.2.2.1 Variant calling 13
1.2.2.2 Copy number and Structural variants 14

Association of cancer with genome alterations, and the potential to design efficient drugs to target the subsequent dysfunctions have been uncovered before any genome sequencing was performed. A frequent translocation in leukemia, the Philadelphia chromosome was detected by microscopic observation of samples with a particular chromosomal preparation in 1960 [Hungerford and Nowell, 1960; Rowley et al., 1977; Rowley, 1973; Larson et al., 1984], and targeted therapies were designed to inhibit the ectopic expression of the protein resulting from the fusion with a tyrosine kinase inhibitor [Druker et al., 1996; Mardis, 2018]. Nonetheless, generalization of this rationale has not met the expected success for a number of reasons, including difficulties for target identification and drug development, but also resistance to treatment, either out of hands, or after a few months of therapy. This work will focus on a transverse phenomenon, tumor evolution, which impacts the different aforementioned aspects, and in particular the elucidation of the genomic characteristics of a tumor, and the acquisition of functional properties. In this chapter, we will explain in details the concepts enabling us to interpret and potentially exploit genomic alterations in cancer, as well as the experimental and computational tools available to measure them.

1.1 Interpretation of genomic features

Over the last decades, DNA sequencing has become more and more efficient, allowing us to deepen our descriptive knowledge of thousands of alterations typically present in a cancer genome. However, detecting the key events that truly influence cancer development is more difficult than it appears.

1.1.1 Driver alterations

Despite the existence of numerous DNA repair mechanisms, a few to a dozen somatic mutations accumulate in cells every year [Werner and Sottoriva, 2018]. In cancer cells, genomic instability leads to an increased mutation rate. A large body of evidence, ranging from mathematical models based on cancer incidence by age [Ashley, 1969], to animal models recapitulating specific forms of cancer [Böck et al., 2014], suggests that only a few key mutations, called "hits" or "drivers" are instrumental in fostering carcinogenesis [Reiter et al., 2019]. This hypothesis has been formalized within the framework of the "hallmarks of cancer", that identify crucial cellular functions that need to be impaired during the development of the disease [Hanahan and Weinberg, 2011]. Identification of driver mutations is an important step towards precision medicine for cancer diagnosis, monitoring and treatment, and several rationales have been developed for that purpose:

- Identification of recurrent alterations, as repeatedly altered genes can be read into being under positive selection for cancer development. The number of alterations of a gene should be corrected for gene length, and background mutation rate [Brown et al., 2019].
- Prediction of functional impact of an alteration, based on structural data of the protein coded by the gene, known regulation sequences or conservation across species information [Tang and Thomas, 2016].
- Pathway analysis to enrich genetic data. Indeed, as briefly mentioned earlier, driver mutations affect key functions in the cell, and are generally part of biological pathways constituted of several genes involved in a given function. We can use this underlying structure to better detect driver genes in several ways: for instance, exclusion patterns can be the sign of driver genes belonging to the same pathway, and hence not mutated together in the same cells, as the second mutation does not provide an additional change [Szczurek and Beerenwinkel, 2014]. Another potential use of known pathways is to consider mutations at the pathway level instead of the gene level to increase the statistical power of detection of drivers [Hofree et al., 2013; Le Morvan et al., 2016].

Recent analyses highlight that a mutation that can be a driver in some tumors is not necessarily one in every context [Reiter et al., 2019; Martincorena et al., 2018]. A way to refine

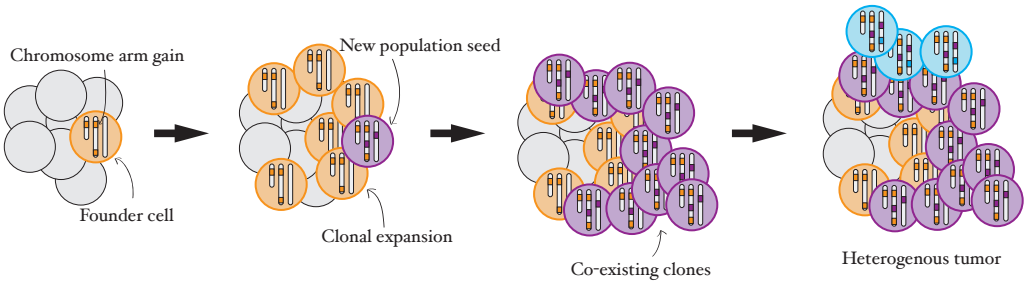

Figure 1.1 – Acquisition of ITH in the course of tumor growth and evolution. A first set of genomic alterations initiate the tumor, which further develops (orange cells) and accumulates more alterations. The acquisition of the purple alterations seeds a new population that develops and coexists along the original population. The process repeats further with the apparition of the blue population.

the detection of driver events lies in the evolutionary history of a tumor. Deconvolution approaches allow us to distinguish early from late events, and more generally retrace mutations order. This order could be indicative of the actual contribution of each alteration. More in-depth analyses involving precise measurements of ITH can also unravel complex patterns such as exclusion patterns or convergent evolution but this time at the patient level (e.g. between distinct independent parts of the same heterogeneous tumor) rather than at the cohort level.

1.1.2 ITH and cancer evolution

1.1.2.1 Origin of ITH

In the 1970s, evolution has appeared as a new framework to study cancer and potential treatments [Nowell, 1976]. Indeed, cancer cells often exhibit genome instability, and accumulate alterations faster than normal cells, so that each cancer cell genome is unique. The alterations that are specific to a cell are called private, and are typically undetectable by bulk sequencing. When new alterations further accumulate, either due to selection or genetic drift, a tumor cell can undergo clonal expansion, and its descendants then represent an increasing proportion of the total tumor population, until they overcome the whole population, or coexist along it, leading to a mosaic structure. Indeed, cancer cells are constrained by their environment for survival, through competition with normal and other tumor cells for access to resources, and the necessity to evade the organism control systems that aim at maintaining tissue homeostasis (e.g. growth factors regulation, cell death, immune system). As a result of clonal expansions, all the descendant cells share the genomic alterations carried by their last common ancestor, which has two consequences: those alterations reach the threshold of detection, and as they are all shared together by this group of cells, they will be detected at similar frequencies by sequencing. This hypothesis is the cornerstone of the ITH methods considered throughout this work. This mechanism is illustrated schematically in Figure 1.1.

1.1.2.2 A few generalities on ITH inference

One can consider several distinct manifestations of ITH:

Functional heterogeneity i.e. all tumor cells may not express the same genes, that pathologists can observe directly on patient samples with immunohistochemistry, through expression or epigenetic assays. Some studies have attempted to link it to genetic heterogeneity [Wen et al., 2016; Park et al., 2016; Kim et al., 2019]

Genetic heterogeneity i.e. all tumor cells have different genomes.

In both cases, ITH can be detected using bulk genomic or functional measurements, that then necessitate to be deconvoluted to identify the different components of the mixture, or using recent approaches for single-cell measurements [Chung et al., 2017; Min et al., 2015;

Navin et al., 2011], that offer a simplification of the deconvolution step. A variety of intermediate settings allow researchers to explore ITH at different resolutions. Experimentally, ITH can be evidenced directly by sequencing of multiple samples (spatially or temporally separated) from the same tumor, but such approaches can be costly (even though sequencing costs keep decreasing), and invasive for the patients. Moreover, those observations can be confounded if samples are not homogeneous. For example in the case of genetic ITH, it can create an illusion that the subclonal mutations present in all samples are actually clonal. Hence, deconvolution is a necessary step for all ITH analysis if the sample is larger than a single cell [Alves et al., 2017]. Unless otherwise stated, heterogeneity will refer to genetic heterogeneity, which is the main focus of this thesis.

We can loosely define the problem of reconstructing ITH as identifying the number and genotypes of the main tumor populations, and infer their phylogenetic relationships. Two main objectives of ITH reconstruction are: (i) to assess whether a tumor is homogeneous or composed of several (detectable) subpopulations with distinct genomes, (ii) to reconstruct the evolutionary relationships between the identified such populations. Without going further into technical details that will be covered in Chapter 2, the main idea behind the first problem is to go beyond the simple detection of somatic alterations from sequencing data, and infer the proportion of cells in the sample that carry the detected alterations, and then try to group them into the correct number of mixture components. This is challenging because several parameters have to be taken into account to go from raw data to a proportion of cells, and measurements are noisy, making the grouping step more difficult. For the second problem, a few intuitive principles are applied to infer phylogenetic relationships between the identified alterations: the pigeonhole principle, or sum-rule, that states that if the sum of the clonal frequencies of two alterations is larger than 100%, at least one cell must have contained both alterations, and the infinite-site assumption, that each alteration occurs only once in the evolutionary history of the tumor, and can not be reversed. Those two principles provide some constraints that allow us to reconstruct potential evolutionary paths from the sequencing of a tumor sample.

1.1.2.3 Clinical implications of ITH

ITH can be leveraged to answer several important questions, including:

Unravel cancer early stages those stages are not directly observed. ITH inference can complete our knowledge of driver events by refining their order of apparition and assess their importance. Better knowledge of tumor evolution patterns and time of growth can have important consequences on future strategies for cancer prevention and screening [Sottoriva et al., 2015a; Fittall and Van Loo, 2019; Dentro et al., 2018].

Inform cancer treatment , as besides identifying driver events to target, reconstructing ITH can be helpful in selecting treatments that could reach all tumor cells, and not only a subset carrying the mutation of interest.

Risk stratification is also an important facet of cancer management, and there are indications that ITH can be a prognostic marker of future malignancy, both from pre-malignant as illustrated on predicting evolution to adenocarcinoma from the Barrett's esophagus [Maley et al., 2006; Martinez et al., 2016], or suggested by the predictability of cancer evolution [Hosseini et al., 2019].

Ideas for new chronic disease management, inspired from what we know about species evolution and population genetics to propose entirely new treatment strategies that are not based on killing cancer cells, but lead the tumor to a stage where it will become extinct by itself, or remain quiescent at a low size [Gatenby et al., 2019].

1.1.3 Mutational signatures

1.1.3.1 Relation with mutational processes

During an individual's lifetime, several processes can cause somatic mutations. Some of these processes are endogenous and inherent to the cellular functions, others are exogenous, such

as exposition to carcinogens. The patterns of exposure can also vary, with lifelong exposure in the case of mutations caused by ageing, or later onset. In that latter situation, a specific mutational activity can either be transient in the case of exposure to an exogenous substance, or permanently acquired, in the case of the advent of an endogenous mechanism like DNA repair defects induced by mutations, though such cases can also be reversible in theory. One can define mutation types to better account for the genomic context of SNVs: six substitutions types (with accounting for reverse complements) and 4 possible 3' and 5' flanking nucleotides, resulting in 96 possibilities, as illustrated in Figure 1.2. There is strong evidence that several mutational processes have different probabilities to produce those 96 mutation types. Some of those particular patterns have been first experimentally observed several decades ago in several particular cases, with the discovery of 1958 of the mechanism by which UV exposition damages DNA [Rörsch et al., 1958], and in 1980 of the spontaneous deamination of methylcytosines at CpG dinucleotides as mutagene [Duncan and Miller, 1980].

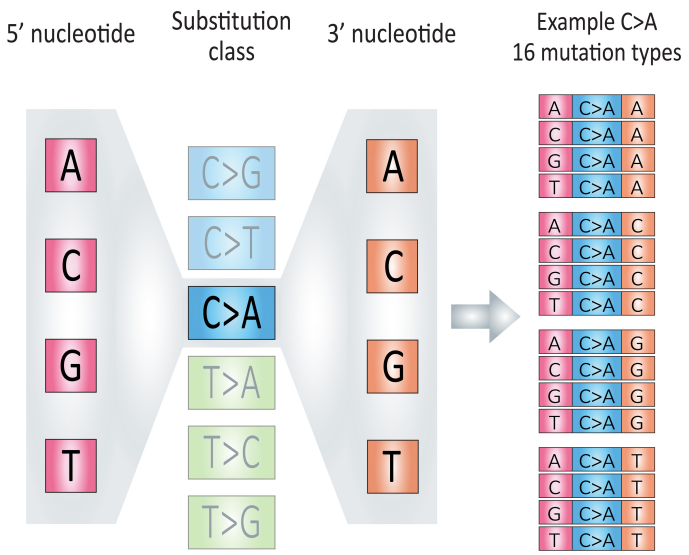

Figure 1.2 – The 96 mutation types. The 16 possible mutation types of the substitution class C>A are shown as an example. source `https://commons.wikimedia.org/wiki/File:MutationTypes_v3.jpg`

First analyses of mutational patterns in cancer genomes in the 1980s were limited to well characterized and famous proteins like p53 [Hollstein et al., 2016], but the increasing availability of cancer genomic sequences allowed Stratton and his team to formalize the concept of mutational signatures [Nik-Zainal et al., 2012] and Alexandrov to propose a first algorithm for their systematic identification and quantification in cancer genomes using non-negative matrix factorization (NMF) on large cohorts (several thousands) of sequenced cancer genomes [Alexandrov et al., 2013]. A mutational signature can be formally defined as a discrete probability distribution over the 96 mutation types. The concept has been further extended to a more refined typology of mutations based on pentanucleotides [Shiraishi et al., 2015; Alexandrov et al., 2018], and also to small insertions and deletions (indels) [Alexandrov et al., 2018], larger structural variants [Nik-Zainal et al., 2016; Macintyre et al., 2018]. A stabilized catalog of signatures is maintained by COSMIC [Forbes et al., 2017; Tate et al., 2019], and was recently updated to include 67 single nucleotide substitutions (SBS) signatures, 11 doublet base substitution (DBS) signatures and 17 small insertion and deletion (ID) signatures. SBS signatures are shown in Figure 1.3.

The underlying assumption of mutational signatures is that each signature represents a mutational process. This principle has been further investigated, both by comparison with known mechanisms (UV, spontaneous C>T mutations by deamination etc), and experimentally, with a first proof of concept illustrating the fact that different genetic alterations in DNA repair pathways induce distinct mutation profiles [Zou et al., 2018b], and systematic

Figure 1.3 – The 67 SBS signatures of COSMIC. Each panel represents a signature extracted by NMF in Alexandrov et al. [2018].

characterization of the effect of known chemical products [Kucab et al., 2019]. Association of signatures to mutational processes is done by confrontation with experimental data, or statistical association of patients' clinical data with signatures. Currently, most signatures are of unknown aetiology. Well-described signature-process associations include ageing (SBS1, SBS5), APOBEC activity (SBS2, SBS13), exposition to Aristolochic acid (SBS22), or to Aflatoxin (SBS24), DNA mismatch repair (MMR) defect (SBS6, 15, 20, 26), homologous recombination (HR) DNA repair defect (SBS3), tobacco smoking (SBS4) or chewing (SBS29), UV exposure (SBS7), somatic hypermutation in lymphoid cells (SBS9), Polymerase epsilon exonuclease domain mutations (SBS10) [Alexandrov et al., 2018].

1.1.3.2 Approaches for signature deconvolution in cancer genomes

There are two ways to approach the detection of mutational signatures in a tumor:

De novo discovery that considers an entire cohort of cancer genomes and extracts mutational signatures without a priori. Several approaches have been implemented: non-negative matrix factorization (NMF) as the first approach, or NMF variants with incorporation of a LASSO penalty to enforce the sparsity of signatures [Ramazzotti et al., 2018] or Bayesian variants, as well as other probabilistic approaches based on Latent Dirichlet Allocation (LDA) [Shiraishi et al., 2015; Matsutani et al., 2019]. Baez-Ortega and Gori [2019] provide a complete mathematical review of those approaches, and a benchmark can be found in Omichessan et al. [2018]. The difficulty lies in the fact that there is no guarantee that each signature corresponds to one mutational mechanism.

Signature refitting is the only applicable approach in the case of small cohorts, and consists in finding the proportions of known signatures in a new sample. We can use the reference signatures from the COSMIC database as input signatures. Common recent approaches rely on linear regression [Rosenthal et al., 2016], quadratic decomposition, or Bayesian approaches [Rubanova et al., 2018], and have also been benchmarked in Omichessan et al. [2018].

1.1.3.3 Future challenges

Mutational signatures provide a unified concept to approach the causes of mutations, and deciphering such processes has promising applications in cancer prevention and patient stratification [Fittall and Van Loo, 2019]. But despite those promises, several open questions remain to be addressed. The consensus around signatures and how to obtain them is still fragile, especially when considering the most recently defined signature for doublet substitutions, indels, and structural variants. A lot of signatures have no known associated mutational process, though systematic screenings are ongoing, both for chemical compounds [Kucab et al., 2019] and cancer drugs [Pich et al., 2019]. Moreover, interactions with the genetic background and intrinsic signature variability between individuals are also considered [Volkova et al., 2019].

Accurate detection and quantification of signature activity is also far from solved with issues of identifiability [Maura et al., 2019; Robinson et al., 2019]. The clinical implications of signature deconvolution for cancer prevention, patient stratification, and therapeutic strategies also remain to be explored. This last question could require to measure the variations of signature activities over the development of the tumor to further unravel the driver forces of carcinogenesis.

1.2 Specificities of sequencing for cancer research

In the previous section we have presented the challenges and some of the main thematics of current research in cancer genomics. All those applications rely on data extracted from the DNA sequencing of tumor samples, and are tailored to its technical particularities. In the rest of this chapter, we will cover in more details the revolutionary advances in sequencing technologies (both in cost and throughput) that have enabled tremendous progress in cancer genomics.

The first two sequencing techniques were described on the same year, with Maxam and Gilbert's chemical chain termination method for DNA sequencing [Maxam and Gilbert, 1977] and the dideoxy method by Sanger et al. [1977], allowing researchers to obtain the first complete human genome sequence in 2001 [Lander et al., 2001; Craig Venter et al., 2001]. This first attempt has required huge investments in human time (13 years) and money (3 billion dollars), but since then, the cost of genome sequencing has dramatically dropped, supporting a broad use in research in various domains. In Europe, 21 countries have committed to transnationally share one million human genomes by 2022 [Saunders et al., 2019].

Sequencing is involved in many aspects of cancer research, which is reflected in the variety of sample preparation techniques and sequencing methods designed to observe the complexity of cancer. A few of those numerous applications are the identification of hereditary risk factors, the identification of genomic (driver) alterations, either at the RNA or the DNA level, useful to isolate potential druggable targets or for patient stratification, and for the reconstruction of the tumor evolution process.

In each case, the sequencing strategy has to be adapted to the desired level of observation. A first specification is the choice of the input biological material:

- a piece of tissue from a biopsy or surgical resection, for a broad but unresolved overview,
- multiple samples from the same patient to improve spatial or temporal resolution,
- single cell sequencing, for easier deconvolution, or
- circulating tumor cells, for non-invasive tumor sampling.

Depending on the biological question to explore, several sequencing settings are available,

- Whole genome sequencing (WGS).
- Whole exome sequencing (WES), where only the protein coding sequences (exons) are captured and sequenced, with the rationale that this will cover the genomic alterations that are the most likely to be involved in the cancer, but only 1 to 2% to the DNA amount.
- Targeted sequencing, either on a selection few dozens to hundreds of gene exons are capture and sequenced, often chosen among known cancer driver genes, or on regions of interest to confirm suspected point mutations (from WES or WGS).
- RNA sequencing, where RNA molecules are captures, and transcribed back to DNA for sequencing.
- More complex settings, like ChIP-seq (Chromatin ImmunoPrecipitation sequencing) to analyze protein interactions with DNA, Atac seq to detect regions of open chromatin, Hi-C to capture genome conformation, bisulfite sequencing to assess DNA methylation.

Finally, the cost is determined by the total amount of sequenced DNA, which depends on the aforementioned total size of the sequenced region, and of the sequencing depth, which can be loosely defined as n, the average number of times each position of the target region is covered by a read, and is denoted nX. The depth typically depends on the objective of the study: to detect subclonal variants, tumor samples are typically sequenced at a depth of 100X for WES, and at least 30X for WGS; the matched normal sample requires a more modest coverage. Targeted sequencing is typically used with a sequencing depth of 500 to 10000X. However, for copy number profile only, under 10X WGS can be used [Raman et al., 2019; Griffith et al., 2015].

1.2.1 Overview of sequencing techniques

A large number of sequencing technologies have been developed over the last two decades, and have been extensively reviewed [Goodwin et al., 2016; Heather and Chain, 2016; van Dijk et al., 2018; Mardis, 2017]. The characteristics of the sequencing have important implications on the genome features that can be detected, and on the specificities of the involved computational pipelines, so we will briefly describe the sequencing techniques landscape. Sanger

sequencing can deal with sequences up to 1000 base pairs (bp) with an accuracy as high as 99.999% [Shendure and Ji, 2008], and relies on a complex setting where the polymerization reaction that elongates DNA is supplied di-deoxynucleotidetriphosphates (ddNTPs) instead of regular deoxynucleotidetriphosphates (dNTPs). The incorporation of a ddNTP prevents further elongation, and the resulting DNA molecules are then separated according to their molecular weight (and hence length) by electrophoresis. The separating power by one nucleotide limits the sequenced length.

The augmentation of sequencing throughput relies on several aspects

- the sequencing of many identical molecules at the same time (after PCR amplification) for robust signal detection,
- the parallelization of the sequencing of many DNA templates in parallel, typically by resorting to spacial resolution.

All approaches proceed with a first step of sequence amplification, and then a second step with the actual sequencing. The 3 main short read sequencing platforms have different solutions for the first step, with either on-bead amplification (454, SOLiD, GeneReader (Quiagen), Ion Torrent) that are then spread on a glass surface or on a plate with wells for the spatial separation, or amplification on a solid phase like Illumina bridge amplification or SOLiD Wildfire template walking. In the second step, ligation sequencing proceeds by successive ligation of fluorescent oligonucleotides with inserted shifts to cover the sequence, and deconvolution of the resulting signal (SOLiD). For sequencing-by-synthesis with cyclic reversible termination (Illumina, Qiagen), similarly to the principle of Sanger sequencing, the incorporation of marked nucleotides with a blocked 3' extremity, preventing elongation. Once the base is identified, the marker and the blocking extremity are removed, and the operation is repeated with the following position. Finally for sequencing-by-synthesis with single-nucleotide addition (454, Ion Torrent), for each position, each of the four possible bases are sequentially added and washed, and their incorporation is detected by the detection of the pyrophosphate molecule that is emitted when DNA elongation occurs. In the case of several identical residues, the quantity of pyrophosphate molecules is estimated.

In all those variations of short-read sequencing, the synchronization of the sequencing of similar sequences is key to reading the signal, and limits the read length. Illumina sequencing offers the longer reads and the higher throughput and now dominates the market, though SOLiD sequencing provides a much lower error rate, comparable to Sanger sequencing [Goodwin et al., 2016].

Short read sequencing has proven instrumental in recent advances in genomics, however, it has several major drawbacks: several types of structural events, in particular involving repetitions can not be resolved, point variations or small indels can not be phased in genome sequencing, and splicing isoforms can not be easily identified in RNA sequencing. Long reads can address those limitations, and constitute the third generation sequencing techniques. Two rationales exist: either truly sequence a single molecule for lengths up to several kilobases, or alter the library preparation step to barcode small sequences originating from the same molecule, then use standard short read sequencing and recover the longer molecule during an extra assembling step. They constitute a promising alternative for higher quality genomic data, but remain expensive compared to short read sequencing, and are still limited in throughput.

1.2.2 Extraction of relevant features

From there we will focus on the problem of reconstructing the tumor history from sequencing data, mostly bulk, but a few remarks will outline the main differences with single cell sequencing. Hence, only genome sequencing approaches will be considered, and alternative signals such as transcriptome, methylome, and epigenome will not be covered, though they have central implications in cancer. The first step towards exploiting genomic data consists in converting the raw sequencing signal to the FASTQ format, containing all the sequenced reads (usually one to several hundreds of billions depending on depth and coverage) and the associated quality. Those short reads are then mapped to the reference genome [Li and

Durbin, 2010; Langmead et al., 2009], and other quality control steps can be taken such as marking or removal or duplicate reads (potential artifacts from the PCR amplification step), filter low mapping quality reads [Broad, 2019], base quality recalibration and realignment around indels [McKenna et al., 2010]. The following steps of analysis aim at detecting small alterations (single nucleotide variants, SNVs and small insertions and deletions, indels), the genome copy number profile, and larger structural variants.

1.2.2.1 Variant calling

Variant calling is a crucial step for subsequent analyses as this determines the list of detected genomic variants in a tumor sample. This is challenging because some variants are barely above the noise level of sequencing data. Several factors contribute to this fact: variants can occur in a small subset of the sample due to ITH and normal contamination, artefactual variations can be generated by polymerases during the amplification step or occur during sequencing itself, some genomic positions are less covered, typically in the GC-poor or GC-rich regions of the genome, some genomic regions are difficult to map (repeated sequences) [Lander and Waterman, 1988], capture and alignment are both biased toward the reference and can also lead to impaired mutation detection.

Over 40 variant callers have been developed in the past decade [Xu, 2018]; we will focus on the ones dedicated to somatic variant calling from a matched tumor-normal pair of sequenced samples. Most variant callers implement a position-based strategy, in which a statistical approach determines the situation best explaining the presence of variant reads. The complexity of the underlying models ranges from simple Fisher's exact test on the 2×2 contingency table of read counts in VarScan2 [Koboldt et al., 2012], or VarDict [Lai et al., 2016] to complex models of normal and tumor allele frequencies accounting for potential subclonal variants in non-diploid regions in MuSE [Fan et al., 2016], or deepSNV [Gerstung et al., 2012]. Sequencing errors are also accounted for differently with either a single threshold [Koboldt et al., 2012], a site-specific estimate [Gerstung et al., 2012], or a sample specific rate, depending on the sequencing depth and contamination of each sample [Fan et al., 2016]. The most recent approaches use a classification model that can incorporate complex features such as strand bias, position along the read, base quality score directly in the model as in SNooPer [Spinella et al., 2016] or DeepVariant [Poplin et al., 2018], and can even additionally aggregate the calls of a group of variant callers to obtain more robust variants, as implemeted in SomaticSeq [Fang et al., 2015], of NeoMutate [Anzar et al., 2019]. A second class of algorithms resort to an haplotype-based strategy like Mutect2 [Cibulskis et al., 2013] in which reads are locally assembled to form candidates haplotypes, that are then confronted to read counts to estimate the likelihood of each haplotype. This strategy is advantageous in regions with poor mapping due to clustered SNVs or indels. More details can be found in the complete review of Xu [2018].

However, measuring the performances of each variant caller remains challenging, as simulated data often fail to fully reproduce sequencing biases, and real data benchmarks are not well suited to evaluate false negative calls. The application of any two variant callers to the same sample will provide discordant outputs in most cases. In practice, researchers often resort to additional ad-hoc filtering strategies to restrain variants, or even manual inspection in IGV, Integrative Genomics Viewer, a visualization tool for aligned reads [Robinson et al., 2017].

Some of the point detection variant callers also include small insertions and deletions, i.e. of a few base pairs size. Other dedicated tools have also been developed. The principle is the same: statistical models comparing the count of reads with or without the alteration in the tumor sample compared to the normal sample. However, this is more challenging as reads including indels are harder to align to the reference genome, so the error rate of indel calling is much higher than for single nucleotide variants, and pre-processing steps like realignment around indels are important.

1.2.2.2 Copy number and Structural variants

Copy number alterations or variations (CNA, CNV) were the first ones detected through the direct observation of karyotypes, and have been associated to cancer and dysfunctional phenotypes quite early. Now the copy number profile of a tumor sample is accessible at higher resolution using comparative genomic hybridization arrays (CGH) or DNA sequencing approaches. The main idea is the same: observe the variations of signal intensity (either hybridization or number of reads) along the genome to distinguish amplified or deleted regions. Due to the coverage biases mentioned before (mappability in repeated regions or GC content), a normalization is necessary, either using a matched normal sample from the same patient, or a pool of normal samples. The resulting profile represents the total copy number profile along the genome. In the case of CGH arrays or WGS, the totality of the genome is covered, while the profile is highly incomplete in the case of WES or targeted sequencing; moreover the capture step induces an additional bias to the data, making it noisier.

The total copy number profile can be refined by focusing on allele-specific copy number. Indeed, the human genome is diploid, so each locus is present in two copies, for the 22 autosomes, and there exists a number of positions known as single nucleotide polymorphisms (SNPs) where each version of the locus has a different nucleotide. There are around 3 to 4 million such positions differing from the reference human genome per individual. Those are genetic variations present in the individual's original genome and are distinct from the somatic SNVs mentioned before that are supplementary genomic alterations that occur during the individual's lifetime. We can also distinguish a third category of alterations beyond SNPs and SNVs, that are the germline "private" alterations of an individual that are not widespread in the population (less than 1%), and hence are not SNPs, and are not considered here. At those SNP positions, one can measure the coverage separately for each allele, and detect allelic imbalance, where one of the alleles (denoted the A allele) is amplified compared to the other (denoted the B allele). Considering the B allele frequency (BAF) allows us to obtain more detailed information about the cancer genomes alterations, and processes at their origin.

To complete the analysis, the signal is segmented, either using only the total copy number or by performing joint segmentation with the BAF signal as implemented in Pierre-Jean et al. [2015], to determine regions of constant copy number, and breakpoints separating those regions. Some methods like Pindel [Ye et al., 2009] or DELLY [Rausch et al., 2012] additionally analyze the split reads covering both ends around a breakpoint to ensure better detection of structural variants, however, WGS is necessary for this step, and long reads exhibit even more power to resolve complex situations that can be incorrectly mapped to the reference genome. Similarly to the variant calling problem, many methods have been developed to uncover the structural variations of tumor genomes, and their precise error rates are hard to evaluate for similar reasons [Pierre-Jean et al., 2015]. Once the genome is segmented, the last stage of CNV calling consists in assigning integer copy number values to each segment, i.e. to determine the ploidy of the tumor. This step is highly confounded by the sample purity, and there exists multiple possible values for the pair (purity, ploidy), i.e. the problem is unidentifiable [Zaccaria and Raphael, 2018; Shen and Seshan, 2016; Favero et al., 2014]. Finally, as in the case of variant calling, the problem is actually further complexified when considering the sample as a mixture of clones with different genomic landscapes; this will be further explored in the next chapter.

Chapter 2

Computational methods to unravel tumor evolution from genomic data

Abstract

The problem of reconstructing the population structure of a tumor sample from genomic sequencing data has raised a lot of interest from the community, and more than eighty methods have been proposed in the last few years to solve it. In this chapter, we provide an overview of those methods, by describing the different types of input data considered in the reconstruction, and the underlying rationales and algorithms. This first outline can be useful for a potential user to select a method adapted to their scientific question and available data, but also highlights the lack of proper evaluation of those methods to choose the right one. This deficiency prevents the identification of the most promising directions for future developments, and keeps the expected accuracy when applying existing methods hidden from non-specialists, which may be misleading when designing experiments or interpreting the obtained results. We focus on some of the difficulties met when considering such a benchmark, which may explain the lack thereof. Finally, this brief review has allowed us to identify potential shortcomings in the field of ITH inference that motivate the contributions presented in the two following chapters, both on methodological developments, and evaluation of the clinical relevance.

Résumé

L'identification des sous-populations cellulaires composant un échantillon tumoral à partir des données de séquençage de leurs génomes est un problème qui a beaucoup intéressé la communauté, et plus de quatre-vingts méthodes ont été développées pour résoudre ce problème. Dans ce chapitre, nous fournissons une vue d'ensemble des méthodes existantes, en nous intéressant aux types de données d'entrée prises en compte pour la reconstruction d'une part, et à la logique et aux algorithmes sous-tendant ces approches d'autre part. Ce premier aperçu peut être utile à l'utilisateur potentiel, en lui permettant d'identifier les méthodes adaptées à sa question scientifique ou aux données disponibles, mais révèle aussi le manque d'évaluations adaptées de ces méthodes, pour pouvoir choisir laquelle appliquer. Cette lacune empêche l'identification des pistes les plus prometteuses à continuer à développer à l'avenir, et réserve aux spécialistes du domaine la connaissance de la véritable exactitude que l'on peut attendre des résultats de ces méthodes, ce qui peut conduire à des erreurs lors de la conception d'expériences ou de l'interprétation des résultats obtenus. Nous présentons ensuite certaines des raisons qui rendent la réalisation de telles évaluations difficiles, expliquant peut-être leur absence. Enfin, cet examen critique nous a permis d'identifier des insuffisances dans le domaine de la mesure de l'hétérogénéité intra-tumorale auxquelles nous nous sommes efforcés de remédier à travers les travaux présentés dans les chapitres suivants, à la fois sur le plan méthodologique que sur celui de l'évaluation de la pertinence clinique de ces mesures.

Contents

2.1 Overview of existing methods **17**
2.1.1 Selection of ITH methods 17
2.1.2 ITH method features, and attribution strategies 17
2.1.2.1 Input description 18
2.1.2.2 Output description 19
2.1.2.3 Preliminary algorithmic characterization 21
2.1.3 Main classes of methods 22
2.2 Challenges for method evaluation **25**
2.2.1 Different inputs, different outputs, different problems . . . 25
2.2.2 Choice of a benchmarking dataset 25
2.2.2.1 Simulated data 26
2.2.2.2 Real data 27
2.2.3 Metrics 27
2.2.4 Previous comparisons of ITH methods 28
2.3 Open questions for ITH inference **30**
2.3.1 Directions for future developments 30
2.3.2 Method evaluation 31

We have briefly introduced the concept of intra-tumor heterogeneity (ITH) in Chapter 1, and its continuous shaping by mutational processes and evolution, and will focus in this chapter on the overview of experimental and computational approaches designed to measure its true extent in tumors. We will first present a comprehensive overview of existing methods, then outline emerging ideas and opportunities for the field, and finally look into the question of the evaluation of their performances.

2.1 Overview of existing methods

Reconstruction of the evolutionary history of a tumor using bulk sequencing data is a problem that has raised interest within the community, and over 80 approaches have been designed to solve a variety of formulations of the question. A large part of those methods (probably not an exhaustive list despite our best efforts), further denoted ITH methods have been reviewed and are summarized in Supplementary Table D.1. Considering the number of methods, we have extracted a number of features to better approach and represent the complex diversity of the developed approach. This first step has allowed us to distinguish broad categories of methods, which can be helpful for the reader or potential user to navigate among methods and identify the one(s) best suited for their needs. We then consider the problem of method evaluation, which is a key issue for further performance improvement, and finally, we outline some challenges for future developments.

2.1.1 Selection of ITH methods

In Figure 2.1, ITH detection steps are represented schematically, and some technical challenges are highlighted. We can distinguish the following main steps of ITH and evolutionary history reconstruction:

1. identification of relevant variants from raw read counts data,
2. estimation of their cancer cell fractions (CCF) in the sample carrying those relevant variants,
3. grouping of variants with similar CCFs (returning either a clustering (denoted 3A), or sets of variants meant to represent actual tumor genotypes (3B)), and
4. reconstruction of a tree recapitulating the tumor evolutionary history?

The surveyed methods vary in the steps that they cover, with several methods taking up only one step; this is a first way to describe ITH methods. A second main distinction is whether the different steps are dealt with sequentially or jointly, with the ambition to integrate and leverage information of one step to inform the others. We do not consider methods that only perform the first step; those are the variant callers presented in the previous chapter. The decision to include a method is sometimes arbitrary, and may reflect the authors' branding, e.g. did they or did they not claim to have developed a CNV caller or an ITH method. ITH methods also vary in the input features they consider: as described in the first chapter, several kinds of descriptors can be extracted from raw sequencing data, and we will detail further how they can be leveraged for ITH deconvolution, and even combined together for the most integrated approaches. Finally, methods are also characterized by their algorithms, and we will describe the main classes convened to solve this problem.

2.1.2 ITH method features, and attribution strategies

Due to the high number of ITH methods to analyze, we have defined 3 groups of criteria to characterize them. This characterization is of course incomplete as each method is unique in its association of input combination, modeling choices, etc, which is not fully captured by our coarse representation, and described in slightly more details in Supplementary Table D.1. We describe here the different criteria chosen to describe the methods. We would like to warn the reader that feature attribution was performed by a human being in a subjective way and

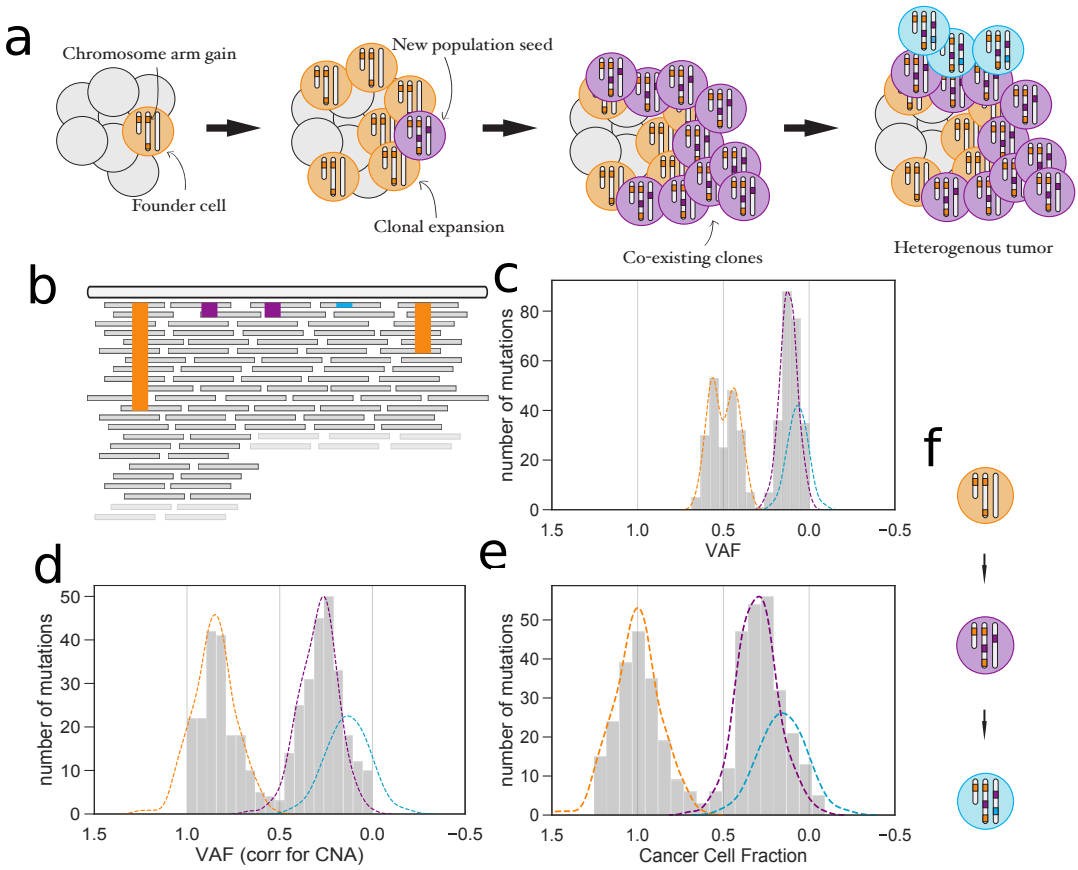

Figure 2.1 – Schematic representation of the ITH reconstruction problem. In panel **a**, we represent the successive clonal expansions leading the tumor to an heterogeneous state, from which a bulk sample is taken, and sequenced, with "raw reads" aligned to the reference genome (panel **b**). The colors represent detected alterations (*step1), with a various proportions of altered reads (VAF). Panels **c**, **d** and **e** represent the VAF histogram, also called site frequency spectrum, with successive normalizations to account for CNVs and sample purity (*step2, for SNVs; other approaches are tailored to deal with CNVs). In each case, the dotted lines represent the envelope for each (true) clone. Inferring those envelopes is the objective of *step3. We can already see that* step2 highly influences the identification of groups of alterations, with CNVs creating an illusion of two distinct clones. *step3 remains challenging, as the blue clone (low frequency, and low number of alterations) is hard to distinguish. Finally,* step4 aims at reconstructing a mutation tree recapitulating the evolutionary history (panel **f**). In the case of one sample, a linear history is always compatible with the data [Beerenwinkel et al., 2015].

can be questioned or further discussed in a number of cases, but we believe that the resulting typology still provides a valuable first perspective on the ITH methods landscape.

2.1.2.1 Input description

Not all ITH methods rely on the same input to provide a description of tumor samples. The underlying rationale of almost all methods considered in this analysis is to estimate and use the CCF of alterations, either SNVs, small indels or CNVs, based on sequencing data. We will here use abusively SNVs to refer to SNVs or small indels, as most methods deal with them in a similar way. Depending on the algorithm and simplifying assumptions, methods can take as input SNVs and/or CNVs, with all possible combination: a method can deal with only one of them, both, and the mandatory character of either input also varies. For instance,

- Some approaches consider only SNVs, such as PhyloSub [Jiao et al., 2014] or PurBayes [Larson and Fridley, 2013].

- SciClone [Miller et al., 2014] and other similar methods accept CNVs as input, in addition to SNVs, but deal with them in a deterministic manner, with exclusion of SNVs from altered regions (SciClone), or a priori normalization as in Palimpsest [Shinde et al., 2018]. Those methods can in theory be run with SNVs only, with a very naive simplifying assumption of unaltered ploidy, though this will impair the results.
- Few methods absolutely require both SNVs and CNVs, and model them jointly such as cloneHD [Fischer et al., 2014], PhyloWGS [Deshwar et al., 2015], PyClone [Roth et al., 2014].
- Other methods, like TITAN [Ha et al., 2014], or THetA [Oesper et al., 2013] solely model CNV abundance.
- Finally, some methods are agnostic to the nature of input alterations, and can work with either of them, like CloneSeeker [Zucker et al., 2019]

This information is partially encoded in the binary variables SNV and CNV, without the information of the mandatory character of those inputs, or the relevance of integration of those data in the subsequent modeling steps. A "yes" value for one of this variable means that the method accepts such input.

Another variation in input requirements between methods is the number of samples: some methods are designed to deal with only one sample, others only with multiple samples, and some are compatible with both settings. We note that the fact that a method is able to provide a result with one setting or the other is not a guarantee that this results is relevant; in particular in the case of phylogeny reconstruction based solely on CCF information from a single sample, a linear evolution is always compatible with observed data, and two samples are necessary to infer a branching pattern [Beerenwinkel et al., 2015]. This information is encoded by the variables `one_sample` and `multiple_samples`, with a "yes" value meaning that the method is compatible with that setting.

Even between methods with the same class of input alterations, each method can adopt a different format and summary information. In the case of SNVs, either raw counts of reference and variant alleles can be required, or directly their ratio (VAF), or even a coarser presence/absence binary pattern. The same diversity is true for CNVs, with methods requiring total or allele-specific integer copy number profiles, or fractional copy numbers, or raw log ratios, or even segmented read counts. In addition, required inputs may include read counts at known germline SNPs in the tumor and paired normal samples. This of course is closely related to the method's algorithm, and will be further discussed in later sections. This is partly encoded in the binary variables `raw_counts_SNV` and `raw_counts_CNV`, indicating if the method accepts some kind of raw data directly based on read counts, or requires already pre-processed data. Complementary information is encoded in the `WGS` and `WES` variables, that indicate if data from WES or WGS can be sufficient to generate input data for the ITH method, that are partly related to required input in the sense that methods relying on split-reads covering a rearrangement can hence only be used with WGS data. There is one point that we could not properly assess for most methods: a recurrent problem of ITH inference the ability to scale to a large number of SNVs (or CNV segments), which might make it challenging to apply some of the methods to WGS data which typically detect several order of magnitude more alterations than WES.

2.1.2.2 Output description

We have established 4 successive steps for ITH reconstruction, that were enumerated in the previous section, and which we will describe in more details here. The first step consists in detecting the alterations from raw sequencing data. This step is typically covered by a variant caller or a CNA caller and then provided as input to ITH methods. However some methods include that step in the ITH pipeline, such as Sclust [Cun et al., 2018], and for other approaches, the calling step is even performed jointly with the deconvolution problem. This is the case only for CNA, where the calling of the copy number profile for each subclone is performed at the same time as subclonal deconvolution, for instance in THetA and THetA2 [Oesper et al., 2013, 2014], TITAN [Ha et al., 2014], cloneHD [Fischer et al., 2014],

HATCHet [Zaccaria and Raphael, 2018]. Intermediate methods, that are not only regular CNA callers but go a little further by associating a CCF at each alteration, such as CHAT [Li and Li, 2014], or Battenberg [Nik-Zainal et al., 2012] also include the first step. Interestingly, this calling step never concerns SNV calling.

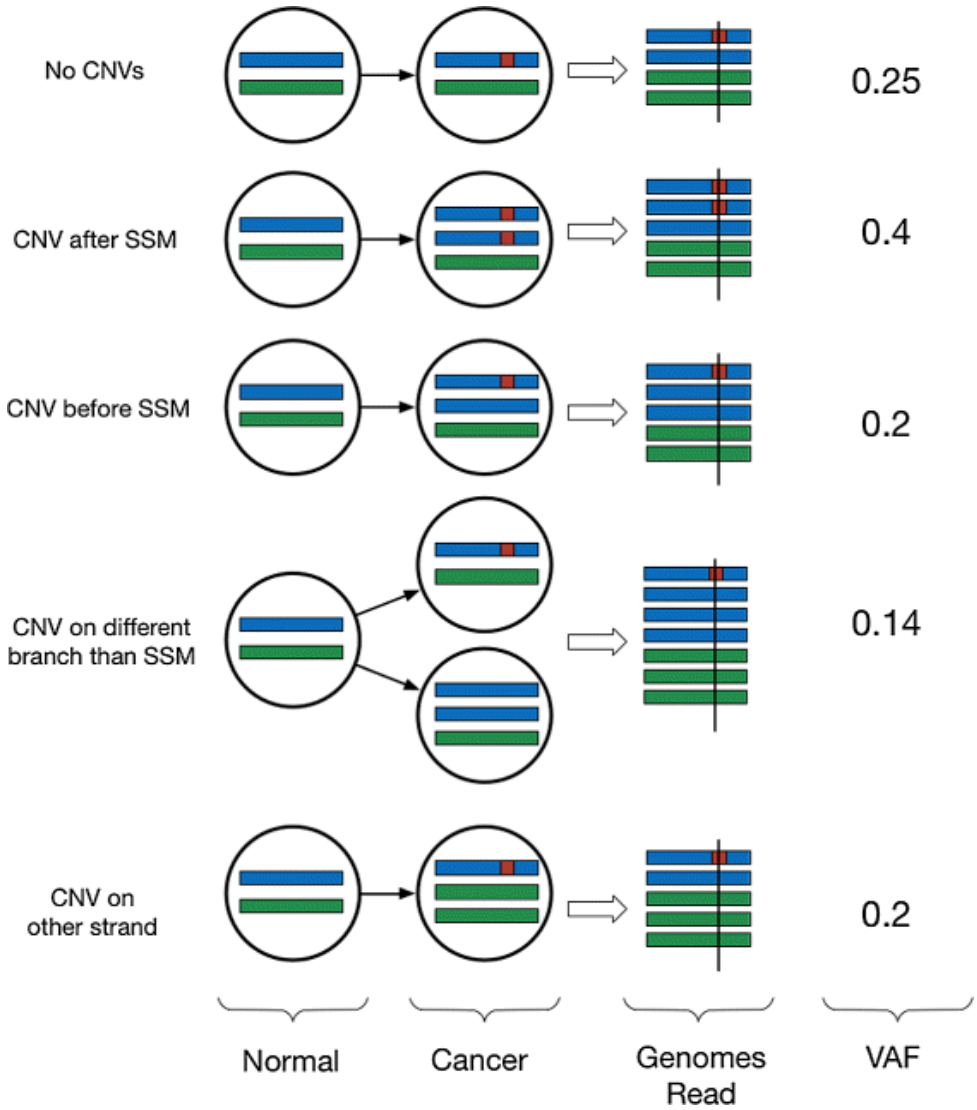

Figure 2.2 – Influence of the cancer cell population CNV at SNV locus, and tumor purity on CCF derivation. Each circle represents a population present in the tumor samples, with each population in equal proportion in the mixture. In each case, the expected measure VAF is computed, illustrating that CNV overlapping SNV locus can be a confounding factor when looking for ITH. Here tumor purity is kept constant to simplify the computation, but it obviously alters the proportion of observed reads from the tumor population, and hence the VAF. This figure is extracted from the publication of PhyloWGS [Deshwar et al., 2015].

A second step we isolated consists in estimating the CCF for each alteration. Let's focus first on the case of SNVs. For each SNV, raw data consists in the variant and total read counts at this position. Their ratio, denoted VAF for Variant Allele Frequency is usually used as a proxy for the actual variable of interest, the Cancer Cell Fraction (CCF). However, as illustrated in Figure 2.2, there are three approaches to obtain one from the other: (i) consider that the VAF is an estimation of the CCF for methods like SciClone, or PurBayes [Miller et al., 2014; Larson and Fridley, 2013], (ii) correct each VAF for purity and copy number in a deterministic way for QuantumClone [Deveau et al., 2018], Palimpsest [Shinde et al., 2018], CliP [Yu et al., 2018], (iii) more sophisticated corrections, that include for instance a joint estimation of ITH and of the copy number correction to apply as in PyClone [Roth

et al., 2014], PhyloWGS [Deshwar et al., 2015], or use of other information, like phased SNPs in OncoPhase [Chedom-Fotso et al., 2016]. An orthogonal way to inform the SNV CCF estimation is to associate it with the clustering step (either 3A or 3B), in particular if some noise level is allowed in CCF estimation.

In the case of CNVs, CCF estimation is more complex, but follows similar principles. Several sources of raw data are available, as detailed in section 1.2.2.2 and 2.1.2.1: total read counts for total copy number inference, heterozygous SNP read counts for allele-specific calling, soft-clipped and discordant reads for more precise structural variant (SV) characterization. A segmentation step is usually needed prior to ITH inference. Methods resort to various simplifying assumptions, such as the existence of only one variant genotype for each segment [Li and Li, 2014], or more complex patterns can be modeled [McPherson et al., 2017; Zaccaria and Raphael, 2018].

The third step of ITH reconstruction consists in grouping alterations with similar characteristics together in a way that is relevant to the tumor evolution. Two ways of addressing this issue have been implemented and are mutually exclusive: a first strategy consists in grouping together alterations with similar CCFs, usually forming peaks in the CCF histogram, and a second strategy aims at identifying genotypes and their mixing proportions. The main approaches for those steps consist in simple clustering approaches, such as k-means in BAMSE [Toosi et al., 2019] or hierarchical clustering in SuperFreq [Flensburg et al., 2018], potentially with some further refinement, or probabilistic mixtures, either finite or infinite (Dirichlet Processes), and either modeling CCFs or raw read counts, with a several choices of distributions (Binomial, Beta-Binomial to account for overdistribution for read counts, and then Gaussian or Beta for CCF estimates). For any algorithmic choice, the difficulty remains the choice of the number of components; classical approaches are used, with either standard criteria like the Bayesian Inference Criterion (BIC) [Schwarz, 1978], or the choice of prior distributions for fully bayesian settings. In the case where genotypes are inferred, probabilistic methods are used in most of the cases when no tree is inferred from those genotypes, with hierarchical probabilistic models specifically designed for this problem, like in Clomial [Zare et al., 2014] or BayClone [Sengupta et al., 2015] for SNV-based approaches, or TITAN [Ha et al., 2014], CloneDeMix [Tai et al., 2018] for CNV-based approaches.

The last step (step4) relies on results from the previous step and reconstructs a clonal tree representing the order acquisition of mutations. It presents similarities with a phylogenetic tree with a major difference: internal nodes in the tree can be observed in tumor samples as new and ancient populations can coexist. Results from step3B are closer to the generation of a tree than step3A where there is no indication of groups of alterations hitting the same tumor cells. Tree inference resorts to two main types of algorithms: probabilistic, and in that case step3B and 4 are jointly performed, as in cloe [Marass et al., 2016], PhyloWGS [Deshwar et al., 2015], or Canopy [Jiang et al., 2016]; or combinatorial, in which case, pre-clustered alterations from step3A are used to reconstruct genotypes and then a tree, as in LICHeE [Popic et al., 2015], or CITUP [Malikic et al., 2015].

2.1.2.3 Preliminary algorithmic characterization

To describe the algorithms used by the different methods, we have very broadly distinguished 3 categories

probabilistic approaches , with various strategies for inference (MCMC, variational inference, EM, maximum likelihood direct computation), and model selection criteria (fully bayesian, BIC or similar alternatives),

graph or combinatorial approaches , where a high number of possible solutions are tested, often with heuristics to reduce the search space, and

optimization algorithms , where an objective function to minimize is defined, and standard descent algorithms are implemented.

Several categories of algorithms can be used sequentially for the different steps of the reconstruction when they are performed independently by a single method.

2.1.3 Main classes of methods

We have characterized each considered ITH method according to the criteria defined in section 2.1.2, and delineated four main classes of ITH algorithms. This typology framework is fully represented in Figure 2.3. Groups are numbered from 1 to 4 from the top to the bottom of the plot.

Figure 2.3 – Typology of ITH methods. For each method, several criteria were evaluated (mostly in a binary way) to characterize each method. Hierarchical clustering was performed to distinguish several classes of approaches. Of course, such a typology is not unique, and is an attempt to provide the reader with a reading grid to approach the complex landscape of ITH methods.

Methods of the first group have in common to return an inferred tree, with or without providing subclonal genotypes, and their proportions in the sample(s), mostly relying on SNVs, with some approaches accounting for CNVs as well, and a few methods relying solely on CNVs. A first distinction that can be done between methods is whether genotype and tree inference are done jointly, which is the most interesting case, as the tree structure can help distinguish subclones that are close in CCF estimations. There are several levels of integration between genotype reconstruction and tree inference, with a fully joint inference, as in Canopy with a complete probabilistic model for both steps [Jiang et al., 2016] or CALDER that proposes a mixt integer linear program [Myers et al., 2019], or CITUP for a small number of mutations [Malikic et al., 2015]. Other methods such as PASTRI rely on a third party ITH clustering algorithm, but consider the posterior CCF distributions of each alteration, instead of hard cluster assignments, allowing for some flexibility while reconstructing the tree [Satas and Raphael, 2017]. Other methods implement heuristics to start from observed genotypes, and alternate clustering and tree reconstruction to provide a result, like TargetClone [Nieboer et al., 2018], and CloneFinder [Miura et al., 2018]. A second category of approaches implement a clustering strategy that is followed by the tree reconstruction step. That provides an advantage in terms of performance and algorithm complexity, though it might be at the expense of deducing constraints from the tree to inform the clustering step. Methods not performing genotype inference usually require grouped alterations as input, like SCHISM [Niknafs et al., 2015], TrAp [Strino et al., 2013] or RecBTP [Hajirasouliha et al., 2014]. Over those tree inference approaches, two stand out by taking the sample origin into account to help the reconstruction: CALDER that leverages the temporal relation between several samples [Myers et al., 2019], and MACHINA that models metastatic seeding and can use the localization of multi-site samples [El-Kebir et al., 2018]. The method Meltos [Ricketts et al., 2019] also proposes an interesting idea: build a high confidence tree from SNVs, and use this tree to help the calling of CNVs, and their placing on the tree. This idea has been already implemented to improve SNV calling [Salari et al., 2013; van Rens et al., 2015], but was not really further used.

The second group contains CNV-based methods that do not output a tree. They have diverse objectives: some like Battenberg are close to CNV callers, and additionally provide a CCF estimate. At the other extreme, methods like ReMixT [McPherson et al., 2017] and RCK [Aganezov and Raphael, 2019] go beyond subclonal inference and CNV profiles, and attempt to reconstruct "assembled" tumor genotypes, with in-between more classical approaches like TITAN [Ha et al., 2014], THetA and THetA2 [Oesper et al., 2013, 2014], p-SCNAClonal [Chu et al., 2018], MixClone [Li and Xie, 2015] and HATCHet [Zaccaria and Raphael, 2018], that return proportions of clones, with CNAs assigned to the various clones. Due to combinatorial complexity, a lot of those methods are limited to two tumor populations. As CNV callers results are often not compatible with ITH, most of these methods require partially raw inputs, usually segmented read counts or segmented log ratios and provide their own inference of integer copy number for the different tumor populations.

The last two groups are focused on SNV-based approaches that provide either genotypes for the third group, or clusters for the last group. Most of those methods are probabilistic, with different a variety of models: finite mixture models or Dirichlet Processes [Ferguson, 1973] that allow for the number of mixture components to be automatically inferred, correction for copy number, considering raw read counts, or directly VAFs, and finally, some of those methods also propose a probabilistic approach to tree inference. In the third group, all methods estimate genotypes, based on probabilistic models. Most of those approaches do not take CNVs as input, and consider only SNVs in copy-neutral regions. Others, like BayClone2 [Lee et al., 2016], while not considering actual measures of CNVs allow the copy number at each position to vary up to a maximum value set by the user. Among methods that only reconstruct genotypes, most methods in this group have similar underlying models, and differ by the inference algorithm: Clomial [Zare et al., 2014] proposes a generative model with parameter inference by an EM algorithm, while BayClone [Sengupta et al., 2015] (and its precursor Bayesian feature allocation [Lee et al., 2015]) are fully Bayesian, and inference is performed by an MCMC algorithm. SeqClone [Ogundijo and Wang, 2019] implements the same model as BayClone, but with a more efficient inference using an Indian Buffet Process [Griffiths and Ghahramani, 2011] and a sequential Monte Carlo approach for inference; it was further ex-

tended to tumor_clones [Ogundijo et al., 2019] with three possibilities for each SNV instead of two (with 0, 1 or 2 copies of the mutation). Finally BayClone2 [Lee et al., 2016] further extended the model to an arbitrary number of mutated copies. For tree reconstruction, two approaches are considered: either a joint process for generating genotypes, their phylogenetic relations and their proportions using a tree-structured stick-breaking (TSSB) process, implemented first in PhyloSub [Jiao et al., 2014] for copy-neutral alterations, and then extended to account for copy-number alterations in PhyloWGS [Deshwar et al., 2015], or a tree-guided latent feature allocation model, close to the models behind Clomial or Bayclone, but with an underlying tree structure enforcing the infinite-site assumption, and the pigeonhole rule, with penalization rather than total impossibility of rule violation. This latter strategy was adopted in cloe [Marass et al., 2016], and TreeClone [Zhou et al., 2019] for SNVs in copy-neutral regions, and further extended to arbitrary CNVs in PhylogenicNDT [Leshchiner et al., 2019] and SIFA [Zeng et al., 2019]. Finally, two methods, PairClone [Zhou et al., 2018], and its extension to a tree structure in TreeClone [Zhou et al., 2019], are leveraging an original and relevant information, of the phasing of SNVs, that of course constraint the space of possible genotypes. In the last group of methods, most approaches are SNV-based methods, and implement probabilistic mixture models, and differ mostly by how they model SNVs (read counts, CCFs), the way they incorporate correction for copy number, the possibility to include several samples, and the exact method for inference (MCMC, EM, variational inference). CloneSig, the original method we propose and further describe in Chapter 4 would belong to this class.

2.2 Challenges for method evaluation

In the previous section, we presented the overwhelming variety of methods designed to resolve ITH. It might already be helpful to the reader or potential user to narrow down the choice depending on the scientific question and/or the data at hand, or be a guide to generate data in an optimal way. Unfortunately, performances of each method, the really useful information needed to choose the best-suited tool is missing from the previous overview. When a new method is published, the reader could expect that the authors provide a complete evaluation of their approach's performances, and convincing evidence that it outperforms existing ones. We have surveyed the evaluations provided for each method presented in the previous section, and reported results in Figure 2.4. A first striking observation is the sparsity of this matrix: very few methods provide a satisfying evaluation of their performances and compare them to existing methods.

We have also extracted from this co-test matrix the number of methods each new methods tests (including the method), and the number each method is used in a benchmark (including its own publication), and results are provided in Figure 2.5. Surprisingly, one third of the methods do not provide any comparison to an independent approach, and one fourth compare themselves to only one other method.

In this section, we will explore the difficulties associated with ITH methods evaluation, and review the existing benchmarks.

2.2.1 Different inputs, different outputs, different problems

A first challenge when trying to evaluate ITH methods is that they all require different inputs, perform different tasks and provide different outputs, which makes the comparison difficult. By relying on the task distinction we elaborated in the previous section, we can however design an evaluation for each task separately. Let us now focus on the technical specifications of the ideal benchmark.

2.2.2 Choice of a benchmarking dataset

An essential ingredient to method evaluation is the data used as input. A proper evaluation would involve simulated data and real data. Indeed, one can distinguish two main causes of failure: (i) the model fails to find the optimal solution, which can be the case for the ITH

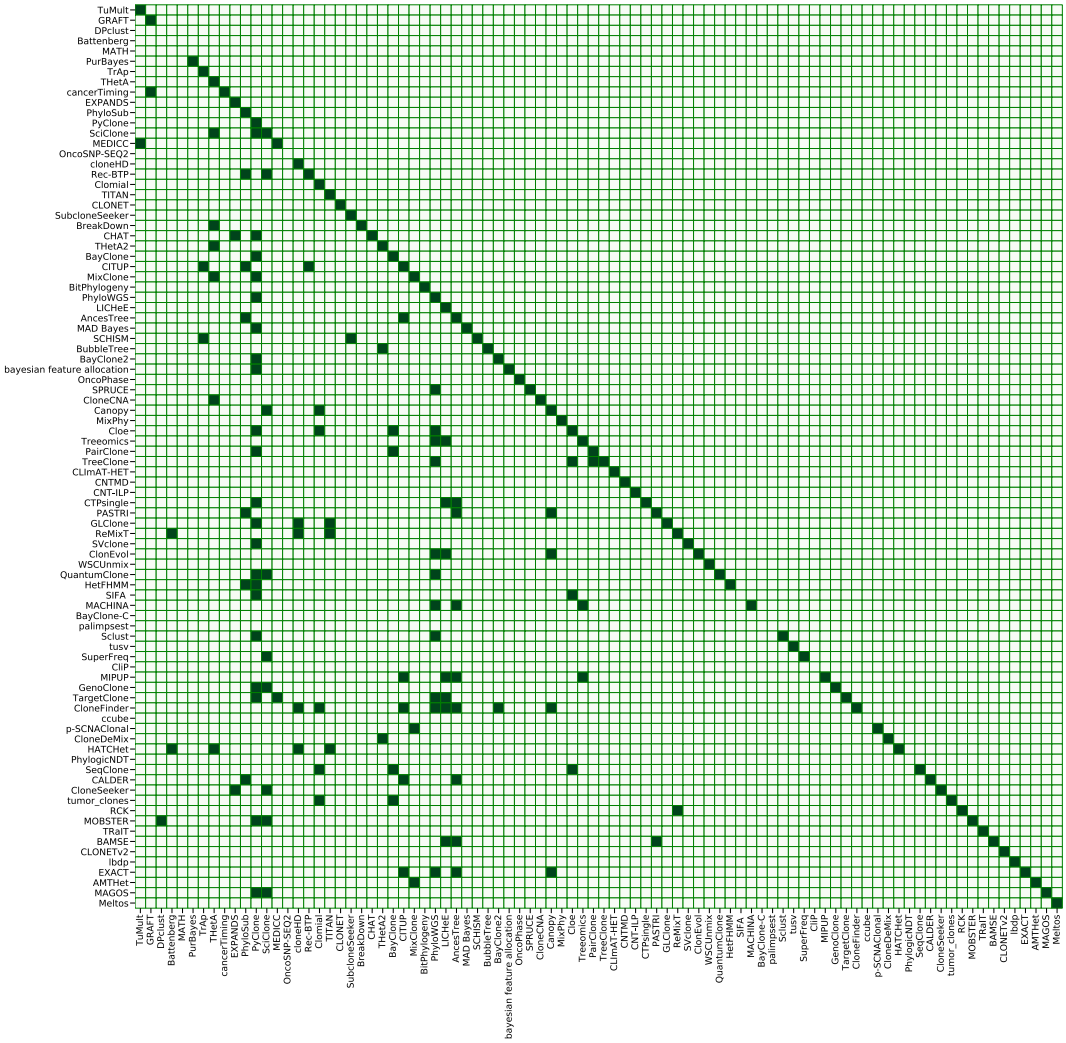

Figure 2.4 – Evaluation of a new published method and comparison with existing ones. ITH methods are ordered by date of publication, from top to bottom and left to right, and a dark cell at position (i, j) indicates that method i provided a comparison with method j on simulated data. Cells on the diagonal (position (i, i)) hence indicate whether results on simulated data for the new method are provided in the publication. As expected, all points are under the diagonal, as no method can compare itself with future methods.

methods as the problem is unidentifiable, and (ii) the model does not capture the real data, which could occur for instance if one applies a CNV-based ITH method to a tumor with only SNVs, or with a less pathological example, if SNVs and CNVs were to occur at different steps of tumor evolution, methods considering only one type of alterations would not be able to reconstruct an accurate picture of the tumor evolution. Simulated data can evaluate the first case, and constitute a sanity check, and real data can be relevant to test whether the method is well-suited to the real-life application.

2.2.2.1 Simulated data

Simulated data is an appealing solution to provide a benchmark for ITH methods, as they provide a controlled environment with the associated ground truth, and allow careful evaluation of methods in a variety of situations that can help researchers disentangle the necessary data to provide robust and accurate results with the different methods, and/or identify the best performing methods. However, the major drawback of simulated data is that one can only simulate hypothetic situations and may lead to a biased view of methods where the best

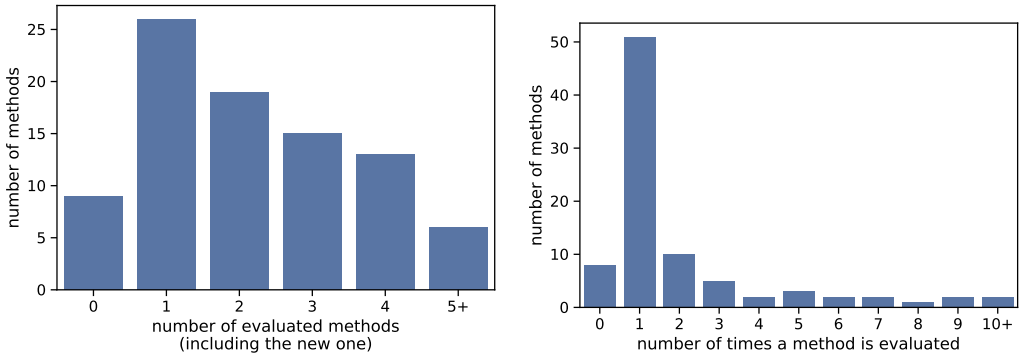

Figure 2.5 – Number of evaluations per method. Left panel presents the number of methods evaluated in each ITH methods publication, with the value "0" meaning that no evaluation on simulated data is provided, "1" that only the presented method is evaluated, etc. The right panel summarizes the number of independent evaluations for each method.

performing methods would actually be the ones with the hypotheses and underlying model closest to the simulated data. Another difficulty is that with the diversity of methods, the simulated data should represent accurately and jointly several aspects of the tumor evolution process, and this is a entirely unsolved problem. Elaborate simulations covering several key aspects exist, like complex CNA acquisition patterns with MASCoTE [Zaccaria and Raphael, 2018], or BAMSurgeon [Ewing et al., 2015], a tool created for variant calling benchmarks. Other aspects are tied with tumor evolution, and are neglected in those previous simulation approaches, such as spatial constraints for tumor growth [Noble et al., 2019], metastatic seeding patterns [El-Kebir et al., 2018] phasing of alterations, mutational signatures for structural variants, and realistic activity ranges and combinations of considered signatures, rate of acquisition of passenger mutations [Caravagna et al., 2019; Dinh et al., 2019], varying tumor sizes. New phenomena will keep being discovered, making it really difficult to provide a static simulated dataset, that could be used as a common reference as new methods are published.

2.2.2.2 Real data

Real data is an obvious solution to the raised issue of simulating complex and poorly-understood aspects of tumor evolution, but it suffers from another very strong limitation: the absence of ground truth. Researchers have resorted to several strategies in the use of such data: the simpler one is to provide a qualitative accuracy assessment with the righteous detection of previously known patterns (detection by another approach, manual reconstruction). This is of course debatable as both assessments can be wrong. Another scheme is to rely on proxys for the ground truth, obtained via a second experimental strategy. One can resort to multi-sample sequencing, that can provide some reliable facts (e.g. 2 SNVs are not in the same clones are they are systematically present in different samples), but each sample must be considered heterogeneous [Alves et al., 2017], and reconstructions from multi-sample are deemed more biased by most available approaches [Caravagna et al., 2019], creating the risk to evaluate methods against an erroneous ground truth. Single cells offer a promising orthogonal approach to unravel the evolutionary history of a tumor, and provides unmixed observed genotypes. However, there are currently serious technical limitations, like an elevated dropout rate that prevents the calling of a subset of SNVs in each cell. New technologies are being developed and may overcome those restrictions in a near future [Laks et al., 2019].

2.2.3 Metrics

The question of evaluation metrics is the counterpart of the fact that each method solves a different flavor of the ITH reconstruction problem. A large number of metrics have been proposed by the authors of the different evaluations, here are some of them

Evaluation of purity estimate: absolute error [Oesper et al., 2013].

Evaluation of step1 (calling of alterations): proportion of the genome with correct copy number estimation [McPherson et al., 2017], global ploidy error [McPherson et al., 2017], median copy number error [Oesper et al., 2013], precision/recall of whole genome duplication calling [Zaccaria and Raphael, 2018].

Evaluation of step2: diverse indicators of the difference between true and inferred CCF, such as percent of alterations where the inferred value is within 10% of the true value, average and maximum values [Fan et al., 2014], correlation between the true and absolute values for all alterations together [Li and Li, 2014].

Evaluation of steps 3A/3B: Salcedo et al. [2018] proposed two scores, one for the number of populations, and one inspired from clustering evaluation (relying on V-measure [Rosenberg and Hirschberg, 2007], and correlation of the co-clustering matrices), proportion of mis-clustered alterations, after identifying a correspondence with the true mutation clusters [Malikic et al., 2015].

Evaluation of step4: Salcedo et al. [2018] proposed to use a metric to compare matrices of ancestor-descendant relationships between pairs of alterations, Malikic et al. [2015] measure whether the proportion of simulations where their method recovers the same exact phylogenetic structure among the top 3 trees, and El-Kebir et al. [2016] report a "recall" metric, that is the proportion of edges in the initial tree correctly recovered in the inference. This is an active field of research as tree distance metrics are also relevant to perform patient stratification [Karpov et al., 2018; DiNardo et al., 2019].

Some metrics might be biased if they measure exactly the quantity optimized by the method, so it could be a good practice to consider several metrics for each aspect.

Other lines of thought for method evaluation could be the agreement between methods, but the chance that the minority is right can not be excluded, especially as a lot of methods have similar simplifying assumptions or similar models, and could create an illusion of consensus. Association with other independent variables could also be used as a proxy for method validation, but no known and well-accepted association could serve such a purpose to our knowledge.

2.2.4 Previous comparisons of ITH methods

Some studies have reflected on the ITH reconstruction problem and existing methods to provide guidance to the community. We consider here three types of such approaches: the reviews, the benchmarks, and a particular case of benchmarks, a Dream Challenge.

Several reviews cover the topic of ITH reconstruction, from distinct points of view:

Mathematical with the work of Beerenwinkel et al. [2015] on mathematics models for cancer or the problem of timing mutations throughout evolution Jolly and Van Loo [2018], and the work of Schwartz and Schäffer [2017] on tumor phylogenies.

Clinical significance several reviews recapitulate the different approaches developed to infer ITH, but with a strong focus on results and new concepts for tumor evolution, and clinical applications [McGranahan and Swanton, 2017; Fittall and Van Loo, 2019; Dagogo-Jack and Shaw, 2018; Turajlic et al., 2019].

Broader views of the problem, with for instance a complete ecological perspective on tumor evolution [Maley et al., 2017].

Though offering the reader some perspectives on the field, none of those reviews was truly able to identify promising methodological avenues of research. This is partly due to the the absence of a proper evaluation of methods, from which the present works also suffers.

To overcome this deficiency, some authors have performed benchmarks of existing methods, that may offer some partial answers. A first benchmark has been published in October 2017 by Farahani et al. [2017], and proposes a very simple setup where two different cell lines are mixed in different proportions, and represent a tumor with two clones. Two pairs of cell lines were selected, a pair with diploid genomes, and a pair with aneuploidies to assess the

influence of CNAs on the reconstruction. A variety of in silico experiments were additionally implemented to measure the impact of the sequencing depth, the number of samples, the number of SNVs. Four methods are evaluated in this framework: PyClone [Roth et al., 2014], SciClone [Miller et al., 2014], Clomial [Zare et al., 2014] and PhyloWGS [Deshwar et al., 2015]. Metrics are standard and measure the absolute error in CCF estimation, and the V-measure [Rosenberg and Hirschberg, 2007] to assess the quality of SNV clustering. The data is available for potential re-use, however there are several limitations to this setting: first it can not be used to evaluate phylogeny reconstruction, as the two cell lines are not related, second there are no details or code available regarding the incorporation of copy number estimates when running the methods, or the way they were obtained. This also probably indicates that this dataset is not appropriate to evaluate CNV-based methods. Low-pass WGS of those mixtures would have been an interesting complement to this dataset. Despite those limitations, this dataset is already relevant for a substantial part of methods of groups 3 and 4 of our proposed typology.

A second team proposes two evaluations, a first one published in 2018 [Miura et al., 2018] is quite broad in the choice of methods and associates methods that return a tree or not, and a second one published as a preprint in 2019 [Miura et al., 2019] that focuses on the phylogeny reconstruction problem. In the first article, the authors also propose a new method, CloneFinder, but we still chose here to consider the benchmark part as it is much more thorough (9 methods are evaluated) and emphasized compared to other ITH method publications. Four datasets are simulated; they differ by the tree shape underlying them, their number of clones, and the number of tumor samples. However, they all have a small number of SNVs (max 100), and a similar read depth (100). They considered only one metric, called "genotype error", consisting in counting the percentage of SNVs wrongly assigned to clones or genotypes after matching inferred clones to the most similar true ones. For method comparison on two real tumor samples sets, the number of clones is additionally reported. Overall, LICHeE [Popic et al., 2015], CloneFinder [Miura et al., 2018] and PhyloWGS [Deshwar et al., 2015] were found to be the best performing methods. The authors also note important disparities in runtimes, and failure of some approaches to run on some of the datasets. In the second benchmark focused on tree inference methods [Miura et al., 2019], the same simulated samples are used, and 7 methods are evaluated, with CloneFinder [Miura et al., 2018], MACHINA [El-Kebir et al., 2018], LICHeE [Popic et al., 2015], MixPhy p[Hujdurovic et al., 2018] and Treeomics[Reiter et al., 2017] being combinatorial approaches, and PhyloWGS [Deshwar et al., 2015] and Cloe [Marass et al., 2016] probabilistic methods. Four metrics are used to cover several aspects of tree reconstruction, in particular the order of mutations, the branching patterns. The authors report overall poor performance for all methods, as the problem is hard and unidentifiable, but found CloneFinder, MACHINA, and LICHeE to show the best performances, and hypothesize that the explicit constraints inferred from multiple samples could be instrumental to their superiority.

A main issue of those benchmarks is that though they are very limited in the number of compared methods, implemented metrics, tested datasets, they represent a huge amount of work, with caveats like installation and setup, specific input and output requirements and formatting for each method, parameter tuning, important computational time, and eventually draw only moderate attention. A natural benchmark alternative is the principle of the challenge, where this overload is distributed among all participants, and is reduced, as each contestant knows well their method. In that spirit, a dream challenge for ITH was organized in Spring 2016.

Unfortunately, very few teams participated to the whole challenge, as reported in Table 2.1. We can think of several reasons for this lack of enrollment:

- Not all methods provide all outputs, hence limiting participation opportunities for most existing methods.
- To overcome installation and environment issues and standardize ITH methods running, an important setup involving Google Compute Engine, Docker and Galaxy was required by each team. Though enforcing such requirements constitutes a strong and admirable commitment towards reproducible science, this may have represented an elevated time investment for some potential participants.

sub-challenge	description	number of teams
1A	Predicting Normal Contamination	8
1B	Predicting Number of Subclones	8
1C	Predicting Subclone Proportions	7
2A	Determining Mutation Assignments to Subclones (hard assignment)	6
2B	Determining Mutation Assignments to Subclones (soft assignment)	3
3	Predicting Subclone Phylogeny	2

Table 2.1 – Sub-Challenges and participation to the ITH Dream Challenge

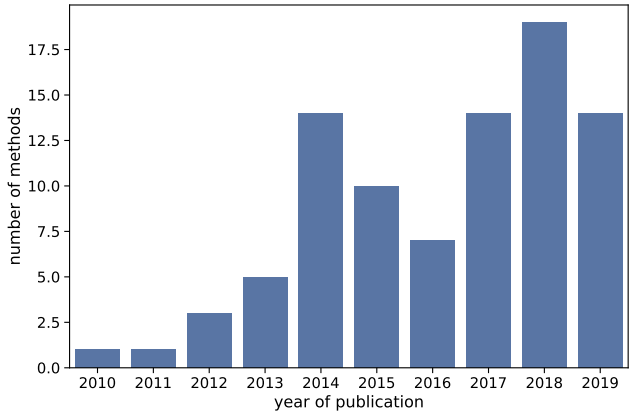

Figure 2.6 – Number of ITH methods published per year.

- The challenge timing seems to have coincided with a period of relative disinterest from the topic, suggested by the publication year distribution of ITH methods (see Figure 2.6).

Overall, though no complete leaderboard has emerges from the Dream Challenge for intatumor heterogeneity has brought an important contribution in the defined metrics, and some interesting insights in metrics design, and relevant characteristics of sequence data to vary for methods evaluation. Additionally, the generated data remains available for future use [Salcedo et al., 2018].

2.3 Open questions for ITH inference

We have presented in the previous sections a large number of methods developed for ITH reconstruction, and considered the difficulties for their evaluation, and hence their broad use by non-specialist bioinformaticians. Here we outline a few ideas for future developments in the field of ITH inference.

2.3.1 Directions for future developments

A great variety of methods has been proposed for solving the problem of ITH inference. Besides the necessity of several methods well adapted to the different possible biological assays (WES, WGS, temporally or spatially diverse samples, single-cell approaches), this abundance has also allowed the different contributors to come up with creative new ways to exploit and combine the raw input data: more and more complex integrations of SNVs and CNVs, combination of several related measures of the CNVs (BAF, average read counts over segments, SNP read counts, split reads, discordant reads), phasing of SNVs, either together or with germline SNPs. Methodological integration is an orthogonal and complementary

direction to improve inference: the structure of the tree can inform the grouping of mutations, and even the detection of alterations, as has been implemented for SNV calling [Salari et al., 2013; van Rens et al., 2015], and CNV detection [Ricketts et al., 2019].

In Chapter 4, we present a novel method CloneSig, that is exactly in the same vein of new evolutionary hints in the data that can be exploited to better reconstruct ITH: the mutation type of SNVs, which is not random but depends on the mutational processes active at the time at their occurrence.

A future challenge, probably more in engineering, would be to achieve the association of all those clues into a single method. In our review, we noticed two tools that tend to integrate most of the steps, from variant calling to tree inference, and further analyses of the clones together: SuperFreq [Flensburg et al., 2018] and PhylogenicNDT [Leshchiner et al., 2019]. Though they do not jointly infer all those steps, this association certainly offers already the possibility to incorporate some dependencies that increase the consistency of the complete analysis.

2.3.2 Method evaluation

To accelerate those future developments, a careful evaluation of the methods could be helpful for several aspects. A first lesson from such results relies in a prioritization of the features that should be implemented, depending on their actual contribution to ITH inference improvement. In parallel, that can also be indicative of the inference algorithms that achieve the best performance. Finally, this is absolutely necessary to truly evaluate which experimental settings would allow a satisfying enough ITH estimation to answer the different scientific questions that researchers will address in the future.

In that respect, the work presented in Chapter 3, though not providing a complete benchmark of methods surveyed in this chapter, or a gold standard evaluation dataset, provides valuable insights on the robustness of conclusions and scientific knowledge of tumors one can truly hope to obtain through the use of one sample per patient with WES.

Chapter 3

Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data

This chapter is published in Plos One [Abécassis et al., 2019]

Abstract

Tumors are made of evolving and heterogeneous populations of cells which arise from successive appearance and expansion of subclonal populations, following acquisition of mutations conferring them a selective advantage. Those subclonal populations can be sensitive or resistant to different treatments, and provide information about tumor aetiology and future evolution. Hence, it is important to be able to assess the level of heterogeneity of tumors with high reliability for clinical applications.

In the past few years, a large number of methods have been proposed to estimate intra-tumor heterogeneity from whole exome sequencing (WES) data, but the accuracy and robustness of these methods on real data remains elusive. Here we systematically apply and compare 6 computational methods to estimate tumor heterogeneity on 1,697 WES samples from the cancer genome atlas (TCGA) covering 3 cancer types (breast invasive carcinoma, bladder urothelial carcinoma, and head and neck squamous cell carcinoma), and two distinct input mutation sets. We observe significant differences between the estimates produced by different methods, and identify several likely confounding factors in heterogeneity assessment for the different methods. We further show that the prognostic value of tumor heterogeneity for survival prediction is limited in those datasets, and find no evidence that it improves over prognosis based on other clinical variables.

In conclusion, heterogeneity inference from WES data on a single sample, and its use in cancer prognosis, should be considered with caution. Other approaches to assess intra-tumoral heterogeneity such as those based on multiple samples may be preferable for clinical applications.

Résumé

Une tumeur est constituée d'un mélange hétérogène de populations cellulaires continuant à évoluer, résultant d'épisodes succesifs d'apparition et de croissance de populations sous-clonals, après l'acquisition de mutations leur conférant un avantage sélectif. Ces populations sous-clonales peuvent être sensibles ou résistantes à des traitements différents, et révèlent certains aspects de l'étiologie et de l'évolution future de la maladie. Il est donc important de

pouvoir mesurer le niveau d'hétérogénéité des tumeurs avec une grande fiabilité en vue de son utilisation médicale.

Au cours des dernières années, un grand nombre de méthodes ont été dévelopées pour évaluer l'hétérogénéité intra-tumorale à partir de données de séquençage d'exome, mais l'exactitude et la robustesse de ces méthodes sur des données réelles restent incertaines. Dans cette étude, nous avons appliqué et comparé de façon systématique six méthodes pour estimer l'hétérogénéité intra-tumorale à partir de données exomiques de 1697 échantillons tumoraux provenant du "Cancer Genome Atlas" (TCGA), représentant trois types de cancer (carcinome invasif du sein, carcinome urothélial de la vessie, et carcinome épidermoïde de la tête ou du cou), et deux ensembles de mutations différents en entrée. Nous avons observé des différences importantes entre les estimations provenant de différentes méthodes, et identifié de possibles facteurs perturbant l'appréciation de l'hétérogénéité intra-tumorale pour ces différentes approches. Nous montrons de plus que la valeur pronostique de l'hétérogénéité intra-tumorale pour la prédiction de la survie est limitée dans ces jeux de données, et n'avons trouvé aucune indication d'une amélioration par rapport au pronostique reposant sur d'autres variables cliniques classiques.

En conclusion, la mesure de l'hétérogénéité à partir de données de séquençage d'exome sur un seul échantillon tumoral, et son utilisation pour évaluer le pronostique des patients devraient être considérés avec précaution. D'autres approches pour évaluer l'hétérogénéité tumorale, par exemple à partir de plusieurs échantillons sont peut-être préférables dans le cadre d'applications médicales.

Contents

3.1 Introduction **36**
3.2 Materials and methods **37**
3.2.1 Data 37
3.2.2 Variant calling filtering 37
3.2.3 ITH methods 38
3.2.3.1 Published methods 38
3.2.3.2 Consensus (CSR) 38
3.2.4 Clinical variables 38
3.2.5 Survival regression 39
3.2.5.1 Model 39
3.2.5.2 Evaluation procedure 39
3.2.6 Immune signatures 39
3.2.7 Correlations 40
3.2.7.1 Comparison metrics 40
3.2.8 WES and single cell paired dataset 40
3.2.8.1 Data availability and preprocessing 40
3.2.8.2 Evaluation metrics 41
3.3 Results **41**
3.3.1 Assessing ITH on TCGA samples 41
3.3.2 Methods quantifying ITH exhibit inconsistent results . . . 43
3.3.3 ITH is a weak and non robust prognosis factor 47
3.3.4 ITH prognosis signal is redundant with other known factors 48
3.4 Discussion **48**
3.4.1 Comparison to similar studies 48
3.4.2 Can we truly measure ITH? 49
3.4.3 Association with survival, link with other variables 50
3.4.4 Can we build a gold standard dataset for benchmark? . . 50

3.1 Introduction

Cancer is characterized by the presence of cells growing and dividing without proper control. In the 1970s, Nowell and colleagues suggested that tumor cells follow evolutionary principles, as any other biological population able to acquire heritable transformations [Nowell, 1976]. This evolutionary framework has proven very useful in deepening our understanding of cancer aetiology [Gerstung et al., 2017].

A consequence of this progressive accumulation of mutations is intra-tumor heterogeneity. Indeed, when a new mutation occurs in a tumor cell and provides an evolutionary advantage, this cell tends to have a higher probability to survive and divide, hence seeding a new clonal population [Dentro et al., 2017]. This new clone may supersede the whole tumor population, or coexist along it. This process results in a tumor made of a mosaic of clones. Next generation sequencing (NGS), in particular whole exome and whole genome sequencing (WES, WGS), can provide new insights into the heterogeneity and evolution of tumors. Indeed, early mutations shared among all cancer cells should be detected in more sequencing reads than mutations acquired later by only a fraction of the tumor cells. Thus it may be possible to estimate the intra-tumor heterogeneity (ITH) and reconstruct the clonal history of tumors from WES or WGS data, as reviewed by Dentro et al. [2017]; Beerenwinkel et al. [2015]; Schwartz and Schäffer [2017], and many computational methods have been developed for that purpose [Roth et al., 2014; Miller et al., 2014; Deshwar et al., 2015; Andor et al., 2014]. We collectively refer to these methods as "ITH methods" in the following. Subclonal reconstruction from single cell sequencing has emerged as a new field, simplifying part of the inference problem, but raising other issues, related to technical limitations (high dropout rate) and high cost, possibly a limitation to the availability of large cohorts [Jahn et al., 2016; Davis and Navin, 2016; Ciccolella et al., 2018; Dentro et al., 2017].

Previous studies have reported that a large proportion of tumors are heterogeneous [Morris et al., 2016; Andor et al., 2016; McGranahan and Swanton, 2017; Dentro et al., 2018], with various consequences for the patient. In particular, high ITH has been associated with treatment resistance and poor prognosis [Dagogo-Jack and Shaw, 2018]. However, those results rely mostly on very detailed case studies involving only a small number of patients, with favorable experimental settings such as high coverage targeted sequencing on top of NGS, multiple sample collection (multi-site or longitudinal studies) [Nik-Zainal et al., 2012; Gerlinger et al., 2014; Navin, 2014] or even single-cell sequencing [Navin et al., 2011]. In the perspective of large-scale application in a clinical context, one needs to consider more accessible data with respect to cost and invasiveness for the patient, like moderate coverage WES on one sample per patient. A precise evaluation of existing ITH methods in this setting is needed to determine whether they allow us to find distinguishable patterns of heterogeneity and evolution of clinical relevance. Several large scale analyses have attempted to depict the evolutionary landscape of ITH in several cancer types [Gerstung et al., 2017], and to assess the prognostic power of ITH. In particular, using data from the cancer genome atlas (TCGA), a significant association between ITH and overall survival was found in at least one of the three studies [Andor et al., 2016; Morris et al., 2016; Noorbakhsh et al., 2018] for 9 cancer types: breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), brain lower grade glioma (LGG), prostate adenocarcinoma (PRAD), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), and colon adenocarcinoma (COAD). However, 5 of them were considered in another study with no significant result. In other cancer types, 2 studies consistently found no significant results for 3 cancer types: bladder urothelial carcinoma (BLAC), lung squamous cell carcinoma (LUSC) and stomach adenocarcinoma (STAD), and all 3 studies found no significant results for lung adenocarcinoma (LUAD) nor for skin cutaneous melanoma (SKM). A possible explanation for this discrepancy is that the studies base their analyses on different computational pipelines, from variant calling to ITH estimation, leading to different and sometimes contradictory results [Noorbakhsh et al., 2018].

To clarify the robustness and consistency of different ITH methods, we perform a systematic benchmark of 18 computational pipelines for ITH estimates from a single WES sample per patient (combining 2 ways to call mutations, and 2 methods to assess copy number varia-

tions (only 3 out of 4 combinations were tested) with 6 ITH methods), using data from 1,697 patients with three types of cancer from the TCGA database (BRCA, BLCA, HNSC). We selected these cancer types following conclusions of Morris et al. [2016], since HNSC, BRCA and BLCA are characterized by respectively high, intermediate and absence of prognostic power of ITH. We show that most existing ITH methods are very sensitive to the choice of mutations and copy number variations called, and that they can give very inconsistent results between each other. We highlight in particular that some methods are influenced by confounding factors such as tumor purity or mutation load. Finally, we show that although ITH measured by some computational pipelines have a weak prognostic power on some cancer types, the prognosis signal is not robust across methods and cancer types, and is confounded with informations available in standard clinical data. To further characterize those inconsistencies, we report results for ITH methods on 7 WES samples associated with single cell sequencing allowing to have an estimate of the ground truth. As a conclusion, we suggest that results of ITH analysis from single sample WES data with current computational pipelines should be manipulated with caution, and that more robust methods or protocols are likely to be needed for clinical applications.

3.2 Materials and methods

3.2.1 Data

We downloaded data from the GDC data portal `https://portal.gdc.cancer.gov/` for 3 cancer types (BLCA - 351 patients, BRCA - 904 patients, HNSC - 442 patients). We gathered annotated somatic mutations, both raw variant calling output, whose access is restricted and public mutations, from the new unified TCGA pipeline `https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/`, with alignment to the GRCh38 assembly, and variant calling using 4 variant callers: MuSe, Mutect2, VarScan2 and SomaticSniper. Instructions for download can be found in the companion Github repository (`https://github.com/judithabk6/ITH_TCGA`). RNAseq data used to compute immune signatures were downloaded through TCGABiolinks [Colaprico et al., 2016], and we downloaded clinical data from the CBIO portal [Gao et al., 2013].

Copy number calling and purity estimation

We obtained copy number alterations (CNA) data from the ASCAT complete results on TCGA data partly reported on the COSMIC database [Martincorena et al., 2017; Forbes et al., 2017]. We then converted ASCAT results on hg19 to GRCh38 coordinates using the `segment_liftover` Python package [Gao et al., 2018]. ASCAT results also provide an estimate of purity, which we used as input to ITH methods when possible. Other purity measures are available [Aran et al., 2015]; however we selected the ASCAT estimate to ensure consistency with CNV data.

The calls of allele-specific copy number and purity from ABSOLUTE [Carter et al., 2012] were downloaded from the GDC data portal `https://gdc.cancer.gov/about-data/publications/pancanatlas` on August 18th 2019. They were converted to GRCh38 as the ones from ASCAT.

3.2.2 Variant calling filtering

Variant calling is known to be a challenging problem. It is common practice to filter variant callers output, as ITH methods are deemed to be highly sensitive to false positive single nucleotide variants (SNVs). We filtered out indels from the public dataset, and considered the union of the 4 variant callers output SNVs. For the protected data, we also removed indels, and then filtered SNVs on the `FILTER` columns output by the variant caller ("PASS" only VarScan2, SomaticSniper, "PASS" or "panel_of_normals" for Mutect2, and "Tier1" to "Tier5" for MuSe). In addition, for all variant callers, we removed SNVs with a frequency in 1000 genomes or Exac greater than 0.01, except if the SNV was reported in COSMIC.

A coverage filter was added, and we kept SNVs with at least 6 reads at the position in the normal sample, of which 1 maximum reports the alternative nucleotide (or with a variant allele frequency (VAF) <0.01), and for the tumor sample, at least 8 reads covering the position, of which at least 3 reporting the variant, or a VAF>0.2. The relative amount of excluded SNVs from protected to public SNV sets varied significantly between the 3 cancer types (see Table B.3). All annotations are the ones downloaded from the TCGA, using VEP v84, and GENCODE v.22, sift v.5.2.2, ESP v.20141103, polyphen v.2.2.2, dbSNP v.146, Ensembl genebuild v.2014-07, Ensembl regbuild v.13.0, HGMD public v.20154, ClinVar v.201601. We further denote the filtered raw mutation set as "Protected SNVs" and the other one, which is publicly available, as "Public SNVs"

3.2.3 ITH methods

3.2.3.1 Published methods

We consider four published ITH methods: SciClone [Miller et al., 2014], PhyloWGS [Deshwar et al., 2015], PyClone [Roth et al., 2014] and EXPANDS [Andor et al., 2014]. In addition, we consider the MATH score [Mroz and Rocco, 2013] as a simple indicator of ITH, as well as a baseline ITH method described below. All computations were stopped after running 15 hours. This threshold was chosen to get results for most samples ($> 95\%$ when time was the limiting factor) for most methods while saving computational resources. Mean and standard deviation (std) of runtimes were computed for each method with each input mutation set separately. All parameters used for each method are detailed in the companion public Github repository containing all the commands `https://github.com/judithabk6/ITH_TCGA`. To ensure comparison, the runtimes were only performed on runs with ASCAT copy number calls.

We performed post-treatment to keep only clones with at least 5 SNVs, except for samples in which all clones were under 5 SNVs when all clones were considered. After running each ITH method we extracted 5 features to characterize ITH in a sample: the number of clones, the proportion of SNVs that belong the the major clone, the minimal cellular prevalence of a subclone, the Shannon index of the clonal distribution, and the cellular prevalence of the largest clone in terms of number of SNVs.

3.2.3.2 Consensus (CSR)

We computed a consensus of several ITH methods using the open source package CSR available at `https://github.com/kaixiany/CSR`. This method relies on matrix factorization to output a consensus clustering. We computed two separate consensus (for protected and public data), using as input the results of PyClone, SciClone, PhyloWGS, EXPANDS and baseline. MATH estimates were not well suited for the consensus. For each run, we ran matrix factorization for a maximum of 500 seconds.

3.2.4 Clinical variables

For each cancer type, we collected clinical variables from the CBIO Portal according to the following conditions: (i) categorical variables were one-hot encoded, and each level was kept if it involved at least 50 patients, and at most 50 patients had another level of the same variable; (ii) we kept numerical variables available for every patient; and (iii) in addition, we only kept the variables (if numerical) or the levels (categorical) which were significantly associated with overall survival by a single-variable cox model estimated with the Python package `lifelines` [Davidson-Pilon et al., 2019] after Benjamini-Hochberg correction for multiple hypothesis testing [Benjamini and Hochberg, 1995]. Tables A.2, A.3, and A.4 summarize the clinical variables retained for each cancer type.

3.2.5 Survival regression

3.2.5.1 Model

To estimate the prognosis power of a set of features, we use a survival SVM model [Van Belle et al., 2011]. Survival SVM maximizes a concave relaxation of the concordance between the predicted survival ranks and the original observed survival, regularized by a Euclidean norm penalty. Formally, given a training set of n patients with survival information $(\mathbf{x}_i, y_i, \delta_i)_{i=1,\dots,n}$, where $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of p features for patient i, $y_i \in \mathbb{R}$ is the time, and $\delta_i \in \{0, 1\}$ indicates the event ($\delta_i = 1$) or censoring ($\delta_i = 0$), a survival SVM learns a linear score of the form $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ for any new patient represented by features $\mathbf{x} \in \mathbb{R}^p$ by solving:

$$\min_{\mathbf{w}} \mathbf{w}^\top \mathbf{w} + \alpha \sum_{i,j \in \mathcal{P}} \max(0, 1 - (\mathbf{w}^\top \mathbf{x}_i - \mathbf{w}^\top \mathbf{x}_j))^2 \,,$$

where $\mathcal{P} = \{(i, j) \in [1, n]^2 \mid y_i \geq y_j \wedge \delta_j = 1\}$ is the set of pairs of patients (i, j) which are comparable, that is, for which we are certain that patient i lived longer than patient j. Intuitively, the loss penalizes the cases where patient i survives longer than patient j but the opposite is predicted by the model. For all computations, we used the function `FastSurvivalSVM` in the Python Package `scikit-survival` [Pölsterl et al., 2017], with default parameters. The model was trained and tested using a 5-fold cross-validation procedure.

3.2.5.2 Evaluation procedure

To assess the accuracy of a survival regression model, we use the concordance index (CI) between the predicted score and the true survival information on a cohort with survival information. Given such a cohort $(\mathbf{x}_i, y_i, \delta_i)_{i=1,\dots,n}$, the CI measures how concordant the predicted survival times $s_i = f(\mathbf{x}_i)$ are with the observed survival times y_i for comparable pairs of patients:

$$CI = \frac{1}{|\mathcal{P}|} \sum_{i,j \in \mathcal{P}} I(s_j - s_i) \,,$$

$$\text{with } I(u) = \begin{cases} 1 & \text{if } u > 0 \,, \\ \frac{1}{2} & \text{if } u = 0 \,, \\ 0 & \text{otherwise.} \end{cases}$$

In practice, we compute an approximation of CI with the function `concordance.index` from the R package `survcomp` [Schröder et al., 2011], using the `noether` method [Pencina and D'Agostino, 2004], and the associated one-sided test to compare CI to 0.5, which is the mean CI obtained with a random predictor. To compare CI's of different methods, we use a paired Student t-test for dependent samples implemented in the function `cindex.comp` from the same package. In both test settings, we aggregate p-values from each of the five cross-validation folds using the Fisher method from Python package `statsmodels`, and apply a Benjamini-Hochberg correction [Benjamini and Hochberg, 1995] to correct for multiple testing.

3.2.6 Immune signatures

We normalized RNAseq raw count data using a variance stabilizing transformation (VST) implemented in the Deseq2 R package [Love et al., 2014], treating each cancer type separately. We mapped genes from Bindea et al. [2013] to Ensembl GeneIds present in the TCGA matrix using EntrezId match table downloaded from Biomart [Zerbino et al., 2018] on March 26th 2018. Out of 681 EntrezId (577 unique), 31 (24 unique) were not matched to an Ensembl Id with associated gene expression in the TCGA RNAseq data. Each signature was then computed by averaging the VST output value for the relevant Ensembl Id for each TCGA sample. The resulting signatures we used can be found as Supplementary Table A.5. For analysis purposes, we use the complementary to the maximal value in the cohort so that the

content in immune cells varies in the same direction as tumor purity and remains a positive quantity. We denote those new variables with the prefix `inv`, e.g., for patient i in the BRCA cohort we define

$$\texttt{inv_T_cells}_i = \left(\max_{j \in \text{BRCA patients}} \texttt{T_cells}_j \right) - \texttt{T_cells}_i \,,$$

where $\texttt{T_cells}_i$ represents the signature for T cells estimated as explained above.

3.2.7 Correlations

We assessed correlations using Pearson's correlation coefficient. We computed the associated significance (for the null hypothesis that the correlation coefficient is 0) using the `scipy.stats.pearsonr` function, and we corrected the significance for multiple testing using the Benjamini Hochberg procedure at $FDR \leq 0.05$.

3.2.7.1 Comparison metrics

In addition to the correlations of the number of clones between methods, we have implemented three metrics derived from Salcedo et al. [2018] to compare ITH methods together:

Score1B measures the adequacy between one number of clones J_1 and another number of clones J_2. It is computed as $\frac{J_1+1-\min(J_1+1,|J_2-J_1|)}{J_1+1}$.

Score1C is the Wasserstein distance between two clusterings, defined by the CCFs of the different clones and their associated weights (proportion of mutations), implemented as the function `stats.wasserstein_distance` in the Python package `scipy`.

Score2A measures the correlation between two binary co-clustering matrices in a vector form, M_1 and M_2. It is the average of 3 correlation coefficients:

- **Pearson correlation coefficient** $PCC = \frac{\text{Cov}(M_1,M_2)}{\sigma_{M_1},\sigma_{M_2}}$, implemented as the function `pearsonr` in the Python package `scipy`,
- **Matthews correlation coefficient** $\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, implemented as the function `metrics.matthews_corrcoef` in the Python package `scikit-learn`,
- **V-measure** is the harmonic mean of a homogeneity score that quantifies the fact that each cluster contains only members of a single class, and a completeness score measuring if all members of a given class are assigned to the same cluster [Rosenberg and Hirschberg, 2007]; here the classes are the true clustering. We used the function `v_measure_score` in the Python package `scikit-learn`.

Before averaging, all those scores were rescaled between 0 and 1 using the score of the minimal score between two "bad scenarios": all mutations are in the same cluster, or all mutations are in their own cluster ($M_{pred} = \mathbf{1}_{N \times N}$ or $M_{pred} = \mathbb{I}_{N \times N}$).

All scores as asymmetrical and were hence computed twice. In the case of score2A, only the mutations present in the two reconstructions were considered.

3.2.8 WES and single cell paired dataset

3.2.8.1 Data availability and preprocessing

The raw data for 7 normal-tumor WES samples analyzed jointly with matching single cell sequencing [Malikic et al., 2019] were downloaded from the NCBI SRA platform `https://www.ncbi.nlm.nih.gov/sra` and processed into fastq format using the tool `fastq-dump` for the two acute lymphoblastic leukemia (ALL) patients (accession numbers: SRR1517761, SRR1517762, SRR1517763, SRR1517764) [Gawad et al., 2014], or directly downloaded in the

fastq format from the EBI ENA platform https://www.ebi.ac.uk/ena for the Triple Negative Breast Cancer patient (TNBC) [Wang et al., 2014] (accession number: SRR1163508 and SRR1298936), and the two samples (primary tumor and liver metastasis) from the two colorectal cancer patients (CRC) [Leung et al., 2017] (accession number: SRR3472566, SRR3472567, SRR3472569, SRR3472571, SRR3472796, SRR3472798, SRR3472799, SRR3472800).

All normal-tumor pairs underwent a pipeline of analysis including alignment with BWA-MEM [Li and Durbin, 2009] with options "-k 19 -T 30 -M", filtering of reads based on target intersection, mapping quality and PCR duplicates removal, using Picard [Broad, 2019], Bedtools [Quinlan and Hall, 2010] and Samtools [Li et al., 2009], and preprocess using GATK [McKenna et al., 2010] for local realignment around indels, and base score recalibration. Variant calling was performed using Mutect2 [Cibulskis et al., 2013], and variants filtered under the same rules as used for the TCGA (only "PASS" variants, and minimal covering rules), and copy number assessed with Facets [Shen and Seshan, 2016]. SNVs used in the analysis with B-SCITE [Malikic et al., 2019], passing the covering filters but not recovered by this pipeline were added to the final variant list. Those variants and the copy number profile were then passed to PyClone, SciClone, PhyloWGS and Expands for ITH deconvolution.

3.2.8.2 Evaluation metrics

To measure the accuracy of subclonal reconstructions from the WES data only using different methods, we compared these reconstructions to the reconstruction obtained by B-SCITE using both WES and single cell sequencing [Malikic et al., 2019]. To quantify the similarity of the different reconstruction results, we compared the number of clones, and for the common mutations, the metric 2A, used in Salcedo et al. [2018] and redefined above.

3.3 Results

3.3.1 Assessing ITH on TCGA samples

We collected somatic mutation information from 1,697 TCGA patients with BLCA ($n = 351$), BRCA ($n = 904$), and HNSC ($n = 442$). We selected these three cancer types following conclusions of Morris et al. [2016], since HNSC, BRCA and BLCA are characterized by respectively high (hazard ratio, HR=3.75, p=0.007 in multivariate Cox model), intermediate (HR=2.5, p=0.15) and absence (HR=1.05, p=0.91) of prognostic power of ITH. For each patient, we collected two sets of mutations based respectively on protected and public SNV sets. The protected set corresponds to raw variant calling outputs, with an extra filtering step described in Methods. The public set corresponds to publicly available SNV calls, filtered from the raw variant calling outputs to only retain somatic mutations with very high confidence, in order to ensure patients' anonymity. Supplementary Table B.3 summarizes some statistics on the number of mutations per sample for each cancer type.

We assess ITH in each sample using 6 representative computational methods: PyClone [Roth et al., 2014], SciClone [Miller et al., 2014], PhyloWGS [Deshwar et al., 2015] EXPANDS [Andor et al., 2014], the mutant-allele tumor heterogeneity (MATH) score [Mroz and Rocco, 2013], and CSR [Dentro et al., 2018], a method providing a consensus of all of the above results (except MATH which is not compatible, see Methods). Table 3.1 summarizes some important properties of the different methods, which might be helpful for designing future studies and selecting the appropriate tool. All methods but MATH take as input the CNA information in addition to a set of somatic mutation VAFs. PyClone and PhyloWGS also take purity as input. All input has to be pre-computed by third-party approaches. While MATH is a single quantitative measure of ITH based on differences in the mutant-allele fractions among mutated loci, all 6 other methods produce more details such as the number of subclones and their respective proportions in the tumor. In particular, PhyloWGS outputs a lineage tree connecting the subclones.

We tested each method three times: on each sample for the two mutation sets combined with ASCAT calls for purity and copy number, and combined with ABSOLUTE calls for the

Method	CNA as input	Purity as input	Outputs tree(s)	Reference	Mean (std) runtime protected in seconds	Mean (std) runtime public in seconds	Success rate (protected)	Success rate (public)
MATH	no	no	no	Mroz and Rocco [2013]	<< 1	<< 1	100%	100%
EXPANDS	yes	no	no	Andor et al. [2014]	891 (604)	267 (258)	89%	**71%**
PyClone	yes	yes	no	Roth et al. [2014]	7,035 (8,464)	1,414 (1,415)	95%	99%
SciClone	yes	no	no	Miller et al. [2014]	62 (48)	41 (51)	92%	**78%**
PhyloWGS	yes	yes	yes	Deshwar et al. [2015]	13,258 (9,058)	4,730 (4,139)	95%	97%

Table 3.1 – Main characteristics of ITH methods tested. The mean runtime is the mean time to process a TCGA sample. The success rate is the fraction of TCGA samples for which the method produced an output without error, with ASCAT calls as input only. The MATH score was computed in one step for all samples, using a table containing all mutations for all samples; the operation lasted 3.21s (std. 47.6 ms) for the protected dataset, and 3.39s (std. 11ms) for the public dataset. All time measurements were measured on a single cluster node with a 2.2 GHz processor and 3GB of RAM.

protected mutation set. We observed that some methods failed to produce an output on some samples, for different reasons (see success rate for each method in Table 3.1). EXPANDS produces an error for 30% of the samples, mostly for tumors with high purity or very few CNAs. SciClone fails to provide an output for samples with an insufficient number of SNV in regions without CNA or LOH event. PyClone and PhyloWGS non completion cases were caused by a too long runtime.

As shown in Figure 3.1, there is little overlap between the samples where each method fails. Out of 1,697 initial TCGA samples, all methods produced an output for the three runs on only 686 samples (296 BRCA, 178 BLCA, 212 HNSC). Those failure cases unveil indications of each method's limitations, in particular EXPANDS and SciClone. In the following we restrict our analysis to those 686 samples. One can note that there is more difference between public and protected results for BRCA samples; this is expected as the number of mutations in those two sets is more different for this cancer type, as shown in B.3.

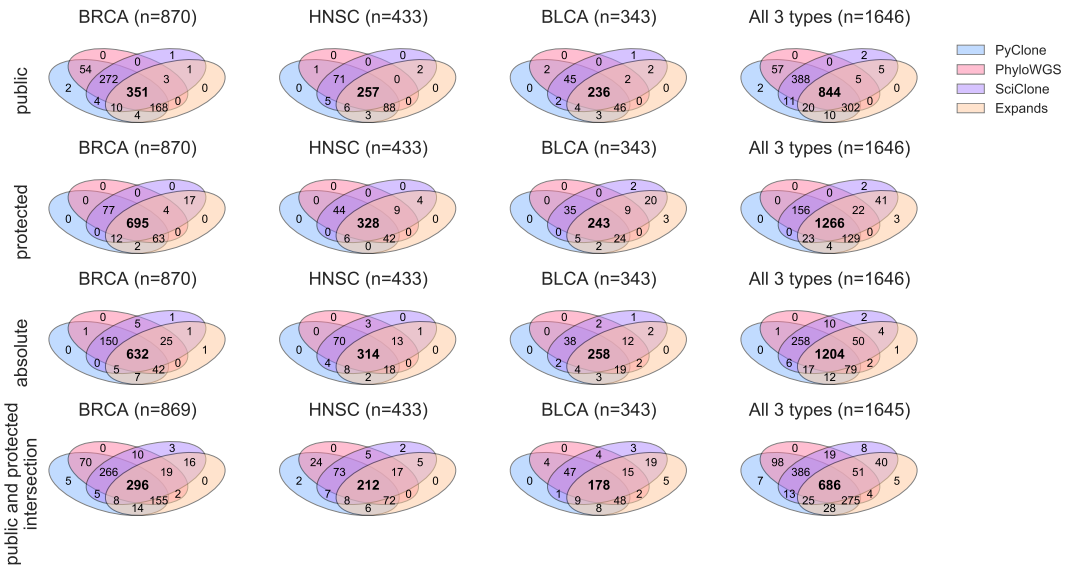

Figure 3.1 – Intersection of successful runs among the 4 considered ITH methods. The upper venn diagrams concern runs with the public input SNV set, the second line with the protected, and the third the overall intersection, as results with both sets are necessary for a proper and rigorous comparison.

In addition to failures, we observed that the runtime varies significantly between methods (Table 3.1). As shown on Figure B.25, the run time of different ITH methods increases with the number of somatic mutations. PyClone and PhyloWGS runtime rises very quickly with the number of mutations in tumor sample, which can be a limitation for applications to

heavily mutated tumors.

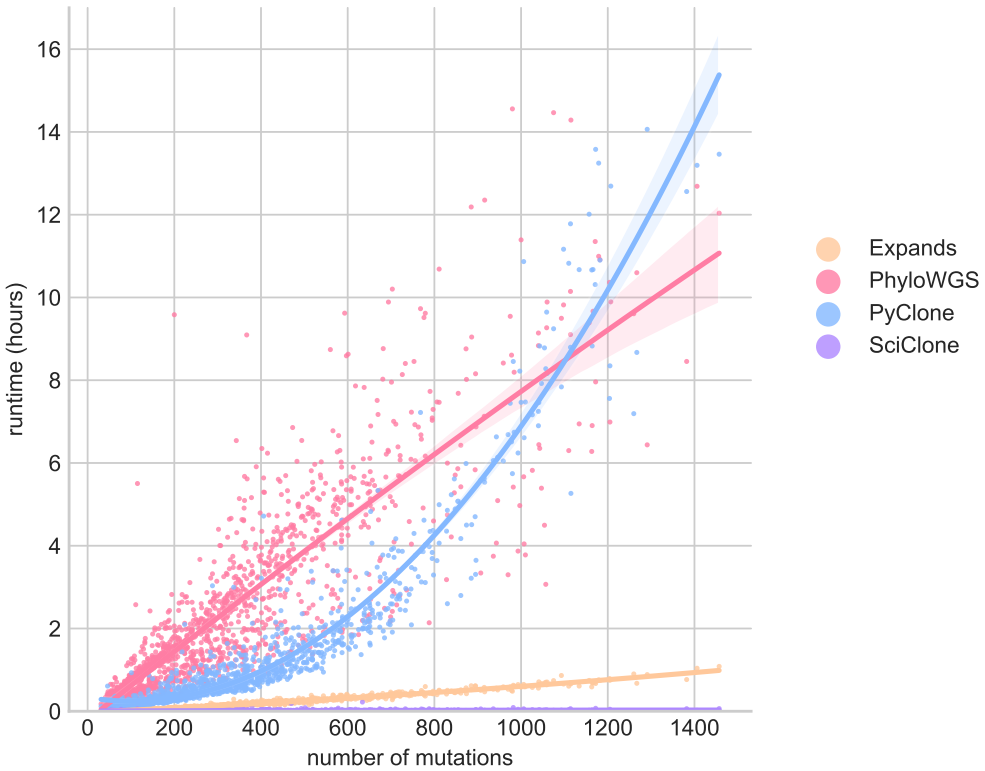

Figure 3.2 – Runtime of the different ITH methods as a function of the number of mutations in each sample. Lines represent second degree polynomial fit with shaded regions are 95% confidence-intervals

3.3.2 Methods quantifying ITH exhibit inconsistent results

As a first evaluation of ITH methods in the absence of ground truth, we assess the agreement between methods, with a focus on the number of clones. Each method except MATH outputs an estimated number S of subclonal populations, ranging from $S = 1$ for an homogeneous, clonal tumor to any positive number for an heterogeneous one. Figure 3.3 presents the distribution of estimated clonality among all samples for each approach and each SNV set, and each copy number calling method. We observe large differences between methods, as well as between SNV sets: for instance, over all samples, the percentage of estimated clonal tumors ($S = 1$) varies from 4% (for PhyloWGS on protected data) to 57 % (for PyClone on public data). Moreover, the number of estimated populations can vary strikingly with the mutation set used, but not really with the different input copy number. There is a clear trend among all methods to yield higher ITH estimates with the protected mutation set. PhyloWGS and EXPANDS (and CSR) are the only methods that detect ITH in almost all tested samples with the protected mutation set.

Another way to compare methods is to consider correlations (Pearson's r) between the estimated numbers of populations. This allows us to include the MATH score in the evaluation, considering it as an increasing function of heterogeneity just like the number of populations. In addition, we add to the comparison 5 measures directly extracted from the NGS analysis, namely, the number of mutations in the protected and in the public sets, the percentage of non-diploid cells (estimated by ASCAT and ABSOLUTE), the purity (estimated by ASCAT and ABSOLUTE), and the *inv_T_cell* (estimated from gene expression signatures). Results are presented in Figure 3.4.

Although a clear and consistent message is hard to extract, a few general trends seem to emerge. First, there is a tendency of results to be more similar for different methods with the same input mutation set, in particular for BRCA, where results for PyClone, SciClone,

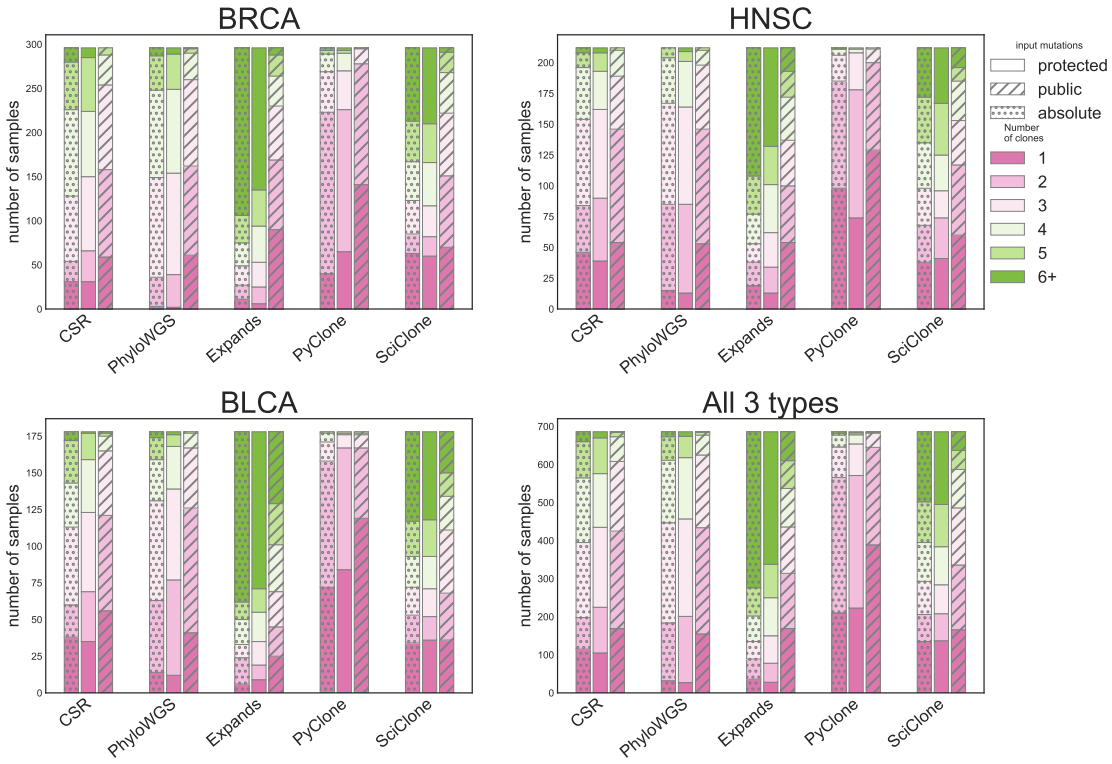

Figure 3.3 – Distribution of number of clones called by ITH methods with public and protected mutations sets as inputs. Distribution of the number of subclones for the tested ITH methods, and 2 alternative input mutation sets for samples in the different cancer types and 2 different copy number methods for the protected mutation set. MATH could not be included in this analysis as this method does not estimate a number of clones.

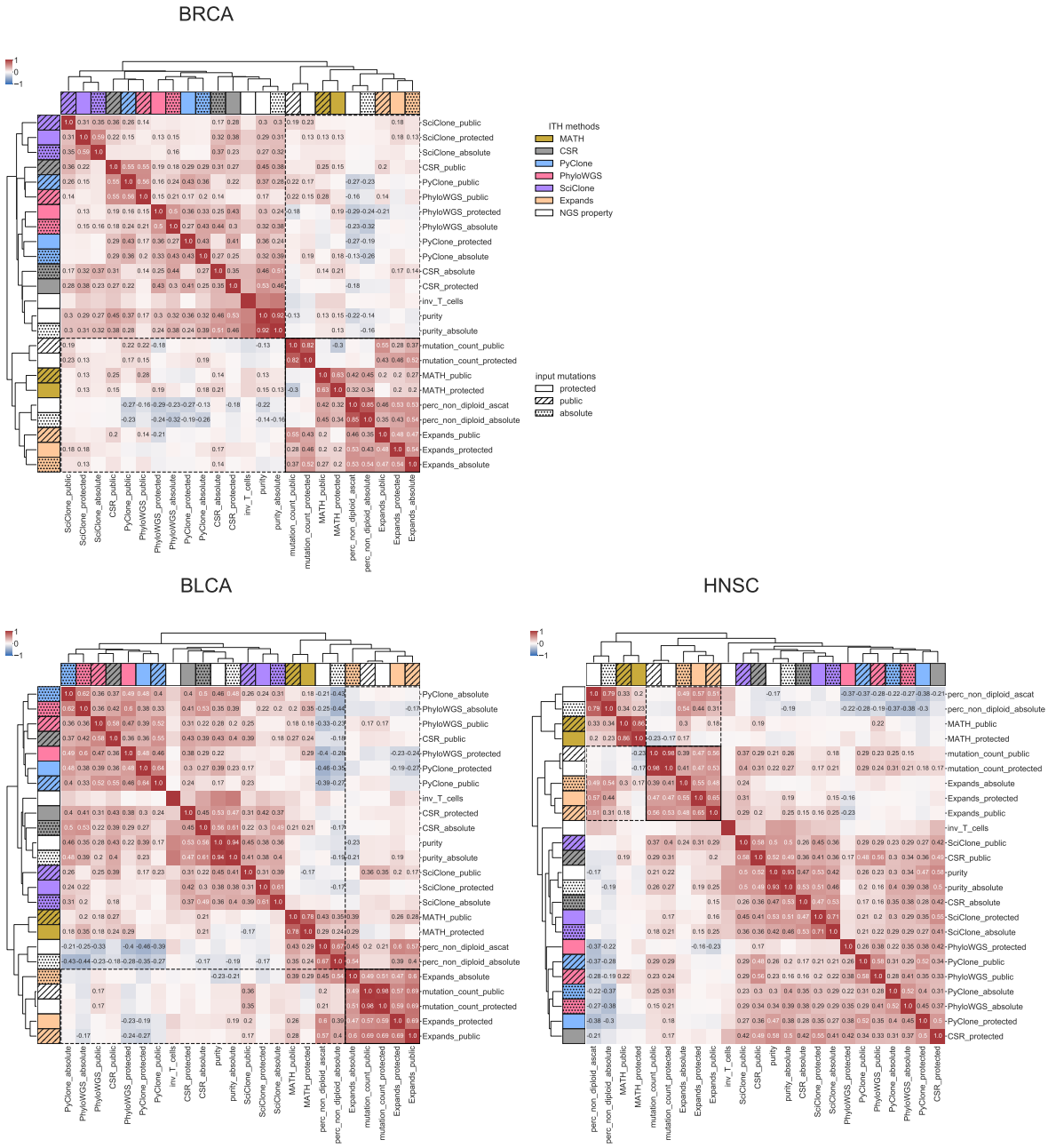

Figure 3.4 – Correlation between various measures of ITH (MATH score, and number of subclones for the other methods), and other potential confounding variables measured using WES and trancriptomics data. Row and color label represent the method used, with white for the genomic measures not involving ITH. Hatches correspond to public mutation sets. Heatmap colors represent the value of the Pearson's r, which is written numerically whenever it is significantly different from 0 ($FDR \leq 0.05$ after Benjamini Hochberg correction for multiple tests). We can observe clustering tendencies stable across the 3 cancer types. One of them is highlighted in black lines.

PhyloWGS and CSR are grouped together for each input set. Second, the really unexpected result is to observe that ITH results with the same input can be uncorrelated, and even significantly negatively correlated. Third, we observe two groups of methods that remain more similar across all three cancer types: EXPANDS and MATH score on the one hand, and PhyloWGS, PyClone, SciClone on the other hand. Those results can be related to the methods themselves. Indeed, PyClone, PhyloWGS and SciClone all define a probabilistic model explaining all observations of copy-number and read counts, based on a mixture model. They differ by the exact nature of the model (choice of distributions, exact definition of parameters), but they have similar structures. SciClone is different from PyClone and PhyloWGS in two ways: it only relies on mutations that are in regions without copy number alterations; and it does not correct for tumor purity. It is therefore not surprising that PyClone and PhyloWGS yield similar results, and that SciClone is a bit more different. CSR performs a consensus of all obtained clusterings; since 3 methods out of 4 have similar results, CSR might be biased towards those 3 methods. Expands makes similar assumptions as PyClone and PhyloWGS. However, the estimation process is very different: Expands estimates a distribution of read number for each position, and then clusters those distributions, while PyClone, SciClone and PhyloWGS attempt to find a common distribution for a group of mutations. The MATH score has an entirely different rationale as it simply ignores CNVs.Similar trends are observed when comparing the methods based on other pairwise comparison metrics (see Supplementary Figure B.11-B.12).

Regarding potential confounding variables, previous studies have reported a correlation between MATH score and CNA abundance [Pereira et al., 2016; Noorbakhsh et al., 2018; Karn et al., 2017], or between purity and ITH, as ITH methods were initially designed to refine purity estimation Carter et al. [2012], and we observe similar behaviors. Association with immune infiltration has also been considered [Karn et al., 2017], though it is worth noting that immune infiltration and tumor purity are not independent, as immune cells are not cancerous. Each group of ITH methods is highly correlated to distinct genomic metrics, mutation load and CNV abundance (`perc_non_diploid`) for the first group (MATH, Expands), and purity (and the opposite of immune cells infiltration (`inv_T_cells`)) for the latter (PyClone, SciClone, PhyloWGS CSR). This might be indicative of systematic biases in the different methods, rather than biological strong signal as previously reported. Indeed, the strength and direction of all correlations vary between the two groups of ITH methods, and is hence hardly reliable or interpretable in terms of clinically actionable information without more data.

Similar results are obtained on an independent dataset of 7 samples from 5 patients where both WES and single cell sequencing was performed. In this dataset, subclonal reconstruction was performed by the method B-SCITE [Malikic et al., 2019] that uses both bulk sequencing and single cell sequencing as input, and provides the most accurate representation possible. To further illustrate the behavior of ITH methods, we have compared results obtained for each sample separately to the B-SCITE result. To evaluate the concordance of each reconstruction to the B-SCITE reconstruction, we compare the number of clones, and the score2A from Salcedo et al. [2018] that evaluates the co-clustering of mutations. The other metrics considered in Salcedo et al. [2018] focus on the distance between the true and reconstructed cancer cell fraction distributions (score 1C), but in this setting, the ground truth does not provide a true CCF distribution estimate, and on the phylogenetic relationships between clones (score 3), but only PhyloWGS provides a tree among the considered methods. For this evaluation, we have left the true estimate for PyClone that provides a lot (several dozens) of clones with a single mutation. The input to ITH methods we have used results from variant calling on the bulk WES data, whereas the input to B-SCITE is more restrictive, and focuses on mutations detected both in the WES and in the single cells; the score2A is computed on the common mutations. Results are presented in Table 3.2. As observed on the TCGA, different methods based on WES data exhibit very different estimates of the number of clones, and none is very close to the estimates of B-SCITE using WES and single-cell data. In terms of clone composition, PyClone is the closest to B-SCITE in terms of score2A correlation in four out or seven samples, although the score2A values remain very modest.

sample	nb clones bscite	nb clones pyclone	nb clones sciclone	nb clones phylowgs	nb clones expands	nb clones CSR	score2A pyclone	score2A sciclone	score2A phylowgs	score2A expands	score2A CSR	MATH score	nb mutations WES	nb mutations metric
CO8_colon	10	20	2	4	6	4	0.000	0.211	0.000	0.146	**0.387**	82.733	272	12
CO8_liver	10	81	6	5	6	3	**0.254**	0.248	0.218	0.241	0.185	80.062	486	17
BRCA_wang_TNBC	13	166	7	3	9	5	**0.465**	0.193	0.269	0.227	0.296	53.939	1458	7
ALL_gawad_P2	7	6	2	4	NA	NA	**0.315**	0.303	0.231	NA	NA	54.918	115	15
CO5_liver	5	29	6	5	3	3	0.060	0.197	0.000	0.143	**0.319**	66.151	271	11
CO5_colon	5	8	6	5	2	3	0.000	**0.180**	0.000	0.093	0.031	70.680	305	10
ALL_gawad_P1	6	18	2	5	NA	NA	**0.345**	0.039	0.138	NA	NA	71.163	110	20

Table 3.2 – Results on the single cell-WES dataset.

3.3.3 ITH is a weak and non robust prognosis factor

To test the prognostic power of each ITH quantification method, we collected survival information for the 686 patients on which all ITH methods ran successfully, and assessed how each ITH method allows to predict survival. Since all ITH methods except MATH output several features related to ITH, we did not test each feature individually but instead estimated a combined score for each method with a survival SVM model (see Methods). More precisely, we extract 5 features from each ITH method: the number of subclonal populations, the proportion of SNVs that belong the the major clone, the minimal cellular prevalence of a subclone, the Shannon index of the clonal distribution, and the cellular prevalence of the largest clone in terms of number of SNVs that enable to distinguish several evolutionary patterns, like early (star-like evolution) or late (tree with a long trunk) clonal diversification (see Methods). We evaluate the performance of each score by 5-fold cross-validation, and prognostic power is assessed on the test fold by computing the concordance index between the SVM prediction and the true patient survival. For MATH, a single feature is computed, so this procedure simply evaluates the concordance index of the MATH score with survival. In addition, we consider a model where all features of all methods (i.e., a total of $6 \times 5 + 1 = 26$ features) are combined together.

Figure 3.5 shows the results for each cancer type, each method, and each set of mutations used.

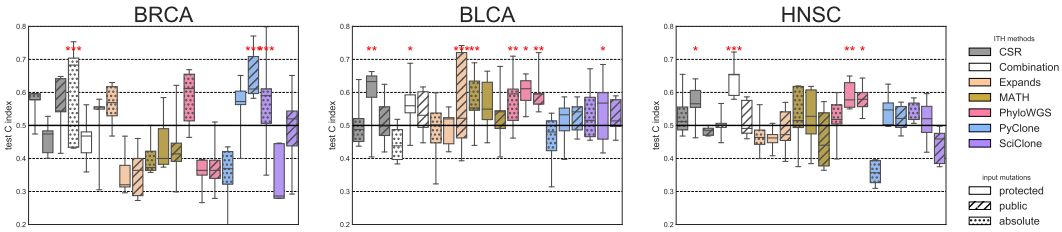

Figure 3.5 – Prognostic power of ITH measured using different ITH method and input mutation set combination. 0.5 corresponds to a random prediction, and stars indicate statistical significance (p-value < 0.001: ***, < 0.01: **, < 0.05: *). Results are presented for 3 cancer types (BRCA, BLCA and HNSC from left to right).

Overall, we observe at least one method achieving significant survival prediction in each cancer type. The combined model is significantly prognostic with both protected and public sets in all three cancer types. Among the three cancer types, in the best case, however, the median concordance index on the test sets barely reaches 0.6 (except with the combination with absolute copy number in BRCA, but with an important variance), which remains modest for any clinical use. This suggests that there may be a weak prognostic signal captured by ITH measurement, but it can not be observed consistently with a single method and a single variant and copy number calling pipeline in the three cancer types, illustrating the frailty of obtained results. The combined model seems to be a robust alternative, as when it is significant, it has a concordance index in the range of the best performing single method; however the case of BRCA seems particular, as many methods perform worse than random.

Some authors, Andor et al. [2016]; Venkatesan and Swanton [2016], have suggested a non-linear relation between survival and ITH, as very high ITH might be damaging for the tumor, while moderate ITH would be associated with aggressive tumors and prone to treatment resistance. To test this hypothesis in our framework we added squared features to the survival model, allowing second order polynomial relations between ITH and survival. However, this

did not significantly impact the results (Supplementary A.1). Indeed, after multiple test correction, only PyClone with the protected mutation set and ABSOLUTE copy number in BRCA prognostic power is increased by adding the squared features ($p = 0.027$, paired t-test), but both CI indexes remain below 0.5. We also assessed whether the relatively poor performance of the different methods was due to the difficulty to learn a prognostic score combining 5 features from limited amounts of training samples, by assessing the prognostic ability of a single feature: the number of clones. A significant improvement was obtained for 7and a significant decrease in performance in 3 of the 36 tested settings (4 methods, 2 mutation sets, 2 copy number methods, 3 cancer types). This suggests that the complexity of the model (polynomial of order 2 instead of linear) and the choice of ITH features have little influence on the results. This might be related to the fact that very little signal can be detected in the first place.

3.3.4 ITH prognosis signal is redundant with other known factors

We have established that in some cases, ITH may exhibit weak prognostic power. It is then very important to assess whether it is complementary to already available prognostic features, like clinical characteristics. To answer this question, we consider relevant clinical features, as described in Methods.

Figure 3.6 presents a comparison between different prediction settings: clinical features without any clonality and clonality associated with clinical features. In all cases, clinical features alone have a significant prognostic power (median CI=0.79 for BRCA, 0.65 for BLCA, 0.65 for HNSC). More importantly, when we combine each ITH feature set with clinical features, we observe no significant improvement over clinical features alone. This suggests that the weak prognostic signal captured by ITH measures is in fact redundant with already available clinical factors.

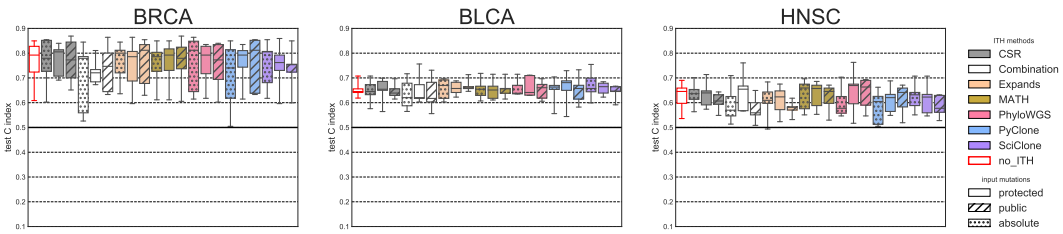

Figure 3.6 – Prognostic power of ITH-derived features compared to other prognostic factors (686 patients in total). ITH-derived features are used in association with clinical features to predict overall survival. Left-most boxplots (with red contour lines) represent results using clinical variables alone, without any ITH, to serve as reference.

3.4 Discussion

3.4.1 Comparison to similar studies

Previous findings report divergent prognostic power for ITH in several pan cancer studies [Andor et al., 2016; Morris et al., 2016; Noorbakhsh et al., 2018]. Andor et al. [2016] analyzed 1,165 patients across 12 cancer types from the TCGA, and found an overall prognostic power by considering all types together, and suggested that this effect might be nonlinear, with a trade-off between ITH and overall survival [Venkatesan and Swanton, 2016]. However, the association between the number of subclones and overall survival was significant with EXPANDS, but not with PyClone results, and no significant association was detected when considering each cancer type separately, except for gliomas. This might be due to the small number of cases of each type (between 33 and 166). Morris et al. [2016] considered 3,383 patients of 9 cancer types from the TCGA and found significant association between the number of subclones found by PyClone in 5 types: HNSC, BRCA, KIRC, LGG, and PRAD. Noorbakhsh et al. [2018] studied 4,722 patients from 11 types from the TCGA, and found

significant prognostic power in 4 types using MATH score and distinct input mutation sets from different variant callers. They obtain significant prognostic association for all variant calling results in only one cancer type: UCEC, and already report some lack of robustness in the results. We have been further in testing up to 7 ITH methods with 2 alternative input mutation sets, in addition to the combination of all methods, and found no significant association, either for the same framework in all considered cancer types, nor for the same cancer type with all frameworks. We have also tested more powerful polynomial models to account for a potential nonlinear relationship, and results were inconclusive. This is an important distinction, because mutation calling can be made robust by additional experiments (targeted sequencing on WXS or WGS candidates), but our results highlight intrinsic limitations of ITH methods.

Considering results in details, there are discrepancies that should be discussed. For BRCA, conclusions are more discordant: Morris et al. [2016] found significant results, Noorbakhsh et al. [2018] did not, and in more specialized studies like METABRIC [Pereira et al., 2016], significant association was found when considering only the upper and lower quartile of MATH score for ER+ tumors. For BLCA, contradictory conclusions were also drawn, as previous studies [Andor et al., 2016; Morris et al., 2016] found no prognostic power and we have with some ITH methods. There are several explanations: each study considered a distinct subset of patients, with a distinct pipeline for calling mutations and measure ITH. This instability with respect to patient selection has been confirmed by our study. All of those studies, including ours, observed ITH prognostic relevance in HNSC. Good prognostic power for HNSC and BLCA might be an indication that the importance of ITH for cancer aetiology differs across cancer types.

3.4.2 Can we truly measure ITH?

Beyond the question of the prognostic power of ITH, our results challenge the very fact that ITH can be measured accurately with one WES sample per patient. Up to 30 methods have been developed to tackle ITH detection and quantification from NGS data in tumor samples, and new ones are still being developed [Eaton et al., 2018]. This analysis has focused on relatively early but among the most widely used ITH methods in order to provide valuable insight on the degree of reliability of provided results. Indeed results presented here show that there is a very weak correlation (and sometimes even a significant negative correlation) between results obtained with different methods on the same patients. Another source of inconsistency is that ITH methods rely on results from previous analysis steps, in particular variant calling. Indeed, all ITH methods rely on the distribution of SNV frequencies, in association or not with structural variants (also called by a variety of dedicated methods). This has already been discussed by Noorbakhsh et al. [2018] for MATH score computation. We show here that this issue is not limited to the MATH score. Some authors have suggested that being very restrictive in variant calling, even resorting to targeted deep sequencing to experimentally validate SNVs [Roth et al., 2014], would exhibit less noisy results. Here we have not observed any evidence that ITH methods estimated more robust results with a restricted input mutation set (i.e. the public mutation set in this study). Overall, lack of agreement between the different ITH measures is a real concern, indicating again that ITH is probably not very accurate. A similar conclusion was recently and independently reached by Bhandari et al. [2018].

Beyond the methods used for ITH inference, the data might also be questioned. Being able to measure ITH to one sample WES with moderate sequencing depth is tempting for future clinical application where the cost and the inconvenience of multiple samples for patients should be limited [Dentro et al., 2017], but it may be unrealistic, as the true heterogeneity of a tumor can be missed by a single biopsy. However, more complex experimental settings have allowed more convincing findings in the field of tumor evolution [Turajlic and Swanton, 2017; Kim et al., 2018], and it may be necessary to further evaluate lack of accuracy due to undersampling from the whole tumor or to use of WES instead of WGS, and the impact of sequencing depth. A recent and broad analysis of ITH with one WGS sample per patient [Dentro et al., 2018] partially answers as the authors could detect ITH in almost every patient, and conduct interesting further analyses as they had confidence in the robustness

of ITH estimates. Most published methods are able to account for multiple samples from the same patient, either sampled at different times or from different regions of the tumor. However, for extension to WGS analysis, our work highlights limitations with respect to the computation time for high numbers of mutation as input.

3.4.3 Association with survival, link with other variables

It is tempting to formulate the hypothesis that higher association with patient survival is a sign of higher accuracy. We have already mentioned some technical issues associated with the setting of one sample WES per patient, as even without measure issues, ITH might just be under-represented in the sample compared to the whole tumor [Shi et al., 2018]. Another limitation is that this does not represent a dynamic measure. For instance a tumor can be clonal because it is not very aggressive, or on the contrary this might be the result of a selective sweep after a phase of new clonal expansion. Moreover, several authors discuss the consequences and the interplay of the presence of distinct subclonal populations, in terms of cooperation [Zhou et al., 2017; McGranahan et al., 2015], competition [Keats et al., 2012; Scott and Marusyk, 2017], or even neutral evolution [Cross et al., 2016; Sottoriva et al., 2017]. Hence, the same level of ITH might uncover very diverse situations, and may not be a prognostic factor by itself.

Moreover, the dataset used in this survival analysis has some particularities: the TCGA has selected patients with criteria allowing high sequencing quality, and ITH analysis itself has further eliminated tumors with no or very high CNA abundance, which may also bias results. Finally, absence of prognosis power in one dataset does not constitute a formal proof that ITH is not associated with survival.

Besides, ITH is likely to be influenced and to interplay with other external factors including tumor micro-environment, immune response, nutrient availability. Recent work has tried to set a full framework for analysis including many factors [Maley et al., 2017]. However, in the case of the TCGA, not all those variables are measurable, but some might be included in further work. In this line of thought, earlier results exhibited correlation of ITH with other factors like CNA abundance, sample purity, immune infiltration [Pereira et al., 2016; Karn et al., 2017; Safonov et al., 2017; Morris et al., 2016]. Our results show that the strength (and even direction in the case of CNA abundance and mutation load) of correlation between those factors and ITH varies between the different tested ITH measures. This again calls for further and more detailed analysis, as results show ambiguity and lack of robustness.

3.4.4 Can we build a gold standard dataset for benchmark?

The main difficulty of ITH estimation is to assess the accuracy of the results. In this work, we have considered two possibilities. The first one on data from the TCGA is to work without any ground truth proxy and measure other features of accuracy: robustness, agreement of results obtained by different methods and association with other clinical variables. The obtained results suggest that the considered ITH methods are relatively robust to changes in the copy number input, but very sensitive to the input mutations. The last two options are more difficult to work with, as one method could be in disagreement with all the others but still provide the most accurate result, and absence or presence of association between ITH and other clinical or genomic variables can be either due to a real biological signal or be an artifact (or bias) of the method. Though the goal of this study is not to provide a formal evaluation of the considered method, the results on the TCGA provide information on systematic trends of each method, and the level of confidence to expect when applying ITH methods.

A second possibility is to try and obtain a proxy for the ground truth. This can be done using single cell sequencing in addition to the bulk sequencing. Though suffering from other issues, single cell sequencing provides true associations or exclusions of mutations, and hence constraints the subclonal reconstruction [Malikic et al., 2019]. However, a large number of cells is necessary. In the 7 samples considered in this study, only a subset of the mutations identified in the bulk sequencing were also identified in single cells, limiting the representativity and the relevance of the extracted accuracy measures. A second possibility is

to rely on several samples from the same tumor to obtain a better ground truth to compare to the result obtained with one sample. However, each sample is a priori heterogeneous itself, requiring a first multi-sample deconvolution. This first step can be challenging, as it is thought that multi-sample reconstruction is subject to a larger statistical bias compared to single sample reconstruction [Caravagna et al., 2019], and the accuracy of this first step will be critical in the final results. A final possibility is to rely on simulated data, which have the major drawback to not be necessarily representative of the true biological data, as recently highlighted for ITH in Caravagna et al. [2019], that point to an aspect of the input data so far overlooked by the community.

Acknowledgments

The authors declare no potential conflicts of interest. We thank Peter Van Loo and Kerstin Haase for sharing ASCAT results on the TCGA, and Christoffer Flensburg for his help with the liftover of ASCAT results. We thank the authors of "Integrative inference of subclonal tumor evolution from single-cell and bulk sequencing data" for help in adapting their results and reprocess the data, and Elodie Girard for her help with NGS data processing. We also thank Alice Schoenauer Sebag for helpful discussions. The results shown here are based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/, under authorization for project 10569 (dbGaP).

Chapter 4

CloneSig: Joint Inference of intra-tumor heterogeneity and signature deconvolution in tumor bulk sequencing data

The content of this chapter has been submitted and is under review.

Abstract

The possibility to sequence DNA in cancer samples has triggered much effort recently to identify the forces at the genomic level that shape tumorigenesis and cancer progression. It has resulted in novel understanding or clarification of two important aspects of cancer genomics: (i) intra-tumor heterogeneity (ITH), as captured by the variability in observed prevalences of somatic mutations within a tumor, and (ii) mutational processes, as revealed by the distribution of the types of somatic mutation and their immediate nucleotide context. These two aspects are not independent from each other, as different mutational processes can be involved in different subclones, but current computational approaches to study them largely ignore this dependency. In particular, sequential methods that first estimate subclones and then analyze the mutational processes active in each clone can easily miss changes in mutational processes if the clonal decomposition step fails, and conversely information regarding mutational signatures is overlooked during the subclonal reconstruction. To address current limitations, we present CloneSig, a new computational method to jointly infer ITH and mutational processes in a tumor from bulk-sequencing data, including whole-exome sequencing (WES) data, by leveraging their dependency. We show through an extensive benchmark on simulated samples that CloneSig is always as good as or better than state-of-the-art methods for ITH inference and detection of mutational processes. We then apply CloneSig to a large cohort of 8,954 tumors with WES data from the cancer genome atlas (TCGA), where we obtain results coherent with previous studies on whole-genome sequencing (WGS) data, as well as new promising findings. This validates the applicability of CloneSig to WES data, paving the way to its use in a clinical setting where WES is increasingly deployed nowadays.

Résumé

La possibilité de séquencer l'ADN des échantillons tumoraux a récemment généré de nombreux efforts dans le but d'identifier les forces qui façonnent la tumorigénèse et la progressin du cancer au niveau génomique. Ces recherches ont permis la compréhension ou l'élucidation de deux apsects importants de la génomique du cancer : (i) l'hétérogénéité intra-tumorale,

que l'on peut déceler par les différences observées dans la fréquence des mutations somatiques d'une tumeur, et (ii) les processus mutationnels, révélés par la distibution des types de mutations somatiques et leur context nucléotidique immédiat. Ces deux aspects ne sont pas indépendants l'un de l'autre, dans la mesure où des processus mutationnels différents peuvent être impliqués dans des sous-clones différents, mais les approches computationnelles qui les étudient ignorent largement cette dépendance. En particulier, les méthodes qui procèdent de façon séquentielle en estimant d'abord la structure sous-clonale de l'échantillon puis en analysant les processus mutationnels à l'œuvre dans chaque clone peuvent facilement passer à côté d'un changement dans les processus mutationnels actifs si la première étape échoue, et inversement, l'information provenant des signatures mutationnelles est ignorées lors de la reconstruction sous-clonale. Pour remédier à ces limitations, nous présentons CloneSig, une nouvelle méthode pour inférer conjointement l'hétérogénéité intra-tumorale et les processus mutationnels dans une tumeur à partir de données de séquençage en masse, y compris d'exome seulement, en mettant à profit leur dépendance. Nous démontrons à l'aide d'une évaluation approfondie sur des données simulées que CloneSig est systématiquement meilleur ou aussi bon que les méthodes de l'état de l'art pour la reconstruction de l'hétérogénéité intra-tumorale et la détection des processus mutationnels. Nous appliquons ensuite CloneSig à une importante cohorte de 8954 échantillons tumoraux pour lesquels des données de séquençage d'exome sont disponibles dans le "Cancer Genome Atlas" (TCGA), pour lesquels nous obtenons des résultats en accord avec de précédentes études menées à partir de données de séquençage de génome complet, mais aussi de nouvelles observations prometteuses. Cela permet de valider l'applicabilité de CloneSig à des données de séquençage d'exome, ouvrant la voie vers son application dans un contexte médical où cette technique est de plus en plus utilisée.

Contents

4.1 Introduction **56**
4.2 Results **57**
4.2.1 Joint estimation of ITH and mutational processes with CloneSig 57
4.2.2 Performance for subclonal reconstruction 58
4.2.3 Performance for signature deconvolution 61
4.2.4 Pan-cancer overview of signature changes 62
4.2.5 Clinical relevance of ITH and signature changes 65
4.3 Discussion **66**
4.3.1 Improved ITH and signature detection in WES 67
4.3.2 Clinical relevance of signature variations 68
4.3.3 Importance of input signatures and challenges 69
4.4 Materials and methods **69**
4.4.1 CloneSig model 69
4.4.2 Parameter estimation 70
4.4.3 Test of mutational signature changes 70
4.4.4 Simulations 70
4.4.4.1 Default simulations 71
4.4.4.2 Simulations for comparison with other ITH and signature methods 71
4.4.4.3 Simulations without signature change between clones 71
4.4.4.4 Simulations to assess the separating power of CloneSig 72
4.4.4.5 Simulations to assess the sensitivity of the statistical test 72
4.4.5 Evaluation metrics 72
4.4.5.1 Metrics evaluating the subclonal decomposition 72
4.4.5.2 Metrics evaluating the identification of mutational signatures 73
4.4.6 Implementation 73
4.4.7 Data 74
4.4.8 Copy number calling and purity estimation 74
4.4.9 Variant calling filtering 74
4.4.10 Construction of a curated list of signatures associated with each cancer type 74
4.4.11 Survival analysis 75

4.1 Introduction

The advent and recent democratization of high-throughput sequencing technologies has triggered much effort recently to identify the genomic forces that shape tumorigenesis and cancer progression. In particular, they have begun to shed light on evolutionary principles happening during cancer progression, and responsible for intra-tumor heterogeneity (ITH). Indeed, as proposed by Nowell [1976], cancer cells progressively accumulate somatic mutations during tumorigenesis and the progression of the disease, following similar evolutionary principles as any biological population able to acquire heritable transformations. As new mutations appear in a tumor, either because they bring a selective advantage or simply through neutral evolution, some cancer cells may undergo clonal expansion until they represent the totality of the tumor or a substantial part of it. This may result in a tumor composed of a mosaic of cell subpopulations with specific mutations. Better understanding these processes can provide valuable insights with implications in cancer detection and monitoring, patient stratification and therapeutic strategy [Dentro et al., 2018; Sottoriva et al., 2015b; Turajlic et al., 2018; Fittall and Van Loo, 2019].

Bulk genome sequencing of a tumor sample allows us in particular to capture two important aspects of ITH. First, by providing an estimate of the proportion of cells harboring each single nucleotide variant (SNV), genome sequencing allows us to assess ITH in terms of presence and proportions of subclonal populations and, to some extent, to reconstruct the evolutionary history of the tumor [Dentro et al., 2017; Roth et al., 2014; Yuan et al., 2018; Deshwar et al., 2015]. This estimation is challenging, both because a unique tumor sample may miss the full extent of the true tumor heterogeneity, and because the computational problem of deconvoluting a bulk sample into subclones is notoriously difficult due to noise and lack of identifiability [Dentro et al., 2017; Shi et al., 2018]. Second, beyond their frequency in the tumor, SNVs also record traces of the mutational processes active at the time of their occurrence through biases in the sequence patterns at which they arise, as characterized with the concept of mutational signature [Alexandrov et al., 2013]. A mutational signature is a probability distribution over possible mutation types, defined by the nature of the substitution and its trinucleotide sequence context, and reflects exogenous or endogenous causes of mutations. Sixty-five such signatures have been outlined [Alexandrov et al., 2018], and are referenced in the COSMIC database, with known or unknown aetiologies. Deciphering signature activities in a tumor sample, and their changes over time, can provide valuable insights about the causes of cancer, the dynamic of tumor evolution and driver events, and finally help us better estimate the patient prognosis and optimize the treatment strategy [Dentro et al., 2018; Fittall and Van Loo, 2019]. A few computational methods have been proposed to estimate the activity of different signatures in a tumor sample from bulk genome sequencing [Alexandrov et al., 2018; Rosenthal et al., 2016].

These two aspects of genome alterations during tumor development are not independent from each other. For example, if a mutation triggers subclonal expansion because it activates a particular mutational process, then new mutations in the corresponding subclone may carry the mark of this process, which may in turn be useful to identify the subclone and its associated mutations from bulk sequencing. Consequently, taking into account mutation types in addition to SNV frequencies may benefit ITH methods. Furthermore, identifying mutational processes specific to particular subclones, and in particular detecting changes in mutational processes during cancer progression, may be of clinical interest since prognosis and treatment options may differ in that case. However, current computational pipelines for ITH and mutational process analysis largely ignore the dependency between these two aspects, and typically treat them independently from each other or sequentially. In the sequential approach, as for example implemented in Palimpsest [Shinde et al., 2018], subclones are first identified by an ITH analysis, and in a second step mutational signatures active in each subclone are investigated. In such a sequential analysis, however, we can not observe changes in mutational signature composition if the initial clonal decomposition step fails to detect correct subclones, and we ignore information regarding mutational signatures during ITH inference. Recently, TrackSig [Rubanova et al., 2018] was proposed to combine these two steps by performing an evolution-aware signature deconvolution, in order to better detect changes in signature activity along tumor evolution. However, while TrackSig overcomes the

need to rely on a previously computed subclonal reconstruction, it does not leverage the possible association between mutation frequency and mutation type to jointly infer ITH and mutation processes active in the tumor. Furthermore, by design TrackSig can only work if a sufficient number of SNV is available, limiting currently its use to whole genome sequencing (WGS) data. This is an important limitation given the popularity of whole exome sequencing (WES) to characterize tumors, particularly in the clinical setting.

In this work, we propose CloneSig, the first method that leverages both the frequency and the mutation type of SNVs to jointly perform ITH reconstruction and decipher the activity of mutational signatures in each subclone. By exploiting the association between subclones and mutational processes to increase its statistical power, we show that CloneSig performs accurate estimations with fewer SNVs than competing methods, and in particular that it can be used with WES data. We show through extensive simulations that CloneSig reaches state-of-the-art performance in subclonal reconstruction and mutation deconvolution from WGS and WES data. We then provide a detailed CloneSig analysis of 8,954 pancancer WES samples from the Cancer Genome Atlas (TCGA), where we recover results coherent with a previous study on WGS [Rubanova et al., 2018] as well as novel promising findings of potential clinical relevance.

4.2 Results

4.2.1 Joint estimation of ITH and mutational processes with CloneSig

We propose CloneSig, a method to jointly infer ITH and estimate mutational processes active in different clones from bulk genome sequencing data of a tumor sample. The rationale behind CloneSig is illustrated in Figure 4.1, which shows a scatter-plot of all SNVs detected by WES in a sarcoma (TCGA patient TCGA-3B-A9HI) along two axes: horizontally, the mutation type of the SNV, and vertically, its cancer cell fraction (CCF) estimated from WES read counts. Following previous work on mutational processes [Alexandrov et al., 2013, 2018], we consider 96 possible mutation types, defined by the nature of the substitution involved and the two flanking nucleotides. Standard methods for ITH assessment and clonal deconvolution only exploit the distribution of CCF values in the sample, as captured by the histogram on the right panel of Figure 4.1, while standard methods for mutational signature analysis only exploit the mutation profiles capturing the distribution of mutation contexts, as represented by the histogram on the bottom panel. However, we clearly see in the scatter-plot that these two parameters are not independent, e.g., C>A mutations tend to occur frequently at low CCF, while C>T mutations occur more frequently at high CCF. CloneSig exploits this association by working directly at the 2D scatter-plot level, in order to jointly infer subclones and mutational processes involved in those subclones. Intuitively, working at this level increases the statistical power of subclone detection when subclones are better separated in the 2D scatter-plot than on each horizontal or vertical axis, i.e., when the activity of mutational processes varies between subclones.

More precisely, CloneSig is based on a probabilistic graphical model [Koller and Friedman, 2009], summarized graphically in Figure 4.2, to model the distribution of allelic counts and trinucleotidic contexts of SNVs in a tumor. These observed variables are statistically associated through shared unobserved latent factors, including the number of clones in the tumor, the CCF of each clone, and the mutational processes active in each clone. CloneSig infers these latent factors for each tumor from the set of SNVs by maximum likelihood estimation, using standard machinery of probabilistic graphical models. Once the parameters of the model are inferred for a given tumor, we can read from them the estimated number of subclones together with their CCF, as well as the set of mutational processes active in each clone along with their strength. In addition, for each individual SNV, CloneSig allows us to estimate the clone and the signature that generated it, in a fully probabilistic manner; for example, in Figure 4.1, each SNV in the scatter-plot is colored according to the most likely mutational signature that generated it, according to CloneSig. Finally, we developed a likelihood ratio-based statistical test to assess whether mutational signatures significantly

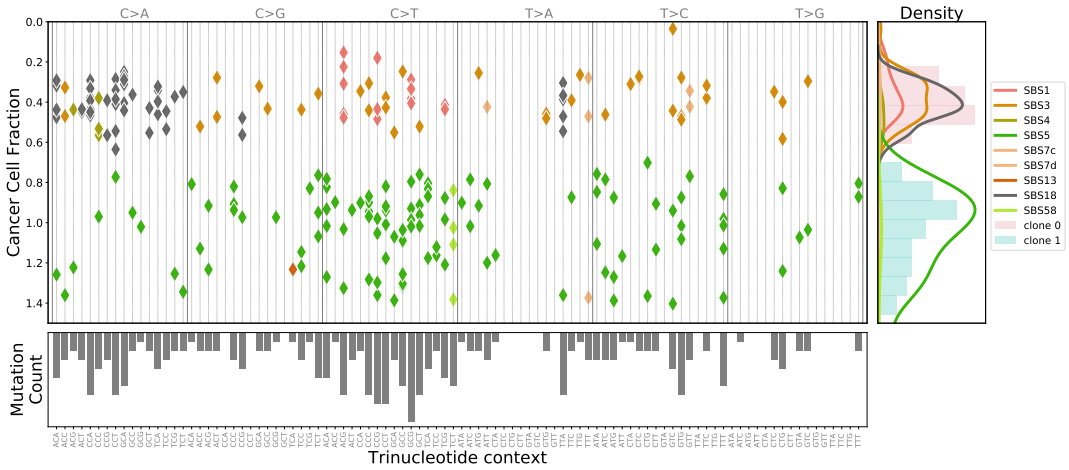

Figure 4.1 – CloneSig analysis of 246 SNVs obtained by WES of a sarcoma sample (patient TCGA-3B-A9HI). The main panel displays all SNVs in 2 dimensions: horizontally the mutation type, which describes the type of substitution together with the flanking nucleotides, and vertically the estimated CCF, as corrected by CloneSig with the estimated mutation multiplicity. From these data CloneSig infers the presence of 2 clones and a number of mutational signatures active in the different clones. Each mutation in the main panel is colored according to the most likely mutational signature according to CloneSig. On the right panel, the CCF histogram is represented and colored with estimated clones, and superimposed with mutational signature density. The bottom panel represents the total mutation type profile. The changing pattern of mutation types with CCF is clearly visible, illustrating the opportunity for CloneSig to perform joint estimation of ITH and signature activity, while most methods so far explore separately those data, considering solely the CCF histogram in the right panel for ITH analysis, or the mutation profile of the bottom panel to infer mutational processes.

differ between subclones, in order to help characterize the evolutionary process involved in the life of the tumor. We refer the reader to the Material and Methods section for all technical details regarding CloneSig.

4.2.2 Performance for subclonal reconstruction

We first assess the ability of CloneSig to correctly reconstruct the subclonal organization of a tumor on simulated data. To simulate data we used the probabilistic graphical model behind CloneSig with a variety of different parameters to investigate different scenarios, leading to a total of 6,300 simulations (see Material and Methods). For each simulation, we run CloneSig and other methods described below, and measure the correctness of the subclonal reconstruction using four different metrics adapted from Salcedo et al. [2018] and described in details in the Material and Method section. Briefly, score1B measures how similar the true and the estimated number of clones are, score1C assesses in addition the correctness of frequency estimates for each subclone, score2A measures the adequacy between the true and predicted co-clustering matrices, and score2C the classification accuracy of clonal and subclonal mutations. We also assess the performance of five other state-of-the-art methods for ITH estimation and compare them to CloneSig. First we evaluate TrackSig [Rubanova et al., 2018], that reconstructs signature activity trajectory along tumor evolution by binning mutations in groups of 100 with decreasing CCFs, and for each group performs signature deconvolution using an expectation-maximization (EM) algorithm. A segmentation algorithm is then applied to determine the number of breakpoints, from which we obtain subclones with different mutational processes. Because of this rationale, the authors recommend to have at least 600 observed mutations to apply TrackSig. For sake of completeness, however, we also apply TrackSig with fewer mutations in order to compare it with other methods in all settings. Second, we test Palimpsest [Shinde et al., 2018], another method which associates mutational signatures and evolutionary history of a tumor. In Palimpsest, a statistical test based on the binomial distribution of variant and reference read counts for each mutation is performed, with correction for copy number, in order to classify mutations as clonal or subclonal. Then,

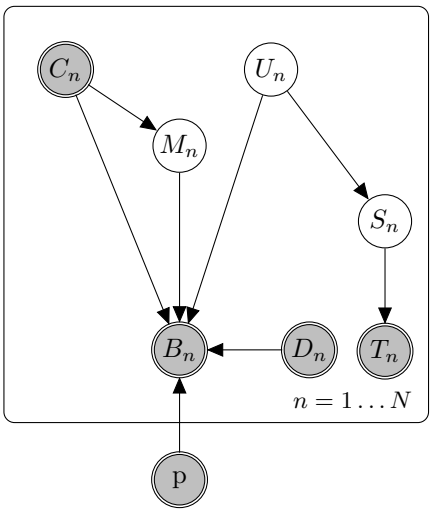

Figure 4.2 – Probabilistic graphical model for CloneSig. This plot summarizes the structure of the probabilistic graphical model underlying CloneSig. Each node represents a random variable, shaded ones being observed, and edges between two nodes describe a statistical dependency encoded as conditional distribution in CloneSig. For a given tumor we observe p, the tumor purity of the sample, and for each SNV, B_n and D_n are respectively the variant and total read counts, C_n is the copy number state, and T_n is the trinucleotide context. Unobserved latent variable include U_n, the clone or subclone where the SNV occurs, S_n, the clone-dependent mutational process that generates the mutation, and M_n, the number of chromosomal copies harboring the mutation. See the main text for details about the distributions and parameters of the model.

for each of the two groups, signature deconvolution is performed using non-negative matrix factorization (NMF). This limitation to two populations can induce a bias in the metrics 1B, 1C and 2A that are inspired from Salcedo et al. [2018], so we introduce the metric 2C to account for the specificity of Palimpsest. Finally, we test three popular methods for ITH reconstruction which do not model mutational processes: PyClone [Roth et al., 2014], a Bayesian clustering model optimized with a Markov Chain Monte Carlo (MCMC) algorithm, Ccube [Yuan et al., 2018], another Bayesian clustering model, optimized with a variational inference method, and SciClone [Miller et al., 2014], also a Bayesian clustering model, optimized with a variational inference method, that only focuses on mutation in copy-number neutral regions.

Figures 4.3 summarize the performance of the different methods according to the different metrics, and under different scenarios, where we vary respectively the number of clones in the simulation (more clones should be more challenging), the number of mutations available (more mutations should help), and the percentage of diploid genome (a higher percentage should be easier). In addition, we provide in Supplementary Section B.2 a more complete benchmark of the different methods when we vary as well the type of mutational signatures used as prior knowledge.

Regarding the estimation of the number of clones (score1B), CloneSig is the best method in all settings, except in the presence of 6 clones. It is in particular the only method achieving a perfect accuracy in identifying samples with one or two clones, and exhibits the best performance for score1B up to 5 clones. Both CloneSig and TrackSig see their performance decrease with the number of clones, as expected, while surprisingly Ccube has the opposite behavior and achieves better results when the number of clones is large. During the experiments we noticed that PyClone tends to find large numbers of clones with only one mutation, so we ignore these clones when we compute score1B in order not to excessively penalize PyClone for this problematic behavior. PyClone, SciClone and Palimpsest have overall a stable performance with varying numbers of clones. Regarding the impact of the number of mutations on score1B, we see that CloneSig outperforms all other methods in all settings. As expected, both CloneSig and TrackSig improve when the number of SNV increases, and

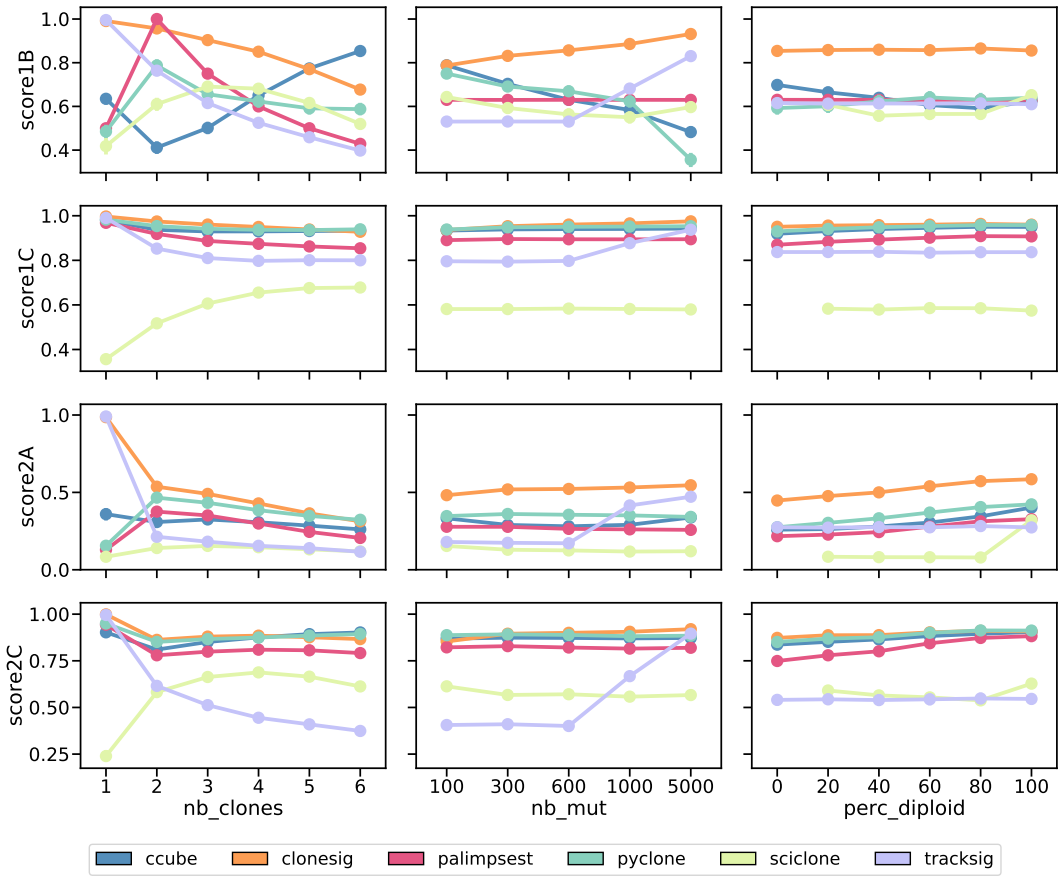

Figure 4.3 – Comparison of CloneSig, TrackSig, Palimpsest, PyClone, SciClone and Ccube for subclonal reconstruction. Each row corresponds to one score, as detailed in the main text. All scores are normalized between 0 and 1, with 1 being the best and 0 the worst. Each column corresponds to a setting where one parameter in the simulation varies: the true number of clones (left), the observed number of mutations (middle), and the diploid proportion of the genome (right). Each point represents the average of the score over all available simulated samples. Bootstrap sampling of the scores was used to compute 95% confidence intervals.

we confirm that TrackSig requires at least 1,000 SNVs to be competitive with other methods in this experiment, while CloneSig reaches the best performance of TrackSig with as few as 100 SNVs. A surprising result is that for PyClone, SciClone and Ccube, score1B decreases with the number of observed mutations, which may suggest a bad calibration of the clone number estimate for large numbers of SNVs; for CloneSig we designed a specific, adaptive estimator for the number of clones since we observed that standard statistical approaches for model selection perform poorly in this setting (see Material and Methods and Supplementary Section B.1.2). The percentage of diploid genome has no visible impact on the performance of any method. Regarding score1C, which focuses not on the number of clones estimated but on their ability to correctly recapitulate the distribution of CCF values, we also see that CloneSig outperforms all other methods in all settings, while PyClone and Ccube are not far behind. TrackSig performs slightly worse, especially as the number of clones increases, but this may be explained by its poor performance when the number of mutations is too low, as performance matches the other methods for 5,000 mutations. Palimpsest has comparatively a relatively poor performance, and seems particularly impacted when the proportion of diploid regions decreases. Indeed, the number of mutated copies in Palimpsest is made under the assumption that the CCF for the mutation is 1, which may jeopardize the correct detection of subclonal mutations. Finally, SciClone is clearly the worse method for score1C, particularly with 1 to 3 clones.

Besides the ability of different methods to reconstruct the correct number of subclones

and their CCF, as assessed by score1B and score1C, we measure with score2A their ability to correctly assign individual mutations to their clones, an important step for downstream analysis of mutations in each subclone. According to score2A, CloneSig outperforms all other methods in all scenarios, illustrating the improved accuracy of accounting for both CCF and mutational signatures when achieving ITH reconstruction. For all methods, score2A decreases when the number of clones increases and when the percentage of diploid genomes decreases, as expected, but the relative order of methods does not change, with CloneSig followed by a group of three methods with similar performances: PyClone, Ccube and Palimpsest. SciClone performs poorly except when the genome is fully diploid, in which case it gets competitive with Palimpsest but still below CloneSig, PyClone and Ccube. The number of mutations has a limited impact on the performance of all methods except for TrackSig, which only becomes competitive after 1,000 mutations. CloneSig with 100 mutations still outperforms TrackSig with 1,000 mutations, though. Finally, when we assess the capacity of each method to simply discriminate clonal from subclonal mutations using score2C, a measure meant not to penalize Palimpsest which only performs that task, we see again that CloneSig is the best in all scenarios, closely followed by Ccube and PyClone, as well as TrackSig with 5,000 mutations. Palimpsest is a bit below these methods, while SciClone and TrackSig with 1,000 mutations or less are clearly not competitive for this metric.

Overall, these experiments show that CloneSig performs as well as or better than the state-of-the-art according to all metrics considered and in all simulated scenarios, confirming that accounting for the mutation type for each mutation, in addition to its CCF, improves the accuracy of subclonal reconstruction. We also confirm that TrackSig, the only existing method that combines CCF and mutational signature information to detect subclones, requires at least 1,000 mutations to obtain results competitive with other methods in our benchmark, while CloneSig reaches good accuracy in all scores with as few as 100 mutations.

CloneSig, like TrackSig, benefits from situations where mutational processes are not similarly active in different subclones to better detect them and assign individual mutations to them. As expected, we observe for example that the improvement of CloneSig over other methods in terms of score2A fades when there is no difference of signature activity between clones, with CloneSig performing as well as PyClone and Ccube in this situation (Supplementary Figure B.12). To further illustrate the interplay between signature change and ability to detect clones, we now test CloneSig on simulations with exactly two clones, and where we vary how the clones differ in terms of CCF, on the one hand, and in terms of mutational processes, on the other hand (quantified in terms of cosine distance between the two profiles of mutation type). Figure 4.4 shows the accuracy of the number of clones detected by CloneSig as a function of these two parameters. We see an increased number of cases where the two clones are correctly distinguished by CloneSig as the distance between the mutation type profiles increases, for a constant CCF difference. For example, when two clones have similar signatures (small cosine distance), they can be detected with a 50% accuracy when the difference between their CCF is around 0.3; when their signatures are very different (large cosine distance), they can be detected with the same accuracy when their CCF only differ by 0.1. We show in Supplementary Figure B.58 how other parameters (number of mutations, sequencing depth, diploid proportion of the genome) also impact the performance of CloneSig in this setting.

4.2.3 Performance for signature deconvolution

In addition to ITH inference in terms of subclones, CloneSig estimates the mutational processes involved in the tumor and in the different subclones. We now assess the accuracy of this estimation on simulated data, using six performance scores detailed in the Material and Methods section. In short, score_sig_1A is the Euclidean distance between the normalized mutation type counts and the reconstructed profile (activity-weighted sum of all signatures); score_sig_1B is the Euclidean distance between the true and the reconstructed profile; score_sig_1C measures the identification of the true signatures; score_sig_1D is the proportion of signatures for which the true causal signature is correctly identified; and score_sig_1E reports the median of the distribution of the cosine distance between the true and the predicted mutation type profile that generated each mutation. We compare Clone-

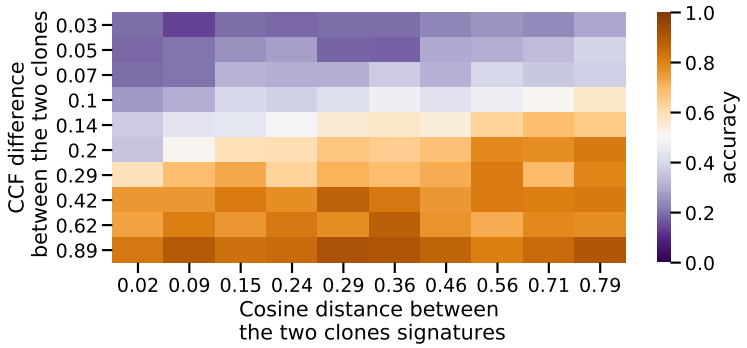

Figure 4.4 – Accuracy of correctly estimating the presence of two clones by CloneSig as a function of the difference in the CCF between the two clones (vertical axis), and of the cosine distance between their mutational profiles. The accuracy denotes the proportion of runs where CloneSig rightfully identifies two clones.

Sig to the two other methods that perform both ITH and mutational process estimation, namely, TrackSig and Palimpsest, and add also deconstructSigs [Rosenthal et al., 2016] in the benchmark, a method that optimizes the mixture of mutational signature of a sample through multiple linear regressions without performing subclonal reconstruction.

Figure 4.5 shows the performance of the different methods according to the different metrics. For Score_sig_1A and Score_sig_1B, all methods exhibit overall similar performances, with a small advantage for CloneSig and TrackSig over Palimpsets and deconstructSigs in several scenarios. For Score_sig_1C, CloneSig and TrackSig exhibit the best AUC to detect present signatures. It may be related to a better sensitivity as CloneSig and TrackSig perform signature deconvolution in smaller subsets of mutations. All methods perform similarly with respect to Score_sig_1D, with CloneSig slightly better than all methods in all settings. The median cosine distance (Score_sig_1E) is also slightly better for CloneSig than for other methods in all settings. Surprisingly, the performance for TrackSig is worse with 5000 mutations; we observed on a few examples that this may be due to the fact that TrackSig tends to find several change points for a single clone change, due to the gradual change in activities along CCF in the overlap zone between two clones.

Overall, as for ITH inference, we conclude that CloneSig is as good as or better than all other methods in all scenarios tested. Further results where we vary other parameters in each methods, notably the set of mutations used as inputs or the set of signatures used as prior knowledge, can be found in Supplementary Section B.2; they confirm the good performance of CloneSig in all settings tested.

4.2.4 Pan-cancer overview of signature changes

We now use CloneSig on real data, to analyze ITH and mutational process changes in a large cohort of 8,954 tumor WES samples from the TCGA spanning 31 cancer types. An overview of the main characteristics of the cohort is presented in Table B.3.

For each sample in the cohort, we estimate with CloneSig the number of subclones present in the tumor, the signatures active in each subclone, and test for the presence of a signature change between clones. Figure 4.6 shows a global summary of the signature changes found in the cohort. For each cancer type, it shows the proportion of samples where a signature change is found, and a visual summary of the proportion of samples where each individual signature is found to increase or to decrease in the largest subclone, compared to the clonal mutations. The thickness of each bar, in addition, indicates the median change of each signature. Overall, CloneSig detects a significant change in signature activity from the protected set of mutations in 32% of all samples, and in 11% when it is trained on the public set of mutations, although these proportions vary between cancer types. In terms of signature changes, we recover patterns already observed in other cohorts, usually using WGS, which confirms that CloneSig is able to detect patterns of ITH and signature activity change using WES data. For example, similarly to the cohort of 2,778 WGS tumors analyzed by the International Cancer Genome

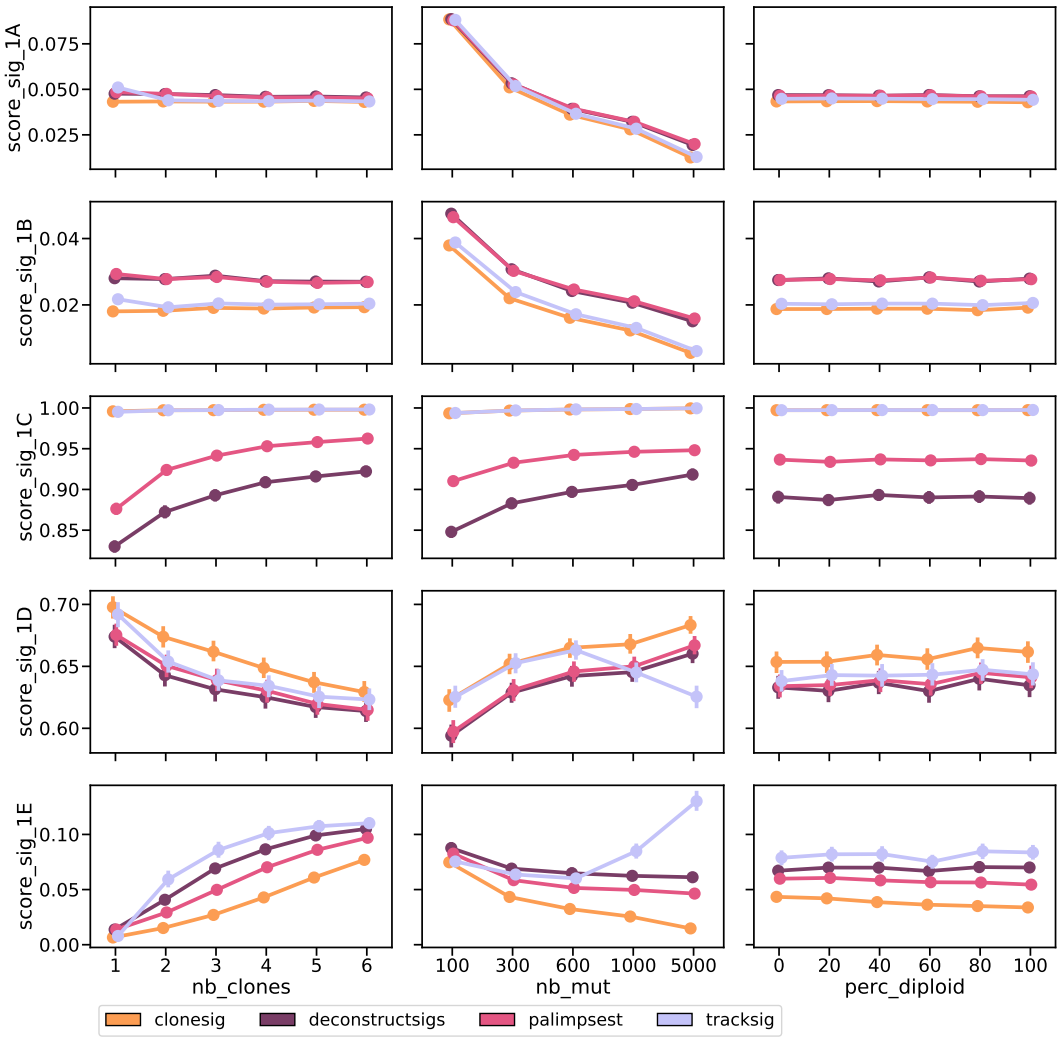

Figure 4.5 – Comparison of CloneSig, TrackSig, Palimpsest, and deconstructSigs for signature deconvolution. Several metrics have been implemented, and are detailed in the main text. Scores 1A, 1B and 1E (respectively first, second and fifth rows) are distance and are better when close to 0, while scores 1C and 1D (respectively third and fourth rows) are normalized between 0 and 1 and are better when close to 1. The results are presented depending on several relevant covariates: the true number of clones (left), the number of mutations (middle), and the diploid proportion of the genome (right). Each point represents the average of the score over all available simulated samples. Bootstrap sampling of the scores was used to compute 95% confidence intervals.

Consortium's Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative which represents the largest dataset of cancer WGS data to date [Dentro et al., 2018], we observe that signature 5, of unknown aetiology, varies in almost all cancer types, and can be both increasing or decreasing. Lifestyle-associated signatures associated with tobacco-smoking (signature 4) and UV light exposure (signature 7) decrease systematically in lung tumors and oral cancers and skin melanoma respectively.

More precisely, patterns of change detected by CloneSig on the TCGA are similar to what was described on the PCAWG cohort for cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), glioblastoma multiforme (GBM), uterine carcinosarcoma (UCS) and uterine corpus endometrial carcinoma (UCEC), kidney chromophobe (KICH), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), skin cutaneous melanoma (SKCM) and stomach adenocarcinoma (STAD). In addition, CloneSig detects several new patterns of variations. In bladder carcinoma (BLCA), signature 3, related to defective homologous recombination-based DNA damage repair is found increasing. In breast cancer

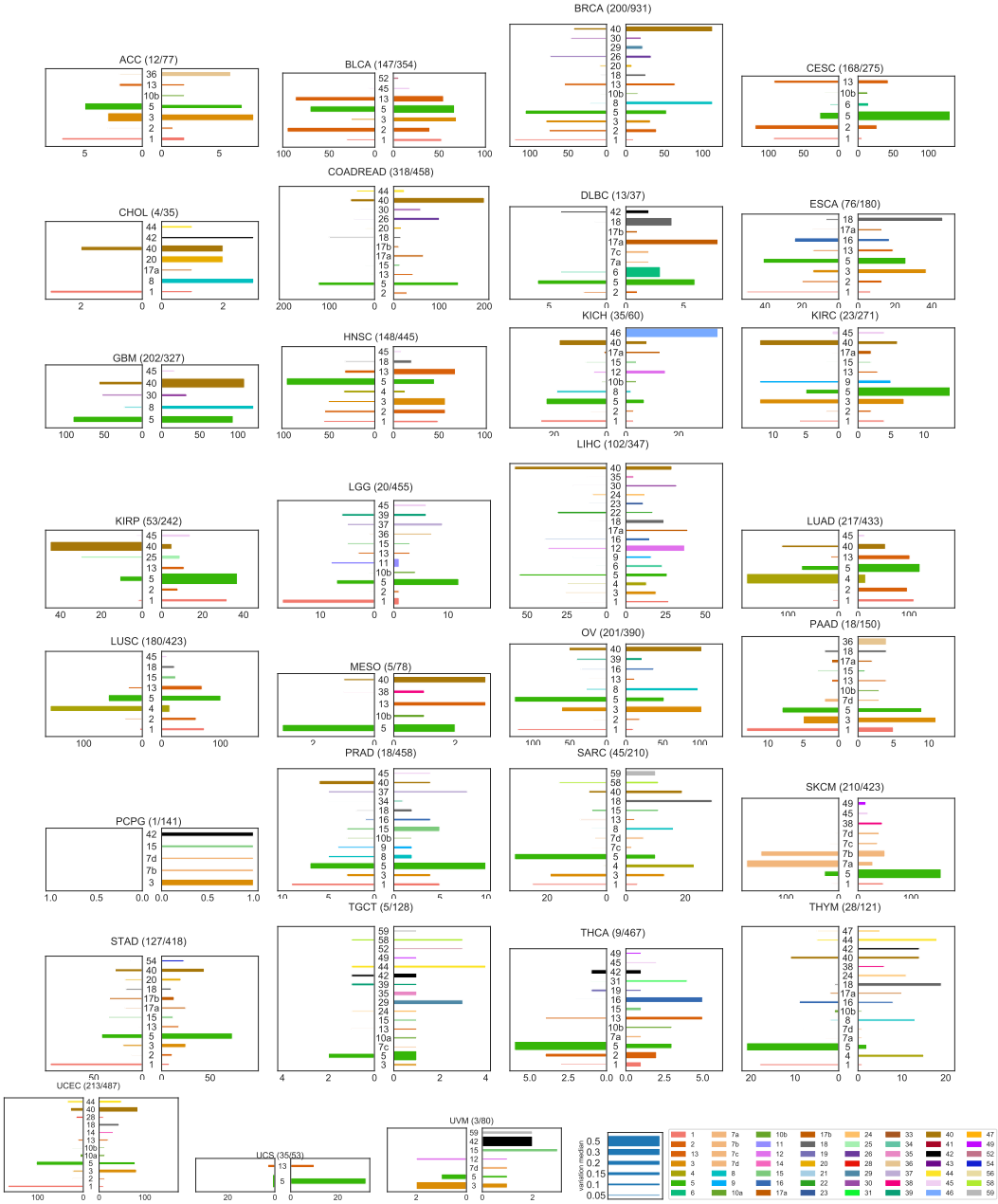

Figure 4.6 – Mutational signature changes in the TCGA cohort. Each plot corresponds to one cancer type, indicates the number of samples with a significant signature change compared to the total number of samples, and shows on the right panel an increase of a signature in the largest subclone, compared to clonal mutations, and on the left panel a decrease. The length of each bar corresponds to the number of patients with such changes, and the thickness to the median observed change.

(BRCA), CloneSig detects three new signature variation patterns: signature 8 is increasing, and signatures 26 and 30 are varying in both directions, while signatures 1 (deamination of 5-methylcytosine to thymine) and 18 (possibly damage by reactive oxygen species) tend to be preferentially decreasing and increasing respectively, instead of varying in both directions according to Dentro et al. [2018]. In prostate adenocarcinoma (PRAD), CloneSig finds signature 3 to be varying in both direction, contrary to solely increasing in Dentro et al. [2018], but similarly to the findings of Espiritu et al. [2018]. Signature 37 is found to vary in both directions instead of decreasing. Additionally to changes identified in Dentro et al. [2018], but already described in Espiritu et al. [2018], CloneSig detects variations in signatures 8, 9 and 16. A new signature seems to exhibit variations along tumor evolution: signature 15

(defective DNA mismatch repair), which was not previously described in PRAD to the best of our knowledge. In lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), we observe the important increase in signature 17 as in Dentro et al. [2018], but no variation of signature 9 (mutations induced during replication by polymerase η), and an undescribed increase in signatures 18 and 6 (defective DNA mismatch repair). In esophageal carcinoma (ESCA), we do not observe the important decrease of signature 17 [Dentro et al., 2018], however, we describe an increase of signature 18 and a variation of signature 16 in both directions. For head-neck squamous cell carcinoma (HNSC), we observe similar patterns for signatures 5, 2 and 13 (related with APOBEC enzymes activity), and 18, but an undescribed increase of signature 3 [Dentro et al., 2018], and a decrease of signature 4 (related to tobacco smoking), probably in relation to the fact that this cohort includes oral tumors. In ovary tumors (OV), increase of signature 40 and decrease of signature 5 are coherent with the findings of Dentro et al. [2018], however, CloneSig finds an important number of samples with an increase of signature 8, while a decrease of this signature was reported in Dentro et al. [2018]. For thyroid carcinoma (THCA), the variations of signatures found are different, however the number of samples with a significant change of signature activity is small. In liver hepatocellular carcinoma (LIHC), and pancreatic adenocarcinoma (PAAD), we report important differences between patterns, in particular with signature 12 reported to decrease systematically in LIHC [Dentro et al., 2018] while we observe an increasing trend, and no variation of signature 40 in PAAD. In colorectal cancer (COADREAD), we observe as described in Dentro et al. [2018] an strong increase in signatures 40 and 17, and a decrease in signature 18, a variation of signature 5 in both direction, and not only an increase, and no variation of signature 1. We also observe an increase in signature 26, observed in one of the three samples analyzed with single cells in Roerink et al. [2018], and an increase in signature 30 that was not previously reported.

In addition, CloneSig detects changes in signature activity in cancer types where they have not yet been characterized to the best of our knowledge, though the number of samples is too low in some cases to detect a strong trend. In adrenocortical carcinoma (ACC), we observe an increase in signature 36 (associated to defective base excision repair) and variations in signature 3. In kidney renal papillary cell carcinoma (KIRP) and kidney renal clear cell carcinoma (KIRC), signature 40 is strongly decreasing, and signature 5 increasing. Additionally CloneSig uncovers variations in signature 3 in most samples with a signature change in KIRC; activity of signature 3 in KIRC was previously outlined in Warsow et al. [2018].

4.2.5 Clinical relevance of ITH and signature changes

We now explore relations between the ITH detected by CloneSig and the potentially associated changes in signature activity and relevant clinical features. Looking first at the pan-cancer scale, we assess whether ITH measured either through the number of detected subclones or the presence of mutational signature changes is associated to overall survival. For that purpose, we split all TCGA samples in three groups using two different strategies, based on CloneSig's output on the protected input mutation set. In the first strategy, the three groups are based on the number of (sub-)clonal populations only (1, 2 or 3+ clones). A multivariate Cox model fitted to the data indicates for 2 clones a hazard ratio (HR) of 1.25 (95% confidence interval (CI): $[1.14, 1.37]$, $p = 2.27\text{e} - 6$), and for 3 clones a HR of 1.41 (CI= $[1.26, 1.58]$, $p = 2.03\text{e} - 9$). A univariate Cox model fitted to compare the populations with 2 or 3+ clones indicates a HR of 1.12 for 3+ clones (CI=$[1.02, 1.23]$, $p = 0.022$). This confirms that the presence of subclones as estimated by CloneSig is associated to survival, but that the difference between 2 and 3+ clones is limited in terms of survival. In the second strategy, we still keep the group of samples with only a single clone, but split the other samples (with 2 or more clonal populations) into two groups based on whether or not CloneSig detects a change in mutational signatures. The Cox results shows a HR of 1.14 without signature change (CI= $[1.04, 1.26]$, $p = 7.11\text{e}-3$), and 1.51 with signature change (CI= $[1.37, 1.67]$, $p = 3.30\text{e}-16$). With a focus on heterogeneous tumors only, the hazard ratio with a signature change compared to those without signature change is 1.33 (CI= $[1.22, 1.44]$, $p = 5.22\text{e} - 11$). As with the first strategy, we observe a significant difference in survival between patients with homogeneous and heterogeneous tumors. However, the presence of a significant change

in signature activity (second strategy) is more strongly associated to survival among heterogeneous tumors, compared to the case when we split the heterogeneous tumors based on the number of clones (Figure 4.7). We get similar results when using the public input mutation set (Supplementary Figure B.59), illustrating CloneSig's robustness to the input signatures, and ability to detect ITH and signature activity changes with a very small number of observed mutations.

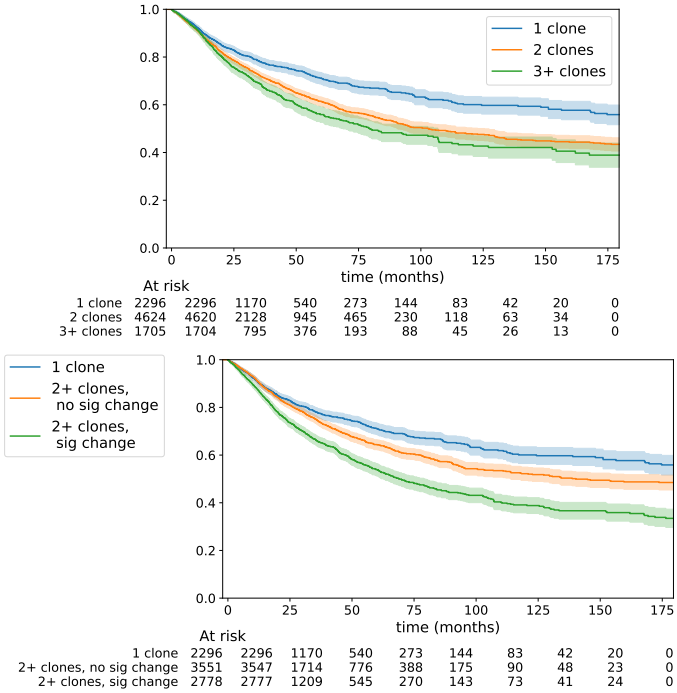

Figure 4.7 – Kaplan-Meier curves for all TCGA samples (8,954 patients with available survival data) distinguishing tumors only along the number of clones (left) or along the number of clones and the presence of a significant change in signatures along tumor evolution (right) using the protected input mutation sets. A multivariate Cox model was fitted in both cases, and indicates for 2 clones, hazard ratio (HR) of 1.25 (95% confidence interval (CI): $[1.14, 1.37]$, $p = 2.27\text{e} - 6$), and 3 clones (HR= 1.41, CI= $[1.26, 1.58]$, $p = 2.03\text{e} - 9$). Considering only heterogeneous tumors, the Cox model results in a HR of 1.12 (CI=$[1.02, 1.23]$, $p = 0.022$) for 3+ clones compared to 2 clones (left). For the distinction based on signature change, without signature change (HR= 1.14, CI= $[1.04, 1.26]$, $p = 7.11\text{e} - 3$), and with signature change (HR= 1.51, CI= $[1.37, 1.67]$, $p = 3.30\text{e} - 16$). For heterogeneous tumors with a signature change, compared to without, the HR is 1.33 (CI= $[1.22, 1.44]$, $p = 5.22\text{e} - 11$) (right)

When considering the same survival analysis for each cancer type separately, we find no significant difference in survival between the different groups (homogeneous and heterogeneous tumors) after correcting for multiple tests. This may be due both to a lack of statistical power in the cancer-specific analysis because of the smaller number of samples available when we split them per cancer types, and to a confounding effect of cancer types where, for example, cancer types with a bad prognosis are enriched in heterogeneous tumors with a significant change in signature activity. Indeed, as shown in Figure 4.8, the proportion of tumors harboring ITH and changes in mutational processes varies a lot between cancer types. Finally, we also investigate whether patient stratification based on CloneSig output, in particular ITH and patterns of signature changes, is correlated with other clinical characteristics such as sex, age, tumor size or grade, but find overall no significant association; for sake of completeness we present detailed results of this analysis in Supplementary Section B.3.

4.3 Discussion

In recent years, a large number of methods have been developed to unravel ITH in tumors [Roth et al., 2014; Miller et al., 2014; Yuan et al., 2018; Turajlic et al., 2015; Dentro et al.,

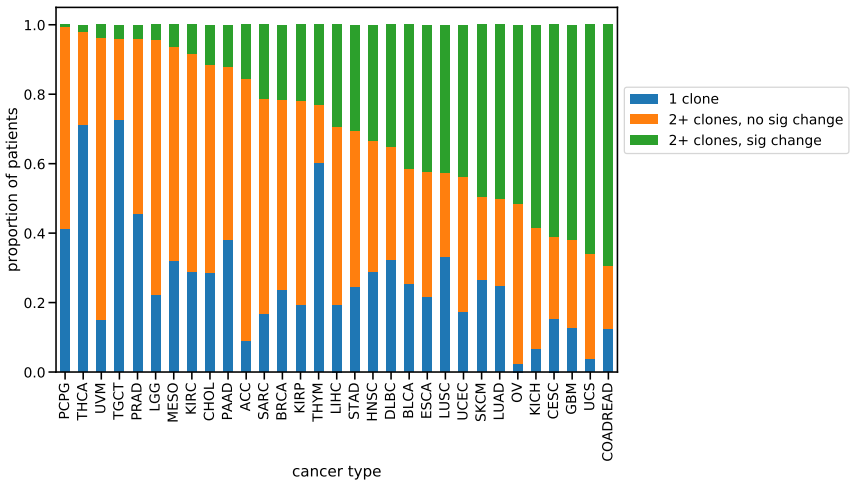

Figure 4.8 – Proportion of patients among the three categories: "1 clone", "2 clones and more without a change in signature activity", and "2 clones and more with a change in signature activity" for the different cancer types considered in the TCGA. Cancer types are sorted from left to right by increasing proportion of heterogeneous samples with change in signature activity.

2017], and have been applied to different cohorts, including the TCGA. Recent analyses illustrate limits encountered when applying those methods to bulk WES [Abécassis et al., 2019; Shi et al., 2018], as the number of observed mutations is small, the variance in read counts can be high, and a unique sample may miss the heterogeneity of the tumor. As sequencing costs are continuously decreasing, WGS, multi-sample sequencing and single cell sequencing will constitute relevant alternatives and simplify the study of ITH. However, to date a much larger number of tumor samples with sufficient clinical annotation (in particular survival data) is available with WES compared to other more advanced technologies, and can lead to interesting insights. Beyond the number of clones present in a tumor, another relevant aspect of tumor evolution is the presence of changes in mutational signatures activities [Fittall and Van Loo, 2019], which could have clinical implications in cancer prevention and treatment, and unravel the evolutionary constraints shaping early tumor development. To the best of our knowledge, TrackSig [Rubanova et al., 2018] and Palimpsest [Shinde et al., 2018] are the only methods addressing the problem of systematic detection of signature changes, but they both present serious limitations: Palimpsest first detects ITH, and then performs signature deconvolution, which has the major drawback that if this first step fails, no signature change can be detected. Moreover, Palimpsest simply aims to distinguish subclonal from clonal mutations, thus ignoring more complex patterns. TrackSig is only applicable to WGS data, and though avoiding the caveat of relying on a previous detection of ITH, the final step of associating signature changes to the subclonal reconstruction is manual. Finally, none of these methods leverages the changes in signature activity to inform and improve the ITH detection step. To overcome these limitations, we have developed CloneSig, the first method to offer joint inference of both subclonal reconstruction and signature deconvolution, which can be applied to WGS as well as to WES data.

4.3.1 Improved ITH and signature detection in WES

CloneSig is a generative probabilistic graphical model that considers somatic mutations as derived from a mixture of clones where different mutational signatures are active. We demonstrated with a thorough simulation study the benefits of the joint inference in detecting ITH, both in WES and WGS samples. We showed that CloneSig is competitive with or outperforms state-of-the art ITH methods, even in the absence of signature activity change between the clones, and is particularly efficient for the detection of samples with one or a few subclones. Interestingly, several other methods we considered including PyClone [Roth et al., 2014], SciClone [Miller et al., 2014] and Ccube [Yuan et al., 2018], are fully Bayesian and

choose the number of clones by maximizing of the posterior probability of the data. In those methods the prior has a regularizing role, and they exhibit a decrease of accuracy as the number of observed mutations increases. This may be related to the fact that the regularizing prior is less influential as more mutations are taken into account. We instead developed a specific adaptive criterion to estimate the number of clones, as we observed that standard statistical tools for model selection performed poorly in preliminary experiments.

When applied to real data, CloneSig's results on the TCGA exhibit a strong association with survival when comparing homogeneous and heterogeneous samples. This effect on survival is stronger than the one reported in Andor et al. [2016], also on the TCGA. This may be due to a better accuracy of CloneSig, as well as to the better statistical power of our analysis with larger sample sizes. Regarding the signature deconvolution problem, results on simulations (Score_sig_1C) suggest that CloneSig exhibits an improved sensitivity. Application to the TCGA also indicates such increased sensitivity: in the TCGA pancreatic ductal adenocarcinoma cohort (PAAD), the original study using deconstructSigs could not detect signature 3 activity in samples with somatic subclonal mutations in genes BRCA1 and BRCA2 [Raphael et al., 2017], while CloneSig reports signature 3 exposure in some PAAD tumors.

4.3.2 Clinical relevance of signature variations

An original result of this study is the ability to further stratify heterogeneous tumors based on the presence of a significant change in signature activity, which seems associated with a worse prognosis. This could be illustrative of a more advanced stage of tumor development where a new generation of driver events supplant the initial drivers of the tumor. However, we could not reproduce those results observed on the whole TCGA cohort in a cancer-type-specific way. There are several possibilities explaining this observation: smaller cohorts may lack statistical power, or there could be a confounding effect where larger proportions of cancer types of bad prognosis are heterogeneous and have a significant change in signature activity compared to cancer types with better prognosis. Even in this latter hypothesis, this stratification can still be the manifestation of a true biological process, and not just an artifact. Indeed, other factors may explain this phenomena, like systematic later diagnosis.

To further assess the clinical relevance of signature changes, we have systematically analyzed whether we could identify an association between the exact pattern of signature change and clinical variables, but found no significant association. However, more refined or complete analyses may be necessary to uncover the full significance of signature activity changes. Previous studies report important signature activity differences between early and metastatic tumors in endometrial and breast cancers [Ashley et al., 2019; Bertucci et al., 2019], with impact on the survival in the breast cancer study [Bertucci et al., 2019]. We could not perform a similar analysis using the TCGA with only untreated primary tumors, but this constitutes new directions and opportunities of research using CloneSig on metastatic cohorts, for instance to refine findings of Bertucci et al. [2019], that compares signatures deconvoluted from the whole metastasis, and could benefit from subclonal analysis to distinguish early and late mutations.

A final potential clinical application could be usage as a marker for personalized treatment. Signature 3 is associated with homologous recombination repair defect (HRD), and a targeted therapy, PARP inhibitors, can successfully target cells with such defect. A first idea is to use detection of signature 3 to identify patients that can benefit from such therapy, and CloneSig exhibits better identification of active signatures, as illustrated in the simulation studies. Indeed, several mutations in genes like BRCA1 and 2, RAD51 are known to cause HRD, but some other mutations are less frequent, or other events may result in HRD and be undetectable using regular genome sequencing, such as epigenetic inactivation [Knijnenburg et al., 2018]. In addition, the intensity of HRD mutational process may be predictive of the treatment response. Pursuing this line of thought, the change in signature activity can also be exploited as an indicator of the current driver status of HRD in tumor development. As the underlying processes of signatures will keep being uncovered, more examples of such applications are likely to arise.

4.3.3 Importance of input signatures and challenges

As illustrated in simulations, and based on our experience with the TCGA, the choice of the input signatures is key to CloneSig's optimal performances. This is related to the unidentifiability of the signature deconvolution problem. Several solutions have been proposed: use of a pre-defined cancer-specific matrix [Alexandrov et al., 2018; Rubanova et al., 2018], selection of signatures based on other genomic information, such as patterns of indels or structural variants, or strand biases [Alexandrov et al., 2018], or with other molecular or clinical covariates [Robinson et al., 2019]. The probabilistic framework of CloneSig is well suited to integrate other mutation types (indels, structural variants), as well as prior knowledge on signature co-occurrence, and a prior based on other molecular and clinical covariates. The difficulty of this approach is the possibility to learn such association patterns. Another direction for further development would be to use CloneSig's model to learn the signatures, or to allow some variations, as suggested in Volkova et al. [2019].

4.4 Materials and methods

4.4.1 CloneSig model

CloneSig is a probabilistic graphical framework, represented in Figure 4.2, to model the joint distribution of SNV frequency and mutational context using several latent variables to capture the subclonal composition of a tumor and the mutational processes involved in each clone. For a given SNV it assumes that we observe the following variables: D, the total number of reads covering the SNV; $B \leq D$, the number of mutated reads; $T \in \{1, \ldots, 96\}$ the index of the mutation type (i.e., the mutation and its flanking nucleotides, up to symmetry by reverse complement); and $C = (C_{normal}, C_{tumor}^{major}, C_{tumor}^{minor})$ the allele-specific copy number at the SNV locus, as inferred using existing tools such as ASCAT [Martincorena et al., 2017]. Here C_{normal} is the total copy number in normal cells, and $(C_{tumor}^{major}, C_{tumor}^{minor})$ are respectively the copy number in the cancer cells of the major and minor allele, respectively. We therefore also observe $C_{tumor} = C_{tumor}^{major} + C_{tumor}^{minor}$, the total copy number in cancer cells. Finally, we assume observed the tumor sample purity p, i.e., the fraction of cancer cells in the sample.

In addition to those observed variables, CloneSig models the following unobserved variables: $U \in \{1, \ldots, J\}$, the index of the clone where the SNV occurs (assuming a total of J clones); $S \in \{1, \ldots L\}$ the index of the mutational signature that generated the SNV (assuming a total of L possible signatures, given *a priori*); and $M \in \{1, \ldots, C_{tumor}^{major}\}$, the number of chromosomes where the SNV is present. Note that here we assume that SNVs can only be present in one of the two alleles, hence the upper bound of M by C_{tumor}^{major}.

Denoting for any integer d by $\Sigma_d = \{u \in \mathbb{R}_+^d, \sum_{i=1}^d u_i = 1\}$ the d-dimensional probability simplex, and for $u \in \Sigma_d$ by $\mathrm{Cat}(u)$ the categorical distribution over $\{1, \ldots, d\}$ with probabilities $u_1, \ldots, u_d$ (i.e., $X \sim \mathrm{Cat}(u)$ means that $P(X = i) = u_i$ for $i = 1, \ldots, d$), let us now describe the probability distribution encoded by CloneSig for a single SNV; its generalization to several SNVs is simply obtained by assuming they are independent and identically distributed (i.i.d.) according to the model for a single SNV. We do not model the law of C and D, which are observed root nodes in Figure 4.2, and therefore only explicit the conditional distribution of (U, S, T, M, B) given (C, D).

Given parameters $\xi \in \Sigma_J$, $\pi \in (\Sigma_L)^J$ and $\mu \in (\Sigma_{96})^L$, we simply model U, S and T as categorical variables:

$$
\begin{aligned}
U &\sim \mathrm{Cat}(\xi)\,, \\
S \,|\, U &\sim \mathrm{Cat}(\pi_U)\,, \\
T \,|\, S &\sim \mathrm{Cat}(\mu_S)\,.
\end{aligned}
$$

Conditionally on C, we assume that the number of mutated chromosomes M is uniformly chosen between 1 and C_{tumor}^{major}, i.e.,

$$
M \,|\, C \sim \mathrm{Cat}(1/C_{tumor}^{major})\,,
$$

where $1/C_{tumor}^{major} \in \Sigma_{C_{tumor}^{major}}$ represents the vector of constant probability. Finally, to define the law of B, the number of mutated reads, we follow a standard approach in previous studies that represent ITH as a generative probabilistic model [Roth et al., 2014; Deshwar et al., 2015; Yuan et al., 2018; Miller et al., 2014] where the law of the mutated read counts for a given SNV must take into account the purity of the tumor, the proportion of cells in the tumor sample carrying that mutation (cancer cell fraction, CCF), as well as the various copy numbers of the normal and tumor cells. More precisely, as reviewed by Dentro et al. [2017], one can show that the expected fraction of mutated reads (variant allele frequency, VAF) satisfies

$$\text{VAF} = \frac{p \times \text{CCF} \times M}{p \times C_{tumor} + (1-p) \times C_{normal}} .$$

Note that this only holds under the classical simplifying assumption that all copy number events are clonal and affect all cells in the sample. If we now denote by $\phi \in [0,1]^J$ the vector of CCF for each clone, and introduce a further parameter $\rho \in \mathbb{R}_+^*$ to characterize the possible overdispersion of mutated read counts compared to their expected values, we finally model the number of mutated reads using a beta binomial distribution as follows:

$$B \mid D, U, C, M \sim \text{BetaBinomial}\left(D, \rho\phi_U\eta(M,C), \rho(1-\phi_U\eta(M,C))\right)$$
$$\text{with } \eta(M,C) = \frac{p \times M}{p \times C_{tumor} + (1-p) \times C_{normal}} .$$

4.4.2 Parameter estimation

Besides the tumor purity p, we assume that the matrix of mutational processes $\mu \in (\Sigma_{96})^L$ is known, as provided by databases like COSMIC and discussed below in Section 4.4.10. We note that we could consider μ unknown and use CloneSig to infer a new set mutational signatures from large cohorts of sequenced tumors, but prefer to build on existing work on mutational processes in order to be able to compare the results of CloneSig to the existing literature. Besides p and μ, the free parameters or CloneSig are J, the number of clones, and $\boldsymbol{\theta} = (\xi, \phi, \pi, \rho)$ which define the distributions of all random variables. On each tumor, we optimize $\boldsymbol{\theta}$ separately for $J = 1$ to $J_{max} = 8$ clones to maximize the likelihood of the observed SNV data in the tumor. The optimization is achieved approximately by an expectation-maximization (EM) algorithm [Dempster et al., 1977] detailed in Supplementary Section B.1.1. The number of clones $J^* \in [1, J_{max}]$ is then estimated by maximizing an adaptive model selection criterion, detailed in Supplementary Section B.1.2.

4.4.3 Test of mutational signature changes

We use a likelihood ratio test to determine the significance of a signature change, by comparing a regular CloneSig fit to a fit with a single mixture of signatures common to all clones. To adapt the test, the parameter of the chi-squared distribution needs a calibration, that we perform on simulated data under the null hypothesis (without change of signatures between clones). We obtain the optimal parameter using a ridge regression model with the number of clones and the degree of freedom of the input signature matrix as covariates. The coefficient values are averaged over 10-fold cross-validation to ensure robustness. We provide more details about this test in Supplementary Section B.1.3.

4.4.4 Simulations

We use several simulation strategies to evaluate the performance of CloneSig and other methods in various situations. We also use simulations to adjust several aspects of CloneSig, in particular the setting of a custom stopping criterion and the calibration of the statistical test to detect a significant signature change along tumor evolution.

4.4.4.1 Default simulations

We implemented a class `SimLoader` to perform data simulation in CloneSig package. The user sets the number of clones J, the number of observed mutations N, and the matrix of L possible signatures μ. She can also specify the desired values for the CCF of each clone $\phi \in [0,1]^J$, the proportion of each clone $\xi \in \Sigma_J$, the exposure of each signature in each clone $\pi \in (\Sigma_L)^J$, and the overdispersion parameter $\rho \in \mathbb{R}^{+*}$ for the beta-binomial distribution, as well as the proportion of the genome that is diploid. If the user does not provide values for one or several parameters, we generate them randomly as follows:

π the number of active signatures follows a $Poisson(7) + 1$ distribution, and the signatures are chosen uniformly among the L available signatures. Then for each subclone, the exposures of active signatures follow a Dirichlet distribution of parameter 1 for each active signature;

ϕ the cancer cell fraction of each clone is set such that the largest clone has a CCF of 1, and each subsequent CCF is uniformly drawn in decreasing order to be greater than 0.1, and at a distance at least 0.05 from the previous clone;

ξ the proportions of clones are drawn from a Dirichlet distribution of parameter 1 for each clone. The proportions are repeatedly drawn until the minimal proportion of a clone is greater than 0.05;

ρ follows a normal distribution of mean 60 and of variance 5.

The same strategy is used for random initialization of the parameters for the EM algorithm.

The total copy number status is drawn for a user-set diploid proportion of the genome with a bell-like distribution centered in 2, and skewed towards the right (see Supplementary Figure B.57 for examples), or from a rounded log-normal distribution of parameters 1 and 0.3. The minor copy number is then drawn as the rounded product between a beta distribution of parameters 5 and 3 and the total copy number. The multiplicity of each mutation n is uniformly drawn between 1 and $C_{n,tumor_{major}}$. The purity is drawn as the minimum between a normal variable of mean 0.7 and of variance 0.1, and 0.99. The other observed variables (T, B, D) are drawn according to CloneSig probabilistic model.

4.4.4.2 Simulations for comparison with other ITH and signature methods

To calibrate the custom stopping criterion and for further evaluation of CloneSig, we simulated $6,300$ datasets using the previously described setting, with a few adjustments: we set the minimal proportion of each clone to 0.1, the minimal difference between 2 successive clone CCFs to 0.1, and we chose the active signatures among the active signatures for each of the 35 cancer types described in the file `signatures_in_samples_and_cancer_types.mat`, extracted from the SigProfiler MATLAB package (version 2.5.1.7, downloaded from Mathworks on May 16th 2019). We draw the number of active signatures as the minimum of a $Pois(7)+1$ distribution and the number of active signatures for this cancer type. We required a cosine distance of at least 0.05 between the mutational profiles of two successive clones.

In total, for each of the 35 cancer types, we generated a simulated sample for each combination of a number of mutations from the set $\{100, 300, 600, 1000, 5000\}$ covering the range observed in WES and WGS data, a percentage of the genome that is diploid from the set $\{0\%, 20\%, 40\%, 60\%, 80\%, 100\%\}$ to assess the impact of copy number variations, and finally, between 1 and 6 clones.

4.4.4.3 Simulations without signature change between clones

We generated a set of simulations similar in all points to the one for comparison with other ITH and signature methods, except that there is a unique signature mixture common to all clones. We used this dataset in two contexts: (i) to evaluate CloneSig in comparison to other methods in the absence of signature change, and (ii) to design a statistical test to assess the significance of a change in mutational signatures. For the latter, the dataset was limited to the first ten cancer types to avoid unnecessary computations.

4.4.4.4 Simulations to assess the separating power of CloneSig

To assess the separating power of CloneSig, we generated a dataset of 5,400 simulated tumor samples with two clones, where each clone represents 50% of the observed SNVs. Our objective was to explore the set of the distance between two clones, in terms of CCF distance, and of cosine distance between the two mutational profiles. For that purpose we first drew ten possible CCF distances evenly on a log scale between 0 and 1, and set to 1 the largest clone CCF. We also generated 30 matrices π with cosine distances covering regularly the possible cosine distances; to obtain them, we first generated 10,000 such π matrices to estimate an empirical distance distribution, and we implemented a rejection sampling strategy to obtain 30 samples from a uniform distribution. For each pair of CCF distance and π matrix, several samples were generated with the number of mutations varying among $\{100, 300, 1000\}$, the diploid proportion of the genome among $\{0.1, 0.5, 0.9\}$, and the sequencing depth among $\{100, 500\}$.

4.4.4.5 Simulations to assess the sensitivity of the statistical test

To measure the sensitivity of the statistical test to detect a significant signature change along tumor evolution, we generated a dataset of 2,700 simulated tumor samples with 2 to 6 clones. We used again a rejection sampling strategy to explore the space of the maximal distance between the profiles between any 2 clones, but the target distribution is here a beta distribution of parameters 1.5 and 8 as a target distribution, as the objective was to sample more thoroughly the small cosine distances. We repeated the sampling of 30 π matrices for 2 to 6 clones, and in each case, and generated several samples with the number of mutations varying among $\{100, 300, 1000\}$, the diploid proportion of the genome among $\{0.1, 0.5, 0.9\}$, and the sequencing depth among $\{100, 500\}$.

4.4.5 Evaluation metrics

We use several evaluation metrics to assess the quality of CloneSig and other comparable methods. Some assess specifically the accuracy of the subclonal decomposition, while others assess the performance of signature deconvolution.

4.4.5.1 Metrics evaluating the subclonal decomposition

The metrics described in this section evaluate the accuracy of the subclonal deconvolution. They are adapted from Salcedo et al. [2018].

Score1B measures the adequacy between the true number of clones J_{true} and the estimated number of clones J_{pred}. It is computed as $\frac{J_{true}+1-\min(J_{true}+1,|J_{pred}-J_{true}|)}{J_{true}+1}$.

Score1C is the Wasserstein similarity, defined as 1 minus the Wasserstein distance between the true and the predicted clustering, defined by the CCFs of the different clones and their associated weights (proportion of mutations), implemented as the function `stats.wasserstein_distance` in the Python package `scipy`.

Score2A measures the correlation between the true and predicted binary co-clustering matrices in a vector form, M_{true} and M_{pred}. It is the average of 3 correlation coefficients:

Pearson correlation coefficient $PCC = \frac{\mathrm{Cov}(M_{true}, M_{pred})}{\sigma_{M_{true}}, \sigma_{M_{pred}}}$, implemented as the function `pearsonr` in the Python package `scipy`,

Matthews correlation coefficient $\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, implemented as the function `metrics.matthews_corrcoef` in the Python package `scikit-learn`,

V-measure is the harmonic mean of a homogeneity score that quantifies the fact that each cluster contains only members of a single class, and a completeness score

measuring if all members of a given class are assigned to the same cluster [Rosenberg and Hirschberg, 2007]; here the classes are the true clustering. We used the function `v_measure_score` in the Python package `scikit-learn`.

Before averaging, all those scores were rescaled between 0 and 1 using the score of the minimal score between two "bad scenarios": all mutations are in the same cluster, or all mutations are in their own cluster ($M_{pred} = \mathbf{1}_{N\times N}$ or $M_{pred} = \mathbb{I}_{N\times N}$).

Score2C quantifies the accuracy of each method prediction of clonal and subclonal mutations. We report the accuracy, and the area under the ROC curve (implemented in function `metrics.roc_auc_score` in the Python package `scikit-learn`), sensitivity and specificity in Supplementary Section B.2

4.4.5.2 Metrics evaluating the identification of mutational signatures

The metrics described in this section evaluate the accuracy of the mutational signature deconvolution.

Score_sig_1A computes the Euclidean distance between normalized mutation type counts (empirical), and the reconstituted profile. This is the objective function of most signature reconstruction approaches (including deconstructSigs[Rosenthal et al., 2016] and Palimpsest [Shinde et al., 2018]).

Score_sig_1B is the Euclidean distance between simulated and estimated signature profiles (weighted sum over all clones). This is closer to the objective of CloneSig and TrackSig [Rubanova et al., 2018].

Score_sig_1C measures the ability of each method to correctly identify present signatures. For CloneSig, no signature has a null contribution to the mixture, so for each clone, the signatures are considered in the decreasing order of their contribution to the mixture, and selected until the cumulative sum reaches 0.95. This rule is applied to all methods. For that metric, the area under the ROC curve (implemented in function `metrics.roc_auc_score` in the Python package `scikit-learn`) is reported, as well as the accuracy, sensitivity, and specificity in Supplementary Section B.2

Score_sig_1D is the percent of mutations with the right signature. For each mutation, the most likely signature is found by taking into account the distribution of each mutation type in each signature, and the contribution of the signature to the mixture.

Score_sig_1E measures for each mutation the cosine distance between the clonal mutation type distribution that generated the mutation and the reconstituted one. We consider a unique global distribution for deconstructSigs. This allows us to measure the relevance of the reconstruction even if the wrong signatures are selected, as several signatures have very similar profiles. The result is a distribution of distances over all mutations, and we report the median of this distribution. We also report in Supplementary Section B.2 more results with the minimum, the maximum, and the standard deviation of this distribution (`max_diff_distrib_mut`, `median_diff_distrib_mut`), as well as the proportions of mutations with a distance below 0.05 or 0.1 (`perc_dist_5` and `perc_dist_10`).

4.4.6 Implementation

CloneSig is implemented in Python, and is available as a Python package at `https://github.com/judithabk6/clonesig`. A wrapper function implements the successive optimization of CloneSig with increasing number of clones. For two clones and more, the model is initialized using results from the precedent run with one fewer clone, by splitting the subclone with the largest contribution to the mixture entropy as described in Baudry and Celeux [2015]. This process is stopped when the maximum number of subclones is reached, or when the selection criterion decreases for two successive runs. A class for simulating data according to the CloneSig model is also implemented, as detailed above.

4.4.7 Data

We downloaded data from the GDC data portal https://portal.gdc.cancer.gov/. We gathered annotated somatic mutations, both raw variant calling output, whose access is restricted and public mutations, from the new unified TCGA pipeline https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/, with alignment to the GRCh38 assembly, and variant calling using 4 variant callers: MuSe, Mutect2, VarScan2 and SomaticSniper. Instructions for download can be found in the companion Github repository (https://github.com/judithabk6/CloneSig_analysis).

4.4.8 Copy number calling and purity estimation

We obtained copy number alterations (CNA) data from the ASCAT complete results on TCGA data partly reported on the COSMIC database [Martincorena et al., 2017; Forbes et al., 2017]. We then converted ASCAT results on hg19 to GRCh38 coordinates using the `segment_liftover` Python package [Gao et al., 2018]. ASCAT results also provide an estimate of purity, which we used as input to ITH methods when possible. Other purity measures are available [Aran et al., 2015]; however we selected the ASCAT estimate to ensure consistency with CNV data.

4.4.9 Variant calling filtering

Variant calling is known to be a challenging problem. It is common practice to filter variant callers output, as ITH methods are deemed to be highly sensitive to false positive SNVs. We filtered out indels from the public dataset, and considered the union of the 4 variant callers output SNVs. For the protected data, we also removed indels, and then filtered SNVs on the `FILTER` columns output by the variant caller ("PASS" only VarScan2, SomaticSniper, "PASS" or "panel_of_normals" for Mutect2, and "Tier1" to "Tier5" for MuSe). In addition, for all variant callers, we removed SNVs with a frequency in 1000 genomes or Exac greater than 0.01, except if the SNV was reported in COSMIC. A coverage filter was added, and we kept SNVs with at least 6 reads at the position in the normal sample, of which 1 maximum reports the alternative nucleotide (or with a variant allele frequency (VAF) <0.01), and for the tumor sample, at least 8 reads covering the position, of which at least 3 reporting the variant, or a VAF>0.2. The relative amount of excluded SNVs from protected to public SNV sets varied significantly between the 3 cancer types (see Table B.3). All annotations are the ones downloaded from the TCGA, using VEP v84, and GENCODE v.22, sift v.5.2.2, ESP v.20141103, polyphen v.2.2.2, dbSNP v.146, Ensembl genebuild v.2014-07, Ensembl regbuild v.13.0, HGMD public v.20154, ClinVar v.201601. We further denote the filtered raw mutation set as "Protected SNVs" and the other one, which is publicly available, as "Public SNVs"

4.4.10 Construction of a curated list of signatures associated with each cancer type

A very important input for CloneSig is the signature matrix. For application to the TCGA data, we restrict ourselves to signatures known to be active in each subtype. To that end, we downloaded the signatures found in the TCGA using SigProfiler [Alexandrov et al., 2018] from synapse table syn11801497. The resulting list was not satisfactory as it lacked important known patterns; for instance signature 3, associated with homologous recombination repair deficiency was not found to be active in any tumor of the prostate cohort, while signature 3 in prostate cancer is well described in the literature [Dentro et al., 2018; Espiritu et al., 2018; Riaz et al., 2017]. We therefore completed the signatures present in each cancer type based on the literature [Dentro et al., 2018; Nik-Zainal et al., 2016; Roerink et al., 2018; Letouzé et al., 2017; Shibata et al., 2018; Ren et al., 2018; Warsow et al., 2018; Royer-Bertrand et al., 2016; Espiritu et al., 2018; Macintyre et al., 2018; Ashley et al., 2019; Liu et al., 2018b; Verhagen et al., 2018], and used the resulting matrix in all CloneSig runs on the TCGA. Our curated list of signatures present in each cancer type is provided in Table B.4.

4.4.11 Survival analysis

We used the Python package `lifelines` to compute the Kaplan-Meier curves and multivariate Cox models.

Chapter 5

Closing remarks

5.1 Conclusion

In this thesis, we have focused our efforts on the field of methods for the inference of tumor evolution, and their potential clinical application, in particular in the setting of one sample per patient, which is currently the standard practice.

In Chapter 2, we have presented the different approaches developed to study and quantify that aspect of cancer genomics, and highlighted a damaging lack of performance evaluations of those methods. The surveyed methods cover diverse aspects of the problem, from the simple detection of several subclonal populations, or the reconstruction of subclonal genotypes, to the most complete view of intra-tumor heterogeneity by reconstructing a mutation tree that recapitulates the history of mutation acquisition. Furthermore, though representing a fair amount of work, evaluations can be beneficial for future developments as they allow researchers to identify input and algorithms that truly impact the performance, and for potential users to choose the best-suited method for their data.

Chapter 3 introduces a first published contribution of this thesis, that provides an analysis of the robustness of ITH estimations by several methods, with different pre-processing pipelines. In addition, we evaluated the association of ITH measures with clinical variables that had previously been found correlated with the number of subclones. We considered three cohorts of patients from the TCGA: Breast cancer patients, Bladder cancer patients, and Head and Neck cancer patients. In all three types, we observed important discordances between the ITH measures from different pipelines, with in some cases a significant positive correlation between the measures from different pipelines, but in other cases an absence of correlation. Similarly, association with clinical variables, in particular the survival, was not robustly recovered with the different ITH measurements. Finally, correlation of the obtained ITH measures with genomic variables, such as mutational burden, copy number abundance and tumor purity, varies significantly between the tested pipelines, and are sometimes of opposite sign. This suggests that such associations may be algorithmic biases rather than a true biological signal. Those biases could be systematically evaluated using simulated approaches.

In Chapter 4, we propose CloneSig, a new method for ITH reconstruction, that also jointly performs signature deconvolution. We demonstrate on simulated data that this joint inference achieves more accurate results than existing methods for subclonal deconvolution, in particular when mutational processes, uncovered by mutational signatures, vary along tumor evolution. We then applied CloneSig to the whole TCGA to observe large-scale trends for each cancer type, and recovered patterns of signature evolution previously observed in WGS data, thus validating the ability of CloneSig to detect significant changes even with a low number of observed SNVs. We also observed new variations of signature activity in some cancer types, including signatures with known associated targeted therapies. Currently, CloneSig is restricted to single nucleotide substitutions, but a lot of work has been dedicated lately to define signatures for other kinds of alterations, and could be further included in CloneSig. This would constitute a new way to integrate several types of alterations, that we have not observed in any of the existing methods surveyed at the beginning of the manuscript.

Another lead to improve the model would be to alter the way signatures depend on clones. Currently, all signatures are equally likely, with the only possible adjustment being the list of considered signatures in input. However, more complex patterns could be modeled, with either a dependency between different signatures, that would be particularly relevant for different types of alteration signatures, or a dependency between signatures and other clinical variables; this is of particular importance if we want to refine the identification of signatures for use in the clinic, as signature deconvolution is an unidentifiable problem. Such association between the occurrence of several types of alterations would be a major step for data integration for ITH reconstruction.

5.2 Perspectives

5.2.1 How relevant is the number of clones to quantify tumor evolution?

An important focus has been dedicated to the number of clones as a measure of intra-tumor heterogeneity. We have demonstrated in Chapter 3 that this was a very non-robust quantity to assess, and that it was not very informative for patient stratification compared to classical clinical variables. Beyond the difficulty to measure it, there could be more fundamental reasons explaining that finding: clone may be an ill-defined concept and there is no true number of clones [Caravagna et al., 2019]. This has been experimentally illustrated by Campbell et al. [2008] as more and more subclonal populations are uncovered as sequencing depth increases. Hence, in the context of sequencing, the number of clones incidentally has a more practical definition, which is the number of populations that can be distinguished given a certain sequencing assay, which may not be identical between samples, notably due to the tumor purity. Moreover, similar subclonal structures can reflect very different evolutionary histories: for instance a single clonal population may be observed because a tumor is young and has not yet undergone clonal diversification, or on the contrary results from a recent selective sweep that has drastically reduced the tumor diversity. This illustrates the naivety of approaching tumor evolution through a single quantity, as tumors are complex populations, characterized by their genomes admittedly, but also by their epigenomes, their transcriptomes, and interactions with the micro-tumor environment, in particular with stromal and immune populations. This was hypothesized by a consortium of researchers [Maley et al., 2017], and recently experimentally described in lung tumors [Sharma et al., 2019], with several tumor characteristics exhibiting important spatial variations, without being reflected by genetic heterogeneity measurements.

However, though being of limited interest for patient stratification, a number of measurements have been developed to quantify tumor evolution, and could inform clinical management of tumors (prevention, detection, treatment), or simply our knowledge of the driving forces of tumor development and aggressiveness. Here are some of those alternative measurements of phenomena closely related to intra-tumor heterogeneity, partly inspired by the field of population genetics

Somatic mutations and heterogeneity in normal tissues, or precancerous lesions are important for two aspects, as they enlighten early steps of cancer development, and can help discriminate between several candidate theories for tumor evolution (continuous, punctuated etc). A second interesting aspect of those data is that it could provide a negative control to better estimate the pathogenic properties of intra-tumor heterogeneity. Indeed, continuous acquisition of alterations, and even clonal expansion phases are not restricted to tumor cells [Maley et al., 2006; Martincorena et al., 2018, 2015; Moore et al., 2018].

Quantification of selection , to answer the underlying question of whether tumor evolution is driven by neutral evolution and genetic drift, or positive selection of clones with higher fitness, or negative selection of clones with low fitness, which is highly debated in the community, and might be indicative of the clinical relevance of ITH [Graham and Sottoriva, 2016; Williams et al., 2016, 2018; Tarabichi et al., 2018].

Age of the tumor and of the successive metastatic seeding events [Hu et al., 2019].

Eelation with selection against neoantigens with again some contradictory results, with evidence for positive selection of clones depleted in neoantigens [Rosenthal et al., 2019], and other reports of lack of such trends [Van den Eynden et al., 2019].

Interestingly, this latter example of investigation of neoantigen selection with contradictory findings is very illustrative of the lack of integration between different alteration types. Indeed, evidence of negative selection in the former study relies on CNV alterations, transcriptional depletion, and associated indications of hypermethylation [Rosenthal et al., 2019], while the latter study conflicting results rely on point mutations [Van den Eynden et al., 2019]. Also, both studies are performed on different datasets.

Overall, the relevant measures to quantify tumor evolution are an active research area, that goes beyond intra-tumor heterogeneity assessments. We hope that the two contributions of this thesis, in assessing the robustness and clinical relevance of ITH, and associating ITH with other manifestations of tumor evolution will provide solid grounds for further developments, and increasing data integration. Indeed, data integration is key to ensure consistency of findings from the different aspects of the data and studied phenomena, and to provide a more accurate detection of subtle effects. This lack of consensus on the right approach may also explain the lack of strong evaluation assays for such quantities, as is can be preemptive to refine and build on existing measures when the real question is what to measure. Moreover, the ideal evaluation should also prioritize the most relevant biological aspects, that are still to be determined.

5.2.2 The necessity to go beyond the TCGA

All the projects of this thesis, and a large proportion of results on tumor evolution rely on data from the Cancer Genome Atlas (TCGA), including landmark results regarding the prognosis power of ITH [Andor et al., 2016], the presence of selection along tumor evolution [Williams et al., 2016], the predictability of tumor evolution [Hosseini et al., 2019], the lack of evidence of selection for neoantigen depletion [Van den Eynden et al., 2019], the establishment of reference mutational signatures [Alexandrov et al., 2018]. Though the TCGA provides an outstanding resource, with a high number of patients, numerous experimental assays performed for each of them, and a great effort in data processing normalization and sharing, the broad use of this dataset presents some weaknesses. A first obvious issue is the lack of validation on independent cohorts, and any bias in the selection of included patients will reflect on the drawn conclusions and generalization. Furthermore, this dataset is not necessarily the best-suited for all analyses: it contains only one sample of untreated primary tumors, with clinical information not always of high enough quality to study overall survival, or event-free survival [Liu et al., 2018a].

In particular for evolution analysis, other assays can be very relevant to fully approach its main characteristics

- One sample per tumor can miss part of the spatial heterogeneity [Opasic et al., 2019].
- Longitudinal studies are necessary to capture the dynamic of a tumor subclonal structure through time, treatments, and other relevant events.
- More comprehensive studies of other types of tumors, like pre-cancerous stages when available, or metastases can provide additional information that helps interpreting findings in primary tumors [Bertucci et al., 2019; Priestley et al., 2019].
- Circulating DNA analyses offer exciting perspectives as they would reduce sampling invasiveness for the patients, but it is not yet clear how much they reflect characteristics from the tumor [Parikh et al., 2019].

Beyond the sampling strategy, the measurement assay can also be leveraged to measure intra-tumor heterogeneity more accurately. Indeed, single cell sequencing can facilitate the inference as genotypes are directly observed, and only the proportions of the different populations and their phylogenetic relationships need to be inferred. Some difficulties remain, as

single cell sequencing remains error prone, and costly, but recent studies are encouraging for its feasibility [Laks et al., 2019]. Finally, long read sequencing may enable to better leverage and detect structural variations, and offers promising perspectives.

The democratization of those more advanced techniques will certainly lead to new formats of data on which CloneSig may not be applicable, as well as the methods examined in our two evaluations. However, we believe and hope that the ideas underlying our contributions, both for robustness analyses on real data, and promotion for joint inference of related phenomena will lay the ground for future work, in the field of tumor evolution or for other problems with similar aspects.

Bibliography

Abécassis J, Hamy A S, Laurent C, Sadacca B, Bonsang-Kitzis H, Reyal F, and Vert J P. Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data. *PLoS ONE*, 14(11):1–22, 2019. ISSN 19326203. doi:10.1371/journal.pone.0224143.

Aganezov S and Raphael B J. Reconstruction of clone- and haplotype-specific cancer genome karyotypes from bulk tumor samples. *bioRxiv*, page 560839, 2019. doi:10.1101/560839.

Ahmadinejad N, Troftgruben S, Maley C C, Wang J, and Liu L. MAGOS: Discovering subclones in tumors sequenced at standard depths. *bioRxiv*, (1):1–31, 2019. doi:10.1101/790386.

Akaike H. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 199–213. Springer, 1998. doi:10.1007/978-1-4612-1694-0_15.

Alexandrov L B, Kim J, Haradhvala N J, et al. The repertoire of mutational signatures in human cancer. *bioRxiv*, page 322859, 2018. ISSN 1936-2692. doi:10.1101/322859.

Alexandrov L B, Nik-Zainal S, Wedge D C, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013. ISSN 0028-0836. doi:10.1038/nature12477.

Alves J M, Prieto T, and Posada D. Multiregional tumor trees are not phylogenies. *Trends in Cancer*, 10(0):e1003703, 2017. ISSN 24058033. doi:10.1016/j.trecan.2017.06.004.

Andor N, Graham T A, Jansen M, Xia L C, Aktipis C A, Petritsch C, Ji H P, and Maley C C. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 22(1):105–113, 2016. ISSN 1078-8956. doi:10.1038/nm.3984.

Andor N, Harness J V, Müller S, Mewes H W, and Petritsch C. Expands: Expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, 30(1):50–60, 2014. ISSN 13674803. doi:10.1093/bioinformatics/btt622.

Anzar I, Sverchkova A, Stratford R, and Clancy T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Medical Genomics*, 12(1):63, 2019. ISSN 1755-8794. doi:10.1186/s12920-019-0508-5.

Aran D, Sirota M, and Butte A J. Systematic pan-cancer analysis of tumour purity. *Nature Communications*, 6(1):8971, 2015. ISSN 2041-1723. doi:10.1038/ncomms9971.

Arlot S. Minimal penalties and the slope heuristics: a survey. *arXiv preprint arXiv:1901.07277*, 2019.

Ashley C W, Da Cruz Paula A, Kumar R, Mandelker D, Pei X, Riaz N, Reis-Filho J S, and Weigelt B. Analysis of mutational signatures in primary and metastatic endometrial cancer reveals distinct patterns of DNA repair defects and shifts during tumor progression. *Gynecologic Oncology*, 152(1):11–19, 2019. ISSN 00908258. doi:10.1016/j.ygyno.2018.10.032.

Ashley D J. The two "hit" and multiple "hit" theories of carcinogenesis. *British Journal of Cancer*, 23(2):313–328, 1969. ISSN 0007-0920. doi:10.1038/bjc.1969.41. URL http://www.nature.com/articles/bjc196941.

Baez-Ortega A and Gori K. Computational approaches for discovery of mutational signatures in cancer. *Briefings in bioinformatics*, 20(1):77–88, 2019. ISSN 14774054. doi:10.1093/bib/bbx082.

Baudry J P and Celeux G. EM for mixtures: Initialization requires special care. *Statistics and Computing*, 25(4):713–726, 2015. ISSN 15731375. doi:10.1007/s11222-015-9561-x.

Beerenwinkel N, Schwarz R F, Gerstung M, and Markowetz F. Cancer evolution: Mathematical models and computational inference. *Systematic Biology*, 64(1):e1–e25, 2015. ISSN 1076836X. doi:10.1093/sysbio/syu081.

Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995. ISSN 00359246. doi:10.2307/2346101.

Bertsekas D P. Projected newton methods for optimization problems with simple constraints. *SIAM Journal on Control and Optimization*, 20(2):221–246, 1982. ISSN 0363-0129. doi:10.1137/0320018.

Bertucci F, Ng C K Y, Patsouris A, et al. Genomic characterization of metastatic breast cancers. *Nature*, 569(7757):560–564, 2019. ISSN 0028-0836. doi:10.1038/s41586-019-1056-z.

Bhandari V, Liu L Y, Salcedo A, Espiritu S M G, Morris Q D, and Boutros P C. The inter and intra-tumoural heterogeneity of subclonal reconstruction. *bioRxiv*, 2018. doi:10.1101/418780.

Biernacki C, Celeux G, and Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000. ISSN 01628828. doi:10.1109/34.865189.

Bindea G, Mlecnik B, Tosolini M, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, 39(4):782–795, 2013. ISSN 10747613. doi:10.1016/j.immuni.2013.10.003.

Böck B C, Stein U, Schmitt C A, and Augustin H G. Mouse models of human cancer. *Cancer Research*, 74(17):4671–4675, 2014. ISSN 0008-5472. doi:10.1158/0008-5472.CAN-14-1424.

Broad I. Picard toolkit. 2019. URL http://broadinstitute.github.io/picard/.

Brown A L, Li M, Goncearenco A, and Panchenko A R. Finding driver mutations in cancer: Elucidating the role of background mutational processes. *PLOS Computational Biology*, 15(4):1–25, 2019. doi:10.1371/journal.pcbi.1006981.

Campbell P J, Pleasance E D, Stephens P J, Dicks E, Rance R, Goodhead I, Follows G A, Green A R, Futreal P A, and Stratton M R. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proceedings of the National Academy of Sciences*, 105(35):13081–13086, 2008. ISSN 0027-8424. doi:10.1073/pnas.0801523105.

Caravagna G, Heide T, Williams M, et al. Model-based tumor subclonal reconstruction. *bioRxiv*, page 586560, 2019. doi:10.1101/586560.

Carter S L, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421, 2012. ISSN 10870156. doi:10.1038/nbt.2203.

Chedom-Fotso D, Ahmed A A, and Yau C. OncoPhase: Quantification of somatic mutation cellular prevalence using phase information. *bioRxiv*, doi:10.1101/046631, 2016. doi:10.1101/046631.

Chu Y, Nie C, and Wang Y. p-SCNAClonal: Somatic copy number alterations based tumor subclonal population inferring method. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1693–1698. IEEE, 2018. ISBN 978-1-5386-5488-0. doi:10.1109/BIBM.2018.8621079.

Chung W, Eum H H, Lee H O, Lee K M, Lee H B, Kim K T, Ryu H S, Kim S, Lee J E, Park Y H, Kan Z, Han W, and Park W Y. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications*, 8(May):1–12, 2017. ISSN 20411723. doi:10.1038/ncomms15081.

Cibulskis K, Lawrence M S, Carter S L, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander E S, and Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, 2013. ISSN 10870156. doi:10.1038/nbt.2514.

Ciccolella S, Soto Gomez M, Patterson M, Della Vedova G, Hajirasouliha I, and Bonizzoni P. Inferring cancer progression from single-cell sequencing while allowing mutation losses. *bioRxiv*, page 268243, 2018. doi:10.1101/268243.

Cmero M, Ong C S, Yuan K, Schröder J, Mo K, Group P E, Working H, Corcoran N M, Papenfuss A T, Hovens C M, Markowetz F, and Macintyre G. SVclone: inferring structural variant cancer cell fraction. *bioRxiv*, page 172486, 2017. doi:10.1101/172486.

Colaprico A, Silva T C, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T S, Malta T M, Pagnotta S M, Castiglioni I, Ceccarelli M, Bontempi G, and Noushmehr H. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71, 2016. ISSN 13624962. doi:10.1093/nar/gkv1507.

Consul S and Vikalo H. Reconstructing intra-tumor heterogeneity via convex optimization and branch-and-bound search. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics - BCB '19*, pages 524–529. ACM Press, New York, New York, USA, 2019. ISBN 9781450366663. doi:10.1145/3307339.3342178.

Craig Venter J, Adams M D, Myers E W, et al. The sequence of the human genome. *Science*, 2001. ISSN 00368075. doi:10.1126/science.1058040.

Cross W C, Graham T A, and Wright N A. New paradigms in clonal evolution: punctuated equilibrium in cancer. *Journal of Pathology*, 240(2):126–136, 2016. ISSN 10969896. doi:10.1002/path.4757.

Cun Y, Yang T p, Achter V, Lang U, and Peifer M. Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. *Nature Protocols*, 13(6):1488–1501, 2018. ISSN 1754-2189. doi:10.1038/nprot.2018.033.

Dagogo-Jack I and Shaw A T. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94, 2018. ISSN 17594782. doi:10.1038/nrclinonc.2017.166.

Dang H X, White B S, Foltz S M, Miller C A, Luo J, Fields R C, and Maher C A. ClonEvol: Clonal ordering and visualization in cancer sequencing. *Annals of Oncology*, 28(12):3076–3082, 2017. ISSN 15698041. doi:10.1093/annonc/mdx517.

Davidson-Pilon C, Kalderstam J, Zivich P, et al. Camdavidsonpilon/lifelines: v0.20.0. 2019. doi:10.5281/zenodo.2584900. URL https://doi.org/10.5281/zenodo.2584900.

Davis A and Navin N E. Computing tumor trees from single cells. *Genome Biology*, 17(1):1–4, 2016. ISSN 1474760X. doi:10.1186/s13059-016-0987-z.

Dempster A P, Laird N M, and Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. ISSN 00359246. doi:10.1111/j.2517-6161.1977.tb01600.x.

Dentro S C, Leshchiner I, Haase K, et al. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *bioRxiv*, page 312041, 2018. doi:10.1101/312041.

Dentro S C, Wedge D C, and Van Loo P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harbor Perspectives in Medicine*, 7(8):a026625, 2017. ISSN 2157-1422. doi:10.1101/cshperspect.a026625.

Deshwar A G, Vembu S, Yung C K, Jang G, Stein L, and Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, 2015. ISSN 1465-6906. doi:10.1186/s13059-015-0602-8.

Deveau P, Colmet Daage L, Oldridge D, et al. QuantumClone: clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction. *Bioinformatics*, 34(11):1808–1816, 2018. ISSN 14602059. doi:10.1093/bioinformatics/bty016.

DiNardo Z, Tomlinson K, Ritz A, and Oesper L. Distance measures for tumor evolutionary trees. *bioRxiv*, page 591107, 2019.

Dinh K N, Kimmel M, Lambert A, and Travaré S. Statistical inference for the evolutionary history of cancer genomes. *bioRxiv*, pages 1–46, 2019.

Donmez N, Malikic S, Wyatt A W, Gleave M E, Collins C C, and Sahinalp S C. Clonality inference from single tumor samples using low-coverage sequence data. *Journal of Computational Biology*, 24(6):515–523, 2017. ISSN 1557-8666. doi:10.1089/cmb.2016.0148.

Druker B J, Tamura S, Buchdunger E, Ohno S, Segal G M, Fanning S, Zimmermann J, and Lydon N B. Effects of a selective inhibitor of the Ab1 tyrosine kinase on the growth of Bcr-Ab1 positive cells. *Nature Medicine*, 1996. ISSN 10788956. doi:10.1038/nm0596-561.

Duncan B K and Miller J H. Mutagenic deamination of cytosine residues in DNA. *Nature*, 287(5782):560–561, 1980. ISSN 0028-0836. doi:10.1038/287560a0.

Eaton J, Wang J, and Schwartz R. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34(13):i357–i365, 2018. ISSN 1367-4803. doi:10.1093/bioinformatics/bty270.

El-Kebir M, Oesper L, Acheson-Field H, and Raphael B J. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015. ISSN 1367-4803. doi:10.1093/bioinformatics/btv261.

El-Kebir M, Raphael B J, Shamir R, Sharan R, Zaccaria S, Zehavi M, and Zeira R. Complexity and algorithms for copy-number evolution problems. *Algorithms for Molecular Biology*, 12(1):13, 2017. ISSN 1748-7188. doi:10.1186/s13015-017-0103-2.

El-Kebir M, Satas G, Oesper L, and Raphael B J. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Systems*, 3(1):43–53, 2016. ISSN 24054712. doi:10.1016/j.cels.2016.07.004.

El-Kebir M, Satas G, and Raphael B J. Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics*, 50(5):718–726, 2018. ISSN 1061-4036. doi:10.1038/s41588-018-0106-z.

Espiritu S M G, Liu L Y, Rubanova Y, et al. The evolutionary landscape of localized prostate cancers drives clinical aggression. *Cell*, 173(4):1003–1013.e15, 2018. ISSN 00928674. doi:10.1016/j.cell.2018.03.029.

Ewing A D, Houlahan K E, Hu Y, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods*, 12(7), 2015. ISSN 1548-7091. doi:10.1038/nmeth.3407.

Fan X, Zhou W, Chong Z, Nakhleh L, and Chen K. Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC Bioinformatics*, 15(1):1–12, 2014. doi:10.1186/1471-2105-15-299.

Fan Y, Xi L, Hughes D S, Zhang J, Zhang J, Futreal P A, Wheeler D A, and Wang W. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome biology*, 17(1):178, 2016. ISSN 1474760X. doi:10.1186/s13059-016-1029-6.

Fang L T, Afshar P T, Chhibber A, Mohiyuddin M, Fan Y, Mu J C, Gibeling G, Barr S, Asadi N B, Gerstein M B, Koboldt D C, Wang W, Wong W H, and Lam H Y. An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biology*, 16(1):197, 2015. ISSN 1474-760X. doi:10.1186/s13059-015-0758-2.

Farahani H, de Souza C P E, Billings R, Yap D, Shumansky K, Wan A, Lai D, Mes-Masson A M, Aparicio S, and P Shah S. Engineered in-vitro cell line mixtures and robust evaluation of computational methods for clonal decomposition and longitudinal dynamics in cancer. *Scientific Reports*, 7(1):13467, 2017. ISSN 2045-2322. doi:10.1038/s41598-017-13338-8.

Favero F, Eklund A C, Joshi T, Marquard a M, Birkbak N J, Krzystanek M, Li Q, Szallasi Z, and Eklund A C. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology*, pages 1–5, 2014. ISSN 0923-7534. doi: 10.1093/annonc/mdu479.

Ferguson T S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973. ISSN 0090-5364. doi:10.1214/aos/1176342360.

Fischer A, Vázquez-García I, Illingworth C J R, and Mustonen V. High-definition reconstruction of clonal composition in cancer. *Cell Reports*, 7(5):1740–1752, 2014. ISSN 22111247. doi:10.1016/j.celrep.2014.04.055.

Fittall M W and Van Loo P. Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome Medicine*, 11(1):20, 2019. ISSN 1756-994X. doi:10.1186/s13073-019-0632-z.

Flensburg C, Sargeant T, Oshlack A, and Majewski I. SuperFreq: Integrated mutation detection and clonal tracking in cancer. *bioRxiv*, page 380097, 2018. doi:10.1101/380097.

Forbes S A, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1121.

Gao B, Huang Q, and Baudis M. segment_liftover : a Python tool to convert segments between genome assemblies. *F1000Research*, 7:319, 2018. ISSN 2046-1402. doi:10.12688/f1000research.14148.1.

Gao J, Aksoy B A, Dogrusoz U, Dresdner G, Gross B, Sumer S O, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, and Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269), 2013. ISSN 19450877. doi:10.1126/scisignal.2004088.

Gatenby R A, Zhang J, and Brown J S. First strike-second strike strategies in metastatic cancer: Lessons from the evolutionary dynamics of extinction. *Cancer Research*, 79(13):3174–3177, 2019. ISSN 15387445. doi:10.1158/0008-5472.CAN-19-0807.

Gawad C, Koh W, and Quake S R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences of the United States of America*, 111(50):17947–17952, 2014. ISSN 10916490. doi:10.1073/pnas.1420822111.

Geng Y, Zhao Z, Xu J, Liu R, Huang Y, Zhang X, Xiao X, Maomao, and Wang J. Identifying heterogeneity patterns of allelic imbalance on germline variants to infer clonal architecture. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10362 LNCS, pages 286–297. 2017. ISBN 9783319633114. doi:10.1007/978-3-319-63312-1_26.

Gerlinger M, Horswell S, Larkin J, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, 46(3):225–233, 2014. ISSN 15461718. doi:10.1038/ng.2891.

Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, Moch H, and Beerenwinkel N. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, 3(1):811, 2012. ISSN 2041-1723. doi:10.1038/ncomms1814.

Gerstung M, Jolly C, Leshchiner I, et al. The evolutionary history of 2,658 cancers. *bioRxiv*, 2017. doi:10.1101/161562.

Goodwin S, Mcpherson J D, and Mccombie W R. Coming of age: ten years of next- generation sequencing technologies. *Nature Publishing Group*, 17(6):333–351, 2016. ISSN 1471-0056. doi:10.1038/nrg.2016.49.

Graham T A and Sottoriva A. Measuring cancer evolution from the genome. *The Journal of pathology*, (November 2016):183–191, 2016. ISSN 1096-9896. doi:10.1002/path.4821.

Greenman C D, Pleasance E D, Newman S, Yang F, Fu B, Nik-Zainal S, Jones D, Lau K W, Carter N, Edwards P A W, Futreal P A, Stratton M R, and Campbell P J. Estimation of rearrangement phylogeny for cancer genomes. *Genome Research*, 22(2):346–361, 2012. ISSN 10889051. doi:10.1101/gr.118414.110.

Griffith M, Miller C A, Griffith O L, et al. Optimizing cancer genome sequencing and analysis. *Cell Systems*, 1(3):210–223, 2015. ISSN 24054712. doi:10.1016/j.cels.2015.08.015.

Griffiths T L and Ghahramani Z. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011. ISSN 15324435.

Ha G, Roth A, and Khattra J. TITAN: Inference of copy number architectures in clonal cell populations from tumor whole genome sequence data. *Genome Research*, 24(11):1881–1893, 2014. doi:10.1101/gr.180281.114.

Hajirasouliha I, Mahmoody A, and Raphael B J. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30:78–86, 2014. ISSN 14602059. doi:10.1093/bioinformatics/btu284.

Hanahan D and Weinberg R a. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011. ISSN 00928674. doi:10.1016/j.cell.2011.02.013.

Heather J M and Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, 2016. ISSN 10898646. doi:10.1016/j.ygeno.2015.11.003.

Hofree M, Shen J P, Carter H, Gross A, and Ideker T. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108–15, 2013. ISSN 1548-7105. doi:10.1038/nmeth.2651.

Hollstein M, Alexandrov L B, Wild C P, Ardin M, and Zavadil J. Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. *Oncogene*, 36(February):1–10, 2016. ISSN 0950-9232. doi:10.1038/onc.2016.192.

Hosseini S R, DIaz-Uriarte R, Markowetz F, and Beerenwinkel N. Estimating the predictability of cancer evolution. *Bioinformatics*, 35(14):i389–i397, 2019. ISSN 14602059. doi:10.1093/bioinformatics/btz332.

Hu Z, Ding J, Ma Z, Sun R, Seoane J A, Scott Shaffer J, Suarez C J, Berghoff A S, Cremolini C, Falcone A, Loupakis F, Birner P, Preusser M, Lenz H J, and Curtis C. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nature Genetics*, 51(7):1113–1122, 2019. ISSN 1061-4036. doi:10.1038/s41588-019-0423-x.

Hujdurovic A, Kacar U, Milanic M, Ries B, and Tomescu A I. Complexity and algorithms for finding a perfect phylogeny from mixed tumor samples. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(1):96–108, 2018. ISSN 1545-5963. doi:10.1109/TCBB.2016.2606620.

Hungerford D A and Nowell P C. A minute chromosome in human chronic granulocytic leukemia. *Science*, 132(3438):1488–1501, 1960. ISSN 0036-8075. doi:10.1126/science.132.3438.1488.

Husić E, Li X, Hujdurović A, Mehine M, Rizzi R, Mäkinen V, Milanič M, and Tomescu A I. MIPUP: Minimum perfect unmixed phylogenies for multi-sampled tumors via branchings and ILP. *Bioinformatics*, 35(5):769–777, 2019. ISSN 14602059. doi:10.1093/bioinformatics/bty683.

Jahn K, Kuipers J, and Beerenwinkel N. Tree inference for single-cell data. *Genome Biology*, 17(1):86, 2016. ISSN 1474760X. doi:10.1186/s13059-016-0936-x.

Jiang Y, Qiu Y, Minn A J, and Zhang N R. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 113(37):E5528–E5537, 2016. ISSN 10916490. doi:10.1073/pnas.1522203113.

Jiao W, Vembu S, Deshwar A G, Stein L, and Morris Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, 15:35, 2014. ISSN 1471-2105. doi:10.1186/1471-2105-15-35.

Jolly C and Van Loo P. Timing somatic events in the evolution of cancer. *Genome Biology*, 19(1):1–9, 2018. ISSN 1474760X. doi:10.1186/s13059-018-1476-3.

Kaiser J. Is genome-guided cancer treatment hyped? *Science*, 360(6387):365–365, 2018. ISSN 0036-8075. doi:10.1126/science.360.6387.365.

Karn T, Jiang T, Hatzis C, Sänger N, El-Balat A, Rody A, Holtrich U, Becker S, Bianchini G, and Pusztai L. Association between genomic metrics and immune infiltration in triple-negative breast cancer. *JAMA Oncology*, 3(12):1707–1711, 2017. ISSN 2374-2437. doi:10.1001/jamaoncol.2017.2140.

Karpov N, Malikic S, Rahman M K, and Sahinalp S C. A multi-labeled tree edit distance for comparing "clonal trees" of tumor progression. *Leibniz International Proceedings in Informatics, LIPIcs*, 113(22):1–19, 2018. ISSN 18688969. doi:10.4230/LIPIcs.WABI.2018.22.

Keats J J, Chesi M, Egan J B, et al. Clonal competition with alternating dominance in multiple myeloma. *Blood*, 120(5):1067–1076, 2012. ISSN 0006-4971. doi:10.1182/blood-2012-01-405985.

Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, Crosetto N, Foukakis T, and Navin N E. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell*, 173(4):879–893.e13, 2018. ISSN 10974172. doi:10.1016/j.cell.2018.03.041.

Kim M, Lee S, Lim S, and Kim S. SpliceHetero: An information theoretic approach for measuring spliceomic intratumor heterogeneity from bulk tumor RNA-seq. *PLoS ONE*, 14(10):1–19, 2019. ISSN 19326203. doi:10.1371/journal.pone.0223520.

Kim S Y, Jacob L, and Speed T P. Combining calls from multiple somatic mutation-callers. *BMC Bioinformatics*, 15(1):154, 2014. ISSN 1471-2105. doi:10.1186/1471-2105-15-154.

Knijnenburg T A, Wang L, Zimmermann M T, et al. Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. *Cell Reports*, 23(1):239–254.e6, 2018. ISSN 22111247. doi:10.1016/j.celrep.2018.03.076.

Koboldt D C, Zhang Q, Larson D E, Shen D, McLellan M D, Lin L, Miller C A, Mardis E R, Ding L, and Wilson R K. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, 2012. ISSN 10889051. doi:10.1101/gr.129684.111.

Koller D and Friedman N. *Probabilistic Graphical Models: Principles and Techniques (Adaptive Computation and Machine Learning series)*. 2009. ISBN 0262013193. doi:10.1016/j.ccl.2010.07.006.

Kucab J E, Zou X, Morganella S, Joel M, Nanda A S, Nagy E, Gomez C, Degasperi A, Harris R, Jackson S P, Arlt V M, Phillips D H, and Nik-Zainal S. A compendium of mutational signatures of environmental agents. *Cell*, 177(4):821–836.e16, 2019. ISSN 10974172. doi:10.1016/j.cell.2019.03.001.

Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett J C, and Dry J R. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research*, 44(11):e108–e108, 2016. ISSN 0305-1048. doi:10.1093/nar/gkw227.

Laks E, McPherson A, Zahn H, et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell*, 179(5):1207–1221.e22, 2019. ISSN 10974172. doi:10.1016/j.cell.2019.10.026.

Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*, 2001. ISSN 00280836. doi:10.1038/35057062.

Lander E S and Waterman M S. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3):231–239, 1988. ISSN 08887543. doi:10.1016/0888-7543(88)90007-9.

Langmead B, Trapnell C, Pop M, and Salzberg S L. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 2009. ISSN 14747596. doi:10.1186/gb-2009-10-3-r25.

Larson N B and Fridley B L. PurBayes: Estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics*, 29(15):1888–1889, 2013. ISSN 13674803. doi:10.1093/bioinformatics/btt293.

Larson R A, Kondo K, Vardiman J W, Butler A E, Golomb H M, and Rowley J D. Evidence for a 15; 17 translocation in every patient with acute promyelocytic leukemia. *The American Journal of Medicine*, 1984. ISSN 00029343. doi:10.1016/0002-9343(84)90994-X.

Le Morvan M, Zinovyev A, and Vert J P. NetNorM: capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. 2016.

Lee J, Müller P, Gulukota K, and Ji Y. A Bayesian feature allocation model for tumor heterogeneity. *The Annals of Applied Statistics*, 9(2):621–639, 2015. ISSN 1932-6157. doi:10.1214/15-AOAS817.

Lee J, Müller P, Sengupta S, Gulukota K, and Ji Y. Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(4):547–563, 2016. ISSN 00359254. doi:10.1111/rssc.12136.

Leshchiner I, Livitz D, Gainor J F, et al. Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. *bioRxiv*, (i):508127, 2019. doi:10.1101/508127.

Letouzé E, Allory Y, Bollet M a, Radvanyi F, and Guyon F. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome biology*, 11(7):R76, 2010. ISSN 1465-6906. doi:10.1186/gb-2010-11-s1-p25.

Letouzé E, Shinde J, Renault V, et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nature Communications*, 8(1):1315, 2017. ISSN 2041-1723. doi:10.1038/s41467-017-01358-x.

Leung M L, Davis A, Gao R, Casasent A, Wang Y, Sei E, Vilar E, Maru D, Kopetz S, and Navin N E. Single-cell DNA sequencing reveals a latedissemination model in metastatic colorectal cancer. *Genome Research*, 27(8):1287–1299, 2017. ISSN 15495469. doi:10.1101/gr.209973.116.

Li B and Li J Z. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biology*, 15:1–23, 2014. ISSN 1465-6906. doi:10.1186/s13059-014-0473-4.

Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Mass Genomics*, 25(14):1754–1760, 2009. ISSN 1367-4811. doi:10.1093/bioinformatics/btp324.

Li H and Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2010. ISSN 13674803. doi:10.1093/bioinformatics/btp698.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. ISSN 13674803. doi:10.1093/bioinformatics/btp352.

Li Y and Xie X. MixClone: a mixture model for inferring tumor subclonal populations. *BMC genomics*, 16(Suppl 2):1–9, 2015. doi:10.1186/1471-2164-16-S2-S1.

Liu J, Lichtenberg T, Hoadley K A, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416.e11, 2018a. ISSN 00928674. doi:10.1016/j.cell.2018.02.052.

Liu Y, Sethi N S, Hinoue T, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell*, 33(4):721–735.e8, 2018b. ISSN 15356108. doi:10.1016/j.ccell.2018.03.010.

Love M I, Huber W, and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 2014. ISSN 1474760X. doi:10.1186/s13059-014-0550-8.

Macintyre G, Goranova T E, De Silva D, et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nature Genetics*, 50(9):1262–1270, 2018. ISSN 1061-4036. doi:10.1038/s41588-018-0179-8.

Maley C C, Aktipis A, Graham T A, et al. Classifying the evolutionary and ecological features of neoplasms. *Nature Reviews Cancer*, 17(10):605–619, 2017. ISSN 1474-175X. doi:10.1038/nrc.2017.69.

Maley C C, Galipeau P C, Finley J C, Wongsurawat V J, Li X, Sanchez C A, Paulson T G, Blount P L, Risques R A, Rabinovitch P S, and Reid B J. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genetics*, 38(4):468–473, 2006. ISSN 10614036. doi:10.1038/ng1768.

Malikic S, Jahn K, Kuipers J, Sahinalp S C, and Beerenwinkel N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature Communications*, 10(1):1–12, 2019. ISSN 20411723. doi:10.1038/s41467-019-10737-5.

Malikic S, McPherson a W, Donmez N, and Sahinalp C S. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(January):1349–1356, 2015. ISSN 1367-4803. doi:10.1093/bioinformatics/btv003.

Marass F, Mouliere F, Yuan K, Rosenfeld N, and Markowetz F. A phylogenetic latent feature model for clonal deconvolution. *The Annals of Applied Statistics*, 10(4):2377–2404, 2016. ISSN 1932-6157. doi:10.1214/16-AOAS986.

Mardis E R. DNA sequencing technologies: 2006-2016. *Nature Protocols*, 12(2):213–218, 2017. ISSN 17502799. doi:10.1038/nprot.2016.182.

Mardis E R. Insights from large-scale cancer genome sequencing. *Annual Review of Cancer Biology*, 2(1):429–444, 2018. ISSN 2472-3428. doi:10.1146/annurev-cancerbio-050216-122035.

Martincorena I, Fowler J C, Wabik A, Lawson A R, Abascal F, Hall M W, Cagan A, Murai K, Mahbubani K, Stratton M R, Fitzgerald R C, Handford P A, Campbell P J, Saeb-Parsy K, and Jones P H. Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917, 2018. ISSN 10959203. doi:10.1126/science.aau3879.

Martincorena I, Raine K M, Gerstung M, Dawson K J, Haase K, Van Loo P, Davies H, Stratton M R, and Campbell P J. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041.e21, 2017. ISSN 00928674. doi:10.1016/j.cell.2017.09.042.

Martincorena I, Roshan A, Gerstung M, et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886, 2015. ISSN 0036-8075. doi:10.1126/science.aaa6806.

Martinez E Z, Achcar J A, and Aragon D C. Parameter estimation of the beta-binomial distribution: an application using the SAS software. *Ciência e Natura*, 37(3):12–19, 2015. ISSN 2179-460X. doi:10.5902/2179460X17512.

Martinez P, Timmer M R, Lau C T, et al. Dynamic clonal equilibrium and predetermined cancer risk in Barrett's oesophagus. *Nature Communications*, 7:1–10, 2016. ISSN 20411723. doi:10.1038/ncomms12158.

Matsutani T, Ueno Y, Fukunaga T, and Hamada M. Discovering novel mutation signatures by latent Dirichlet allocation with variational bayes inference. *Bioinformatics*, 35(22):4543–4552, 2019. ISSN 1367-4803. doi:10.1093/bioinformatics/btz266.

Maugis C and Michel B. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15:41–68, 2011. ISSN 1292-8100. doi:10.1051/ps/2009004.

Maura F, Degasperi A, Nadeu F, Leongamornlert D, Davies H, Moore L, Royo R, Ziccheddu B, Puente X S, Avet-Loiseau H, Campbell P J, Nik-Zainal S, Campo E, Munshi N, and Bolli N. A practical guide for mutational signature analysis in hematological malignancies. *Nature Communications*, 10(1):2969, 2019. ISSN 2041-1723. doi:10.1038/s41467-019-11037-8.

Maxam A M and Gilbert W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 1977. ISSN 00278424. doi:10.1073/pnas.74.2.560.

McGranahan N, Favero F, De Bruin E C, Birkbak N J, Szallasi Z, and Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*, 7(283):283ra54–283ra54, 2015. ISSN 19466242. doi:10.1126/scitranslmed.aaa1408.

McGranahan N and Swanton C. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell*, 168(4):613–628, 2017. ISSN 10974172. doi:10.1016/j.cell.2017.01.018.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, and DePristo M. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):254–260, 2010. ISSN 1088-9051. doi:10.1101/gr.107524.110.20.

McPherson A W, Roth A, Ha G, Chauve C, Steif A, de Souza C P, Eirew P, Bouchard-Côté A, Aparicio S, Sahinalp S C, and Shah S P. ReMixT: Clone-specific genomic structure estimation in cancer. *Genome Biology*, 18(1):1–14, 2017. ISSN 1474760X. doi:10.1186/s13059-017-1267-2.

Miller C A, White B S, Dees N D, et al. SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Computational Biology*, 10(8):e1003665, 2014. ISSN 1553-7358. doi:10.1371/journal.pcbi.1003665.

Min J W, Kim W J, Han J A, Jung Y J, Kim K T, Park W Y, Lee H O, and Choi S S. Identification of distinct tumor subpopulations in lung adenocarcinoma via single-cell RNA-seq. *PLoS ONE*, 10(8):1–17, 2015. ISSN 19326203. doi:10.1371/journal.pone.0135817.

Miura S, Gomez K, Murillo O, Huuki L A, Vu T, Buturla T, and Kumar S. Predicting clone genotypes from tumor bulk sequencing of multiple samples. *Bioinformatics (Oxford, England)*, 34(23):4017–4026, 2018. ISSN 13674811. doi:10.1093/bioinformatics/bty469.

Miura S, Vu T, Deng J, Buturla T, Choi J, and Kumar S. Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data. *bioRxiv*, page 697318, 2019. doi:10.1101/697318.

Moore L, Leongamornlert D, Coorens T H H, Sanders M A, Brunner S F, Lee-six H, Rahbari R, and Moody S. The mutational landscape of normal human endometrial epithelium. 2018.

Morris L G, Riaz N, Desrichard A, Şenbabaoğlu Y, Hakimi A A, Makarov V, Reis-Filho J S, and Chan T A. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*, 7(9), 2016. ISSN 1949-2553. doi:10.18632/oncotarget.7067.

Mroz E A and Rocco J W. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncology*, 49(3):211–215, 2013. ISSN 13688375. doi:10.1016/j.oraloncology.2012.09.007.

Myers M A, Satas G, and Raphael B J. CALDER: Inferring phylogenetic trees from longitudinal tumor samples. *Cell Systems*, 8(6):514–522.e5, 2019. ISSN 24054720. doi:10.1016/j.cels.2019.05.010.

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie W R, Hicks J, and Wigler M. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–95, 2011. ISSN 00280836. doi:10.1038/nature09807.

Navin N E. Tumor evolution in response to chemotherapy: Phenotype versus genotype. *Cell Reports*, 6(3):417–419, 2014. ISSN 22111247. doi:10.1016/j.celrep.2014.01.035.

Neyman J and Pearson E S. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 231(694-706):289–337, 1933. ISSN 1364-503X. doi:10.1098/rsta.1933.0009.

Nieboer M M, Dorssers L C, Straver R, Looijenga L H, and De Ridder J. TargetClone: A multi-sample approach for reconstructing subclonal evolution of tumors. *PLoS ONE*, 13(11):1–22, 2018. ISSN 19326203. doi:10.1371/journal.pone.0208002.

Nik-Zainal S, Davies H, Staaf J, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016. ISSN 0028-0836. doi:10.1038/nature17676.

Nik-Zainal S, Van Loo P, Wedge D C, et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012. ISSN 00928674. doi:10.1016/j.cell.2012.04.023.

Niknafs N, Beleva-Guthrie V, Naiman D Q, and Karchin R. Subclonal hierarchy inference from somatic mutations: Automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *PLoS Computational Biology*, 11(10):1–26, 2015. ISSN 15537358. doi:10.1371/journal.pcbi.1004416.

Noble R, Burri D, Kather J N, and Beerenwinkel N. Spatial structure governs the mode of tumour evolution. *bioRxiv*, (Figure 1):1–18, 2019. doi:10.1101/586735.

Noorbakhsh J, Kim H, Namburi S, and Chuang J H. Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power. *Scientific Reports*, 8(1):1–12, 2018. ISSN 20452322. doi:10.1038/s41598-018-29154-7.

Nowell P. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976. ISSN 0036-8075. doi:10.1126/science.959840.

Oesper L, Mahmoody A, and Raphael B J. THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7821 LNBI(7):171–172, 2013. ISSN 03029743. doi:10.1007/978-3-642-37195-0_14.

Oesper L, Satas G, and Raphael B J. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24):3532–3540, 2014. ISSN 1367-4803. doi:10.1093/bioinformatics/btu651.

Ogundijo O E and Wang X. SeqClone: Sequential Monte Carlo based inference of tumor subclones. *BMC Bioinformatics*, 20(1):1–15, 2019. ISSN 14712105. doi:10.1186/s12859-018-2562-y.

Ogundijo O E, Zhu K, Wang X, and Anastassiou D. A sequential Monte Carlo algorithm for inference of subclonal structure in cancer. *PLoS ONE*, 14(1), 2019. ISSN 19326203. doi:10.1371/journal.pone.0211213.

Omichessan H, Severi G, and Perduca V. Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. *bioRxiv*, page 483982, 2018. doi:10.1101/483982.

Opasic L, Zhou D, Werner B, Dingli D, and Traulsen A. How many samples are needed to infer truly clonal mutations from heterogenous tumours? *BMC Cancer*, 19(1):1–11, 2019. ISSN 14712407. doi:10.1186/s12885-019-5597-1.

Parikh A R, Leshchiner I, Elagina L, et al. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nature Medicine*, 25(September), 2019. ISSN 1546-170X. doi:10.1038/s41591-019-0561-9.

Park Y, Lim S, Nam J W, and Kim S. Measuring intratumor heterogeneity by network entropy using RNA-seq data. *Scientific Reports*, 6(October):37767, 2016. ISSN 2045-2322. doi:10.1038/srep37767.

Pencina M J and D'Agostino R B. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123, 2004. ISSN 0277-6715. doi:10.1002/sim.1802.

Pereira B, Chin S F, Rueda O M, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications*, 7(May):11479, 2016. ISSN 20411723. doi:10.1038/ncomms11479.

Pich O, Muiños F, Lolkema M P, Steeghs N, Gonzalez-Perez A, and Lopez-Bigas N. The mutational footprints of cancer therapies. *Nature Genetics*, 2019. ISSN 1061-4036. doi:10.1038/s41588-019-0525-5.

Pierre-Jean M, Rigaill G, and Neuvial P. Performance evaluation of DNA copy number segmentation methods. *Briefings in Bioinformatics*, 16(4):600–615, 2015. ISSN 1467-5463. doi:10.1093/bib/bbu026.

Pölsterl S, Gupta P, Wang L, Conjeti S, Katouzian A, and Navab N. Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients. *F1000Research*, 5:2676, 2017. ISSN 2046-1402. doi:10.12688/f1000research.8231.3.

Popic V, Salari R, Hajirasouliha I, Kashef-Haghighi D, West R B, and Batzoglou S. Fast and scalable inference of multi-sample cancer lineages. *Genome Biology*, 16(1):91, 2015. ISSN 1465-6906. doi:10.1186/s13059-015-0647-8.

Poplin R, Chang P C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar P T, Gross S S, Dorfman L, McLean C Y, and DePristo M A. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, 2018. ISSN 1087-0156. doi:10.1038/nbt.4235.

Prandi D, Baca S C, Romanel A, Barbieri C E, Mosquera J M, Fontugne J, Beltran H, Sboner A, Garraway L A, Rubin M A, and Demichelis F. Unraveling the clonal hierarchy of somatic genomic aberrations. *Genome Biology*, 15(8):439, 2014. ISSN 1474-760X. doi:10.1186/s13059-014-0439-6.

Prandi D and Demichelis F. Ploidy- and purity-adjusted allele-specific DNA analysis using CLONETv2. *Current Protocols in Bioinformatics*, 67(1):1–23, 2019. ISSN 1934-3396. doi:10.1002/cpbi.81.

Priestley P, Baber J, Lolkema M P, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*, 575(7781):210–216, 2019. ISSN 14764687. doi:10.1038/s41586-019-1689-y.

Purdom E, Ho C, Grasso C S, Quist M J, Cho R J, and Spellman P. Methods and challenges in timing chromosomal abnormalities within cancer samples. *Bioinformatics (Oxford, England)*, 29(24):3113–20, 2013. ISSN 1367-4811. doi:10.1093/bioinformatics/btt546.

Qiao Y, Quinlan A R, Jazaeri A a, Verhaak R, Wheeler D a, and Marth G T. SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biology*, 15(8):443, 2014. ISSN 1465-6906. doi:10.1186/PREACCEPT-1691129011290725.

Quinlan A R and Hall I M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. ISSN 13674803. doi:10.1093/bioinformatics/btq033.

Rahman M S, Nicholson A E, and Haffari G. HetFHMM: A novel approach to infer tumor heterogeneity using factorial hidden markov models. *Journal of Computational Biology*, 25(2):182–193, 2018. ISSN 1557-8666. doi:10.1089/cmb.2017.0101.

Raman L, Dheedene A, De Smet M, Van Dorpe J, and Menten B. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Research*, 47(4):1605–1614, 2019. ISSN 0305-1048. doi:10.1093/nar/gky1263.

Ramazzotti D, Graudenzi A, De Sano L, Antoniotti M, and Caravagna G. Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data. *BMC Bioinformatics*, 20(1):210, 2019. ISSN 1471-2105. doi:10.1186/s12859-019-2795-4.

Ramazzotti D, Lal A, Liu K, Tibshirani R, and Sidow A. De novo mutational signature discovery in tumor genomes using SparseSignatures. *bioRxiv*, page 384834, 2018. doi:10.1101/384834.

Raphael B J, Hruban R H, Aguirre A J, et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell*, 32(2):185–203.e13, 2017. ISSN 18783686. doi:10.1016/j.ccell.2017.07.007.

Rausch T, Zichner T, Schlattl A, Stütz A M, Benes V, and Korbel J O. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18), 2012. ISSN 13674803. doi:10.1093/bioinformatics/bts378.

Ray S, Jia B, Safavi S, van Opijnen T, Isberg R, Rosch J, and Bento J. Exact inference under the perfect phylogeny model. 2019.

Reiter J G, Baretti M, Gerold J M, Makohon-Moore A P, Daud A, Iacobuzio-Donahue C A, Azad N S, Kinzler K W, Nowak M A, and Vogelstein B. An analysis of genetic heterogeneity in untreated cancers. *Nature Reviews Cancer*, pages 16–21, 2019. ISSN 1474-175X. doi: 10.1038/s41568-019-0185-x.

Reiter J G, Makohon-Moore A P, Gerold J M, Bozic I, Chatterjee K, Iacobuzio-Donahue C A, Vogelstein B, and Nowak M A. Reconstructing metastatic seeding patterns of human cancers. *Nature Communications*, 8(1):14114, 2017. ISSN 2041-1723. doi: 10.1038/ncomms14114.

Ren W, Ye X, Su H, et al. Genetic landscape of hepatitis B virus–associated diffuse large B-cell lymphoma. *Blood*, 131(24):2670–2681, 2018. ISSN 0006-4971. doi: 10.1182/blood-2017-11-817601.

Riaz N, Blecua P, Lim R S, Shen R, Higginson D S, Weinhold N, Norton L, Weigelt B, Powell S N, and Reis-Filho J S. Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nature Communications*, 8(1):857, 2017. ISSN 2041-1723. doi:10.1038/s41467-017-00921-w.

Ricketts C, Seidman D, Popic V, Hormozdiari F, Batzoglou S, and Hajirasouliha I. Meltos: multi-sample tumor phylogeny reconstruction for structural variants. *Bioinformatics*, 2019. ISSN 1367-4803. doi:10.1093/bioinformatics/btz737.

Robinson J T, Thorvaldsdóttir H, Wenger A M, Zehir A, and Mesirov J P. Variant review with the Integrative Genomics Viewer. *Cancer Research*, 77(21):e31–e34, 2017. ISSN 0008-5472. doi:10.1158/0008-5472.CAN-17-0337.

Robinson W, Sharan R, and Leiserson M D M. Modeling clinical and molecular covariates of mutational process activity in cancer. *Bioinformatics*, 35(14):i492–i500, 2019. ISSN 1367-4803. doi:10.1093/bioinformatics/btz340.

Roerink S F, Sasaki N, Lee-Six H, et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature*, 556(7702):457–462, 2018. ISSN 0028-0836. doi:10.1038/s41586-018-0024-3.

Roman T, Xie L, and Schwartz R. Automated deconvolution of structured mixtures from heterogeneous tumor genomic data. *PLOS Computational Biology*, 13(10):e1005815, 2017. ISSN 1553-7358. doi:10.1371/journal.pcbi.1005815.

Rörsch A, Beukers R, Ijlstra J, and Berends W. The effect of U.V.-light on some components of the nucleic acids: I. Uracil, thymine. *Recueil des Travaux Chimiques des Pays-Bas*, 77(5):423–429, 1958. ISSN 01650513. doi:10.1002/recl.19580770506.

Rosenberg A and Hirschberg J. V-Measure: A conditional entropy-based external cluster evaluation measure. *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1(June):410–420, 2007.

Rosenthal R, Cadieux E L, Salgado R, et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature*, 567(7749):479–485, 2019. ISSN 14764687. doi:10.1038/s41586-019-1032-7.

Rosenthal R, McGranahan N, Herrero J, Taylor B S, and Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1):31, 2016. ISSN 1474-760X. doi:10.1186/s13059-016-0893-4.

Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Côté A, and Shah S P. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, 2014. ISSN 1548-7091. doi:10.1038/nmeth.2883.

Rowley J D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243(5405):290–293, 1973. ISSN 00280836. doi:10.1038/243290a0.

Rowley J D, Golomb H M, Vardiman J, Fukuhara S, Dougherty C, and Potter D. Further evidence for a non-random chromosomal abnormality in acute promyelocytic leukemia. *International Journal of Cancer*, 1977. ISSN 10970215. doi:10.1002/ijc.2910200608.

Royer-Bertrand B, Torsello M, Rimoldi D, et al. Comprehensive genetic landscape of uveal melanoma by whole-genome sequencing. *The American Journal of Human Genetics*, 99(5):1190–1198, 2016. ISSN 00029297. doi:10.1016/j.ajhg.2016.09.008.

Rubanova Y, Shi R, Li R, Wintersinger J, Sahin N, and Morris Q. TrackSig: reconstructing evolutionary trajectories of mutations in cancer. *bioRxiv*, page 260471, 2018. doi:10.1101/260471.

Safonov A, Jiang T, Bianchini G, Győrffy B, Karn T, Hatzis C, and Pusztai L. Immune gene expression is associated with genomic aberrations in breast cancer. *Cancer Research*, 77(12):3317–3324, 2017. ISSN 0008-5472. doi:10.1158/0008-5472.CAN-16-3478.

Salari R, Saleh S S, Kashef-Haghighi D, Khavari D, Newburger D E, West R B, Sidow A, and Batzoglou S. Inference of tumor phylogenies with improved somatic mutation discovery. *Journal of Computational Biology*, 20(11):933–944, 2013. ISSN 1066-5277. doi:10.1089/cmb.2013.0106.

Salcedo A, Tarabichi M, Espiritu S M G, et al. Creating standards for evaluating tumour subclonal reconstruction. Technical report, 2018.

Sanger F, Nicklen S, and Coulson A R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 1977. ISSN 00278424. doi:10.1073/pnas.74.12.5463.

Satas G and Raphael B J. Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics*, 33(14):i152–i160, 2017. ISSN 14602059. doi:10.1093/bioinformatics/btx270.

Saunders G, Baudis M, Becker R, et al. Leveraging european infrastructures to access 1 million human genomes by 2022. *Nature Reviews Genetics*, 2019. ISSN 1471-0056. doi:10.1038/s41576-019-0156-9.

Schröder M S, Culhane A C, Quackenbush J, and Haibe-Kains B. survcomp: An R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, 27(22):3206–3208, 2011. ISSN 13674803. doi:10.1093/bioinformatics/btr511.

Schwartz R and Schäffer A A. The evolution of tumour phylogenetics: Principles and practice. *Nature Reviews Genetics*, 18(4):213–229, 2017. ISSN 14710064. doi:10.1038/nrg.2016.170.

Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. ISSN 0090-5364. doi:10.1214/aos/1176344136.

Schwarz R F, Trinh A, Sipos B, Brenton J D, Goldman N, and Markowetz F. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Computational Biology*, 10(4), 2014. ISSN 15537358. doi:10.1371/journal.pcbi.1003535.

Scott J and Marusyk A. Somatic clonal evolution: A selection-centric perspective. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1867(2):139–150, 2017. ISSN 18792561. doi: 10.1016/j.bbcan.2017.01.006.

Sengupta S, Wang J, Lee J, Müller P, Gulukota K, Banerjee A, and Ji Y. BayClone: Bayesian nonparametric inference of tumor subclones using ngs data. *Proceedings of The Pacific Symposium on Biocomputing (PSB)*, 20:20, 2015. ISSN 2335-6936.

Sharma A, Merritt E, Hu X, Cruz A, Jiang C, Sarkodie H, Zhou Z, Malhotra J, Riedlinger G M, and De S. Non-genetic intra-tumor heterogeneity is a major predictor of phenotypic heterogeneity and ongoing evolutionary dynamics in lung tumors. *Cell Reports*, 29(8):2164–2174.e5, 2019. ISSN 22111247. doi:10.1016/j.celrep.2019.10.045.

Shen R and Seshan V E. FACETS: Allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research*, 44(16):1–9, 2016. ISSN 13624962. doi:10.1093/nar/gkw520.

Shendure J and Ji H. Next-generation DNA sequencing. 2008. doi:10.1038/nbt1486.

Shi W, Ng C K, Lim R S, Jiang T, Kumar S, Li X, Wali V B, Piscuoglio S, Gerstein M B, Chagpar A, Weigelt B, Pusztai L, Reis-Filho J S, and Hatzis C. Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity. *SSRN Electronic Journal*, 25(6):1446–1457, 2018. ISSN 1556-5068. doi:10.2139/ssrn.3155634.

Shibata T, Arai Y, and Totoki Y. Molecular genomic landscapes of hepatobiliary cancer. *Cancer Science*, 109(5):1282–1291, 2018. ISSN 13479032. doi:10.1111/cas.13582.

Shinde J, Bayard Q, Imbeaud S, Hirsch T Z, Liu F, Renault V, Zucman-Rossi J, and Letouzé E. Palimpsest: an r package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics*, 34(19):3380–3381, 2018. ISSN 1367-4803. doi:10.1093/bioinformatics/bty388.

Shiraishi Y, Tremmel G, Miyano S, and Stephens M. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genetics*, 11(12):1–21, 2015. ISSN 15537404. doi:10.1371/journal.pgen.1005657.

Sottoriva A, Barnes C P, and Graham T A. Catch my drift? making sense of genomic intra-tumour heterogeneity. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1867(2):95–100, 2017. ISSN 0304419X. doi:10.1016/j.bbcan.2016.12.003.

Sottoriva A, Kang H, Ma Z, Graham T A, Salomon M P, Zhao J, Marjoram P, Siegmund K, Press M F, Shibata D, and Curtis C. A Big Bang model of human colorectal tumor growth. *Nature Genetics*, 47(3):209–216, 2015a. ISSN 1061-4036. doi:10.1038/ng.3214.

Sottoriva A, Kang H, Ma Z, Graham T A, Salomon M P, Zhao J, Marjoram P, Siegmund K, Press M F, Shibata D, and Curtis C. A Big Bang model of human colorectal tumor growth. *Nature Genetics*, 47(3):209–216, 2015b. ISSN 1061-4036. doi:10.1038/ng.3214.

Spinella J F, Mehanna P, Vidal R, Saillour V, Cassart P, Richer C, Ouimet M, Healy J, and Sinnett D. SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics*, 17(1):912, 2016. ISSN 1471-2164. doi:10.1186/s12864-016-3281-2.

Strino F, Parisi F, Micsinai M, and Kluger Y. TrAp: A tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research*, 41(17):1–15, 2013. ISSN 03051048. doi:10.1093/nar/gkt641.

Szczurek E and Beerenwinkel N. Modeling mutual exclusivity of cancer mutations. *PLoS Computational Biology*, 10(3), 2014. doi:10.1371/journal.pcbi.1003503.

Tai A S, Peng C H, Peng S C, and Hsieh W P. Decomposing the subclonal structure of tumors with two-way mixture models on copy number aberrations. *PLOS ONE*, 13(12):e0206579, 2018. ISSN 1932-6203. doi:10.1371/journal.pone.0206579.

Tang H and Thomas P D. Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics*, 203(2):635–647, 2016. ISSN 0016-6731. doi:10.1534/genetics.116.190033.

Tarabichi M, Martincorena I, Gerstung M, et al. Neutral tumor evolution? 2018. doi:10.1038/s41588-018-0258-x.

Tate J G, Bamford S, Jubb H C, et al. COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2019. ISSN 13624962. doi:10.1093/nar/gky1015.

Toosi H, Moeini A, and Hajirasouliha I. BAMSE: Bayesian model selection for tumor phylogeny inference among multiple samples. *BMC Bioinformatics*, 20(S11):282, 2019. ISSN 1471-2105. doi:10.1186/s12859-019-2824-3.

Turajlic S, McGranahan N, and Swanton C. Inferring mutational timing and reconstructing tumour evolutionary histories. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1855(2):264–275, 2015. ISSN 0304419X. doi:10.1016/j.bbcan.2015.03.005.

Turajlic S, Sottoriva A, Graham T, and Swanton C. Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics*, 20(7):404–416, 2019. ISSN 1471-0056. doi:10.1038/s41576-019-0114-6.

Turajlic S and Swanton C. TRACERx renal: tracking renal cancer evolution through therapy. *Nature Reviews Urology*, 14(10):575–576, 2017. ISSN 1759-4812. doi:10.1038/nrurol.2017.112.

Turajlic S, Xu H, Litchfield K, et al. Deterministic evolutionary trajectories influence primary tumor growth: TRACERx renal. *Cell*, 173(3):595–610.e11, 2018. ISSN 00928674. doi:10.1016/j.cell.2018.03.043.

Van Belle V, Pelckmans K, Van Huffel S, and Suykens J A. Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011. ISSN 09333657. doi:10.1016/j.artmed.2011.06.006.

Van den Eynden J, Jiménez-Sánchez A, Miller M L, and Larsson E. Lack of detectable neoantigen depletion signals in the untreated cancer genome. *Nature Genetics*, page 478263, 2019. ISSN 1061-4036. doi:10.1038/s41588-019-0532-6.

van Dijk E L, Jaszczyszyn Y, Naquin D, and Thermes C. The third revolution in sequencing technology. *Trends in Genetics*, 34(9):666–681, 2018. ISSN 13624555. doi:10.1016/j.tig.2018.05.008.

van Rens K E, Mäkinen V, and Tomescu A I. SNV-PPILP: refined SNV calling for tumor data using perfect phylogenies and ILP. *Bioinformatics*, 31(7):1133–1135, 2015. ISSN 1460-2059. doi:10.1093/bioinformatics/btu755.

Venkatesan S and Swanton C. Tumor evolutionary principles: How intratumor heterogeneity influences cancer treatment and outcome. *American Society of Clinical Oncology Educational Book*, 36:e141–e149, 2016. ISSN 1548-8748. doi:10.14694/EDBK_158930.

Verhagen C V, Vossen D M, Borgmann K, et al. Fanconi anemia and homologous recombination gene variants are associated with functional DNA repair defects in vitro and poor outcome in patients with advanced head and neck squamous cell carcinoma. *Oncotarget*, 9(26):18198–18213, 2018. ISSN 19492553. doi:10.18632/oncotarget.24797.

Volkova N V, Meier B, González-Huici V, Bertolini S, Gonzalez S, Abascal F, Martincorena I, Campbell P J, Gartner A, and Gerstung M. Mutational signatures are jointly shaped by DNA damage and repair. *bioRxiv*, 44(0):686295, 2019. doi:10.1101/686295.

Wang Y, Waters J, Leung M L, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, 2014. ISSN 14764687. doi:10.1038/nature13600.

Warsow G, Hübschmann D, Kleinheinz K, et al. Genomic features of renal cell carcinoma with venous tumor thrombus. *Scientific Reports*, 8(1):7477, 2018. ISSN 2045-2322. doi:10.1038/s41598-018-25544-z.

Wen Y, Wei Y, Zhang S, Li S, Liu H, Wang F, Zhao Y, Zhang D, and Zhang Y. Cell subpopulation deconvolution reveals breast cancer heterogeneity based on DNA methylation signature. *Briefings in bioinformatics*, (February):1–15, 2016. ISSN 1477-4054. doi:10.1093/bib/bbw028.

Werner B and Sottoriva A. Variation of mutational burden in healthy human tissues suggests non-random strand segregation and allows measuring somatic mutation rates. *PLOS Computational Biology*, 14(6):e1006233, 2018. ISSN 1553-7358. doi:10.1371/journal.pcbi.1006233.

Wilks S S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938. ISSN 0003-4851. doi:10.1214/aoms/1177732360.

Williams M J, Werner B, Barnes C P, Graham T A, and Sottoriva A. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48(3):1–9, 2016. ISSN 1061-4036. doi:10.1038/ng.3489.

Williams M J, Werner B, Heide T, Curtis C, Barnes C P, Sottoriva A, and Graham T A. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 50(6):895–903, 2018. ISSN 15461718. doi:10.1038/s41588-018-0128-6.

Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16:15–24, 2018. ISSN 20010370. doi:10.1016/j.csbj.2018.01.003.

Xu Y, Müller P, Yuan Y, Gulukota K, and Ji Y. MAD bayes for tumor heterogeneity—feature allocation with exponential family sampling. *Journal of the American Statistical Association*, 110(510):503–514, 2015. ISSN 0162-1459. doi:10.1080/01621459.2014.995794.

Yau C. Accounting for sources of bias and uncertainty in copy number-based statistical deconvolution of heterogeneous tumour samples. *Yonago Acta Medica*, pages 1–25, 2014. doi:10.1101/004655.

Ye K, Schulz M H, Long Q, Apweiler R, and Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, 2009. ISSN 13674803. doi:10.1093/bioinformatics/btp394.

Yu K, Shin S J, Zhu H, and Wang W. CliP: fast subclonal architecture reconstruction from whole-genome sequencing data. 2018.

Yu Z, Li A, and Wang M. CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data. *BMC Bioinformatics*, 17(1):310, 2016. ISSN 1471-2105. doi:10.1186/s12859-016-1174-7.

Yu Z, Li A, and Wang M. CLImAT-HET: detecting subclonal copy number alterations and loss of heterozygosity in heterogeneous tumor samples from whole-genome sequencing data. *BMC Medical Genomics*, 10(1):15, 2017. ISSN 1755-8794. doi:10.1186/s12920-017-0255-4.

Yuan K, Macintyre G, Liu W, Group P E, Working H, and Markowetz F. Ccube: A fast and robust method for estimating cancer cell fractions. *bioRxiv*, page 484402, 2018. doi: 10.1101/484402.

Yuan K, Sakoparnig T, Markowetz F, and Beerenwinkel N. BitPhylogeny: A probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biology*, 16(1):36, 2015. ISSN 1474760X. doi:10.1186/s13059-015-0592-6.

Zaccaria S, El-Kebir M, Klau G W, and Raphael B J. The copy-number tree mixture deconvolution problem and applications to multi-sample bulk sequencing tumor data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10229 LNCS, pages 318–335. 2017. ISBN 9783319569697. doi:10.1007/978-3-319-56970-3_20.

Zaccaria S and Raphael B J. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *bioRxiv*, page 496174, 2018. doi:10.1101/496174.

Zare H, Wang J, Hu A, Weber K, Smith J, Nickerson D, Song C Z, Witten D, Blau C A, and Noble W S. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Computational Biology*, 10(7), 2014. ISSN 15537358. doi:10.1371/journal.pcbi.1003703.

Zeng L, Warren J L, and Zhao H. Phylogeny-based tumor subclone identification using a Bayesian feature allocation model. *The Annals of Applied Statistics*, 13(2):1212–1241, 2019. ISSN 1932-6157. doi:10.1214/18-AOAS1223.

Zerbino D R, Achuthan P, Akanni W, et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018. ISSN 13624962. doi:10.1093/nar/gkx1098.

Zhou H, Neelakantan D, and Ford H L. Clonal cooperativity in heterogenous cancers. *Seminars in Cell & Developmental Biology*, 64:79–89, 2017. ISSN 10849521. doi: 10.1016/j.semcdb.2016.08.028.

Zhou T, Müller P, Sengupta S, and Ji Y. PairClone: a Bayesian subclone caller based on mutation pairs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):rssc.12328, 2018. ISSN 0035-9254. doi:10.1111/rssc.12328.

Zhou T, Sengupta S, Müller P, and Ji Y. TreeClone: Reconstruction of tumor subclone phylogeny based on mutation pairs using next generation sequencing data. *The Annals of Applied Statistics*, 13(2):874–899, 2019. ISSN 1932-6157. doi:10.1214/18-AOAS1224.

Zhu W, Kuziora M, Creasy T, Lai Z, Morehouse C, Guo X, Sebastian Y, Shen D, Huang J, Dry J R, Xue F, Jiang L, Yao Y, and Higgs B W. BubbleTree: an intuitive visualization to elucidate tumoral aneuploidy and clonality using next generation sequencing data. *Nucleic Acids Research*, 44(4):e38–e38, 2016. ISSN 0305-1048. doi:10.1093/nar/gkv1102.

Zou M, Jin R, and Au K F. Revealing tumor heterogeneity of breast cancer by utilizing the linkage between somatic and germline mutations. *Briefings in Bioinformatics*, 00(April):1–10, 2018a. ISSN 1467-5463. doi:10.1093/bib/bby084.

Zou X, Owusu M, Harris R, Jackson S P, Loizou J I, and Nik-zainal S. Validating the concept of mutational signatures with isogenic cell models. *Nature Communications*, 9(1):1744, 2018b. ISSN 2041-1723. doi:10.1038/s41467-018-04052-8.

Zucker M R, Abruzzo L V, Herling C D, Barron L L, Keating M J, Abrams Z B, Heerema N, and Coombes K R. Inferring clonal heterogeneity in cancer using SNP arrays and whole genome sequencing. *Bioinformatics*, 35(17):2924–2931, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz057.

Appendix A

Supplementary materials for the ITH methods comparison

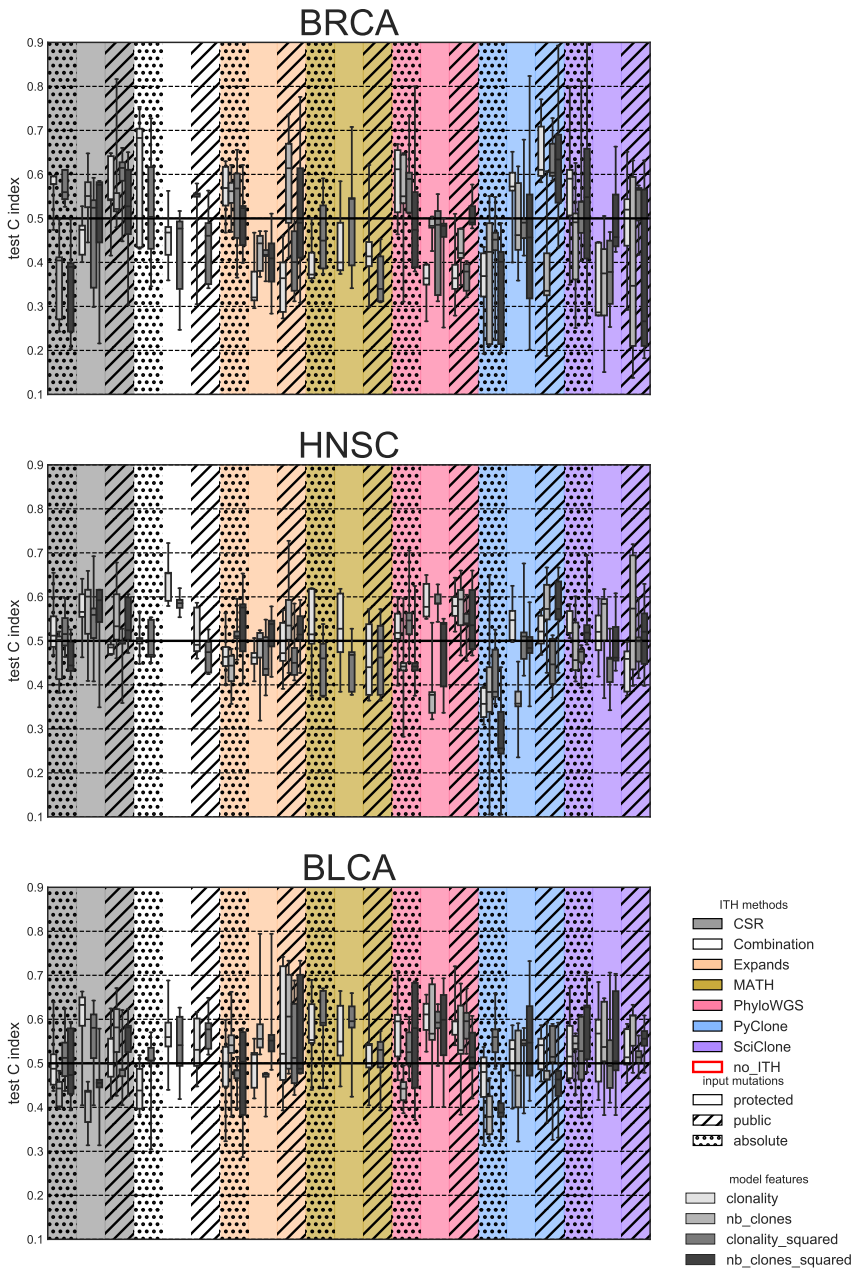

Figure A.1 – Prognostic power of diverse combination of ITH-derived features, on the three cancer types (respectively BRCA, HNSC and BLCA from top to bottom). In each plot, the background color indicates the ITH method used. Each method is tested on protected or public mutations (hashed). For each method, we assess the ability to predict survival with a survival SVM using 4 sets of features: (i) the number of clones alone, (ii) the five custom features which include the number of clones, and (iii) and (iv) the concatenations of features in (i) and (ii) with their squares, to account for possible nonlinear quadratic effects. We observe no clear trend of one of the two sets performs systematically better than the other, and the squared features have not significantly improved results either.

Variable	Hazard ratio	P-value	Corrected P-value
age_at_diagnosis	1.032023	4.328538e-05	0.000577
PROSPECTIVE_COLLECTION_NO	0.885701	4.431127e-01	0.631390
PROSPECTIVE_COLLECTION_YES	1.129050	4.431127e-01	0.631390
RETROSPECTIVE_COLLECTION_NO	1.138738	4.114412e-01	0.631390
RETROSPECTIVE_COLLECTION_YES	0.885701	4.431127e-01	0.631390
SEX_Female	1.114329	5.131653e-01	0.651638

SEX_Male	0.897401	5.131653e-01	0.651638
RACE_WHITE	1.144930	5.054529e-01	0.651638
HISTORY_OTHER_MALIGNANCY_No	1.100403	5.660149e-01	0.696634
HISTORY_OTHER_MALIGNANCY_Yes	0.908758	5.660149e-01	0.696634
NONINVASIVE_BLADDER_HISTORY_NO	0.878835	3.891088e-01	0.610367
NONINVASIVE_BLADDER_HISTORY_YES	0.868563	4.940541e-01	0.651638
NONINVASIVE_BLADDER_HISTORY_[Not Available]	1.298013	1.113381e-01	0.329891
NONINVASIVE_BLADDER_CA_TX_TYPE_[Not Applicable]	0.878835	3.891088e-01	0.610367
NONINVASIVE_BLADDER_CA_TX_TYPE_[Not Available]	1.225435	2.146545e-01	0.490639
TX_90DAYS_POST_RESECTION_[Not Applicable]	0.878835	3.891088e-01	0.610367
TX_90DAYS_POST_RESECTION_[Not Available]	1.329106	6.724015e-02	0.244510
TX_COMPLETE_RESPONSE_[Not Applicable]	0.878835	3.891088e-01	0.610367
TX_COMPLETE_RESPONSE_[Not Available]	1.183784	2.774378e-01	0.539748
TX_INDUCTION_COURSES_INDICATOR_[Not Applicable]	0.878835	3.891088e-01	0.610367
TX_INDUCTION_COURSES_INDICATOR_[Not Available]	1.234697	1.748396e-01	0.436750
TX_MAINTENANCE_COURSES_INDICATOR_[Not Applicable]	0.878835	3.891088e-01	0.610367
TX_MAINTENANCE_COURSES_INDICATOR_[Not Available]	1.233463	1.768826e-01	0.436750
OCCUPATION_CURRENT_Retired	0.917140	6.255270e-01	0.733401
OCCUPATION_CURRENT_[Not Available]	1.271108	1.467738e-01	0.404893
OCCUPATION_CURRENT_retired	0.960885	8.446695e-01	0.889126
OCCUPATION_PRIMARY_[Not Available]	1.295025	9.419336e-02	0.301419
OCCUPATION_PRIMARY_CHEMICAL_EXPOSURE_[Not Avail...	1.765620	4.091600e-03	0.032733
OCCUPATION_PRIMARY_INDUSTRY_[Not Available]	1.160094	3.247413e-01	0.590439
FAMILY_HISTORY_CANCER_RELATIONSHIP_[Not Available]	1.226713	1.856187e-01	0.436750
FAMILY_HISTORY_CANCER_TYPE_[Not Available]	1.226713	1.856187e-01	0.436750
RADIATION_TREATMENT_ADJUVANT_NO	0.712057	2.632089e-02	0.131604
RADIATION_TREATMENT_ADJUVANT_[Not Available]	1.354145	5.230845e-02	0.210985
PHARMACEUTICAL_TX_ADJUVANT_NO	0.986814	9.296012e-01	0.941368
PHARMACEUTICAL_TX_ADJUVANT_YES	0.676359	4.539278e-02	0.201746
PHARMACEUTICAL_TX_ADJUVANT_[Not Available]	1.387004	3.427365e-02	0.161288
HISTOLOGICAL_SUBTYPE_Non-Papillary	1.364821	7.158132e-02	0.248979
HISTOLOGICAL_SUBTYPE_Papillary	0.673148	2.630851e-02	0.131604
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Other meth...	0.813722	4.457771e-01	0.631390
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Transureth...	1.234871	2.692175e-01	0.539748
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_OTHER_[Not	1.290044	3.459972e-01	0.610367
AJCC_STAGING_EDITION_6th	0.931002	6.427039e-01	0.733401
AJCC_STAGING_EDITION_7th	1.061071	7.031650e-01	0.760178
ANGIOLYMPHATIC_INVASION_NO	0.485392	3.934074e-05	0.000577
ANGIOLYMPHATIC_INVASION_YES	1.818002	6.992848e-05	0.000799
ANGIOLYMPHATIC_INVASION_[Not Available]	1.282364	1.626703e-01	0.433787
LYMPH_NODES_EXAMINED_NO	0.861162	4.914651e-01	0.651638
LYMPH_NODES_EXAMINED_YES	1.144075	4.498650e-01	0.631390
EXTRACAPSULAR_EXTENSION_NO	1.101253	5.846915e-01	0.708717
EXTRACAPSULAR_EXTENSION_YES	1.411088	5.538363e-02	0.210985
EXTRACAPSULAR_EXTENSION_[Not Available]	0.748263	5.291460e-02	0.210985
EXTRACAPSULAR_EXTENSION_PRESENT_[Not Available]	0.804289	2.901143e-01	0.539748
METASTATIC_SITE_Lymph node only	1.687527	6.731283e-03	0.041832
METASTATIC_SITE_None	0.641938	5.729977e-03	0.041673
METASTATIC_SITE_[Not Available]	1.050773	7.417814e-01	0.791234
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage II	0.457221	3.329395e-05	0.000577
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage III	0.840677	2.814664e-01	0.539748
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IV	2.264256	4.728912e-08	0.000004
INCIDENTAL_PROSTATE_CANCER_NO	1.002834	9.852841e-01	0.985284
INCIDENTAL_PROSTATE_CANCER_YES	0.926590	6.685831e-01	0.742870
INCIDENTAL_PROSTATE_CANCER_[Not Available]	1.107850	6.188245e-01	0.733401
AJCC_INCIDENTAL_PROSTATE_CANCER_[Not Available]	1.135748	4.866109e-01	0.651638
PRIMARY_SITE_Bladder - NOS	0.839975	2.465470e-01	0.533075
CLIN_T_STAGE_T2	1.072482	6.867865e-01	0.752643
CLIN_T_STAGE_[Not Available]	0.801825	1.408015e-01	0.402290
ICD_10_C67.2	0.981491	9.276451e-01	0.941368
ICD_10_C67.9	0.934338	6.508937e-01	0.733401
ICD_O_3_HISTOLOGY_8120/3	1.264392	2.765995e-01	0.539748
ICD_O_3_HISTOLOGY_8130/3	0.790498	2.838725e-01	0.539748
ICD_O_3_SITE_C67.2	0.981491	9.276451e-01	0.941368
ICD_O_3_SITE_C67.9	0.934338	6.508937e-01	0.733401

TISSUE_SOURCE_SITE_DK	0.563206	2.178870e-02	0.124507
TISSUE_SOURCE_SITE_XF	1.263603	2.246627e-01	0.499250
AJCC_TUMOR_PATHOLOGIC_PT_simple_T2	0.515628	4.470852e-04	0.004471
AJCC_TUMOR_PATHOLOGIC_PT_simple_T3	1.271201	1.094625e-01	0.329891
AJCC_TUMOR_PATHOLOGIC_PT_simple_T4	1.880901	9.411876e-04	0.008366
AJCC_NODES_PATHOLOGIC_PN_simple_N0	0.450186	1.450583e-07	0.000006
AJCC_NODES_PATHOLOGIC_PN_simple_N2	2.130213	5.222366e-06	0.000139
AJCC_METASTASIS_PATHOLOGIC_PM_simple_M0	0.663211	6.797682e-03	0.041832
AJCC_METASTASIS_PATHOLOGIC_PM_simple_MX	1.305408	7.476644e-02	0.249221

Table A.2 – Clinical variables significance for single-variable cox model for BLCA (409 patients). Variable significantly associated with survival are shaded.

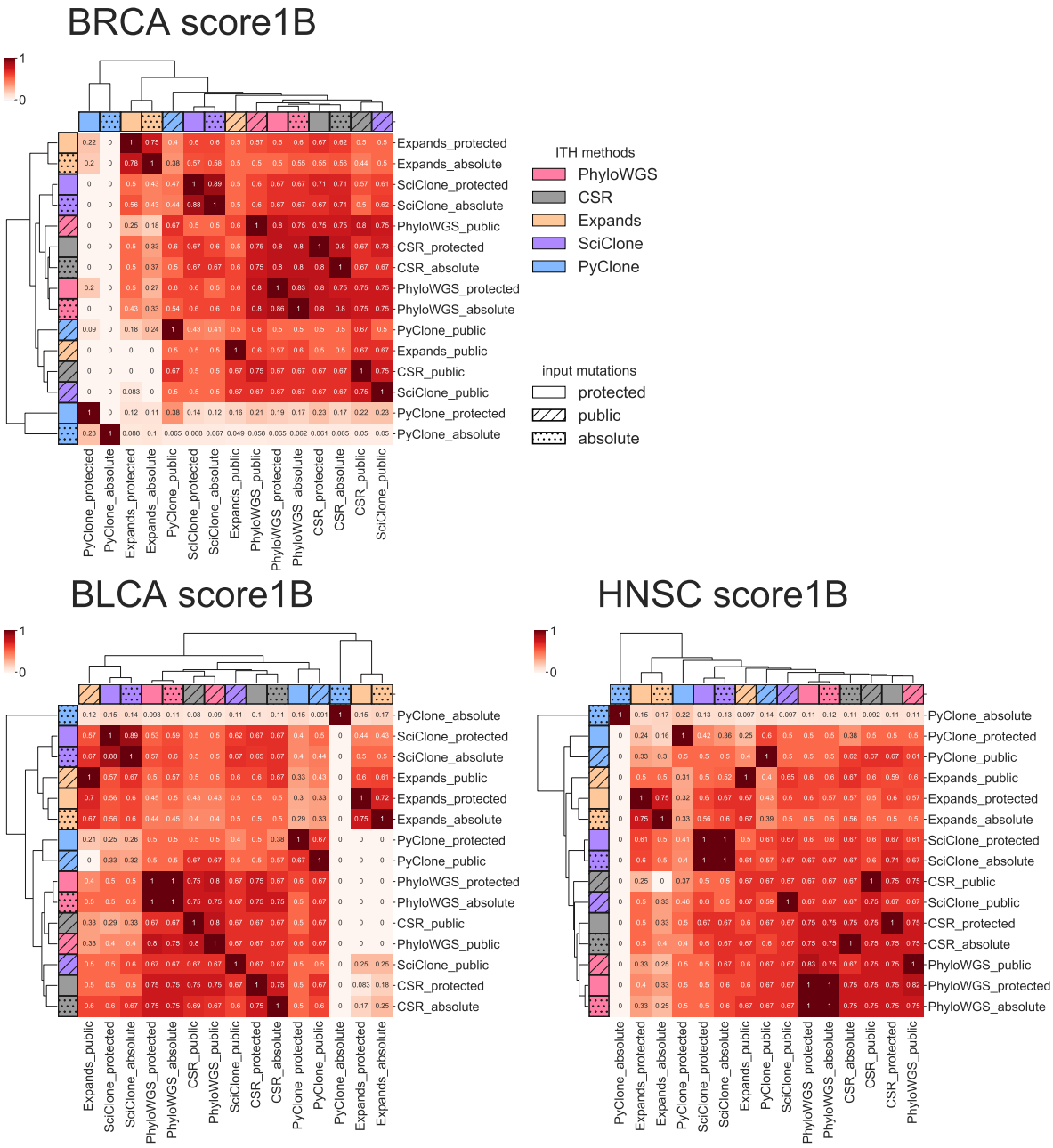

Figure A.2 – Pairwise computation of score1B for the different ITH methods and inputs. Score1B is a metric designed in Salcedo et al. [2018] penalizes differences between the number of clones inferred in each case in a symmetric way (only the difference matters, either more or fewer clones are detected), following the formula $\frac{J_1+1-\min(J_1+1,|J_2-J_1|)}{J_1+1}$, with J_1 and J_2 the numbers of clones found by each method. The score was computed for all patients, and this heatmap represents the median score. We observe a particular feature of PyClone, which tends to find a lot (sometimes several dozens) of clones with only one mutation. They were discarded when comparing the number of clones, but not for the computation of metric 1B to ensure consistency with the other metrics.

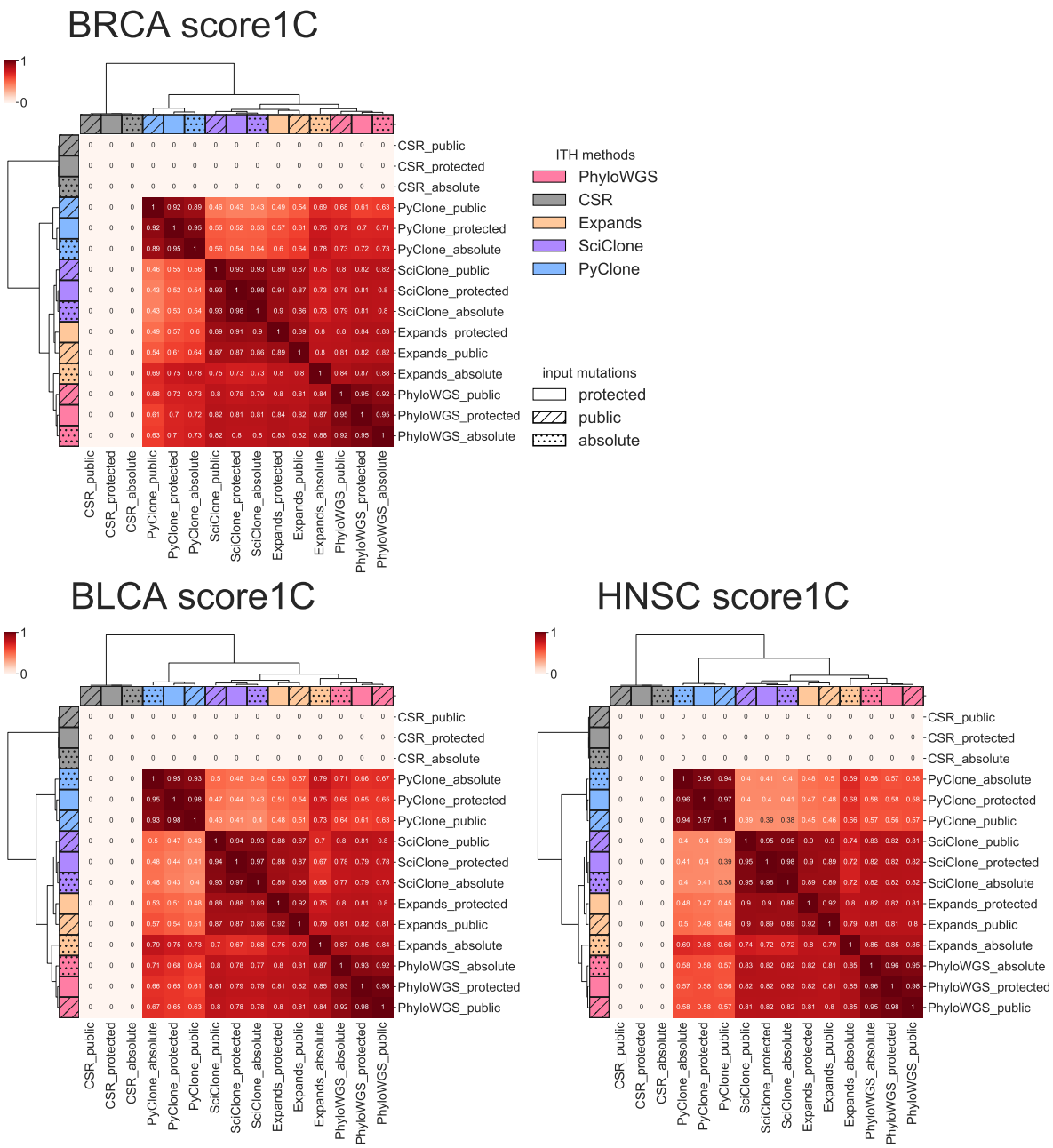

Figure A.3 – Pairwise computation of score1C for the different ITH methods and inputs. Score1C is a metric designed in Salcedo et al. [2018] that represents the Wasserstein distance between the cancer cell fraction (CCF) distribution resulting from each clone's mean CCF and number of mutations. Due to the number of single-mutation clones of PyClone, the resulting distribution is quite different from the other cases. As CSR only takes as input the mutation attribution to clones by other methods, without taking into account their CCF, we did not compute score1C for that method. The score was computed for all patients, and this heatmap represents the median score.

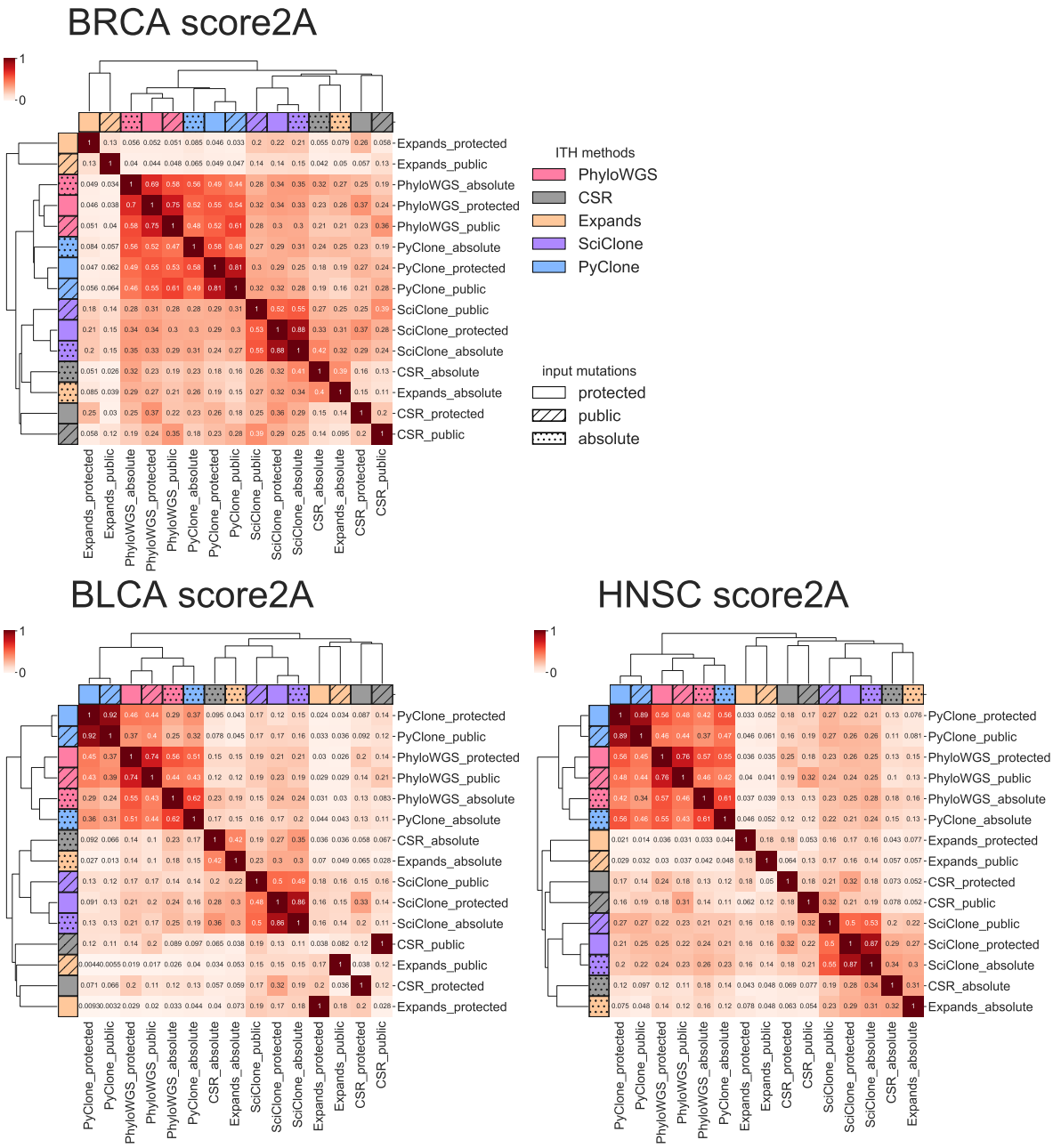

Figure A.4 – Pairwise computation of score2A for the different ITH methods and inputs. Score2A is a metric designed in Salcedo et al. [2018] that assesses the similarity of the mutation clustering resulting from subclonal reconstruction (see Methods for details). We recover the previously observed pattern that PyClone and PhyloWGS are the closest methods. The score was computed for all patients, and this heatmap represents the median score.

	BRCA	BLCA	HNSC
number of samples	962	351	445
Protected mutations			
average	530	735	456
std	1,189	890	544
min	79	61	31
median	342	525	340
max	21,821	12,774	7,941
Public mutations			
average	121	352	202
std	375	426	271
min	1	2	1
median	63	241	140
max	7,919	5,478	3,935

Table A.1 – Summary statistics of the number of protected and public mutations per sample for BRCA, BLCA and HNSC samples. The protected set corresponds to raw variant calling outputs. The public set corresponds to publicly available SNV calls.

Variable	Hazard ratio	P-value	Corrected P-value
age_at_diagnosis	1.032118	4.566898e-07	1.077788e-05
PROSPECTIVE_COLLECTION_NO	0.602422	4.530180e-02	1.243166e-01
PROSPECTIVE_COLLECTION_YES	1.686136	3.888498e-02	1.176520e-01
RETROSPECTIVE_COLLECTION_NO	1.686136	3.888498e-02	1.176520e-01
RETROSPECTIVE_COLLECTION_YES	0.602422	4.530180e-02	1.243166e-01
MENOPAUSE_STATUS_Post (prior bilateral ovariect...	1.324224	9.572180e-02	2.026486e-01
MENOPAUSE_STATUS_Pre (<6 months since LMP AND n...	0.431889	8.809889e-04	5.887468e-03
MENOPAUSE_STATUS_[Not Available]	2.943808	9.817809e-07	1.930836e-05
RACE_BLACK OR AFRICAN AMERICAN	1.215466	3.390936e-01	5.275821e-01
RACE_WHITE	0.825045	2.987240e-01	4.828689e-01
RACE_[Not Available]	1.274439	5.361448e-01	6.741716e-01
ETHNICITY_NOT HISPANIC OR LATINO	2.207632	1.176302e-02	5.552148e-02
ETHNICITY_[Not Available]	0.603636	1.259999e-01	2.447704e-01
HISTORY_OTHER_MALIGNANCY_No	0.669493	2.469654e-01	4.047489e-01
HISTORY_OTHER_MALIGNANCY_Yes	1.501028	2.411757e-01	4.016651e-01
RADIATION_TREATMENT_ADJUVANT_NO	0.825072	6.454644e-01	7.693415e-01
RADIATION_TREATMENT_ADJUVANT_YES	1.184551	6.240900e-01	7.525332e-01
RADIATION_TREATMENT_ADJUVANT_[Not Available]	0.991267	9.744419e-01	9.936324e-01
PHARMACEUTICAL_TX_ADJUVANT_YES	0.893189	7.194462e-01	8.162947e-01
PHARMACEUTICAL_TX_ADJUVANT_[Not Available]	1.001332	9.961223e-01	9.961223e-01
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Core needl...	0.545734	3.371119e-04	2.841372e-03
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Fine needl...	1.669895	1.358907e-02	5.938927e-02
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Other meth...	1.009070	9.765300e-01	9.936324e-01
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Tumor rese...	0.740284	3.397987e-01	5.275821e-01
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_[Not Avail...	1.944929	3.630026e-02	1.157684e-01
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_OTHER_[Not...	0.994946	9.868232e-01	9.952576e-01
SURGICAL_PROCEDURE_FIRST_Lumpectomy	0.690083	9.704508e-02	2.026486e-01
SURGICAL_PROCEDURE_FIRST_Modified Radical Maste...	1.476196	2.170723e-02	8.004541e-02
SURGICAL_PROCEDURE_FIRST_Other	0.882243	5.027363e-01	6.448140e-01
SURGICAL_PROCEDURE_FIRST_Simple Mastectomy	0.657380	9.788958e-02	2.026486e-01
SURGICAL_PROCEDURE_FIRST_[Not Available]	2.480055	2.729308e-03	1.463901e-02
FIRST_SURGICAL_PROCEDURE_OTHER_Surgical Resection	1.534372	3.107861e-01	4.955778e-01
FIRST_SURGICAL_PROCEDURE_OTHER_[Not Available]	1.084261	6.601793e-01	7.790115e-01

PATH_MARGIN_Negative	0.328745	2.487080e-11	1.467377e-09
PATH_MARGIN_Positive	1.456750	1.265338e-01	2.447704e-01
PATH_MARGIN_[Not Available]	4.295169	1.280242e-14	1.510685e-12
SURGERY_FOR_POSITIVE_MARGINS_[Not Available]	0.881873	7.014933e-01	8.141488e-01
MARGIN_STATUS_REEXCISION_Negative	1.189978	5.498957e-01	6.830283e-01
MARGIN_STATUS_REEXCISION_[Not Available]	0.788031	3.970790e-01	5.714063e-01
STAGING_SYSTEM_Axillary lymph node dissection a...	1.147628	4.478192e-01	6.047185e-01
STAGING_SYSTEM_Sentinel lymph node biopsy plus ...	0.471890	9.803200e-04	6.088303e-03
STAGING_SYSTEM_Sentinel node biopsy alone	0.517762	9.212199e-03	4.529331e-02
STAGING_SYSTEM_[Not Available]	2.306054	2.789049e-06	3.502342e-05
MICROMET_DETECTION_BY_IHC_NO	0.983191	9.205373e-01	9.860025e-01
MICROMET_DETECTION_BY_IHC_YES	0.330809	1.172534e-05	1.257809e-04
MICROMET_DETECTION_BY_IHC_[Not Available]	2.187550	2.968086e-06	3.502342e-05
LYMPH_NODES_EXAMINED_YES	1.127012	4.842736e-01	6.279592e-01
LYMPH_NODES_EXAMINED_[Not Available]	0.789955	1.767459e-01	3.160002e-01
AJCC_STAGING_EDITION_5th	2.522972	2.425021e-06	3.502342e-05
AJCC_STAGING_EDITION_6th	0.391254	3.279076e-07	9.673274e-06
AJCC_STAGING_EDITION_7th	1.553473	5.971021e-02	1.455199e-01
AJCC_STAGING_EDITION_[Not Available]	0.905993	7.048542e-01	8.141488e-01
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage I	0.623674	1.050290e-01	2.136796e-01
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IA	0.246046	1.650590e-02	6.956059e-02
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IIA	0.697084	6.472348e-02	1.510602e-01
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IIB	0.828285	3.501038e-01	5.365227e-01
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IIIA	1.293391	2.416799e-01	4.016651e-01
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IIIC	1.898479	6.528872e-02	1.510602e-01
ER_STATUS_BY_IHC_Negative	1.272320	1.906432e-01	3.308221e-01
ER_STATUS_BY_IHC_Positive	0.662619	1.727163e-02	7.027765e-02
ER_STATUS_IHC_PERCENT_POSITIVE_90-99%	0.370064	2.554685e-03	1.435490e-02
ER_STATUS_IHC_PERCENT_POSITIVE_<10%	0.771901	4.520339e-01	6.047185e-01
ER_STATUS_IHC_PERCENT_POSITIVE_[Not Available]	2.276830	1.936838e-05	1.758053e-04
ER_POSITIVITY_SCALE_USED_3 Point Scale	0.369684	5.031369e-02	1.349322e-01
ER_POSITIVITY_SCALE_USED_[Not Available]	2.435285	3.311614e-02	1.116487e-01
ER_POSITIVITY_SCALE_OTHER_[Not Available]	1.595260	1.827359e-01	3.218334e-01
BRACHYTHERAPY_TOTAL_DOSE_POINT_A_[Not Available]	0.994099	9.767911e-01	9.936324e-01
PR_STATUS_BY_IHC_Negative	1.257745	1.746637e-01	3.160002e-01
PR_STATUS_BY_IHC_Positive	0.717066	4.336782e-02	1.243166e-01
PR_STATUS_IHC_PERCENT_POSITIVE_90-99%	0.082244	1.281943e-02	5.818049e-02
PR_STATUS_IHC_PERCENT_POSITIVE_<10%	0.987074	9.609962e-01	9.936324e-01
PR_STATUS_IHC_PERCENT_POSITIVE_[Not Available]	2.440652	1.832734e-05	1.758053e-04
PR_POSITIVITY_SCALE_USED_3 Point Scale	0.401865	7.305251e-02	1.626452e-01
PR_POSITIVITY_SCALE_USED_[Not Available]	2.235079	5.428402e-02	1.411211e-01
PR_POSITIVITY_IHC_INTENSITY_SCORE_3+	0.127355	4.012324e-02	1.183636e-01
PR_POSITIVITY_IHC_INTENSITY_SCORE_[Not Available]	3.001282	2.553857e-03	1.435490e-02
PR_POSITIVITY_SCALE_OTHER_[Not Available]	1.327000	3.966578e-01	5.714063e-01
PR_POSITIVITY_DEFINE_METHOD_[Not Available]	1.008404	9.676018e-01	9.936324e-01
IHC_HER2_Equivocal	0.601537	5.547120e-02	1.411211e-01
IHC_HER2_Negative	0.791443	1.711526e-01	3.155626e-01
IHC_HER2_Positive	1.504160	7.472326e-02	1.632842e-01
IHC_HER2_[Not Available]	1.460324	3.585853e-02	1.157684e-01
HER2_IHC_PERCENT_POSITIVE_<10%	0.408510	2.110215e-02	8.004541e-02
HER2_IHC_PERCENT_POSITIVE_[Not Available]	1.864321	2.281919e-02	8.159588e-02
HER2_POSITIVITY_METHOD_TEXT_[Not Available]	2.593860	6.042777e-02	1.455199e-01
HER2_FISH_STATUS_Negative	0.739080	1.150908e-01	2.301816e-01
HER2_FISH_STATUS_Positive	0.880007	7.106553e-01	8.141488e-01
HER2_FISH_STATUS_[Not Available]	1.385428	6.801151e-02	1.543338e-01
HER2_COPY_NUMBER_[Not Available]	1.645350	1.487810e-01	2.831638e-01
CENT17_COPY_NUMBER_[Not Available]	1.548415	2.047798e-01	3.502032e-01
PRIMARY_SITE_Left	1.945749	7.591896e-04	5.599023e-03
PRIMARY_SITE_Left Upper Inner Quadrant	0.734898	4.268913e-01	6.047185e-01
PRIMARY_SITE_Left Upper Outer Quadrant	0.826630	4.397503e-01	6.047185e-01
PRIMARY_SITE_Right	0.892605	6.249852e-01	7.525332e-01
PRIMARY_SITE_Right Upper Outer Quadrant	1.138461	6.005181e-01	7.381368e-01
HISTOLOGICAL_DIAGNOSIS_Infiltrating Ductal Carc...	0.969817	8.637923e-01	9.351146e-01
HISTOLOGICAL_DIAGNOSIS_Infiltrating Lobular Car...	0.823077	3.912033e-01	5.714063e-01
ICD_O_3_HISTOLOGY_8500/3	0.949261	7.692879e-01	8.563771e-01
ICD_O_3_HISTOLOGY_8520/3	0.871304	5.370519e-01	6.741716e-01
METASTATIC_TUMOR_INDICATOR_NO	0.641384	2.830420e-02	9.823221e-02
METASTATIC_TUMOR_INDICATOR_[Not Available]	1.149966	4.502190e-01	6.047185e-01
TISSUE_SOURCE_SITE_A2	0.816065	4.561012e-01	6.047185e-01
TISSUE_SOURCE_SITE_A8	1.386201	4.396658e-01	6.047185e-01

TISSUE_SOURCE_SITE_AR	0.426279	7.036257e-03	3.609906e-02
TISSUE_SOURCE_SITE_B6	1.053316	8.202897e-01	9.046185e-01
TISSUE_SOURCE_SITE_BH	2.772661	1.247824e-08	4.908109e-07
TISSUE_SOURCE_SITE_D8	0.947428	9.275108e-01	9.860025e-01
TISSUE_SOURCE_SITE_E2	0.555754	1.612908e-01	3.021003e-01
TISSUE_SOURCE_SITE_E9	1.350339	4.798622e-01	6.279592e-01
AJCC_TUMOR_PATHOLOGIC_PT_simple_T1	0.686002	5.620924e-02	1.411211e-01
AJCC_TUMOR_PATHOLOGIC_PT_simple_T2	0.951778	7.620026e-01	8.563458e-01
AJCC_TUMOR_PATHOLOGIC_PT_simple_T3	1.220149	3.652398e-01	5.525423e-01
AJCC_NODES_PATHOLOGIC_PN_simple_N0	0.428539	2.290188e-06	3.502342e-05
AJCC_NODES_PATHOLOGIC_PN_simple_N1	1.159871	3.741624e-01	5.588754e-01
AJCC_NODES_PATHOLOGIC_PN_simple_N2	1.714021	1.986213e-02	7.812439e-02
AJCC_NODES_PATHOLOGIC_PN_simple_N3	2.564601	6.195610e-04	4.873880e-03
AJCC_METASTASIS_PATHOLOGIC_PM_simple_M0	0.509375	8.980883e-04	5.887468e-03
AJCC_METASTASIS_PATHOLOGIC_PM_simple_MX	1.062431	8.307274e-01	9.076466e-01

Table A.3 – Clinical variables significance for single-variable cox model for BRCA (1080 patients). Variable significantly associated with survival are shaded.

Variable	Hazard ratio	P-value	Corrected P-value
INITIAL_PATHOLOGIC_DX_YEAR	0.944473	0.000002	0.000085
age_at_diagnosis	1.019742	0.001594	0.012355
PRIMARY_SITE_Floor of mouth	1.633189	0.009866	0.048599
PRIMARY_SITE_Larynx	0.870495	0.392949	0.589424
PRIMARY_SITE_Oral Cavity	1.216594	0.257959	0.436185
PRIMARY_SITE_Oral Tongue	1.027363	0.864976	0.935381
LATERALITY_Left	0.997223	0.985683	0.985683
LATERALITY_Right	0.596169	0.001229	0.010391
LATERALITY_[Not Available]	1.487693	0.003515	0.021791
PROSPECTIVE_COLLECTION_NO	1.275968	0.136430	0.309463
PROSPECTIVE_COLLECTION_YES	0.796014	0.163125	0.337124
RETROSPECTIVE_COLLECTION_NO	0.796014	0.163125	0.337124
RETROSPECTIVE_COLLECTION_YES	1.275968	0.136430	0.309463
SEX_Female	1.382324	0.024394	0.087255
SEX_Male	0.723420	0.024394	0.087255
RACE_WHITE	0.742605	0.111804	0.273433
ETHNICITY_NOT HISPANIC OR LATINO	0.853619	0.414453	0.611812
LYMPH_NODE_NECK_DISSECTION_INDICATOR_NO	1.236683	0.203896	0.403454
LYMPH_NODE_NECK_DISSECTION_INDICATOR_YES	0.814286	0.215391	0.412907
LYMPH_NODE_DISSECTION_METHOD_Functional (Limite...	1.086460	0.540168	0.749786
LYMPH_NODE_DISSECTION_METHOD_Modified Radical N...	0.741422	0.093637	0.241895
LYMPH_NODE_DISSECTION_METHOD_[Not Available]	1.090921	0.583249	0.763974
LYMPH_NODES_EXAMINED_NO	1.470161	0.061855	0.179766
LYMPH_NODES_EXAMINED_YES	0.837805	0.290092	0.481761
PATH_MARGIN_Negative	0.777605	0.079280	0.216854
PATH_MARGIN_Positive	1.727697	0.002205	0.014645
PATH_MARGIN_[Not Available]	0.699698	0.158253	0.337124
AJCC_STAGING_EDITION_6th	1.010909	0.944167	0.955798
AJCC_STAGING_EDITION_7th	0.973421	0.855035	0.935381
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage II	0.668524	0.054444	0.168777
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage III	0.726051	0.114665	0.273433
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IVA	1.532143	0.001807	0.012924
AJCC_PATHOLOGIC_TUMOR_STAGE_[Not Available]	0.962185	0.849093	0.935381
EXTRACAPSULAR_SPREAD_PATHOLOGIC_Microscopic Ext...	2.254494	0.000001	0.000085
EXTRACAPSULAR_SPREAD_PATHOLOGIC_No Extranodal E...	0.548915	0.000017	0.000392
EXTRACAPSULAR_SPREAD_PATHOLOGIC_[Not Available]	0.990168	0.945521	0.955798
GRADE_G1	0.657592	0.059833	0.179498
GRADE_G2	1.431246	0.010999	0.050448
GRADE_G3	0.957962	0.783689	0.915850
ANGIOLYMPHATIC_INVASION_NO	0.693371	0.009929	0.048599
ANGIOLYMPHATIC_INVASION_YES	1.484568	0.008174	0.047509
ANGIOLYMPHATIC_INVASION_[Not Available]	1.057948	0.692475	0.847371
PERINEURAL_INVASION_NO	0.516015	0.000030	0.000565
PERINEURAL_INVASION_YES	1.588730	0.000670	0.006230
PERINEURAL_INVASION_[Not Available]	1.147166	0.339620	0.535334
HPV_STATUS_P16_Negative	0.856744	0.477818	0.673289
HPV_STATUS_P16_[Not Available]	1.469555	0.052450	0.168202
HPV_STATUS_ISH_Negative	0.842503	0.448772	0.642089
HPV_STATUS_ISH_[Not Available]	1.564703	0.043660	0.145013
TOBACCO_SMOKING_HISTORY_INDICATOR_1	0.853226	0.347479	0.538592
TOBACCO_SMOKING_HISTORY_INDICATOR_2	1.349280	0.033815	0.116475
TOBACCO_SMOKING_HISTORY_INDICATOR_3	0.708845	0.088463	0.235059
TOBACCO_SMOKING_HISTORY_INDICATOR_4	0.949831	0.733565	0.874635
ALCOHOL_HISTORY_DOCUMENTED_NO	1.064729	0.660954	0.841584
ALCOHOL_HISTORY_DOCUMENTED_YES	0.941649	0.669647	0.841584
RADIATION_TREATMENT_ADJUVANT_NO	1.062557	0.787828	0.915850
RADIATION_TREATMENT_ADJUVANT_YES	0.905136	0.549419	0.751412
RADIATION_TREATMENT_ADJUVANT_[Not Available]	1.055628	0.712535	0.860594
PHARMACEUTICAL_TX_ADJUVANT_NO	0.774111	0.144766	0.320553
PHARMACEUTICAL_TX_ADJUVANT_YES	1.454739	0.067970	0.191551
PHARMACEUTICAL_TX_ADJUVANT_[Not Available]	1.032903	0.825634	0.935381
CLIN_N_STAGE_N0	0.853501	0.240473	0.436185
CLIN_N_STAGE_N1	0.915126	0.621348	0.802575
CLIN_N_STAGE_N2b	0.980360	0.916449	0.955798

CLIN_T_STAGE_T2	0.838379	0.251047	0.436185
CLIN_T_STAGE_T3	1.269859	0.104375	0.262348
CLIN_T_STAGE_T4a	1.014869	0.918479	0.955798
CLINICAL_STAGE_Stage II	0.872315	0.435405	0.632698
CLINICAL_STAGE_Stage III	0.978740	0.898501	0.955798
CLINICAL_STAGE_Stage IVA	1.138939	0.338664	0.535334
ICD_10_C02.9	0.970829	0.852993	0.935381
ICD_10_C04.9	2.068995	0.000137	0.001823
ICD_10_C14.8	1.218218	0.254981	0.436185
ICD_10_C32.9	0.914839	0.578172	0.763974
ICD_O_3_HISTOLOGY_8070/3	1.344115	0.221993	0.412907
ICD_O_3_HISTOLOGY_8071/3	0.898413	0.682330	0.846089
ICD_O_3_SITE_C02.9	0.970829	0.852993	0.935381
ICD_O_3_SITE_C04.9	2.068995	0.000137	0.001823
ICD_O_3_SITE_C14.8	1.218218	0.254981	0.436185
ICD_O_3_SITE_C32.9	0.914839	0.578172	0.763974
TISSUE_SOURCE_SITE_CN	1.213414	0.297016	0.484604
TISSUE_SOURCE_SITE_CR	0.431810	0.009643	0.048599
TISSUE_SOURCE_SITE_CV	1.663811	0.000385	0.003978
AJCC_TUMOR_PATHOLOGIC_PT_simple_T2	0.659617	0.011391	0.050448
AJCC_TUMOR_PATHOLOGIC_PT_simple_T3	1.459494	0.016649	0.070379
AJCC_TUMOR_PATHOLOGIC_PT_simple_T4	1.388501	0.017779	0.071889
AJCC_NODES_PATHOLOGIC_PN_simple_N0	0.556349	0.000185	0.002146
AJCC_NODES_PATHOLOGIC_PN_simple_N1	0.576861	0.022426	0.086899
AJCC_NODES_PATHOLOGIC_PN_simple_N2	1.834907	0.000009	0.000292
AJCC_NODES_PATHOLOGIC_PN_simple_NX	1.165686	0.381317	0.581352
AJCC_METASTASIS_PATHOLOGIC_PM_simple_M0	0.818485	0.176511	0.356858
AJCC_METASTASIS_PATHOLOGIC_PM_simple_MX	0.979428	0.928250	0.955798
AJCC_METASTASIS_PATHOLOGIC_PM_simple_[N	1.185902	0.220529	0.412907

Table A.4 – Clinical variables significance for single-variable cox model for HNSC (526 patients). Variable significantly associated with survival are shaded.

CellType	Symbol	AffymetrixID	EntrezGene	Gene Ensembl ID
B cells	MS4A1	217418_x_at	931	ENSG00000156738
B cells	TCL1A	209995_s_at	8115	ENSG00000100721
B cells	MS4A1	210356_x_at	931	ENSG00000156738
B cells	TCL1A	39318_at	8115	ENSG00000100721
B cells	HLA-DOB	205671_s_at	3112	ENSG00000241106
B cells	HLA-DOB	205671_s_at	3112	ENSG00000239457
B cells	HLA-DOB	205671_s_at	3112	ENSG00000243496
B cells	HLA-DOB	205671_s_at	3112	ENSG00000241386
B cells	HLA-DOB	205671_s_at	3112	ENSG00000241910
B cells	HLA-DOB	205671_s_at	3112	ENSG00000243612
B cells	PNOC	205901_at	5368	ENSG00000168081
B cells	KIAA0125	206478_at	9834	ENSG00000277059
B cells	KIAA0125	206478_at	9834	ENSG00000226777
B cells	CD19	206398_s_at	930	ENSG00000177455
B cells	CR2	205544_s_at	1380	ENSG00000117322
B cells	IGHG1	213674_x_at	3500	NaN
B cells	FCRL2	221239_s_at	79368	ENSG00000132704
B cells	BLK	206255_at	640	ENSG00000136573
B cells	IGHG1	222285_at	3500	NaN
B cells	COCH	205229_s_at	1690	ENSG00000100473
B cells	OSBPL10	219073_s_at	114884	ENSG00000144645
B cells	IGHA1	215118_s_at	3493	NaN
B cells	TNFRSF17	206641_at	608	ENSG00000048462
B cells	ABCB4	207819_s_at	5244	ENSG00000005471
B cells	BLNK	207655_s_at	29760	ENSG00000095585
B cells	GLDC	204836_at	2731	ENSG00000178445
B cells	MEF2C	209200_at	4208	ENSG00000081189
B cells	MEF2C	209199_s_at	4208	ENSG00000081189
B cells	IGHM	209374_s_at	3507	NaN
B cells	FAM30A	220377_at	29064	NaN
B cells	SPIB	205861_at	6689	ENSG00000269404
B cells	BCL11A	219497_s_at	53335	ENSG00000119866
B cells	GNG7	206896_s_at	2788	ENSG00000176533
B cells	IGKC	215217_at	3514	NaN
B cells	CD72	215925_s_at	971	ENSG00000137101
B cells	MICAL3	212715_s_at	57553	ENSG00000243156
B cells	BCL11A	210347_s_at	53335	ENSG00000119866
B cells	BACH2	221234_s_at	60468	ENSG00000112182
B cells	IGL@	217138_x_at	3535	NaN
B cells	CCR9	207445_s_at	10803	ENSG00000173585
B cells	QRSL1	218948_at	55278	ENSG00000130348
B cells	DTNB	215565_at	1838	ENSG00000138101
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000236418
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000232062
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000223793
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000231526
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000257473
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000228284
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000225890
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000231823
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000206301
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000206305
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000225103
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000233192
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000196735
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000237541
B cells	SCN3A	210432_s_at	6328	ENSG00000153253
B cells	QRSL1	218949_s_at	55278	ENSG00000130348
B cells	SLC15A2	205316_at	6565	ENSG00000163406
T cells	PRKCQ	210038_at	5588	ENSG00000065675
T cells	CD3D	213539_at	915	ENSG00000167286
T cells	CD3G	206804_at	917	ENSG00000160654
T cells	CD28	206545_at	940	ENSG00000178562
T cells	LCK	204891_s_at	3932	ENSG00000182866
T cells	TRAT1	217147_s_at	50852	ENSG00000163519
T cells	PRKCQ	210039_s_at	5588	ENSG00000065675
T cells	BCL11B	219528_s_at	64919	ENSG00000127152
T cells	CD2	205831_at	914	ENSG00000116824
T cells	LCK	204890_s_at	3932	ENSG00000182866

T cells	TRBC1	213193_x_at	28639	NaN
T cells	TRBC1	210915_x_at	28639	NaN
T cells	TRA@	209670_at	28755	NaN
T cells	ITM2A	202747_s_at	9452	ENSG00000078596
T cells	SH2D1A	210116_at	4068	ENSG00000183918
T cells	CD6	213958_at	923	ENSG00000013725
T cells	CD96	206761_at	10225	ENSG00000153283
T cells	NCALD	211685_s_at	83988	ENSG00000104490
T cells	GIMAP5	218805_at	55340	ENSG00000196329
T cells	TRA@	209671_x_at	6955	NaN
T cells	CD3E	205456_at	916	ENSG00000198851
T cells	SKAP1	205790_at	8631	ENSG00000141293
T cells	TRA@	213830_at	6955	NaN
T cells	TRA@	216191_s_at	6955	NaN
T helper cells	ICOS	210439_at	29851	ENSG00000163600
T helper cells	LRBA	214109_at	987	ENSG00000198589
T helper cells	ITM2A	202746_at	9452	ENSG00000078596
T helper cells	FAM111A	218248_at	63901	ENSG00000166801
T helper cells	PHF10	219126_at	55274	ENSG00000130024
T helper cells	NUP107	218768_at	57122	ENSG00000111581
T helper cells	SEC24C	202361_at	9632	ENSG00000176986
T helper cells	NAP1L4	201414_s_at	4676	ENSG00000205531
T helper cells	NAP1L4	201414_s_at	4676	ENSG00000273562
T helper cells	BATF	205965_at	10538	ENSG00000156127
T helper cells	ASF1A	203428_s_at	25842	ENSG00000111875
T helper cells	FRYL	212546_s_at	285527	ENSG00000075539
T helper cells	FUSIP1	213594_x_at	10772	ENSG00000188529
T helper cells	TRA@	215524_x_at	10730	ENSG00000136758
T helper cells	TRA@	217412_at	6955	NaN
T helper cells	RPA1	201528_at	6117	ENSG00000132383
T helper cells	UBE2L3	200683_s_at	7332	ENSG00000185651
T helper cells	ANP32B	201306_s_at	10541	ENSG00000136938
T helper cells	DDX50	221699_s_at	79009	ENSG00000107625
T helper cells	C13orf34	219544_at	79866	ENSG00000136122
T helper cells	PPP2R5C	213305_s_at	5527	ENSG00000078304
T helper cells	SLC25A12	203340_s_at	8604	ENSG00000115840
T helper cells	ATF2	205446_s_at	1386	ENSG00000115966
T helper cells	CD28	211856_x_at	940	ENSG00000178562
T helper cells	GOLGA8A	208798_x_at	23015	ENSG00000175265
Tcm	CDC14A	210441_at	8556	ENSG00000079335
Tcm	ATM	208442_s_at	472	ENSG00000149311
Tcm	USP9Y	206624_at	8287	ENSG00000114374
Tcm	PCNX	215175_at	22990	ENSG00000100731
Tcm	ATM	210858_x_at	472	ENSG00000149311
Tcm	FOXP1	215221_at	27086	ENSG00000114861
Tcm	KLF12	206965_at	11278	ENSG00000118922
Tcm	ST3GAL1	215874_at	6482	ENSG00000008513
Tcm	INPP4B	215864_at	8821	ENSG00000109452
Tcm	CASP8	207686_s_at	841	ENSG00000064012
Tcm	MLL	216624_s_at	4297	ENSG00000118058
Tcm	PCM1	209997_x_at	5108	ENSG00000078674
Tcm	RP11-74E24.2	205787_x_at	441155	ENSG00000215817
Tcm	PHC3	215521_at	80012	ENSG00000173889
Tcm	NFATC3	210556_at	4775	ENSG00000072736
Tcm	LOC202134	215133_s_at	202134	ENSG00000182230
Tcm	TIMM8A	210800_at	1678	ENSG00000126953
Tcm	ATF7IP	216197_at	55729	ENSG00000171681
Tcm	REPS1	215201_at	85021	ENSG00000135597
Tcm	PSPC1	215083_at	55269	ENSG00000121390
Tcm	RPP38	215743_at	9397	ENSG00000152465
Tcm	HNRPH1	213472_at	3187	ENSG00000169045
Tcm	STX16	221638_s_at	8675	ENSG00000124222
Tcm	CYLD	214272_at	1540	ENSG00000083799
Tcm	SNRPN	216850_at	6638	ENSG00000128739
Tcm	TRAF3IP3	215275_at	80342	ENSG00000009790
Tcm	NEFL	221805_at	4747	ENSG00000277586
Tcm	POLR2J2	217610_at	246721	ENSG00000228049
Tcm	AQP3	203747_at	360	ENSG00000165272
Tcm	CG030	215105_at	116828	ENSG00000281026
Tcm	PDXDC2	215920_s_at	283970	ENSG00000196696
Tcm	CLUAP1	204576_s_at	23059	ENSG00000103351
Tcm	DOCK9	215041_s_at	23348	ENSG00000088387

Tcm	CYorf15B	214131_at	84663	NaN
Tcm	CREBZF	213584_s_at	58487	ENSG00000137504
Tcm	CEP68	207971_s_at	23177	ENSG00000011523
Tcm	TXK	206828_at	7294	ENSG00000074966
Tcm	SLC7A6	203578_s_at	9057	ENSG00000103064
Tcm	FYB	205285_s_at	2533	ENSG00000082074
Tcm	MAP3K1	214786_at	4214	ENSG00000095015
Tem	TRA@	217397_at	6955	NaN
Tem	PRKY	206279_at	5616	ENSG00000099725
Tem	VIL2	217230_at	7430	ENSG00000092820
Tem	GDPD5	32502_at	81544	ENSG00000158555
Tem	CCR2	206978_at	1231	NaN
Tem	MEFV	208262_x_at	4210	ENSG00000103313
Tem	C7orf54	210109_at	27099	ENSG00000279078
Tem	FLI1	210786_s_at	2313	ENSG00000151702
Tem	TBC1D5	201815_s_at	9779	ENSG00000131374
Tem	DDX17	208719_s_at	10521	ENSG00000100201
Tem	AKT3	212609_s_at	10000	ENSG00000275199
Tem	AKT3	212609_s_at	10000	ENSG00000117020
Tem	EWSR1	211825_s_at	2130	ENSG00000182944
Tem	TBCD	201759_at	6904	ENSG00000278759
Tem	TBCD	201759_at	6904	ENSG00000141556
Tem	CCR2	207794_at	1231	NaN
Tem	NFATC4	205897_at	4776	ENSG00000100968
Tem	LTK	207106_s_at	4058	ENSG00000062524
Th1 cells	IFNG	210354_at	3458	ENSG00000111537
Th1 cells	LTA	206975_at	4049	ENSG00000238130
Th1 cells	LTA	206975_at	4049	ENSG00000173503
Th1 cells	LTA	206975_at	4049	ENSG00000231408
Th1 cells	LTA	206975_at	4049	ENSG00000226979
Th1 cells	LTA	206975_at	4049	ENSG00000230279
Th1 cells	LTA	206975_at	4049	ENSG00000226275
Th1 cells	LTA	206975_at	4049	ENSG00000223919
Th1 cells	APBB2	213419_at	323	ENSG00000163697
Th1 cells	DOK5	214844_s_at	55816	ENSG00000101134
Th1 cells	IL12RB2	206999_at	3595	ENSG00000081985
Th1 cells	APBB2	40148_at	323	ENSG00000163697
Th1 cells	APOD	201525_at	347	ENSG00000189058
Th1 cells	ZBTB32	220118_at	27033	ENSG00000011590
Th1 cells	CD38	205692_s_at	952	ENSG00000004468
Th1 cells	CSF2	210229_s_at	1437	ENSG00000164400
Th1 cells	CTLA4	221331_x_at	1493	ENSG00000163599
Th1 cells	CD70	206508_at	970	ENSG00000125726
Th1 cells	DPP4	211478_s_at	1803	ENSG00000197635
Th1 cells	EGFL6	219454_at	25975	ENSG00000198759
Th1 cells	BST2	201641_at	684	ENSG00000130303
Th1 cells	DUSP5	209457_at	1847	ENSG00000138166
Th1 cells	LRP8	205282_at	7804	ENSG00000157193
Th1 cells	IL22	221165_s_at	50616	ENSG00000127318
Th1 cells	DGKI	206806_at	9162	ENSG00000157680
Th1 cells	CCL4	204103_at	6351	ENSG00000275824
Th1 cells	CCL4	204103_at	6351	ENSG00000275302
Th1 cells	CCL4	204103_at	6351	ENSG00000277943
Th1 cells	DPP4	203716_s_at	1803	ENSG00000197635
Th1 cells	GGT1	211417_x_at	2678	ENSG00000100031
Th1 cells	LRRN3	209840_s_at	54674	ENSG00000173114
Th1 cells	SYNGR3	205691_at	9143	ENSG00000127561
Th1 cells	ATP9A	212062_at	10079	ENSG00000054793
Th1 cells	BTG3	205548_s_at	10950	ENSG00000281484
Th1 cells	BTG3	205548_s_at	10950	ENSG00000154640
Th1 cells	CMAH	210571_s_at	8418	NaN
Th1 cells	HBEGF	38037_at	1839	ENSG00000113070
Th1 cells	SGCB	205120_s_at	6443	ENSG00000163069
Th2 cells	PMCH	206942_s_at	5367	ENSG00000183395
Th2 cells	AHI1	220841_s_at	54806	ENSG00000135541
Th2 cells	PTGIS	208131_s_at	5740	ENSG00000124212
Th2 cells	AHI1	220842_at	54806	ENSG00000135541
Th2 cells	CXCR6	211469_s_at	10663	ENSG00000172215
Th2 cells	EVI5	209717_at	7813	ENSG00000067208
Th2 cells	AHI1	221569_at	54806	ENSG00000135541
Th2 cells	IL26	221111_at	55801	ENSG00000111536
Th2 cells	MB	204179_at	4151	ENSG00000198125

Th2 cells	NEIL3	219502_at	55247	ENSG00000109674
Th2 cells	GSTA4	202967_at	2941	ENSG00000170899
Th2 cells	PHEX	210617_at	5251	ENSG00000102174
Th2 cells	SMAD2	203076_s_at	4087	ENSG00000175387
Th2 cells	CENPF	209172_s_at	1063	ENSG00000117724
Th2 cells	ANK1	208353_x_at	286	ENSG00000029534
Th2 cells	ADCY1	213245_at	107	ENSG00000164742
Th2 cells	AI582773	214373_at	728210	NaN
Th2 cells	LAIR2	207509_s_at	3904	ENSG00000277335
Th2 cells	LAIR2	207509_s_at	3904	ENSG00000274084
Th2 cells	LAIR2	207509_s_at	3904	ENSG00000167618
Th2 cells	LAIR2	207509_s_at	3904	ENSG00000275819
Th2 cells	SNRPD1	202691_at	6632	ENSG00000167088
Th2 cells	CXCR6	206974_at	10663	ENSG00000172215
Th2 cells	MICAL2	212472_at	9645	ENSG00000133816
Th2 cells	DHFR	202534_x_at	1719	ENSG00000228716
Th2 cells	SMAD2	203077_s_at	4087	ENSG00000175387
Th2 cells	WDHD1	204728_s_at	11169	ENSG00000198554
Th2 cells	BIRC5	210334_x_at	332	ENSG00000089685
Th2 cells	DHFR	48808_at	1719	ENSG00000228716
Th2 cells	SLC39A14	212110_at	23516	ENSG00000104635
Th2 cells	HELLS	220085_at	3070	ENSG00000119969
Th2 cells	LIMA1	217892_s_at	51474	ENSG00000050405
Th2 cells	CDC25C	205167_s_at	995	ENSG00000158402
Th2 cells	CDC7	204510_at	8317	ENSG00000097046
Th2 cells	GATA3	209602_s_at	2625	ENSG00000107485
TFH	CHI3L2	213060_s_at	1117	ENSG00000064886
TFH	CXCL13	205242_at	10563	ENSG00000156234
TFH	MYO7A	33197_at	4647	ENSG00000137474
TFH	CHGB	204260_at	1114	ENSG00000089199
TFH	MYO7A	208189_s_at	4647	ENSG00000137474
TFH	ICA1	210547_x_at	3382	ENSG00000003147
TFH	HEY1	218839_at	23462	ENSG00000164683
TFH	CDK5R1	204995_at	8851	ENSG00000176749
TFH	ST8SIA1	210073_at	6489	ENSG00000111728
TFH	PDCD1	207634_at	5133	ENSG00000276977
TFH	PDCD1	207634_at	5133	ENSG00000188389
TFH	BLR1	216734_s_at	643	ENSG00000160683
TFH	KIAA1324	221874_at	57535	ENSG00000116299
TFH	PVALB	205336_at	5816	ENSG00000100362
TFH	PVALB	205336_at	5816	ENSG00000274665
TFH	ICA1	207949_s_at	3382	ENSG00000003147
TFH	TSHR	210055_at	7253	ENSG00000165409
TFH	C18orf1	209574_s_at	753	ENSG00000168675
TFH	HEY1	44783_s_at	23462	ENSG00000164683
TFH	TOX	204529_s_at	9760	ENSG00000198846
TFH	BLR1	206126_at	643	ENSG00000160683
TFH	SLC7A10	220868_s_at	56301	ENSG00000130876
TFH	SMAD1	210993_s_at	4086	ENSG00000170365
TFH	POMT1	218476_at	10585	ENSG00000130714
TFH	PASK	216945_x_at	23178	ENSG00000115687
TFH	MKL2	218259_at	57496	ENSG00000186260
TFH	PTPN13	204201_s_at	5783	ENSG00000163629
TFH	PASK	213534_s_at	23178	ENSG00000115687
TFH	KCNK5	219615_s_at	8645	ENSG00000164626
TFH	C18orf1	207996_s_at	753	ENSG00000168675
TFH	ZNF764	57516_at	92595	ENSG00000169951
TFH	MAF	206363_at	4094	ENSG00000178573
TFH	MYO6	210480_s_at	4646	ENSG00000196586
TFH	SIRPG	220485_s_at	55423	ENSG00000089012
TFH	THADA	54632_at	63892	ENSG00000115970
TFH	THADA	220212_s_at	63892	ENSG00000115970
TFH	MAGEH1	218573_at	28986	ENSG00000187601
TFH	B3GAT1	219521_at	27087	ENSG00000109956
TFH	MAF	209348_s_at	4094	ENSG00000178573
TFH	SH3TC1	219256_s_at	54436	ENSG00000125089
TFH	HIST1H4K	214463_x_at	8362	ENSG00000273542
TFH	STK39	202786_at	27347	ENSG00000198648
Th17 cells	IL17A	208402_at	3605	ENSG00000112115
Th17 cells	IL17A	216876_s_at	3605	ENSG00000112115
Th17 cells	IL17RA	205707_at	23765	ENSG00000177663
Th17 cells	RORC	206419_at	6097	ENSG00000143365

TReg	FOXP3	221333_at	50943	ENSG00000049768
TReg	FOXP3	221334_s_at	50943	ENSG00000049768
CD8 T cells	CD8B	207979_s_at	926	ENSG00000172116
CD8 T cells	CD8A	205758_at	925	ENSG00000153563
CD8 T cells	CD8B	215332_s_at	926	ENSG00000172116
CD8 T cells	PF4	206390_x_at	5196	ENSG00000163737
CD8 T cells	PRR5	47069_at	55615	ENSG00000186654
CD8 T cells	SF1	210172_at	7536	ENSG00000168066
CD8 T cells	LIME1	219541_at	54923	ENSG00000203896
CD8 T cells	DNAJB1	200664_s_at	3337	ENSG00000132002
CD8 T cells	ARHGAP8	219168_s_at	55615	ENSG00000186654
CD8 T cells	GZMM	207460_at	3004	ENSG00000197540
CD8 T cells	SLC16A7	207057_at	9194	ENSG00000118596
CD8 T cells	SFRS7	213649_at	6432	ENSG00000115875
CD8 T cells	APBA2	209871_s_at	321	ENSG00000034053
CD8 T cells	APBA2	209871_s_at	321	ENSG00000276495
CD8 T cells	C4orf15	210054_at	79441	ENSG00000214367
CD8 T cells	LEPROTL1	202595_s_at	23484	ENSG00000104660
CD8 T cells	ZFP36L2	201367_s_at	678	ENSG00000152518
CD8 T cells	GADD45A	203725_at	1647	ENSG00000116717
CD8 T cells	ZFP36L2	201369_s_at	678	ENSG00000152518
CD8 T cells	MYST3	216361_s_at	7994	ENSG00000083168
CD8 T cells	ZEB1	208078_s_at	6935	ENSG00000148516
CD8 T cells	ZNF609	212620_at	23060	ENSG00000180357
CD8 T cells	C12orf47	64432_at	51275	ENSG00000234608
CD8 T cells	THUMPD1	206555_s_at	55623	ENSG00000066654
CD8 T cells	VAMP2	201557_at	6844	ENSG00000220205
CD8 T cells	ZNF91	206059_at	7644	ENSG00000167232
CD8 T cells	ZNF22	218006_s_at	7570	ENSG00000165512
CD8 T cells	TMC6	214958_s_at	11322	ENSG00000141524
CD8 T cells	DNAJB1	200666_s_at	3337	ENSG00000132002
CD8 T cells	FLT3LG	210607_at	2323	ENSG00000090554
CD8 T cells	CDKN2AIP	218929_at	55602	ENSG00000168564
CD8 T cells	TSC22D3	207001_x_at	1831	ENSG00000157514
CD8 T cells	TBCC	202495_at	6903	ENSG00000124659
CD8 T cells	RBM3	208319_s_at	5935	ENSG00000102317
CD8 T cells	ABT1	218405_at	29777	ENSG00000146109
CD8 T cells	C19orf6	212574_x_at	91304	ENSG00000182087
CD8 T cells	CAMLG	203538_at	819	ENSG00000164615
CD8 T cells	PPP1R2	202165_at	5504	ENSG00000184203
CD8 T cells	AES	217729_s_at	166	ENSG00000104964
CD8 T cells	KLF9	203543_s_at	687	ENSG00000119138
CD8 T cells	PRF1	214617_at	5551	ENSG00000180644
Tgd	TRD@	217143_s_at	6964	NaN
Tgd	TARP	211144_x_at	445347	ENSG00000211689
Tgd	C1orf61	205103_at	10485	ENSG00000125462
Tgd	TRGV9	209813_x_at	6983	NaN
Tgd	CD160	207840_at	11126	ENSG00000117281
Tgd	TARP	216920_s_at	445347	ENSG00000211689
Tgd	FEZ1	203562_at	9638	ENSG00000149557
Cytotoxic cells	KLRD1	210606_x_at	3824	ENSG00000134539
Cytotoxic cells	KLRF1	220646_s_at	51348	ENSG00000150045
Cytotoxic cells	GNLY	37145_at	10578	ENSG00000115523
Cytotoxic cells	GNLY	205495_s_at	10578	ENSG00000115523
Cytotoxic cells	CTSW	214450_at	1521	ENSG00000172543
Cytotoxic cells	KLRB1	214470_at	3820	ENSG00000111796
Cytotoxic cells	KLRD1	207795_s_at	3824	ENSG00000134539
Cytotoxic cells	KLRK1	205821_at	22914	ENSG00000213809
Cytotoxic cells	NKG7	213915_at	4818	ENSG00000105374
Cytotoxic cells	GZMH	210321_at	2999	ENSG00000100450
Cytotoxic cells	KLRD1	207796_x_at	3824	ENSG00000134539
Cytotoxic cells	SIGIRR	218921_at	59307	ENSG00000185187
Cytotoxic cells	ZBTB16	205883_at	7704	ENSG00000109906
Cytotoxic cells	RUNX3	204198_s_at	864	ENSG00000020633
Cytotoxic cells	APOL3	221087_s_at	80833	ENSG00000128284
Cytotoxic cells	RORA	210426_x_at	6095	ENSG00000069667
Cytotoxic cells	APBA2	209870_s_at	321	ENSG00000034053
Cytotoxic cells	APBA2	209870_s_at	321	ENSG00000276495
Cytotoxic cells	SIGIRR	52940_at	59307	ENSG00000185187
Cytotoxic cells	WHDC1L1	213908_at	339005	NaN
Cytotoxic cells	DUSP2	204794_at	1844	ENSG00000158050
Cytotoxic cells	GZMA	205488_at	3001	ENSG00000145649

NK cells	LOC643313	211050_x_at	643313	NaN
NK cells	GAGE2	207739_s_at	2574	ENSG00000236362
NK cells	ZNF747	206180_x_at	65988	ENSG00000169955
NK cells	XCL1	206366_x_at	6375	ENSG00000143184
NK cells	XCL2	214567_s_at	6846	ENSG00000143185
NK cells	AF107846	217058_at	2778	ENSG00000087460
NK cells	SLC30A5	220181_x_at	64924	ENSG00000145740
NK cells	NM_014114	220691_at	259230	ENSG00000198964
NK cells	MCM3AP	215582_x_at	8888	ENSG00000160294
NK cells	TBXA2R	207555_s_at	6915	ENSG00000006638
NK cells	CDC5L	209057_x_at	988	ENSG00000096401
NK cells	LOC730096	215182_x_at	730096	NaN
NK cells	FUT5	211225_at	2527	ENSG00000130383
NK cells	FGF18	206986_at	8817	ENSG00000156427
NK cells	MRC2	209280_at	9902	ENSG00000011028
NK cells	RP5-886K2.1	208014_x_at	27308	NaN
NK cells	SPN	216981_x_at	6693	ENSG00000197471
NK cells	PSMD4	210459_at	5710	ENSG00000159352
NK cells	PRX	220024_s_at	57716	ENSG00000105227
NK cells	FZR1	209415_at	51343	ENSG00000105325
NK cells	ZNF205	206416_at	7755	ENSG00000122386
NK cells	AL080130	212972_x_at	323	ENSG00000163697
NK cells	ZNF528	215019_x_at	84436	ENSG00000167555
NK cells	MAPRE3	203842_s_at	22924	ENSG00000084764
NK cells	BCL2	207004_at	596	ENSG00000171791
NK cells	NM_017616	221068_at	25959	ENSG00000197256
NK cells	ARL6IP2	217580_x_at	64225	ENSG00000119787
NK cells	SPN	206057_x_at	6693	ENSG00000197471
NK cells	FZR1	211865_s_at	51343	ENSG00000105325
NK cells	PDLIM4	214174_s_at	8572	ENSG00000131435
NK cells	NM_014274	206827_s_at	55503	ENSG00000165125
NK cells	NM_014274	206827_s_at	55503	ENSG00000276971
NK cells	LDB3	216888_at	11155	ENSG00000122367
NK cells	ADARB1	209979_at	104	ENSG00000197381
NK cells	SMEK1	215607_x_at	55671	ENSG00000100796
NK cells	TCTN2	206438_x_at	79867	ENSG00000168778
NK cells	TINAGL1	219058_x_at	64129	ENSG00000142910
NK cells	IGFBP5	203426_s_at	3488	ENSG00000115461
NK cells	ALDH1B1	209646_x_at	219	ENSG00000137124
NK cells	NCR1	217095_x_at	9437	ENSG00000278362
NK cells	NCR1	217095_x_at	9437	ENSG00000275156
NK cells	NCR1	217095_x_at	9437	ENSG00000277442
NK cells	NCR1	217095_x_at	9437	ENSG00000276450
NK cells	NCR1	217095_x_at	9437	ENSG00000278025
NK cells	NCR1	217095_x_at	9437	ENSG00000189430
NK cells	NCR1	217095_x_at	9437	ENSG00000275521
NK cells	NCR1	217095_x_at	9437	ENSG00000277334
NK cells	NCR1	217095_x_at	9437	ENSG00000275637
NK cells	NCR1	217095_x_at	9437	ENSG00000273535
NK cells	NCR1	217095_x_at	9437	ENSG00000277824
NK cells	NCR1	217095_x_at	9437	ENSG00000275822
NK cells	NCR1	217095_x_at	9437	ENSG00000274053
NK cells	NCR1	217095_x_at	9437	ENSG00000273506
NK cells	NCR1	217095_x_at	9437	ENSG00000277629
NK cells	NCR1	217095_x_at	9437	ENSG00000273916
NK cells	NCR1	217095_x_at	9437	ENSG00000284113
NK cells	NCR1	217095_x_at	9437	ENSG00000284208
NK cells	NCR1	217088_s_at	9437	ENSG00000278362
NK cells	NCR1	217088_s_at	9437	ENSG00000275156
NK cells	NCR1	217088_s_at	9437	ENSG00000277442
NK cells	NCR1	217088_s_at	9437	ENSG00000276450
NK cells	NCR1	217088_s_at	9437	ENSG00000278025
NK cells	NCR1	217088_s_at	9437	ENSG00000189430
NK cells	NCR1	217088_s_at	9437	ENSG00000275521
NK cells	NCR1	217088_s_at	9437	ENSG00000277334
NK cells	NCR1	217088_s_at	9437	ENSG00000275637
NK cells	NCR1	217088_s_at	9437	ENSG00000273535
NK cells	NCR1	217088_s_at	9437	ENSG00000277824
NK cells	NCR1	217088_s_at	9437	ENSG00000275822
NK cells	NCR1	217088_s_at	9437	ENSG00000274053
NK cells	NCR1	217088_s_at	9437	ENSG00000273506
NK cells	NCR1	217088_s_at	9437	ENSG00000277629

NK cells	NCR1	217088_s_at	9437	ENSG00000273916
NK cells	NCR1	217088_s_at	9437	ENSG00000284113
NK cells	NCR1	217088_s_at	9437	ENSG00000284208
NK cells	NCR1	207860_at	9437	ENSG00000278362
NK cells	NCR1	207860_at	9437	ENSG00000275156
NK cells	NCR1	207860_at	9437	ENSG00000277442
NK cells	NCR1	207860_at	9437	ENSG00000276450
NK cells	NCR1	207860_at	9437	ENSG00000278025
NK cells	NCR1	207860_at	9437	ENSG00000189430
NK cells	NCR1	207860_at	9437	ENSG00000275521
NK cells	NCR1	207860_at	9437	ENSG00000277334
NK cells	NCR1	207860_at	9437	ENSG00000275637
NK cells	NCR1	207860_at	9437	ENSG00000273535
NK cells	NCR1	207860_at	9437	ENSG00000277824
NK cells	NCR1	207860_at	9437	ENSG00000275822
NK cells	NCR1	207860_at	9437	ENSG00000274053
NK cells	NCR1	207860_at	9437	ENSG00000273506
NK cells	NCR1	207860_at	9437	ENSG00000277629
NK cells	NCR1	207860_at	9437	ENSG00000273916
NK cells	NCR1	207860_at	9437	ENSG00000284113
NK cells	NCR1	207860_at	9437	ENSG00000284208
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278656
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278361
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000277982
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275626
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000276357
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275511
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000273735
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275262
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278442
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278707
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278758
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000277181
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275838
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000276739
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000277709
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000273911
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000276004
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278403
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275083
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278809
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278726
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275629
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000276882
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000240403
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278710
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000274722
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275416
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278474
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275566
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278850
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000276424
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284295
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284384
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284466
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284101
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284213
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284046
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284063
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000283975
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284528
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284053
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284192
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000283951
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278656
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278361
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000277982
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275626
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000276357
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275511
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000273735
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275262

NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278442
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278707
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278758
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000277181
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275838
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000276739
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000277709
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000273911
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000276004
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278403
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275083
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278809
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278726
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275629
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000276882
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000240403
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278710
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000274722
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275416
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278474
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275566
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278850
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000276424
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284295
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284384
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284466
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284101
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284213
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284046
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284063
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000283975
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284528
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284053
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284192
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000283951
NK CD56dim cells	SPON2	218638_s_at	10417	ENSG00000159674
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000275623
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000275008
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000274952
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000276459
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000277554
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000276590
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000274410
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000275658
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000274830
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000278327
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000274402
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000277484
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000277317
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000273887
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000243772
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000274108
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000273947
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000277924
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000278369
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000276218
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284132
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284504
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284510
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000283996
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000283708
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000283702
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000283790
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284236
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284044
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284333
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284241
NK CD56dim cells	GZMB	210164_at	3002	ENSG00000100453
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000274283
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000275608
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000276534
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000275434

NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000276498
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000275037
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000274465
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000274146
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000277175
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000274920
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000278427
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000278856
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000273518
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000276501
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000277272
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000274036
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000274948
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275717
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275486
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000276379
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275288
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275786
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000167633
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000278368
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000276423
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000278079
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275545
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275659
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000273775
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000276329
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000284426
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000283954
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000283729
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000284589
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000284177
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000283731
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000284093
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000283827
NK CD56dim cells	FLJ20699	216434_at	55020	ENSG00000075234
NK CD56dim cells	TMEPAI	217875_s_at	56937	ENSG00000124225
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278656
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278361
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000277982
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275626
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000276357
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275511
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000273735
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275262
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278442
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278707
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278758
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000277181
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275838
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000276739
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000277709
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000273911
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000276004
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278403
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275083
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278809
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278726
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275629
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000276882
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000240403
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278710
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000274722
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275416
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278474
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275566
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278850
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000276424
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284295
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284384
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284466
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284101
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284213

NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284046
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284063
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000283975
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284528
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284053
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284192
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000283951
NK CD56dim cells	IL21R	221658_s_at	50615	ENSG00000103522
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278656
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278361
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000277982
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275626
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000276357
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275511
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000273735
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275262
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278442
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278707
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278758
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000277181
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275838
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000276739
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000277709
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000273911
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000276004
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278403
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275083
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278809
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278726
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275629
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000276882
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000240403
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278710
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000274722
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275416
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278474
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275566
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278850
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000276424
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284295
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284384
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284466
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284101
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284213
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284046
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284063
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000283975
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284528
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284053
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284192
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000283951
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000278490
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274763
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000275513
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000275062
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000278723
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274254
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274639
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276086
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274511
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276433
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276328
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000277620
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274786
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276196
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000275172
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276572
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276875
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274556
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276806
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276084
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274394

NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274696
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276930
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000277028
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000277552
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000242019
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000277392
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274480
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000275433
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000278729
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000277596
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000273502
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274724
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000283875
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000284104
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000284371
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000283823
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000283915
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000284086
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000283966
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000284480
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000284127
NK CD56dim cells	KIR2DS5	208203_x_at	3810	ENSG00000274739
NK CD56dim cells	KIR2DS5	208203_x_at	3810	ENSG00000276676
NK CD56dim cells	KIR2DS2	211532_x_at	3807	NaN
NK CD56dim cells	GTF3C1	202320_at	2975	ENSG00000077235
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000278304
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000273603
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000273517
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000278120
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000275421
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000275306
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000276327
NK CD56dim cells	EDG8	221417_x_at	53637	ENSG00000180739
NK CD56bright cells	DUSP4	204014_at	1846	ENSG00000120875
NK CD56bright cells	RRAD	204803_s_at	6236	ENSG00000166592
NK CD56bright cells	XCL1	206365_at	6375	ENSG00000143184
NK CD56bright cells	PLA2G6	215938_s_at	8398	ENSG00000184381
NK CD56bright cells	PLA2G6	204691_x_at	8398	ENSG00000184381
NK CD56bright cells	NIBP	221672_s_at	83696	ENSG00000167632
NK CD56bright cells	FOXJ1	205906_at	2302	ENSG00000129654
NK CD56bright cells	03/06/09	215908_at	10299	ENSG00000145495
NK CD56bright cells	DUSP4	204015_s_at	1846	ENSG00000120875
NK CD56bright cells	PLA2G6	210647_x_at	8398	ENSG00000184381
NK CD56bright cells	MADD	38398_at	8567	ENSG00000110514
NK CD56bright cells	BG255923	215409_at	254531	ENSG00000176454
NK CD56bright cells	MPPED1	206436_at	758	ENSG00000186732
NK CD56bright cells	MUC3B	214676_x_at	57876	NaN
DC	CD209	207277_at	30835	ENSG00000090659
DC	CCL17	207900_at	6361	ENSG00000102970
DC	HSD11B1	205404_at	3290	ENSG00000117594
DC	CCL13	206407_s_at	6357	ENSG00000181374
DC	CCL22	207861_at	6367	ENSG00000102962
DC	PPFIBP2	212841_s_at	8495	ENSG00000166387
DC	NPR1	32625_at	4881	ENSG00000169418
iDC	CD1B	206749_at	910	ENSG00000158485
iDC	VASH1	203940_s_at	22846	ENSG00000071246
iDC	F13A1	203305_at	2162	ENSG00000124491
iDC	CD1E	215784_at	913	ENSG00000158488
iDC	MMP12	204580_at	4321	ENSG00000262406
iDC	FABP4	203980_at	2167	ENSG00000170323
iDC	CLEC10A	206682_at	10462	ENSG00000132514
iDC	SYT17	205613_at	51760	ENSG00000103528
iDC	MS4A6A	219666_at	64231	ENSG00000110077
iDC	CTNS	204925_at	1497	ENSG00000040531
iDC	GUCA1A	206062_at	2978	ENSG00000048545
iDC	CARD9	220162_s_at	64170	ENSG00000187796
iDC	CD1E	208592_s_at	913	ENSG00000158488
iDC	ABCG2	209735_at	9429	ENSG00000118777
iDC	CD1A	210325_at	909	ENSG00000158477
iDC	PPARG	208510_s_at	5468	ENSG00000132170
iDC	RAP1GAP	203911_at	5909	ENSG00000076864
iDC	SLC7A8	216604_s_at	23428	ENSG00000092068

iDC	GSTT1	203815_at	2952	ENSG00000277656
iDC	NM_021941	218019_s_at	8566	ENSG00000160209
iDC	FZD2	210220_at	2535	ENSG00000180340
iDC	CSF1R	203104_at	1436	ENSG00000182578
iDC	HS3ST2	219697_at	9956	ENSG00000122254
iDC	CH25H	206932_at	9023	ENSG00000138135
iDC	LMAN2L	221274_s_at	81562	ENSG00000114988
iDC	SLC26A6	221572_s_at	65010	ENSG00000225697
iDC	BLVRB	202201_at	645	ENSG00000090013
iDC	NUDT9	218375_at	53343	ENSG00000170502
iDC	PREP	204117_at	5550	ENSG00000085377
iDC	TM7SF4	221266_s_at	81501	ENSG00000164935
iDC	TACSTD2	202286_s_at	4070	ENSG00000184292
iDC	CD1C	205987_at	911	ENSG00000158481
aDC	CCL1	207533_at	6346	ENSG00000108702
aDC	EBI3	219424_at	10148	ENSG00000105246
aDC	INDO	210029_at	3620	ENSG00000131203
aDC	LAMP3	205569_at	27074	ENSG00000078081
aDC	OAS3	218400_at	4940	ENSG00000111331
pDC	IL3RA	206148_at	3563	ENSG00000185291
Eosinophils	IL5RA	211517_s_at	3568	ENSG00000091181
Eosinophils	KCNH2	205262_at	3757	ENSG00000055118
Eosinophils	TKTL1	216370_s_at	8277	ENSG00000007350
Eosinophils	IL5RA	210744_s_at	3568	ENSG00000091181
Eosinophils	EMR1	207111_at	2015	ENSG00000174837
Eosinophils	KCNH2	210036_s_at	3757	ENSG00000055118
Eosinophils	CCR3	208304_at	1232	ENSG00000183625
Eosinophils	ACACB	49452_at	32	ENSG00000076555
Eosinophils	THBS1	201108_s_at	7057	ENSG00000137801
Eosinophils	GALC	211810_s_at	2581	ENSG00000054983
Eosinophils	TKTL1	214183_s_at	8277	ENSG00000007350
Eosinophils	RNU2	210230_at	728965	NaN
Eosinophils	CLC	206207_at	1178	ENSG00000105205
Eosinophils	THBS1	201109_s_at	7057	ENSG00000137801
Eosinophils	HIST1H1C	209398_at	3006	ENSG00000187837
Eosinophils	CYSLTR2	220813_at	57105	ENSG00000152207
Eosinophils	HRH4	221170_at	59340	ENSG00000134489
Eosinophils	RNASE2	206111_at	6036	ENSG00000169385
Eosinophils	CAT	211922_s_at	847	ENSG00000121691
Eosinophils	LRP5L	214873_at	91355	ENSG00000100068
Eosinophils	SYNJ1	207594_s_at	8867	ENSG00000159082
Eosinophils	SYNJ1	212990_at	8867	ENSG00000159082
Eosinophils	THBS4	204776_at	7060	ENSG00000113296
Eosinophils	GPR44	206361_at	11251	ENSG00000183134
Eosinophils	KBTBD11	204301_at	9920	ENSG00000176595
Eosinophils	KBTBD11	204301_at	9920	ENSG00000273645
Eosinophils	HES1	203394_s_at	3280	ENSG00000114315
Eosinophils	ABHD2	205566_at	11057	ENSG00000140526
Eosinophils	TIPARP	212665_at	25976	ENSG00000163659
Eosinophils	SMPD3	219695_at	55512	ENSG00000103056
Eosinophils	MYO15B	59375_at	80022	NaN
Eosinophils	TGIF1	203313_s_at	7050	ENSG00000177426
Eosinophils	RRP12	216360_x_at	23223	ENSG00000052749
Eosinophils	ACACB	43427_at	32	ENSG00000076555
Eosinophils	IGSF2	207167_at	9398	ENSG00000134256
Eosinophils	HES1	203395_s_at	3280	ENSG00000114315
Eosinophils	RCOR3	218344_s_at	55758	ENSG00000117625
Eosinophils	EPN2	203463_s_at	22905	ENSG00000072134
Eosinophils	C9orf156	222195_s_at	51531	ENSG00000136932
Eosinophils	SIAH1	202981_x_at	6477	ENSG00000196470
Eosinophils	ACACB	221928_at	32	ENSG00000076555
Macrophages	MARCO	205819_at	8685	ENSG00000019169
Macrophages	CXCL5	214974_x_at	6374	ENSG00000163735
Macrophages	SCG5	203889_at	6447	ENSG00000166922
Macrophages	SCG5	203889_at	6447	ENSG00000277614
Macrophages	SCG5	203889_at	6447	ENSG00000281931
Macrophages	SULT1C2	205342_s_at	6819	ENSG00000198203
Macrophages	SULT1C2	211470_s_at	6819	ENSG00000198203
Macrophages	MSR1	214770_at	4481	ENSG00000038945
Macrophages	CTSK	202450_s_at	1513	ENSG00000143387
Macrophages	PTGDS	212187_x_at	5730	ENSG00000107317
Macrophages	COLEC12	221019_s_at	81035	ENSG00000158270

Macrophages	GPC4	204984_at	2239	ENSG00000076716
Macrophages	MSR1	208423_s_at	4481	ENSG00000038945
Macrophages	PCOLCE2	219295_s_at	26577	ENSG00000163710
Macrophages	CHIT1	208168_s_at	1118	ENSG00000133063
Macrophages	PTGDS	211748_x_at	5730	ENSG00000107317
Macrophages	KAL1	205206_at	3730	ENSG00000011201
Macrophages	CLEC5A	219890_at	23601	ENSG00000258227
Macrophages	GPC4	204983_s_at	2239	ENSG00000076716
Macrophages	ME1	204058_at	4199	ENSG00000065833
Macrophages	DNASE2B	220380_at	58511	ENSG00000137976
Macrophages	CCL7	208075_s_at	6354	ENSG00000108688
Macrophages	FN1	214701_s_at	2335	ENSG00000115414
Macrophages	CD163	203645_s_at	9332	ENSG00000177575
Macrophages	GM2A	215891_s_at	2760	ENSG00000196743
Macrophages	SCARB2	201647_s_at	950	ENSG00000138760
Macrophages	BCAT1	214452_at	586	ENSG00000060982
Macrophages	BCAT1	214390_s_at	586	ENSG00000060982
Macrophages	RAI14	202052_s_at	26064	ENSG00000039560
Macrophages	MSR1	211887_x_at	4481	ENSG00000038945
Macrophages	COL8A2	52651_at	1296	ENSG00000171812
Macrophages	CD163	215049_x_at	9332	ENSG00000177575
Macrophages	APOE	203381_s_at	348	ENSG00000130203
Macrophages	CHI3L1	209396_s_at	1116	ENSG00000133048
Macrophages	ATG7	218673_s_at	10533	ENSG00000197548
Macrophages	CD84	211190_x_at	8832	ENSG00000066294
Macrophages	FDX1	203646_at	2230	ENSG00000137714
Macrophages	MS4A4A	219607_s_at	51338	ENSG00000110079
Macrophages	SGMS1	212989_at	259230	ENSG00000198964
Macrophages	EMP1	201324_at	2012	ENSG00000134531
Macrophages	CYBB	203922_s_at	1536	ENSG00000165168
Macrophages	CD68	203507_at	968	ENSG00000129226
Mast cells	PRG2	211743_s_at	5553	ENSG00000186652
Mast cells	CTSG	205653_at	1511	ENSG00000100448
Mast cells	TPSAB1	215382_x_at	7177	ENSG00000172236
Mast cells	SLC18A2	205857_at	6571	ENSG00000165646
Mast cells	TPSAB1	205683_x_at	7177	ENSG00000172236
Mast cells	MS4A2	207497_s_at	2206	ENSG00000149534
Mast cells	CPA3	205624_at	1359	ENSG00000163751
Mast cells	TPSB2	207134_x_at	64499	ENSG00000197253
Mast cells	TPSAB1	216474_x_at	7177	ENSG00000172236
Mast cells	NM_003293	207741_x_at	64499	ENSG00000197253
Mast cells	TPSAB1	210084_x_at	7177	ENSG00000172236
Mast cells	MS4A2	207496_at	2206	ENSG00000149534
Mast cells	TPSAB1	217023_x_at	7177	ENSG00000172236
Mast cells	GATA2	209710_at	2624	ENSG00000179348
Mast cells	HDC	207067_s_at	3067	ENSG00000140287
Mast cells	LOH11CR2A	210102_at	4013	ENSG00000110002
Mast cells	SIGLEC6	210796_x_at	946	ENSG00000105492
Mast cells	ELA2	206871_at	1991	ENSG00000277571
Mast cells	ELA2	206871_at	1991	ENSG00000197561
Mast cells	LOH11CR2A	205011_at	4013	ENSG00000110002
Mast cells	CMA1	214533_at	1215	ENSG00000092009
Mast cells	SIGLEC6	206520_x_at	946	ENSG00000105492
Mast cells	PGDS	206726_at	27306	ENSG00000163106
Mast cells	MLPH	218211_s_at	79083	ENSG00000115648
Mast cells	ADCYAP1	206281_at	116	ENSG00000141433
Mast cells	SIGLEC6	206519_x_at	946	ENSG00000105492
Mast cells	SLC24A3	57588_at	57419	ENSG00000185052
Mast cells	CALB2	205428_s_at	794	ENSG00000172137
Mast cells	CALB2	205428_s_at	794	ENSG00000282830
Mast cells	SLC24A3	219090_at	57419	ENSG00000185052
Mast cells	KIT	205051_s_at	3815	ENSG00000157404
Mast cells	TAL1	206283_s_at	6886	ENSG00000162367
Mast cells	ABCC4	203196_at	10257	ENSG00000125257
Mast cells	PPM1H	212686_at	57460	ENSG00000111110
Mast cells	MAOB	204041_at	4129	ENSG00000069535
Mast cells	HPGD	211549_s_at	3248	ENSG00000164120
Mast cells	SCG2	204035_at	7857	ENSG00000171951
Mast cells	PTGS1	205127_at	5742	ENSG00000095303
Mast cells	CEACAM8	206676_at	1088	ENSG00000124469
Mast cells	MPO	203949_at	4353	ENSG00000005381
Mast cells	NR0B1	206645_s_at	190	ENSG00000169297

Mast cells	LOC339524	215039_at	339524	ENSG00000267272
Neutrophils	CSF3R	203591_s_at	1441	ENSG00000119535
Neutrophils	CYP4F3	206515_at	4051	ENSG00000186529
Neutrophils	VNN3	220528_at	55350	ENSG00000093134
Neutrophils	FPRL1	210773_s_at	2358	ENSG00000171049
Neutrophils	KCNJ15	216782_at	3772	ENSG00000157551
Neutrophils	MME	203434_s_at	4311	ENSG00000196549
Neutrophils	IL8RA	207094_at	3577	ENSG00000163464
Neutrophils	IL8RB	207008_at	3579	ENSG00000180871
Neutrophils	MME	203435_s_at	4311	ENSG00000196549
Neutrophils	FCGR3B	204007_at	2215	ENSG00000162747
Neutrophils	DYSF	218660_at	8291	ENSG00000135636
Neutrophils	KCNJ15	211806_s_at	3772	ENSG00000157551
Neutrophils	FCAR	211816_x_at	2204	ENSG00000278415
Neutrophils	FCAR	211816_x_at	2204	ENSG00000274580
Neutrophils	FCAR	211816_x_at	2204	ENSG00000275136
Neutrophils	FCAR	211816_x_at	2204	ENSG00000275970
Neutrophils	FCAR	211816_x_at	2204	ENSG00000276985
Neutrophils	FCAR	211816_x_at	2204	ENSG00000186431
Neutrophils	FCAR	211816_x_at	2204	ENSG00000276858
Neutrophils	FCAR	211816_x_at	2204	ENSG00000275269
Neutrophils	FCAR	211816_x_at	2204	ENSG00000273738
Neutrophils	FCAR	211816_x_at	2204	ENSG00000275564
Neutrophils	FCAR	211816_x_at	2204	ENSG00000283953
Neutrophils	FCAR	211816_x_at	2204	ENSG00000283750
Neutrophils	FCAR	211816_x_at	2204	ENSG00000284245
Neutrophils	FCAR	211816_x_at	2204	ENSG00000284004
Neutrophils	FCAR	211816_x_at	2204	ENSG00000284061
Neutrophils	FCAR	211307_s_at	2204	ENSG00000278415
Neutrophils	FCAR	211307_s_at	2204	ENSG00000274580
Neutrophils	FCAR	211307_s_at	2204	ENSG00000275136
Neutrophils	FCAR	211307_s_at	2204	ENSG00000275970
Neutrophils	FCAR	211307_s_at	2204	ENSG00000276985
Neutrophils	FCAR	211307_s_at	2204	ENSG00000186431
Neutrophils	FCAR	211307_s_at	2204	ENSG00000276858
Neutrophils	FCAR	211307_s_at	2204	ENSG00000275269
Neutrophils	FCAR	211307_s_at	2204	ENSG00000273738
Neutrophils	FCAR	211307_s_at	2204	ENSG00000275564
Neutrophils	FCAR	211307_s_at	2204	ENSG00000283953
Neutrophils	FCAR	211307_s_at	2204	ENSG00000283750
Neutrophils	FCAR	211307_s_at	2204	ENSG00000284245
Neutrophils	FCAR	211307_s_at	2204	ENSG00000284004
Neutrophils	FCAR	211307_s_at	2204	ENSG00000284061
Neutrophils	CEACAM3	210789_x_at	1084	ENSG00000170956
Neutrophils	FPRL1	210772_at	2358	ENSG00000171049
Neutrophils	HIST1H2BC	214455_at	8347	ENSG00000180596
Neutrophils	HPSE	219403_s_at	10855	ENSG00000173083
Neutrophils	FLJ11151	218610_s_at	55313	ENSG00000103381
Neutrophils	CREB5	205931_s_at	9586	ENSG00000146592
Neutrophils	S100A12	205863_at	6283	ENSG00000163221
Neutrophils	FCGR3B	204006_s_at	2215	ENSG00000162747
Neutrophils	TNFRSF10C	211163_s_at	8794	ENSG00000173535
Neutrophils	SLC22A4	205896_at	6583	ENSG00000197208
Neutrophils	KIAA0329	204307_at	9895	ENSG00000196663
Neutrophils	SLC25A37	218136_s_at	51312	ENSG00000147454
Neutrophils	BST1	205715_at	683	ENSG00000109743
Neutrophils	FCAR	207674_at	2204	ENSG00000278415
Neutrophils	FCAR	207674_at	2204	ENSG00000274580
Neutrophils	FCAR	207674_at	2204	ENSG00000275136
Neutrophils	FCAR	207674_at	2204	ENSG00000275970
Neutrophils	FCAR	207674_at	2204	ENSG00000276985
Neutrophils	FCAR	207674_at	2204	ENSG00000186431
Neutrophils	FCAR	207674_at	2204	ENSG00000276858
Neutrophils	FCAR	207674_at	2204	ENSG00000275269
Neutrophils	FCAR	207674_at	2204	ENSG00000273738
Neutrophils	FCAR	207674_at	2204	ENSG00000275564
Neutrophils	FCAR	207674_at	2204	ENSG00000283953
Neutrophils	FCAR	207674_at	2204	ENSG00000283750
Neutrophils	FCAR	207674_at	2204	ENSG00000284245
Neutrophils	FCAR	207674_at	2204	ENSG00000284004
Neutrophils	FCAR	207674_at	2204	ENSG00000284061
Neutrophils	CEACAM3	208052_x_at	1084	ENSG00000170956

Neutrophils	CRISPLD2	221541_at	83716	ENSG00000103196
Neutrophils	TNFRSF10C	206222_at	8794	ENSG00000173535
Neutrophils	G0S2	213524_s_at	50486	ENSG00000123689
Neutrophils	SIGLEC5	220000_at	8778	ENSG00000105501
Neutrophils	CD93	202878_s_at	22918	ENSG00000125810
Neutrophils	MGAM	206522_at	8972	ENSG00000257335
Neutrophils	MGAM	206522_at	8972	ENSG00000282607
Neutrophils	ALPL	215783_s_at	249	ENSG00000162551
Neutrophils	FPR1	205119_s_at	2357	ENSG00000171051
Neutrophils	CD93	202877_s_at	22918	ENSG00000125810
Neutrophils	PDE4B	222326_at	5142	ENSG00000184588
Neutrophils	LILRB2	210146_x_at	10288	ENSG00000274513
Neutrophils	LILRB2	210146_x_at	10288	ENSG00000131042
Neutrophils	LILRB2	210146_x_at	10288	ENSG00000275463
Neutrophils	LILRB2	210146_x_at	10288	ENSG00000277751
Neutrophils	LILRB2	210146_x_at	10288	ENSG00000276146
SW480 cancer cells	KRT5	201820_at	3852	ENSG00000186081
SW480 cancer cells	RBP1	203423_at	5947	ENSG00000114115
SW480 cancer cells	TRIM29	202504_at	23650	ENSG00000137699
SW480 cancer cells	DEFA5	207529_at	1670	ENSG00000164816
SW480 cancer cells	BMP4	211518_s_at	652	ENSG00000125378
SW480 cancer cells	EEF1A2	204540_at	1917	ENSG00000101210
SW480 cancer cells	VSNL1	203797_at	7447	ENSG00000163032
SW480 cancer cells	ASPSCR1	218908_at	79058	ENSG00000169696
SW480 cancer cells	IGF2	202409_at	3481	ENSG00000167244
SW480 cancer cells	IGF2	202409_at	3481	ENSG00000129965
SW480 cancer cells	MFAP2	203417_at	4237	ENSG00000117122
SW480 cancer cells	FGF3	214571_at	2248	ENSG00000186895
SW480 cancer cells	S100A2	204268_at	6273	ENSG00000196754
SW480 cancer cells	INHBB	205258_at	3625	ENSG00000163083
SW480 cancer cells	JAG2	209784_s_at	3714	ENSG00000184916
SW480 cancer cells	LOC89944	213713_s_at	89944	ENSG00000149328
SW480 cancer cells	BAMBI	203304_at	25805	ENSG00000095739
SW480 cancer cells	JAG2	32137_at	3714	ENSG00000184916
SW480 cancer cells	BMP7	209591_s_at	655	ENSG00000101144
SW480 cancer cells	RPP25	219143_s_at	54913	ENSG00000178718
SW480 cancer cells	RHOD	209885_at	29984	ENSG00000173156
SW480 cancer cells	DHRS2	214079_at	10202	ENSG00000100867
SW480 cancer cells	ITGB4	204989_s_at	3691	ENSG00000132470
SW480 cancer cells	NTSR1	207360_s_at	4923	ENSG00000101188
SW480 cancer cells	STRA6	221701_s_at	64220	ENSG00000137868
SW480 cancer cells	SLC1A5	208916_at	6510	ENSG00000105281
SW480 cancer cells	VSNL1	203798_s_at	7447	ENSG00000163032
SW480 cancer cells	FKBP4	200894_s_at	2288	ENSG00000004478
SW480 cancer cells	S100A3	206027_at	6274	ENSG00000188015
SW480 cancer cells	TEAD4	41037_at	7004	ENSG00000197905
SW480 cancer cells	KLK6	204733_at	5653	ENSG00000167755
SW480 cancer cells	CCND1	208711_s_at	595	ENSG00000110092
SW480 cancer cells	SLC27A5	219733_s_at	10998	ENSG00000083807
SW480 cancer cells	HOXA9	214651_s_at	3205	ENSG00000078399
SW480 cancer cells	F12	205774_at	2161	ENSG00000131187
SW480 cancer cells	LRFN4	219491_at	78999	ENSG00000173621
SW480 cancer cells	NM_024609	218678_at	10763	ENSG00000132688
SW480 cancer cells	SLC6A8	210854_x_at	6535	ENSG00000130821
SW480 cancer cells	PCTK1	207239_s_at	5127	ENSG00000102225
SW480 cancer cells	SLC6A8	213843_x_at	6535	ENSG00000130821
SW480 cancer cells	KRT13	207935_s_at	3860	ENSG00000171401
Normal mucosa	TSPAN8	203824_at	7103	ENSG00000127324
Normal mucosa	LGALS4	204272_at	3960	ENSG00000171747
Normal mucosa	LGALS4	204272_at	3960	ENSG00000282992
Normal mucosa	DCN	201893_x_at	1634	ENSG00000011465
Normal mucosa	COL3A1	215076_s_at	1281	ENSG00000168542
Normal mucosa	COL3A1	201852_x_at	1281	ENSG00000168542
Normal mucosa	CEACAM5	201884_at	1048	ENSG00000105388
Normal mucosa	TAGLN	205547_s_at	6876	ENSG00000149591
Normal mucosa	KRT20	213953_at	54474	ENSG00000263057
Normal mucosa	KRT20	213953_at	54474	ENSG00000171431
Normal mucosa	DCN	211896_s_at	1634	ENSG00000011465
Normal mucosa	MYH11	201497_x_at	4629	ENSG00000133392
Normal mucosa	MYH11	201497_x_at	4629	ENSG00000276480
Normal mucosa	FXYD3	202489_s_at	5349	ENSG00000089356
Normal mucosa	ACTG2	202274_at	72	ENSG00000163017

Normal mucosa	MYLK	202555_s_at	4638	ENSG00000065534
Normal mucosa	TPM1	210987_x_at	7168	ENSG00000140416
Normal mucosa	CDH17	209847_at	1015	ENSG00000079112
Normal mucosa	NFIB	209289_at	4781	ENSG00000147862
Normal mucosa	MGP	202291_s_at	4256	ENSG00000111341
Normal mucosa	SPARCL1	200795_at	8404	ENSG00000152583
Normal mucosa	RGS5	209071_s_at	8490	ENSG00000232995
Normal mucosa	RGS5	209071_s_at	8490	ENSG00000143248
Normal mucosa	MYH11	207961_x_at	4629	ENSG00000133392
Normal mucosa	MYH11	207961_x_at	4629	ENSG00000276480
Normal mucosa	PPAP2B	212226_s_at	8613	ENSG00000162407
Normal mucosa	COL3A1	211161_s_at	1281	ENSG00000168542
Normal mucosa	IGFBP7	201162_at	3490	ENSG00000163453
Normal mucosa	PPAP2B	209355_s_at	8613	ENSG00000162407
Normal mucosa	CALD1	201616_s_at	800	ENSG00000122786
Normal mucosa	C1R	212067_s_at	715	ENSG00000159403
Normal mucosa	CALD1	212077_at	800	ENSG00000122786
Normal mucosa	AGR2	209173_at	10551	ENSG00000106541
Normal mucosa	GNA11	564_at	2767	ENSG00000088256
Normal mucosa	HEPH	203903_s_at	9843	ENSG00000089472
Normal mucosa	GNA11	213944_x_at	2767	ENSG00000088256
Normal mucosa	GNG12	212294_at	55970	ENSG00000172380
Normal mucosa	ADH1B	209612_s_at	125	ENSG00000196616
Normal mucosa	TPM1	206116_s_at	7168	ENSG00000140416
Normal mucosa	CRIM1	202551_s_at	51232	ENSG00000277354
Normal mucosa	CRIM1	202551_s_at	51232	ENSG00000150938
Normal mucosa	FBLN1	202994_s_at	2192	ENSG00000077942
Normal mucosa	IGFBP5	211959_at	3488	ENSG00000115461
Normal mucosa	LAMA4	202202_s_at	3910	ENSG00000112769
Normal mucosa	CAV1	212097_at	857	ENSG00000105974
Normal mucosa	WASL	205809_s_at	8976	ENSG00000106299
Blood vessels	CDH5	204677_at	1003	ENSG00000179776
Lymph vessels	PDPN	204879_at	10630	ENSG00000162493
Lymph vessels	VEGFC	209946_at	7424	ENSG00000150630
Lymph vessels	FIGF	206742_at	2277	ENSG00000165197

Table A.5 – Signatures adapted from Bindea et al. [2013]. Genes Id were matched from tables available at https://github.com/judithabk6/ITH_TCGA/tree/master/external_data

Appendix B

Supplementary materials for CloneSig

B.1 Supplementary methods

B.1.1 EM algorithm for parameter estimation

In this section we detail the EM algorithm used to estimate the parameters $\boldsymbol{\theta} = (\xi, \phi, \pi, \rho)$ of CloneSig, for a given number of clones J. To lighten notations, we use in this section the notation $M_{max_n} = (C_{tumor}^{major})_n$ for the maximum value that M_n can take. We do not model the distributions of the observed variables C_n (copy number information) and D_n (total read count), and therefore only consider the following complete conditional log-likelihood:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \log \left[\prod_{n=1}^{N} \mathbb{P}(B_n, T_n, U_n, S_n, M_n | D_n, C_n; \boldsymbol{\theta}, p, \mu) \right] \\
&= \log \left[\prod_{n=1}^{N} \prod_{u=1}^{J} \prod_{s=1}^{L} \prod_{m=1}^{M_{max_n}} (\mathbb{P}(U_n = u; \boldsymbol{\theta}) \mathbb{P}(S_n = s | U_n = u; \boldsymbol{\theta}) \mathbb{P}(T_n = t | S_n = s; \mu) \right. \\
&\quad \left. \mathbb{P}(M_n = m | C_n) \mathbb{P}(B_n | D_n, C_n, M_n = m, U_n = u; \boldsymbol{\theta}, p))^{\mathbb{I}(S_n = s, U_n = u, M_n = m)} \right] \\
&= \sum_{n=1}^{N} \sum_{u=1}^{J} \sum_{s=1}^{L} \sum_{m=1}^{M_{max_n}} \mathbb{I}(S_n = s, U_n = u, M_n = m) \\
&\log \left[\xi_u \pi_{us} \mu_{st} M_{max_n}^{-1} \, \mathrm{BB}(B_n \, ; \, D_n, \rho \phi_u \eta_{nm}, \rho(1 - \phi_u \eta_{nm})) \right] ,
\end{aligned}$$

where BB is the beta-binomial density:

$$\mathrm{BB}(k \, ; \, n, \alpha, \beta) = \binom{n}{k} \frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)\Gamma(\alpha+\beta)}{\Gamma(n+\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)} ,$$

and

$$\eta_{nm} = \frac{pm}{p \times (C_{tumor})_n + (1-p) \times (C_{normal})_n} .$$

To maximize $\mathcal{L}(\boldsymbol{\theta})$, we introduce the auxiliary function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta'})$ as the expected value of the loglikelihood function of $\boldsymbol{\theta}$ when the latent variables follow the law with parameters $\boldsymbol{\theta'}$, that will be alternatively computed and maximized in the two steps of the EM algorithm. For that purpose, let us denote by $\mathbf{X}_n = (C_n, T_n, B_n, D_n)$ the set observed variables for the n-th

SNV, and $\mathcal{D} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$ the totality of observed variables. Then we define:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta'}) &= \mathbb{E}(\mathcal{L}(\boldsymbol{\theta})|\mathcal{D}; \boldsymbol{\theta'}, p, \mu) \\ &= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} q_{nu} r_{nus} v_{mnu} \log\left[\xi_u \pi_{us} \mu_{st} M_{max_n}^{-1} \,\mathrm{BB}(B_n\,;\, D_n, \rho\phi_u\eta_{nm}, \rho(1-\phi_u\eta_{nm}))\right],\end{aligned} \tag{B.1}$$

with

$$q_{nu} = \mathbb{P}(U_n = u|\mathbf{X}_n; \boldsymbol{\theta'}), \tag{B.2}$$
$$r_{nus} = \mathbb{P}(S_n = s|U_n = u, \mathbf{X}_n; \boldsymbol{\theta'}), \tag{B.3}$$
$$v_{mnu} = \mathbb{P}(M_n = m|U_n = u, \mathbf{X}_n; \boldsymbol{\theta'}). \tag{B.4}$$

The EM algorithm iteratively builds a sequence of estimate $\boldsymbol{\theta^1}, \boldsymbol{\theta^2}, \ldots$ by solving recursively

$$\boldsymbol{\theta^i} = \mathrm{argmax}_{\boldsymbol{\theta}}\, \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}}).$$

For that purpose, at each iteration i, the expectation (E) step first consists in computing the function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})$ with the current parameters $\boldsymbol{\theta^{i-1}}$. In other words, we must estimate the variables (B.2)-(B.4). Given the conditional independence relationships encoded in the graphical model (Figure 4.2), one easily gets:

$$q_{nu} = \frac{\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} \xi_u^{i-1}\,\mathrm{BB}(B_n|D_n, \rho^{i-1}\phi_u^{i-1}\eta_{nm}^{i-1}, \rho^{i-1}(1-\phi_u^{i-1}\eta_{nm}^{i-1}))\mu_{sT_n}\pi_{us}^{i-1}}{\sum_{u'=1}^{J}\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} \xi_{u'}^{i-1}\,\mathrm{BB}(B_n|D_n, \rho^{i-1}\phi_{u'}^{i-1}\eta_{nm}^{i-1}, \rho^{i-1}(1-\phi_{u'}^{i-1}\eta_{nm}^{i-1}))\mu_{sT_n}\pi_{u's}^{i-1}}, \tag{B.5}$$

$$r_{nus} = \frac{\mu_{sT_n}\pi_{us}^{i-1}}{\sum_{s'=1}^{M}\mu_{s'T_n}\pi_{us'}^{i-1}}, \tag{B.6}$$

$$v_{mnu} = \frac{\sum_{s=1}^{L}\xi_u^{i-1}\,\mathrm{BB}(B_n|D_n, \rho^{i-1}\phi_u^{i-1}\eta_{nm}^{i-1}, \rho^{i-1}(1-\phi_u^{i-1}\eta_{nm}^{i-1}))\mu_{sT_n}\pi_{us}^{i-1}}{\sum_{s=1}^{L}\sum_{m'=1}^{M_{max_n}} \xi_u^{i-1}\,\mathrm{BB}(B_n|D_n, \rho^{i-1}\phi_u^{i-1}\eta_{nm'}^{i-1}, \rho^{i-1}(1-\phi_u^{i-1}\eta_{nm'}^{i-1}))\mu_{sT_n}\pi_{us}^{i-1}}. \tag{B.7}$$

In the maximization (M) step, we compute $\boldsymbol{\theta^i}$ by plugging the estimates of the E-step (B.5)-(B.7) onto (B.1) and maximizing $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})$ separately for each component of $\boldsymbol{\theta}$. The maximization in ξ and π are easily obtained as:

$$\forall u \in (1\ldots J), \xi_u^i = \sum_{n=1}^{N}\frac{q_{nu}}{N},$$

$$\forall u \in (1\ldots J), \forall s \in (1\ldots L), \pi_{us}^i = \frac{\sum_{n=1}^{N} r_{nus} q_{nu}}{\sum_{n'=1}^{N} q_{n'u}}.$$

The optimization of ϕ and ρ inside the beta-binomial density term are not computable using a close formula. We therefore resort to numerical optimization and use a projected Newton method, with line search to set the Newton step at each iteration [Bertsekas, 1982], in order to compute approximations of ϕ^i and ρ^i that respect constraints on their domain. Indeed, ρ must be non-negative and ϕ is a proportion so in the unit interval. For that purpose,

we now compute the first and second derivatives of $\mathcal{Q}$ with respect to ϕ and $\tau = 1/\rho$:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}}) = \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} r_{nus} q_{nu} v_{mnu} \Bigg[&\log(\xi_u \mu_{st} \pi_{us} M_{max_n}^{-1}) + \log(\binom{d_n}{b_n}) \\ &+ \log(\Gamma(b_n + \frac{\phi_u \eta_{nm}}{\tau})) + \log(\Gamma(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n)) + \log(\Gamma(\frac{1}{\tau})) \\ &- \log(\Gamma(\frac{1}{\tau} + d_n)) - \log(\Gamma(\frac{\phi_u \eta_{nm}}{\tau})) - \log(\Gamma(\frac{1-\phi_u\eta_{nm}}{\tau}))\Bigg]\end{aligned}$$

Let's now compute derivatives. ψ_0 and ψ_1 denote the digamma and trigamma functions respectively.

$$\begin{aligned}\frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \tau} = &\sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{m=1}^{M_{max_n}} \frac{q_{nu} v_{mnu}}{\tau^2} \Bigg[-\eta_{nm}\phi_u \psi_0(b_n + \frac{\phi_u \eta_{nm}}{\tau}) \\ &-(1-\eta_{nm}\phi_u)\psi_0(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) - \psi_0(\frac{1}{\tau}) + \psi_0(\frac{1}{\tau} + d_n) + \eta_{nm}\phi_u\psi_0(\frac{\eta_{nm}\phi_u}{\tau}) \\ &+(1-\eta_{nm}\phi_u)\psi_0(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg] \\ \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \tau^2} = &\sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \Bigg[\frac{2}{\tau^3}\Bigg(\eta_{nm}\phi_u\psi_0(b_n + \frac{\phi_u\eta_{nm}}{\tau}) \\ &+(1-\eta_{nm}\phi_u)\psi_0(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) + \psi_0(\frac{1}{\tau}) - \psi_0(\frac{1}{\tau} + d_n) - \eta_{nm}\phi_u\psi_0(\frac{\eta_{nm}\phi_u}{\tau}) \\ &-(1-\eta_{nm}\phi_u)\psi_0(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg) + \frac{1}{\tau^4}\Bigg(\eta_{nm}^2\phi_u^2\psi_1(b_n + \frac{\phi_u\eta_{nm}}{\tau}) \\ &+(1-\eta_{nm}\phi_u)^2\psi_1(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) + \psi_1(\frac{1}{\tau}) - \psi_1(\frac{1}{\tau} + d_n) - \eta_{nm}^2\phi_u^2\psi_1(\frac{\eta_{nm}\phi_u}{\tau}) \\ &-(1-\eta_{nm}\phi_u)^2\psi_1(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg)\Bigg] \\ \frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u} = &\sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \frac{\eta_{nm}}{\tau} \Bigg[\psi_0(b_n + \frac{\phi_u\eta_{nm}}{\tau}) - \psi_0(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) \\ &-\psi_0(\frac{\eta_{nm}\phi_u}{\tau}) + \psi_0(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg] \\ \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u^2} = &\sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \frac{\eta_{nm}^2}{\tau^2} \Bigg[\psi_1(b_n + \frac{\phi_u\eta_{nm}}{\tau}) + \psi_1(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) \\ &-\psi_1(\frac{\eta_{nm}\phi_u}{\tau}) - \psi_1(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg] \\ \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u \partial \tau} = &\sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \frac{\eta_{nm}}{\tau^2} \Bigg[-\psi_0(b_n + \frac{\phi_u\eta_{nm}}{\tau}) - \frac{\eta_{nm}\phi_u}{\tau}\psi_1(b_n + \frac{\phi_u\eta_{nm}}{\tau}) \\ &+\psi_0(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) + \frac{(1-\eta_{nm}\phi_u)}{\tau}\psi_1(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) + \psi_0(\frac{\eta_{nm}\phi_u}{\tau}) \\ &+\frac{\phi_u\eta_{nm}}{\tau}\psi_1(\frac{\eta_{nm}\phi_u}{\tau}) - \psi_0(\frac{1-\eta_{nm}\phi_u}{\tau}) - \frac{1-\phi_u\eta_{nm}}{\tau}\psi_1(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg] \\ \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u \partial \phi_{u'}} = &\, 0\end{aligned}$$

For sake of completeness, we provide below a second, equivalent computation using another

formulation following Martinez et al. [2015].

$$
\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}}) &= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} r_{nus} q_{nu} v_{mnu} \\
&\log\left[\xi_u \mu_{st} \pi_{us} M_{max_n}^{-1} \binom{d_n}{b_n} \frac{\Gamma(b_n + \rho\phi_u\eta_{nm})\Gamma(\rho(1-\phi_u\eta_{nm}) + d_n - b_n)}{\Gamma(\rho + d_n)}\right. \\
&\left.\frac{\Gamma(\rho)}{\Gamma(\rho\phi_u\eta_{nm})\Gamma(\rho(1-\phi_u\eta_{nm})}\right] \\
&= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} r_{nus} q_{nu} v_{mnu} \\
&\log\left[\xi_u \mu_{st} \pi_{us} M_{max_n}^{-1} \binom{d_n}{b_n} \frac{\prod_{i=0}^{b_n-1}(\phi_u\eta_{nm} + \frac{i}{\rho}) \prod_{i=0}^{d_n-b_n-1}(1-\phi_u\eta_{nm} + \frac{i}{\rho})}{\prod_{i=0}^{d_n-1}(1+\frac{i}{\rho})}\right] \\
&= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} r_{nus} q_{nu} v_{mnu} \left[\log(\xi_u \mu_{st} \pi_{us} M_{max_n}^{-1}) + \log(\binom{d_n}{b_n})\right. \\
&\left.+ \sum_{i=0}^{b_n-1}\left[\log(\phi_u\eta_{nm} + \frac{i}{\rho})\right] + \sum_{i=0}^{d_n-b_n-1}\left[\log(1-\phi_u\eta_{nm} + \frac{i}{\rho})\right] - \sum_{i=0}^{d_n-1}\left[\log(1+\frac{i}{\rho})\right]\right]
\end{aligned}
$$

Let's set $\tau = \frac{1}{\rho}$. We are trying to compute maximum likelihood estimates for ϕ_u and τ.

$$
\begin{aligned}
\frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \tau} &= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \left[\sum_{i=0}^{b_n-1}\left[\frac{i}{\phi_u\eta_{nm} + i\tau}\right] + \sum_{i=0}^{d_n-b_n-1}\left[\frac{i}{1-\phi_u\eta_{nm} + i\tau}\right]\right. \\
&\left.- \sum_{i=0}^{d_n-1}\left[\frac{i}{1+i\tau}\right]\right] \\
\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \tau^2} &= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \left[-\sum_{i=0}^{b_n-1}\left[\frac{i^2}{(\phi_u\eta_{nm} + i\tau)}\right] - \sum_{i=0}^{d_n-b_n-1}\left[\frac{i^2}{(1-\phi_u\eta_{nm} + \tau)^2}\right]\right. \\
&\left.+ \sum_{i=0}^{d_n-1}\left[\frac{i^2}{(1+\tau)^2}\right]\right] \\
\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u \partial \tau} &= \sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \left[-\sum_{i=0}^{b_n-1}\left[\frac{i\eta_{nm}}{(\phi_u\eta_{nm} + i\tau)^2}\right] + \sum_{i=0}^{d_n-b_n-1}\left[\frac{i\eta_{nm}}{(1-\phi_u\eta_{nm} + i\tau)^2}\right]\right] \\
\frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u} &= \sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \left[\sum_{i=0}^{b_n-1}\left[\frac{\eta_{nm}}{\phi_u\eta_{nm} + i\tau}\right] + \sum_{i=0}^{d_n-b_n-1}\left[\frac{-\eta_{nm}}{1-\phi_u\eta_{nm} + i\tau}\right]\right] \\
\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u^2} &= \sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \left[-\sum_{i=0}^{b_n-1}\left[\frac{\eta_{nm}}{\phi_u\eta_{nm} + i\tau}\right]^2 - \sum_{i=0}^{d_n-b_n-1}\left[\frac{\eta_{nm}}{1-\phi_u\eta_{nm} + i\tau}\right]^2\right] \\
\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_{u'} \partial \phi_u} &= 0
\end{aligned}
$$

We can then plug these formulas in the projected Newton algorithm to estimate ϕ^i and ρ^i. We repeat the E and M steps until $\|\boldsymbol{\theta^i} - \boldsymbol{\theta^{i-1}}\| < 10^{-5} \times J \times L$.

B.1.2 Selecting the number of clones

As explained in Supplementary Section B.1.1, the EM algorithm allows us to optimize all parameters of the CloneSig model for a given number of clones J. Here we explain how to estimate J.A first idea to automatize that choice is to rely on a model selection heuristics,

such as the widely used Bayesian Information Criterion (BIC) [Schwarz, 1978], an asymptotic Bayesian criterion aiming at selecting the model best supported by the data. BIC is defined as

$$BIC(J) = \ell(\mathcal{D}; \boldsymbol{\theta}_J) - \frac{D_J}{2} \log N,$$

where $\ell(\mathcal{D}; \boldsymbol{\theta}_J)$ is the maximum log-likelihood as estimated by the EM procedure with J clones, and D_J is the degree of freedom of the model; by default, we take it equal to the number of free parameters, namely, $D_J = J * (L - 1 + 2)$ for J clones, where L is the number of signatures. Indeed, for each clone, we have $L - 1$ parameters for the signature proportions (π), the frequency of the clone (ϕ_u), and the proportion of the clone ξ_u. We have to remove 1 because $\sum_{u=1}^{J} \xi_u = 1$, and add 1 for the overdispersion parameter τ.

On simulations, however, we found that while BIC correctly identifies the number of clones when the number of SNVs is large, it tends to performs poorly when the number of mutations is low (a few hundreds) in which case it quasi systematically selects a single clone. On the other hand, when we observe the variation of the log-likelihood with the number of components J as for example in Supplementary Figure B.1, we clearly see an "elbow" for some $J > 1$, suggesting that the information about J is properly captured by CloneSig's likelihood but not by BIC.

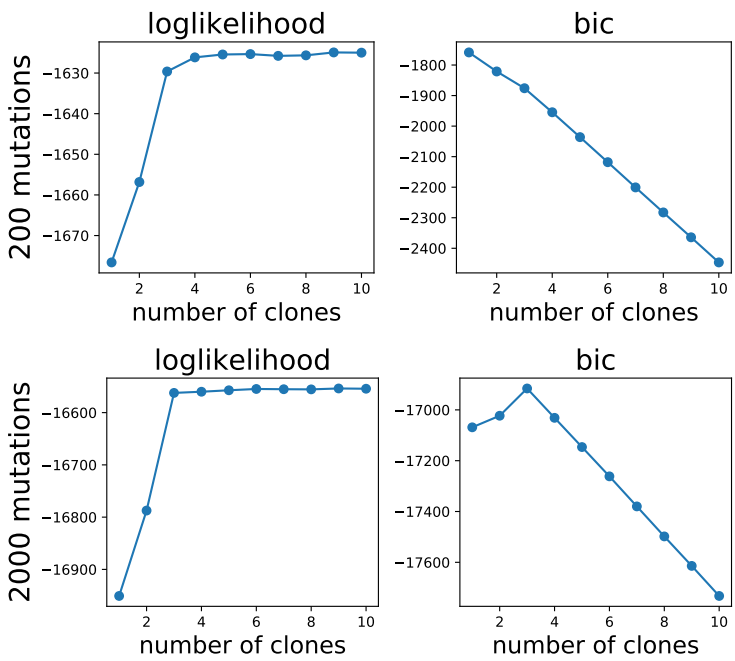

Figure B.1 – Evolution of the loglikelihood and BIC criterion for 2 simulated samples, with the same parameters and 200 mutations (up panels), and 2000 mutations (bottom panels). In both cases, the loglikelihood has an "elbow" at 3 clones indicating that the likelihood of the data increases much less at the addition of an additional mixture component beyond 3 components. The BIC criterion is maximal at 3 clones when 2000 mutations are observed, but at 1 clone in the case with 200 mutations.

We observed similar behaviors with other classical criteria such as the Akaike Information Criterion (AIC) [Akaike, 1998], the Integrated Classification Likelihood (ICL) [Biernacki et al., 2000]), or the slope heuristics as described in Maugis and Michel [2011]. This difficulty can be related to results from statistical theory of model selection and penalization suggesting that asymptotic results are known up to a factor when applied to smaller datasets [Arlot, 2019], and therefore propose now as an alternative an empirical criterion that can be fit on data with known model, such as simulations. More precisely, we consider the following criterion:

$$BIC_\alpha(J) = \ell(\mathcal{D}; \boldsymbol{\theta}_J) - \alpha D_J \log N\,. \tag{B.8}$$

with $\alpha > 0$ is a free parameter to be user-defined or estimated, and D_J is a measure complexity

of the model.

While we leave α as a user-defined parameter in the CloneSig software, we now propose a systematic approach to estimate it when we can simulate samples. For each simulated sample, we fit CloneSig for 1 to 10 clones. The objective is to estimate a parameter α such that $BIC_{\alpha,J}$ is maximal for the true number of clones J_{true} on all or most simulations. To achieve that, we formulate it as a standard supervised classification problem where for each simulation and each $J \neq J_{true}$, we want $BIC_\alpha(J_{true}) > BIC_\alpha(J)$; since $BIC_\alpha(J)$ is itself a linear function of α, we estimate α by minimizing a convex proxy to the number of errors, namely,

$$\min_\alpha \sum_{\mathcal{D}} \sum_{J \neq J_{true}} \phi\left(BIC_\alpha(J_{true}) - BIC_\alpha(J)\right) , \tag{B.9}$$

where $\phi(u) = \max(0, 1-u)$ is the hinge loss that pushes its argument to be larger than one when minimized; solving (B.9) is a simple support vector machine (SVM) problem that we solve with a standard SVM solver.

The second important aspect of (B.8) is D_J, that measures the complexity of the model with J clones. The original BIC penalizes the "dimension of the model" [Schwarz, 1978], that can be interpreted as the degree of freedom of the model, and we now discuss different possible definitions for it. The parameters ϕ, ξ and ρ determining the CCFs and proportions of the different clones in the mixture must clearly be counted as in BIC. Regarding the signatures however, one can notice that the signatures are neither orthogonal (some signatures are very similar), nor independent (some signatures are associated with the same underlying biological process). Instead of just counting the number of signatures, we therefore propose to estimate the degree of freedom dof_L of the matrix with L signatures by the number of eigenvalues of the cosine similarity matrix greater than 0.5 in absolute value. As shown in Figure B.2, dof_L is roughly proportional to L, at least for L up to 20. Another source of degree of freedom is the copy number. Indeed, for each observed mutation, several values of the number of mutated copies are considered, so if the maximal average multiplicity for mutations in the sample is $M_{max_{avg}}$, a unique clone CCF corresponds in average to $M_{max_{avg}}$ possible VAFs, adding some freedom to the model. We therefore consider four possible definitions for D_J, indexed with letters A to D.

$$D_J^A = J \times (L+1) \times M_{max_{avg}} , \tag{B.10}$$

$$D_J^B = J \times (L+1) , \tag{B.11}$$

$$D_J^C = J \times (\text{dof}_L + 1) \times M_{max_{avg}} , \tag{B.12}$$

$$D_J^D = J \times (\text{dof}_L + 1) . \tag{B.13}$$

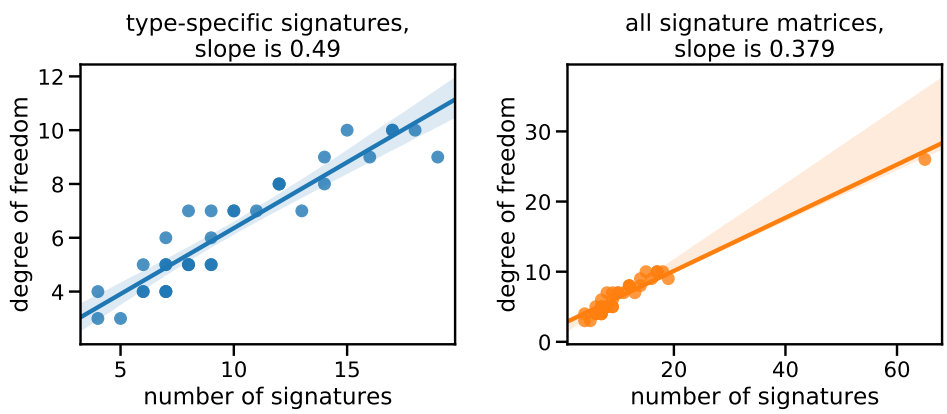

Figure B.2 – Variation of the degree of freedom of a subset of cancer type-specific signatures (35 distinct types) or for all available signatures depending on the number of signatures. The left panel shows the dependence for subsets of the 65 signatures only, and the right panel for the 65 signatures additionally. We see that the dependence with the number of signature (slope) is different in the two cases.

Moreover, if we consider the variations of the degree of freedom associated with L signatures, dof_L, as a function of L for the 35 available cancer types, and for the all 65 signatures, we note that there is a gap, as the maximal number of signatures in one cancer type is 19, and that the slope seems different for a subset or for all the signatures (see Supplementary Figure B.2). The dependency being quite different, this raises the question of whether we should estimate a single α for all situations (i.e., a unique BIC model), or whether we should fit two BIC models: one for the cases where CloneSig is run with only cancer type-specific signatures, and one for the case where CloneSig is run with all the 65 signatures.

For each possible definition of D_J (B.10)-(B.13), and for each setting (estimating a unique or two separate BIC models), we ran simulations to estimate the value of α such that $BIC_{\alpha,j}, j \in \{1, \ldots, 10\}$ is maximal for the true value of J, by solving (B.9). To evaluate the results, we split the dataset into a train (80% of data) and a test set (20%), and assess the accuracy of J estimation on the test set. To evaluate the stability of the learnt parameter α, we compute the 95% confidence interval over 10 independent train-test splits. The values for learnt coefficients, averaged over 10 independent train/test splits for each case are presented in Table B.1. The test accuracies for different criteria and different learning settings

	D_J^A	D_J^B	D_J^C	D_J^D
separate model (subset)	-0.037 ± 0.000215	-0.061 ± 0.000268	-0.056 ± 0.000336	-0.092 ± 0.000404
unique model	-0.014 ± 0.000072	-0.023 ± 0.000101	-0.034 ± 0.000173	-0.055 ± 0.000233
separate model (65 signatures)	-0.012 ± 0.000060	-0.020 ± 0.000087	-0.030 ± 0.000146	$-0.0490.000214\pm$

Table B.1 – Values for the coefficients α for different penalty shapes and training subset. We see that the coefficients for the whole dataset and for the 65 signatures examples are close. Overall, the confidence interval for the coefficients are small.

are presented in Figure B.3. We first see that, as mentioned earlier, standard model selection

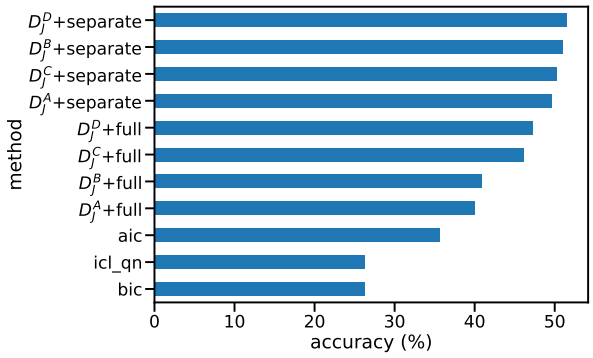

Figure B.3 – Test accuracy of various model selection criteria. BIC, AIC and ICL are standard model selection. The others are attempts to learn a valid criterion on simulated data.

criteria (BIC, AIC, ICL) perform overall poorly. Second, we notice that the "separate" strategy is usually slightly better than the "full" strategy, i.e., learning a single α for CloneSig with all 65 signatures or only a subset is not as good as learning two different α's. As for the definition of D_J, we see in both cases that using the degree of freedom of the signature matrix is better than counting the number of columns, and that taking into account the variations in copy numbers through $M_{max_{avg}}$ does not bring any benefit. A complete overview of the number of clones found over the test set for each penalization strategy is given in Figure B.4. In conclusion, we use in all our experiments an adaptive BIC criterion based on D_j^D as a measure of degree of freedom, and α estimated separately when CloneSig is fitted with 65 signatures or with a cancer-specific subset.

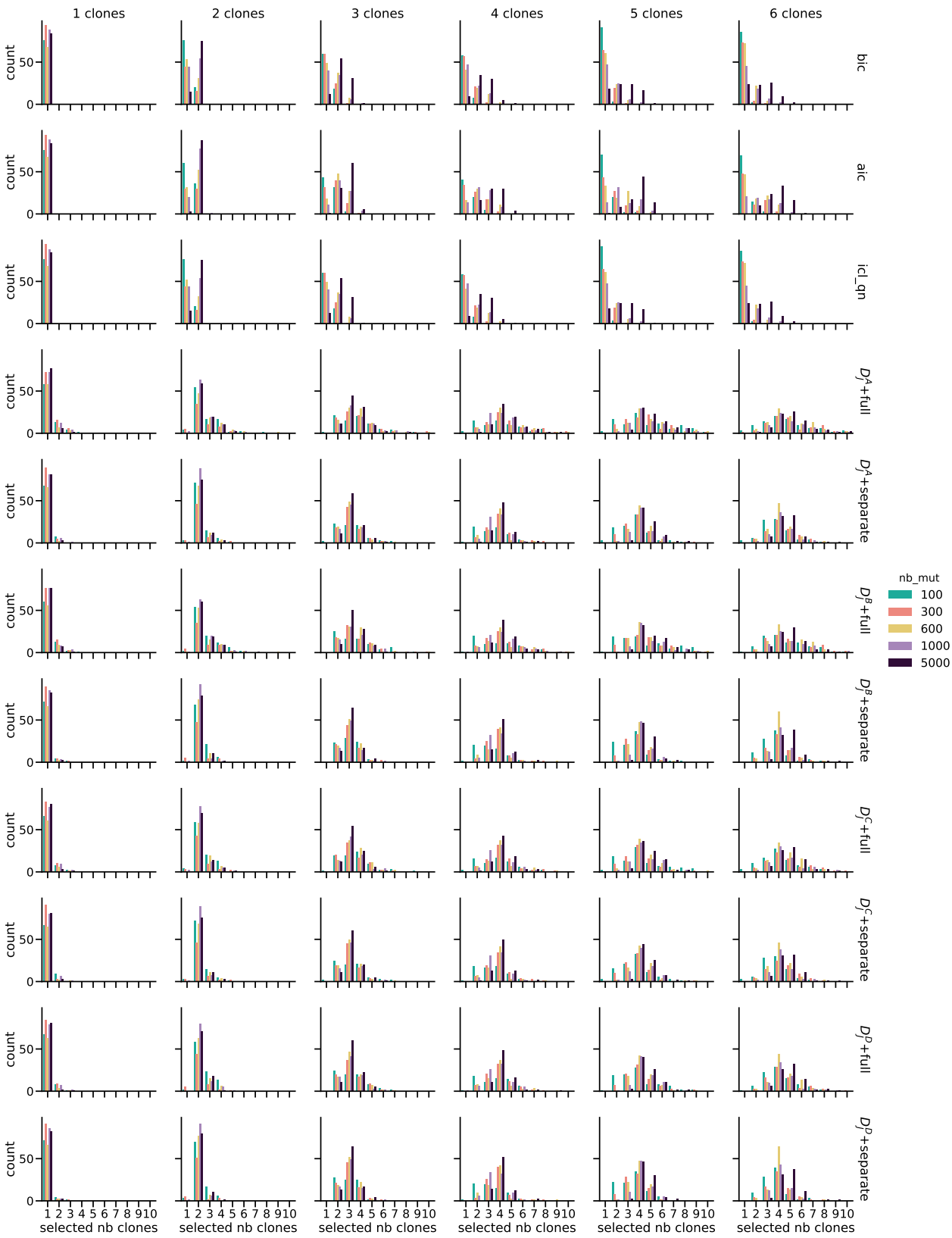

Figure B.4 – Number of clones found with different model selection criteria on the test set (not used to fit the model selection criteria). This illustrates the improved accuracy of the adapted BIC criterion compared to classical criteria

B.1.3 Statistical test for signature change

To assess whether a signature change between clones is statistically significant, we design and calibrate a statistical test. To that end, we compare the likelihood of a CloneSig model with J clones as determined by the model selection criterion, and the likelihood of a model with the same clones but a single mixture of signatures common to all the clones (and found by fitting all observed mutations together). The objective of the test is to determine whether the difference between the two likelihoods is significant. To that end, we implement a likelihood-ratio test based on the statistics:

$$\lambda = \frac{\ell_{sigCst}}{\ell_{sigChange}} .$$

Following Neyman-Pearson lemma [Neyman and Pearson, 1933] one can set a threshold c to reject the null hypothesis that there is no signature change if λ is lower or equal to c with a certain level of significance α determined by the distributions of the likelihood of the model. As this distribution is unknown, we apply the Wilks theorem stating that asymptotically, $-2\log(\lambda)$ follows a chi-squared distribution of parameter the difference in dimensionality between the two alternative models [Wilks, 1938].

As previously illustrated for the model selection criterion, the number of parameters is different from the degree of freedom in the case of CloneSig, so we resort to simulations to fit the degree of freedom of the test. We simulate a dataset with a similar mixture of signatures for all clones of each sample, and focused on samples with at least 2 clones, as described in Material and Methods. For the purpose of calibration, we use the true number of clones to fit the two alternative models. The objective of this approach is to fit a chi-squared distribution on the empirical distribution of $-2\log(\lambda)$ obtained in simulations. This is achieved again in two settings: fitting with all 65 signatures or with a cancer type-specific subset of signatures. In both cases, the distribution for each number of clones J evokes indeed a chi-squared distribution (Figure B.5)

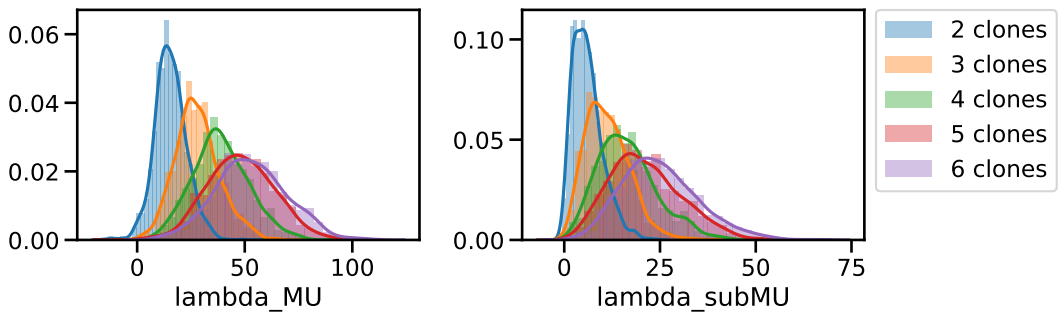

Figure B.5 – Empirical distribution of $-2\log(\lambda)$, with $\lambda = \frac{\ell_{sigCst}}{\ell_{sigChange}}$ obtained by fitting CloneSig with the true number of clones on simulated data, either with all 65 signatures (left), or with a subset of cancer type-specific signatures. The distribution is estimated separately for each number of clones.

To fit the degree of freedom to use in the implementation of the test, as the degree of freedom of a chi-squared-distributed variable is its mean, we train a linear ridge regression model to fit $-2\log(\lambda)$ to relevant covariates. Four covariates were initially considered: the number of clones, the degree of freedom of the input signature matrix, the number of mutations, and the diploid proportion of the genome. We found that the last two variables have no visible correlation with the target variable (see Supplementary Figure B.6). Additionally, when added to the model, with standard scaling of input variables, they have coefficients more than ten times smaller than the ones of the number of clones, and the signature degree of freedom. We therefore compute the final model on the two retained (unscaled) variables, and we average the values of the coefficients over 10-fold cross-validation. The resulting coefficients are reported in Table B.2.

To finally ascertain the validity of the test, we now check the uniform distribution of the p-values for negative samples in Figure B.7. There is a slight deviation from the uniform distribution, probably due to the fact that CloneSig does not necessarily converge to the true

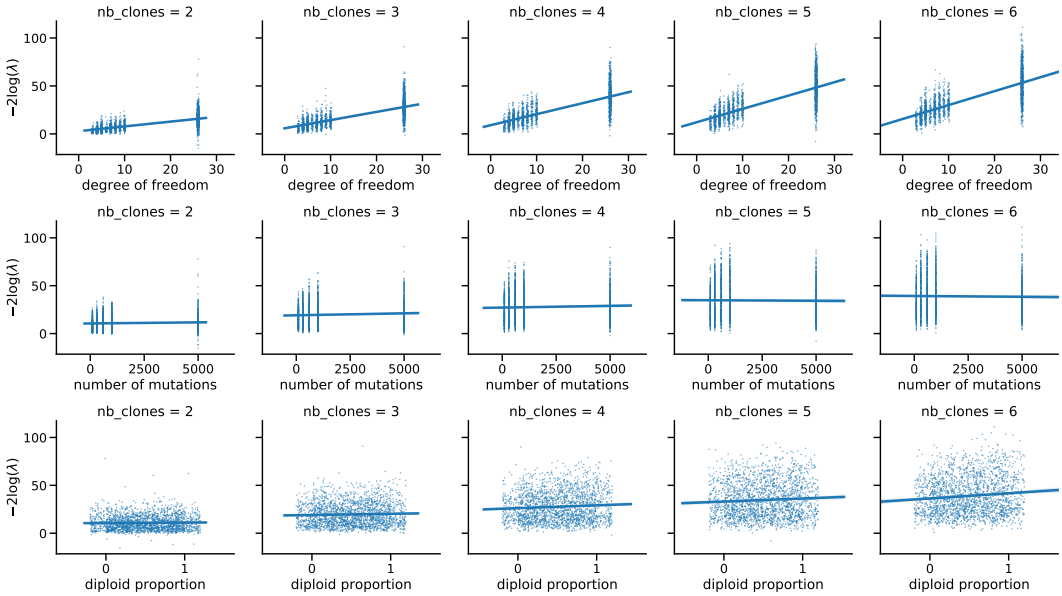

Figure B.6 – Correlation of $-2\log(\lambda)$, with $\lambda = \frac{\ell_{sigCst}}{\ell_{sigChange}}$ with potentially relevant covariates.

	Intercept	Number of Clones coefficient	Degree of freedom coefficient
separate model (subset)	-13.677 ± 0.0778	4.777 ± 0.0117	1.662 ± 0.00991
unique model	-19.420 ± 0.0589	7.124 ± 0.0169	1.069 ± 0.00210
separate model (65 signatures)	-1.156 ± 0.107	9.470 ± 0.0279	0 ± 0

Table B.2 – Values for the coefficient α for different penalty shapes and training subset. We see that the coefficients for the whole dataset and for the 65 signatures examples are close. Overall, the confidence interval for the coefficients are small.

model likelihood (and instead to a local maxima), and thus does not respect the conditions of application of Wilks theorem.

We finally explore the sensitivity of the test on the maximum cosine distance between signatures. The dataset used for that purpose consists of 2,700 samples with the number of clones varying between 2 and 6. For each number of clones, we drew 30 distinct π matrices with distinct maximal cosine distances between the mutation type profiles. For each number of clones and π matrix, we generated a sample with varying number of observed mutations, diploid percent of the genome, and sequencing depth. Figure B.8 illustrates the proportion of samples where the test p-value is below 0.05 depending on the maximal distance between two subclones. We observe that detection is more efficient as the distance between clones becomes larger. Dependence on other variables is explored in Supplementary Figure B.9.

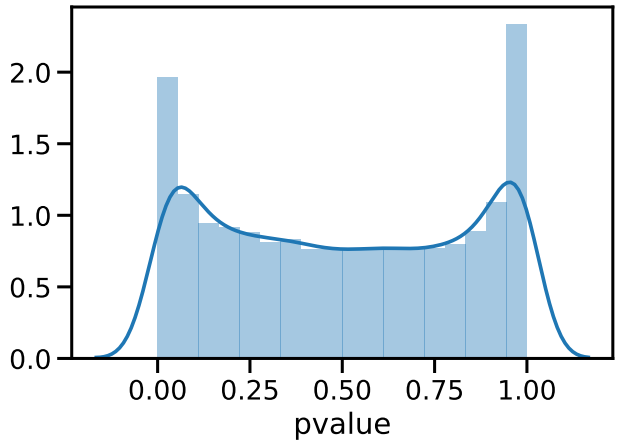

Figure B.7 – Empirical distribution of the p-values of the calibrated test of significance of signature change for negative simulated samples.

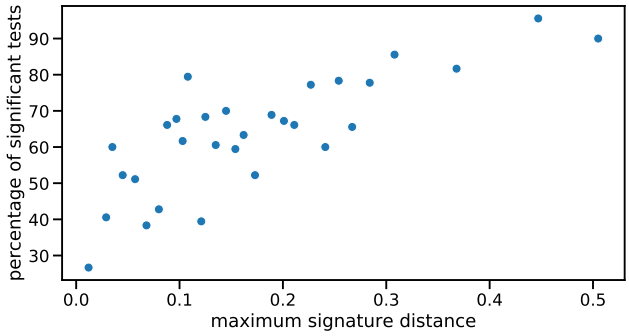

Figure B.8 – Percentage of significant tests depending on the max distance between 2 clones, quantized in 30 bins.

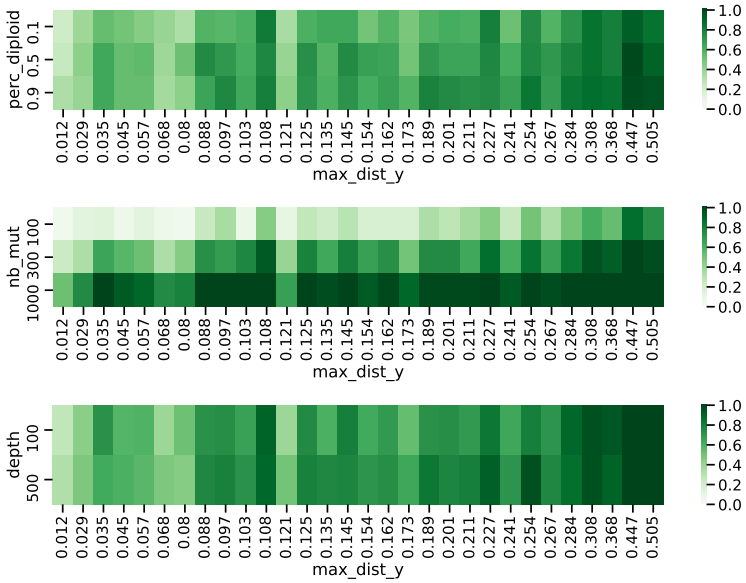

Figure B.9 – Percentage of significant tests depending several variables: number of mutations, and percentage of diploid genome, and sequencing depth

B.1.4 Several "modes" to run CloneSig

A crucial difficulty in performing mutational signature deconvolution is the identifiability of the problem. Indeed, several mixtures of signatures may provide satisfying results. The most common approach to address this issue is to reduce the number of candidate signatures, in particular by using only signatures known to be active in the cancer type of the considered tumor sample [Alexandrov et al., 2018] (approach `cancer_type`). An alternative approach is to perform two successive fits, the first one on all mutations in the sample in order to select potentially active signatures by keeping those with a contribution greater than a threshold, and the second one to refit those selected signatures with varying number of clones. This avoids the situation where a lot of signatures have very small contributions to the final mixture [Rubanova et al., 2018] (approach `prefit`). Those two alternatives are implemented in CloneSig (see Figure B.10) and also tested for all methods tested (see supplementary Figures B.11-B.24). For the subclonal reconstruction problem, we see that the two approaches that limit the number of signatures have similar performance and improve the accuracy of CloneSig, especially in cases with few mutations. However, for the signature deconvolution problem, even though the `prefit` approach exhibits improved performance compared to taking all signatures, the `cancer_type` approach shows significantly better results. The results were similar for the other signature deconvolution methods, so for the rest of the analysis, we retain the `cancer_type` approach, and report only one result per method, to simplify the interpretation.

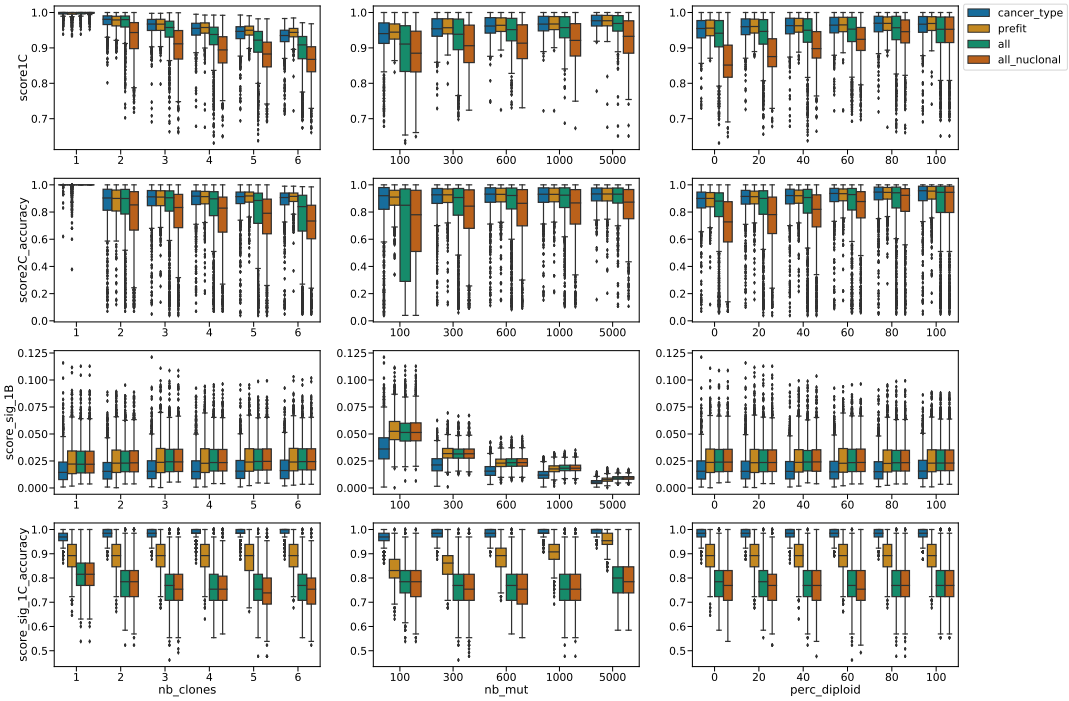

Figure B.10 – CloneSig's performance for 3 different input signature strategies: use all available signature (all), a subset of cancer type-specific signatures (cancer_type), or proceed in two steps by first fitting all mutations together to select potential signatures, and then actally run CloneSig with the selected subset (prefit). Additionally, the contribution of CloneSig'a approach for accounting for copy number was evaluated, by implementing the simpler approach from Palimpsest [Letouzé et al., 2017] (all_nuclonal).

B.2 Full benchmarking results

To fully assess CloneSig's performance in simulations, in comparison with other state-of-the-art approaches for subclonal reconstruction and signature deconvolution, we report here the full results with all tested "modes" (all signatures, a subset of cancer-type-specific signatures, or a pre-fit step where only the most prominent signatures found on the whole set of mutations are then retained for the true signature deconvolution for CloneSig, TrackSig and Palimpsest). In this extensive version of the results, we report all metrics used to create score2C (AUC, specificity, sensitivity), score_sig_1C and score_sig_1E (`max_diff_distrib_mut`, `median_diff_distrib_mut`, `perc_dist_5` and `perc_dist_10`).

Regarding the subclonal reconstruction problem, for all metrics, there is little difference between the different modes of each signature-aware method, except for `score2C_sensitivity` for CloneSig, where the use of the cancer-type-specific subset exhibits better results. For signature deconvolution, there is a higher variability of results with respect to the run mode. CloneSig is the best performing method, except for one metric: `max_diff_distrib_mut`. For `Score_sig_1C`, the mode cancer-type-specific subset for CloneSig achieves a very good specificity, but the other modes have a high proportion of false positive signatures.

Additionally, we conduct a similar benchmark in the case where there is no signature change between subclones, and present results in Supplementary Figures B.11 to B.25, panels b, c, f. The improvement of CloneSig over other methods in subclonal reconstruction is partially lost in this setting, but CloneSig remains competitive, and the best performing method for score 1B up to 3 clones. A similar trend is visible for all scores for the subclonal reconstruction problem, with slightly worse scores, and higher inter-quartile space when there is no signature variation between clones. For the signature deconvolution problem, most metrics are unaffected, except for score_sig_1E, where all methods perform better and close the gap with CloneSig. Overall, CloneSig performs better than other methods when there are differences of signature activities between subclones, and remains competitive with other approaches in the absence of signature change.

The runtimes of all methods for those simulations are presented in Figure B.25. The main determinant of runtime is the number of input mutations for all methods. CloneSig is slower than methods involving variational inference for the subclonal reconstruction problem, but is significantly faster than PyClone, especially for high numbers of mutations, thus illustrating its scalability to both WES and WGS data.

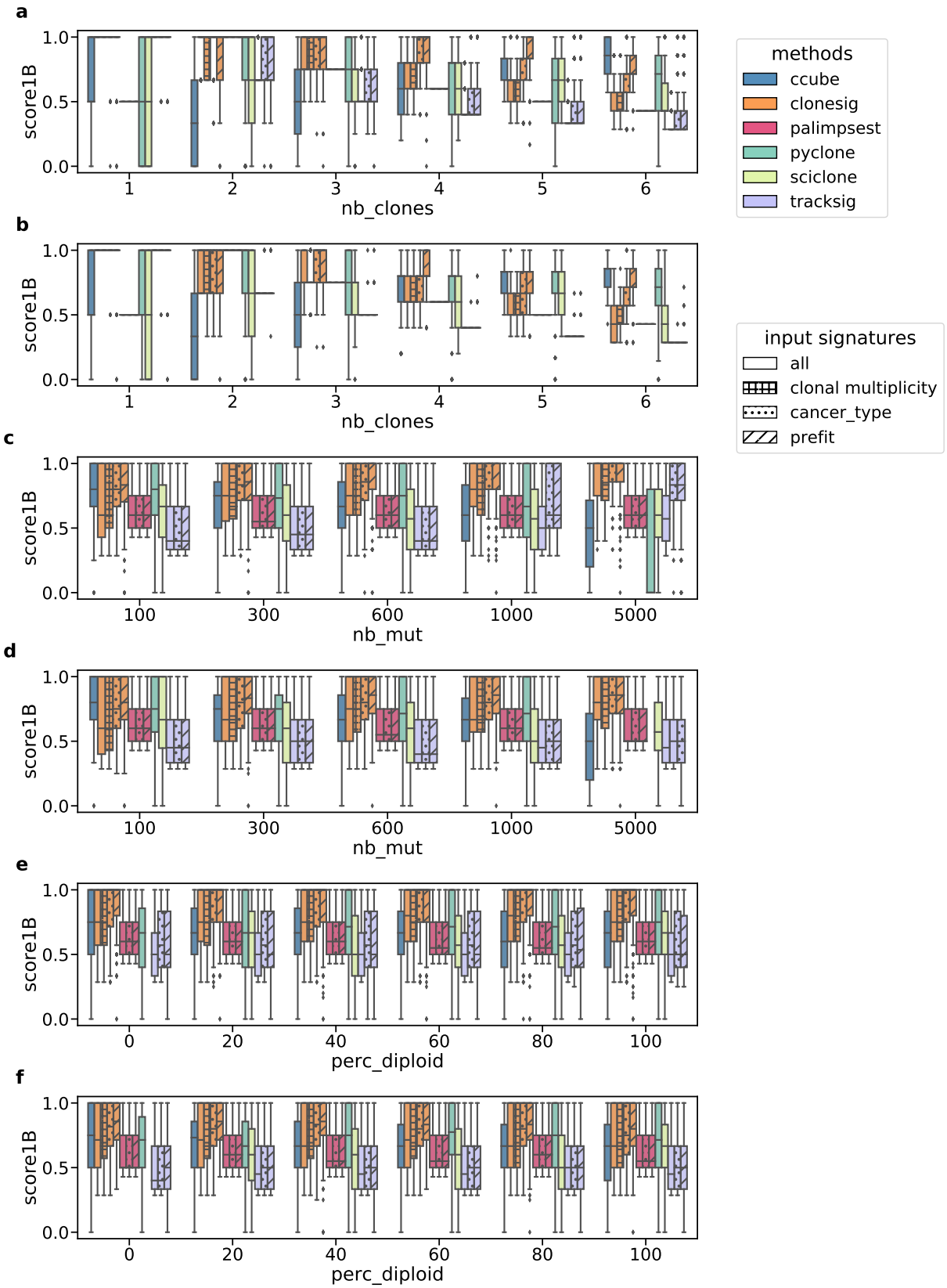

Figure B.11 – Score_1B for ITH methods on simulated data, with varying number of clones (a,b), number of observed mutations (c,d) and diploid percent of the genome (e,f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

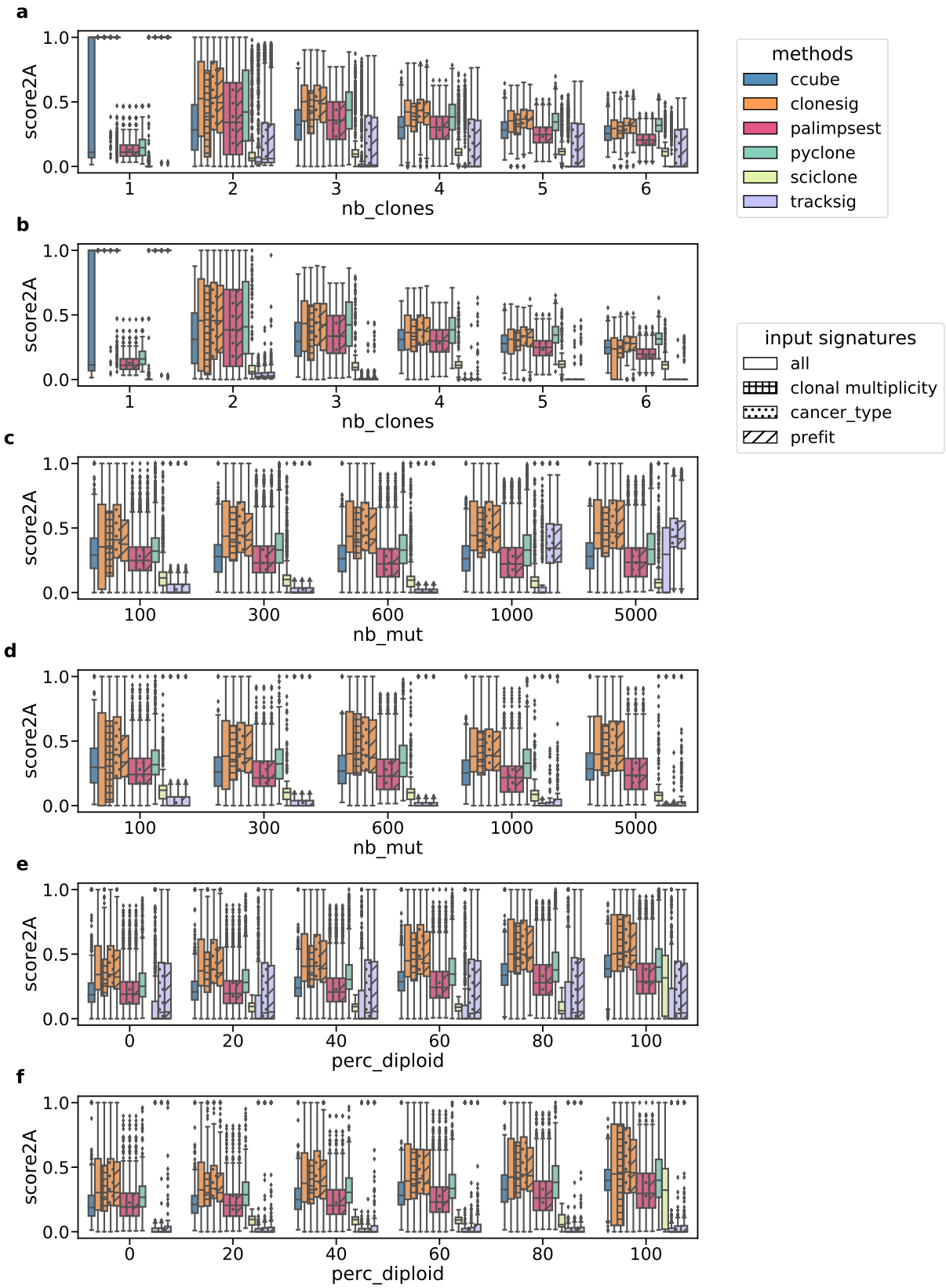

Figure B.12 – Score_2A for ITH methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

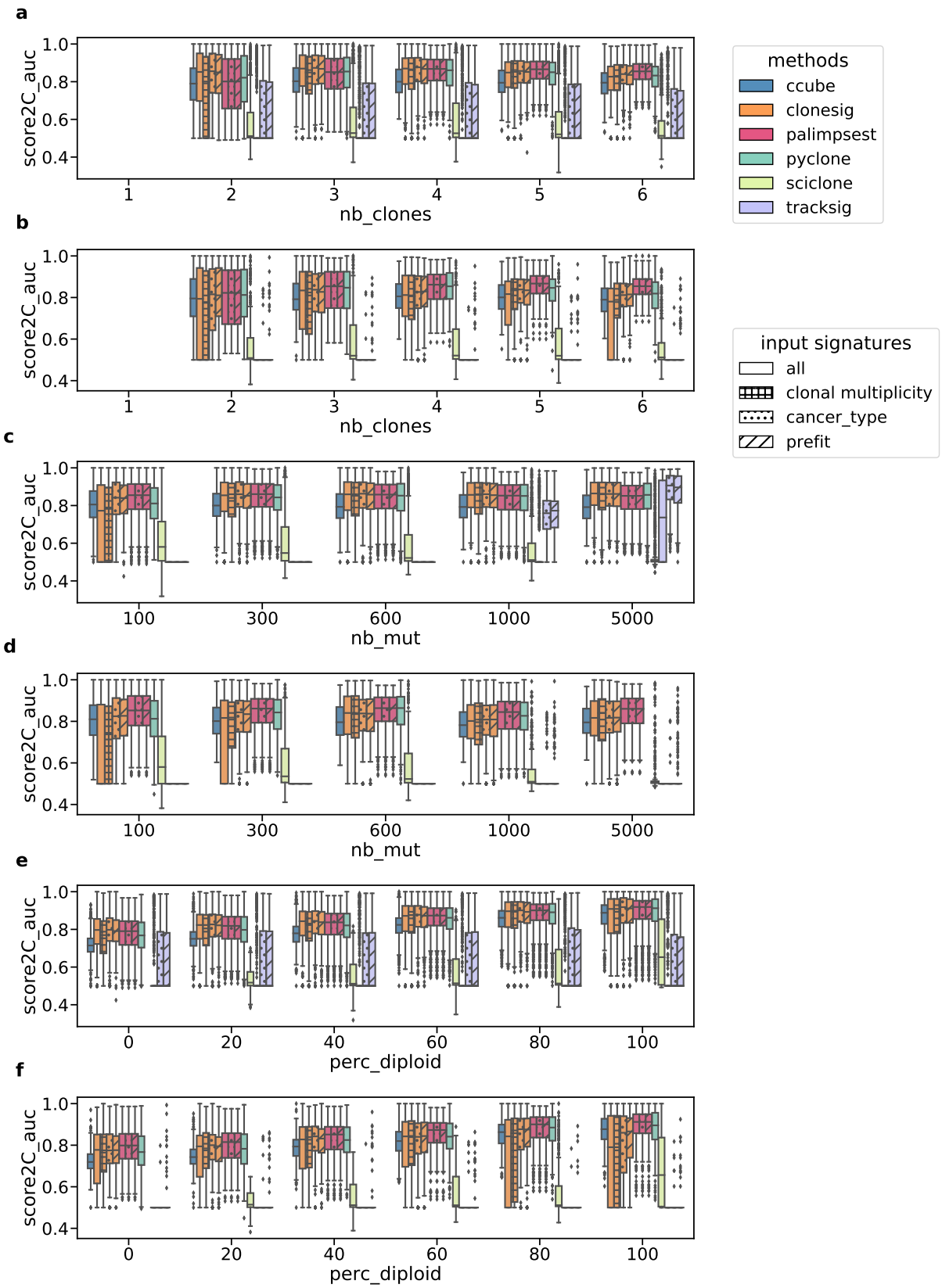

Figure B.13 – Score_2C (area under the curve) for ITH methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

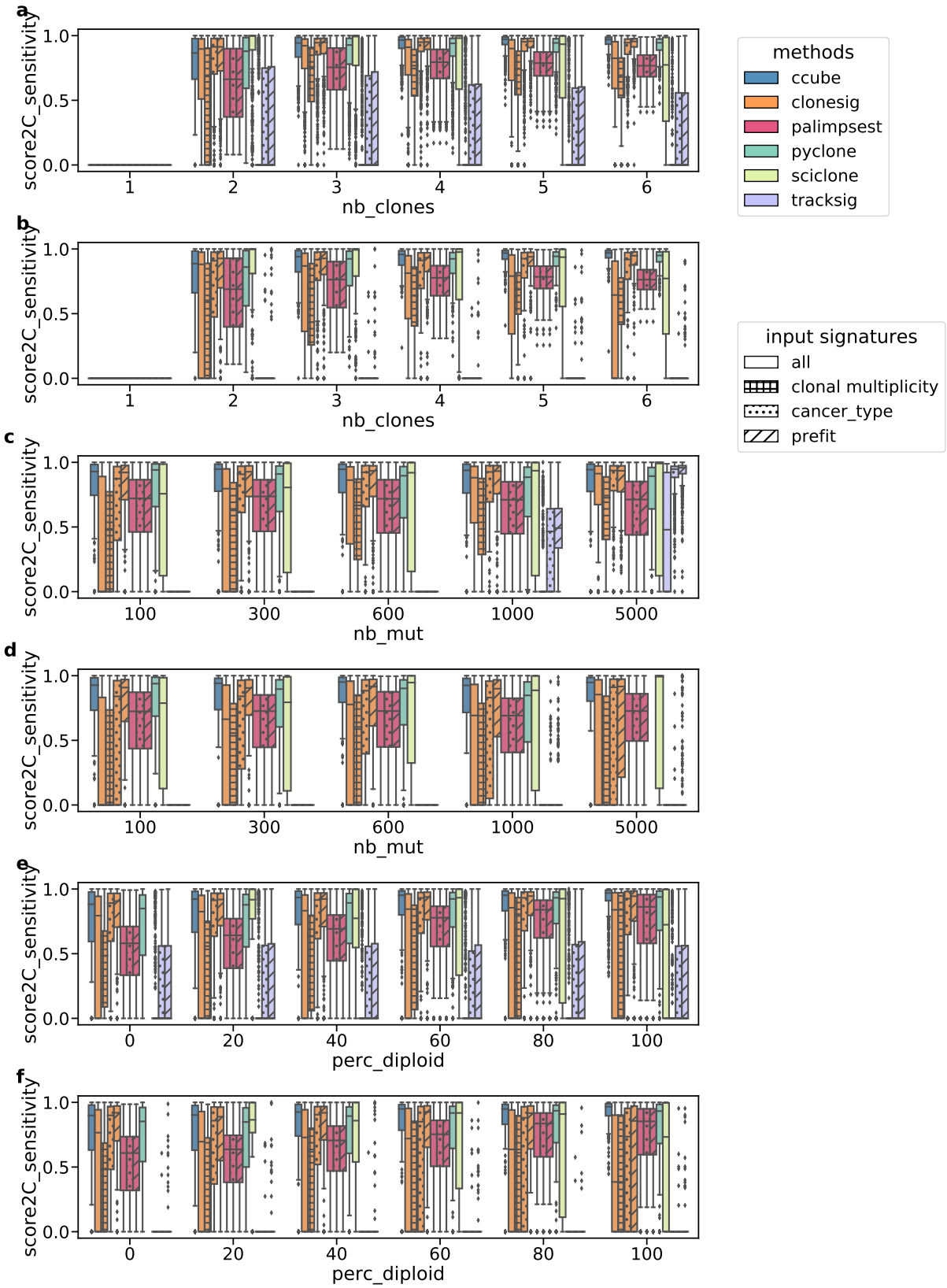

Figure B.14 – Score_2C (sensitivity) for ITH methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

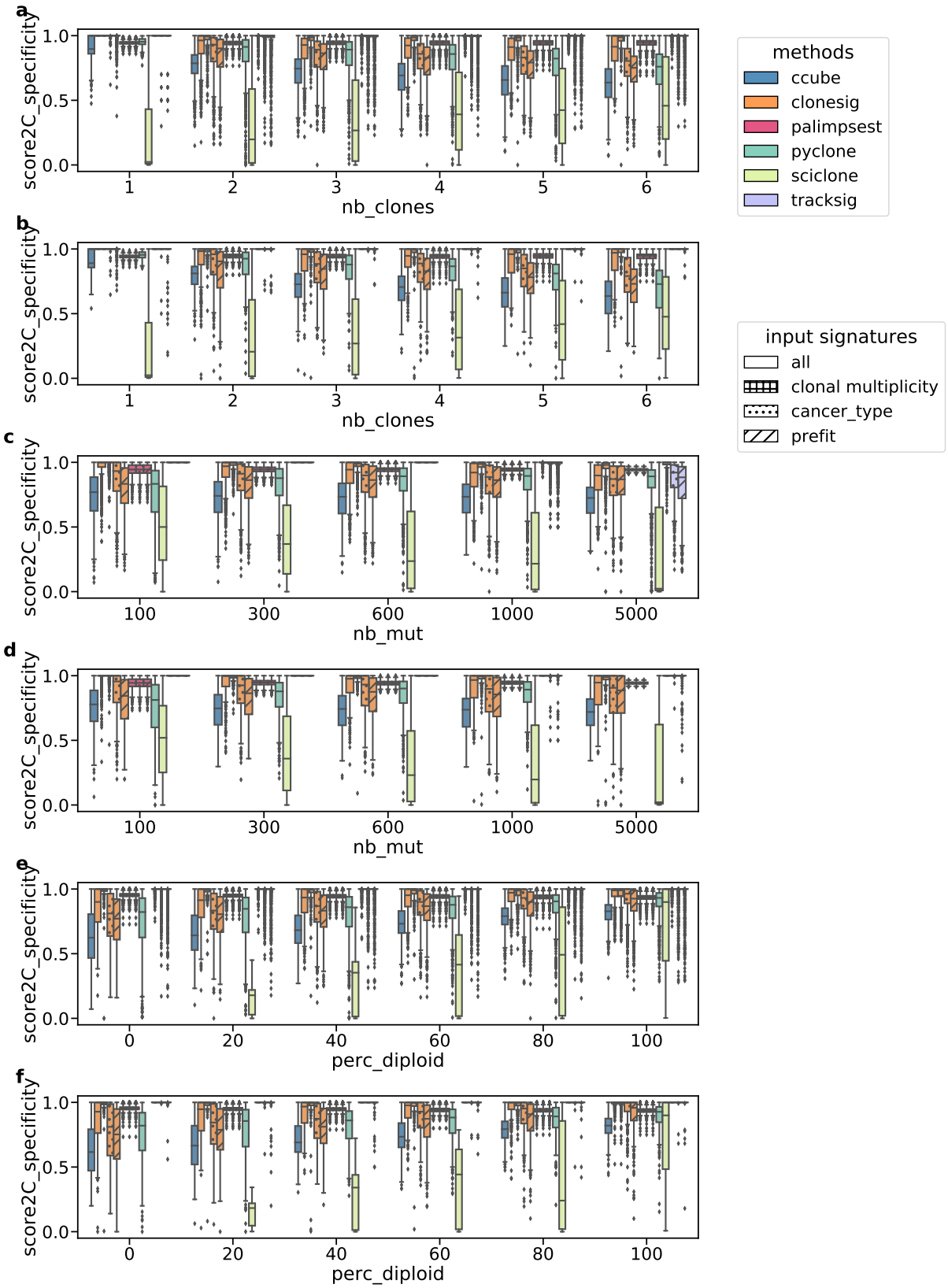

Figure B.15 – Score_2C (specificity) for ITH methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

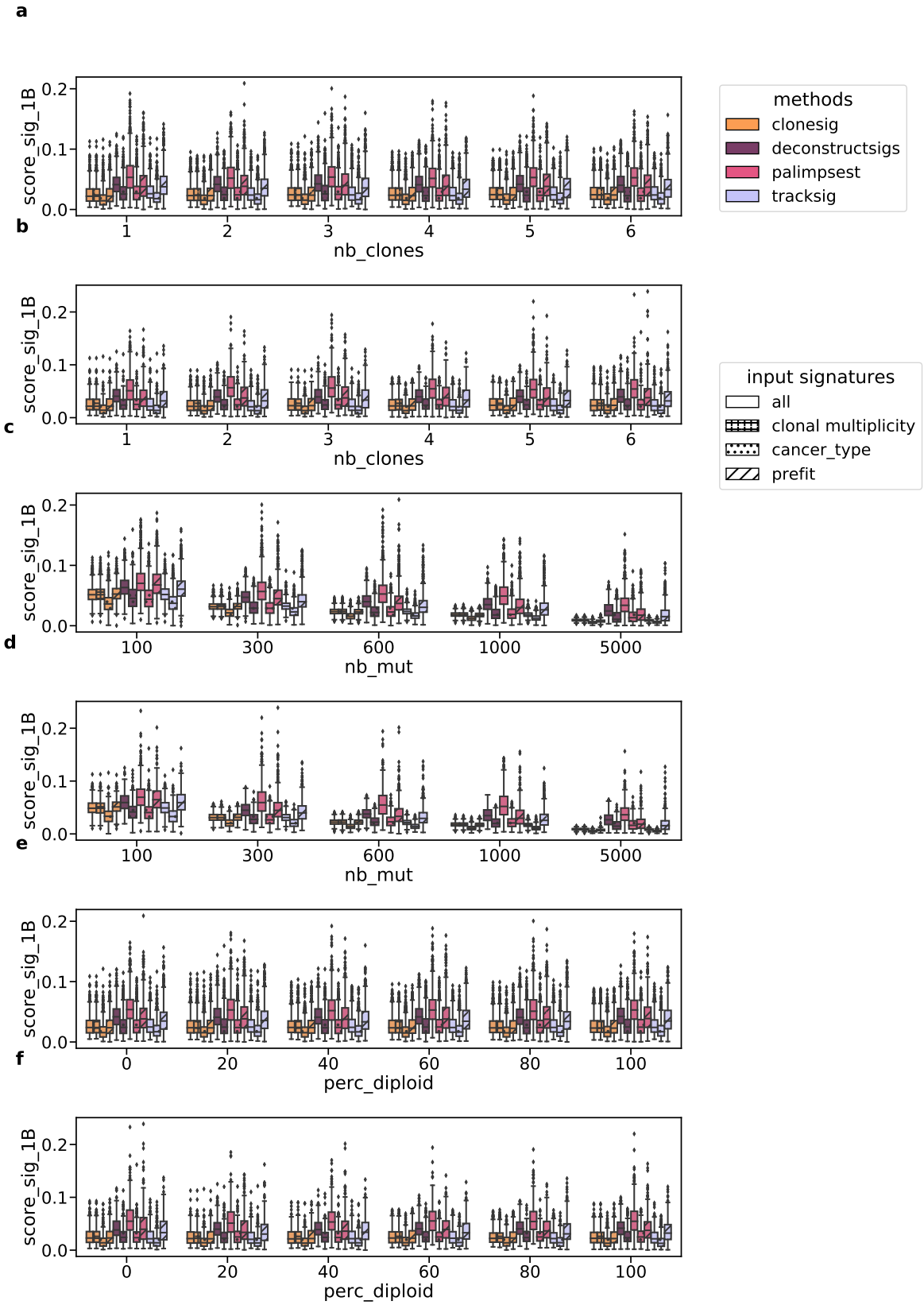

Figure B.16 – Score_sig_1B for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

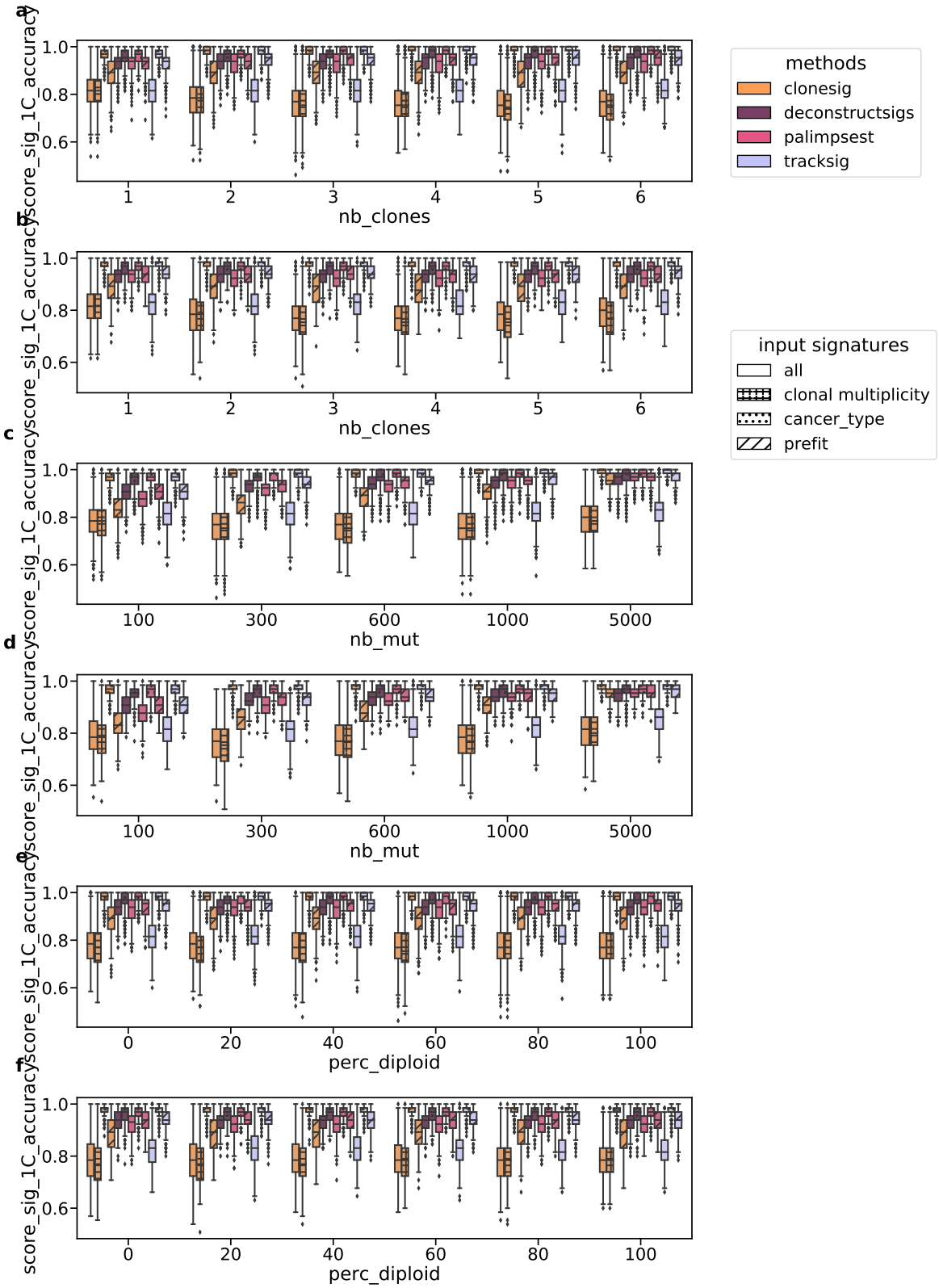

Figure B.17 – Score_sig_1C (accuracy) for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

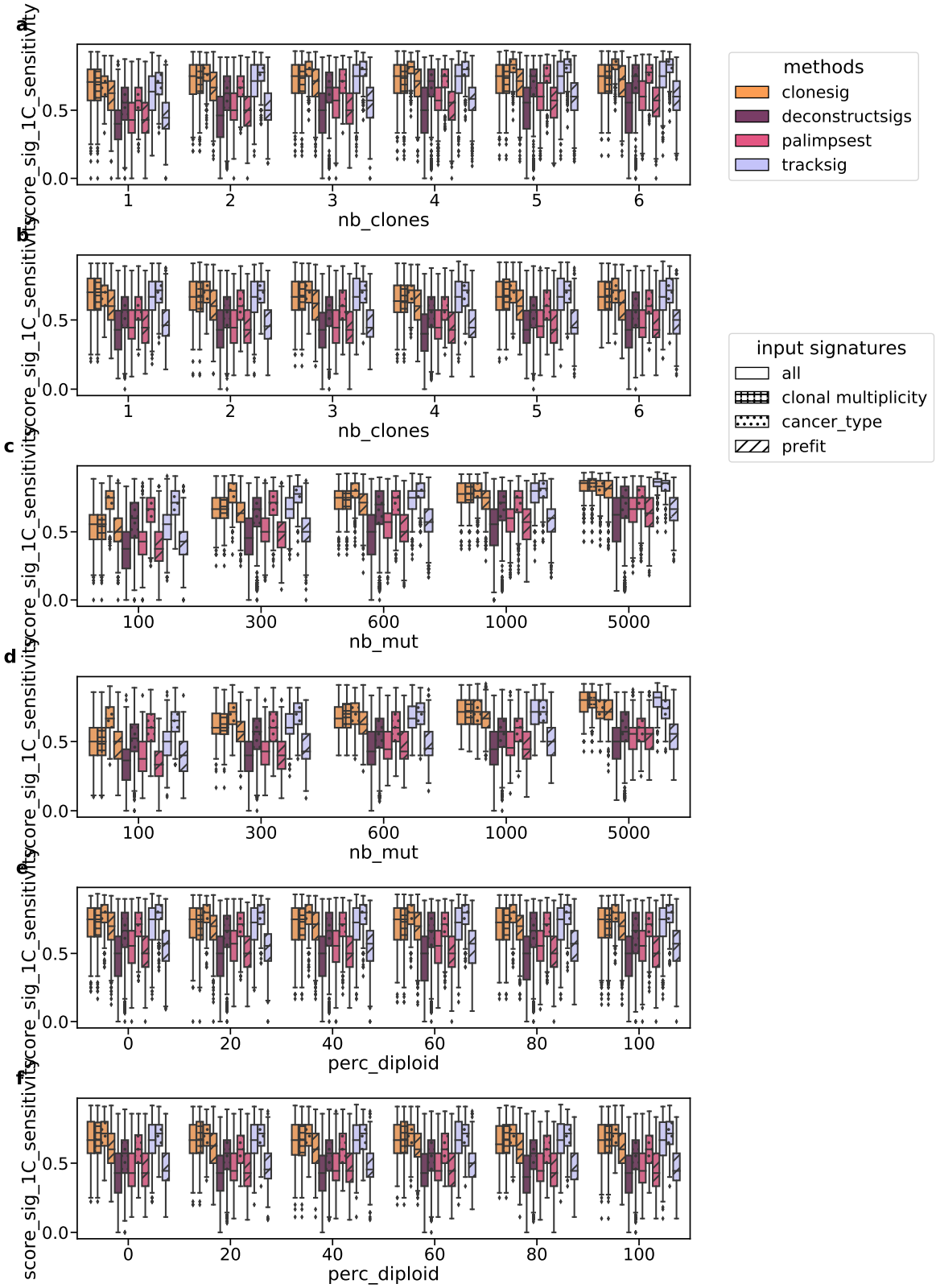

Figure B.18 – Score_sig_1C (sensitivity) for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

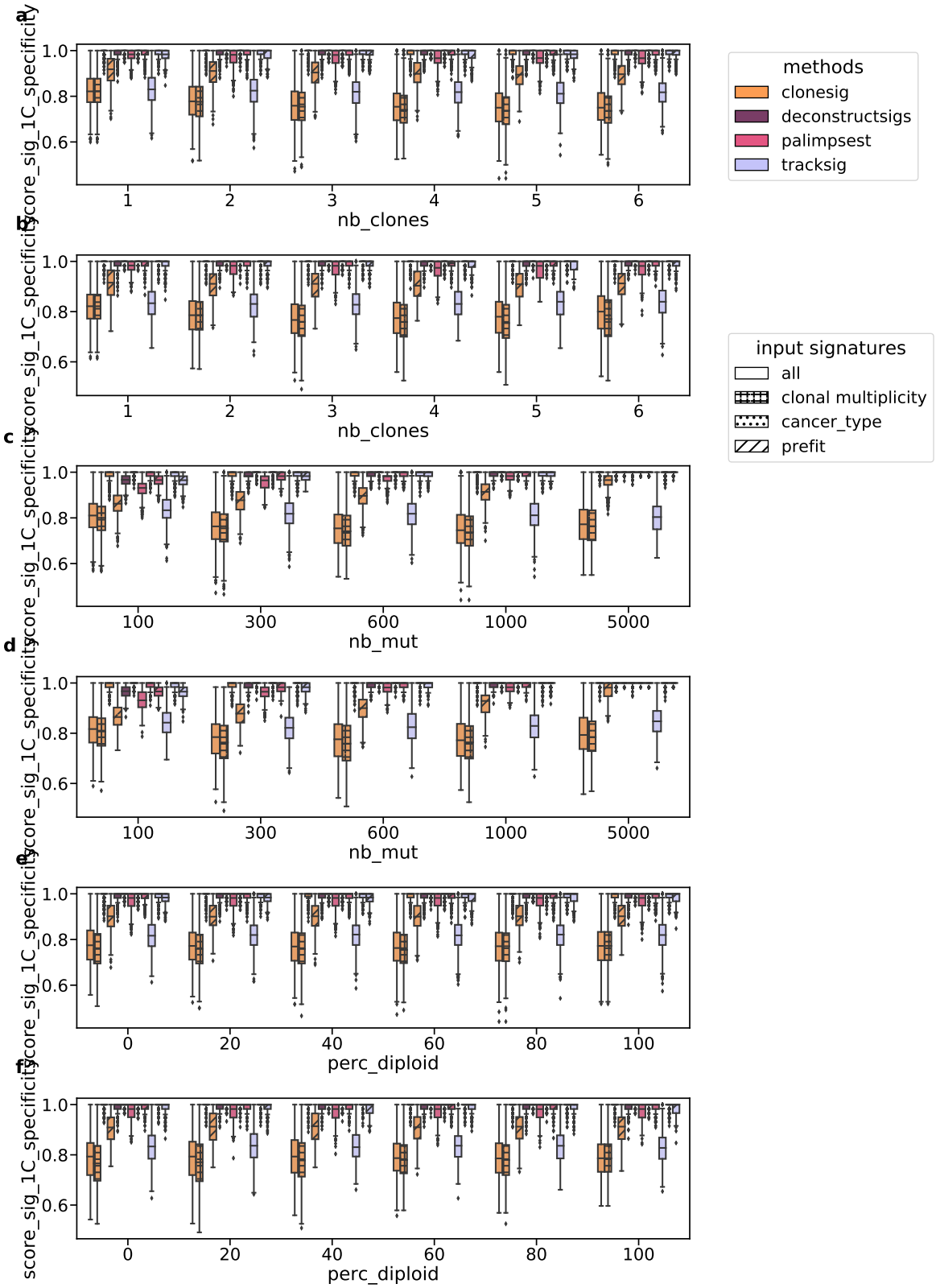

Figure B.19 – Score_sig_1C (specificity) for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

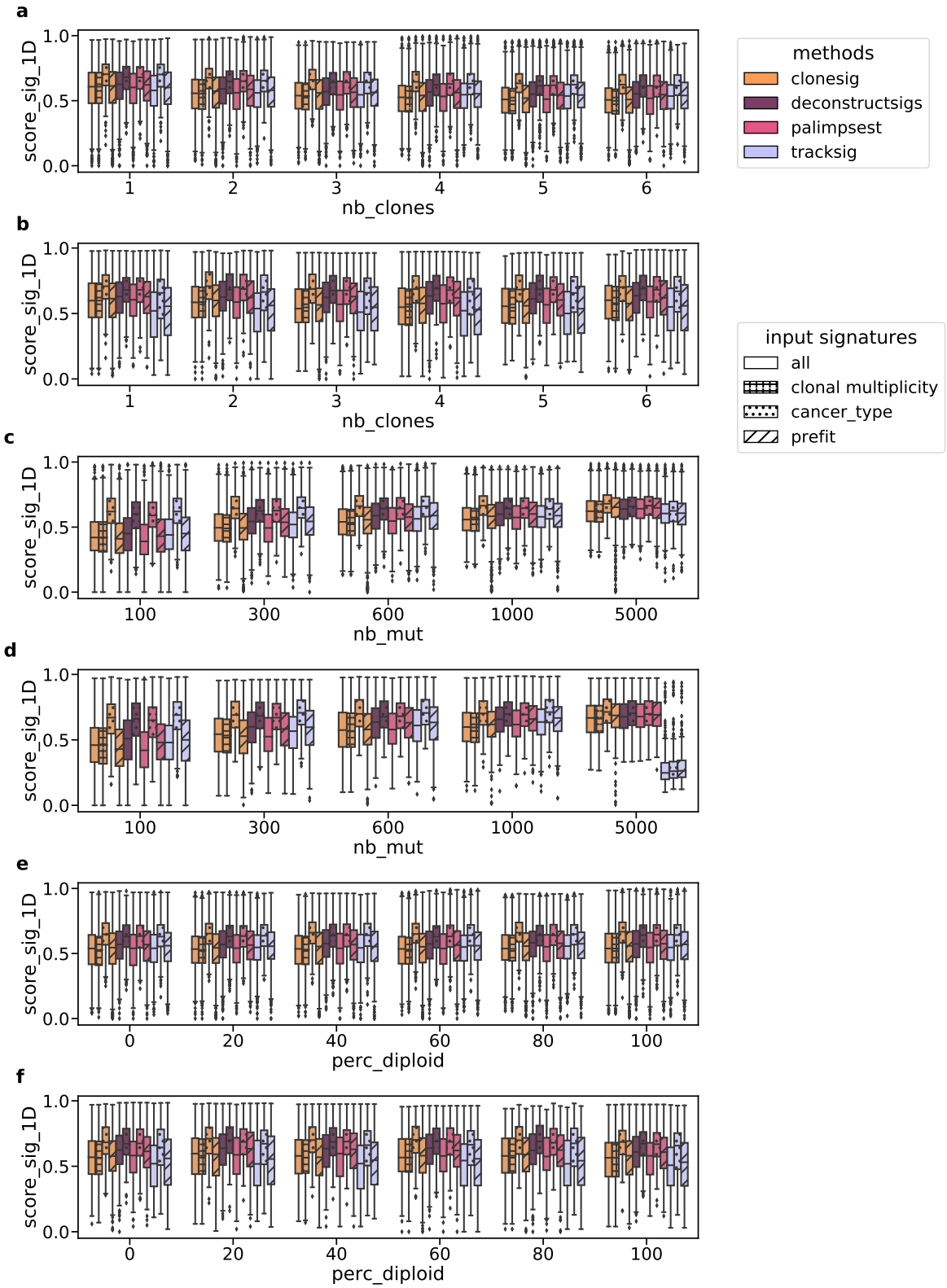

Figure B.20 – Score_sig_1D for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

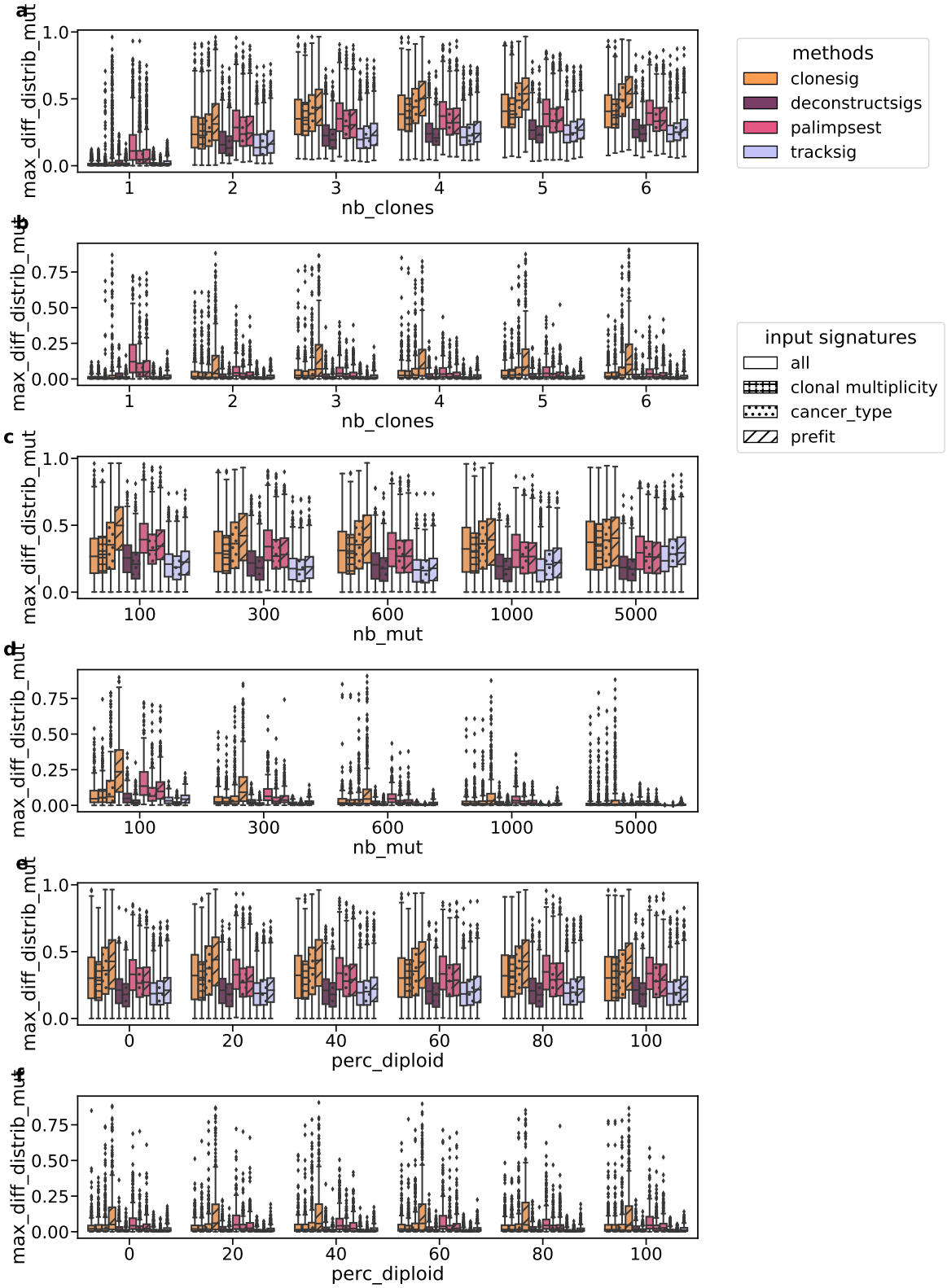

Figure B.21 – Maximal cosine distance between the true and estimated mutation type profile for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

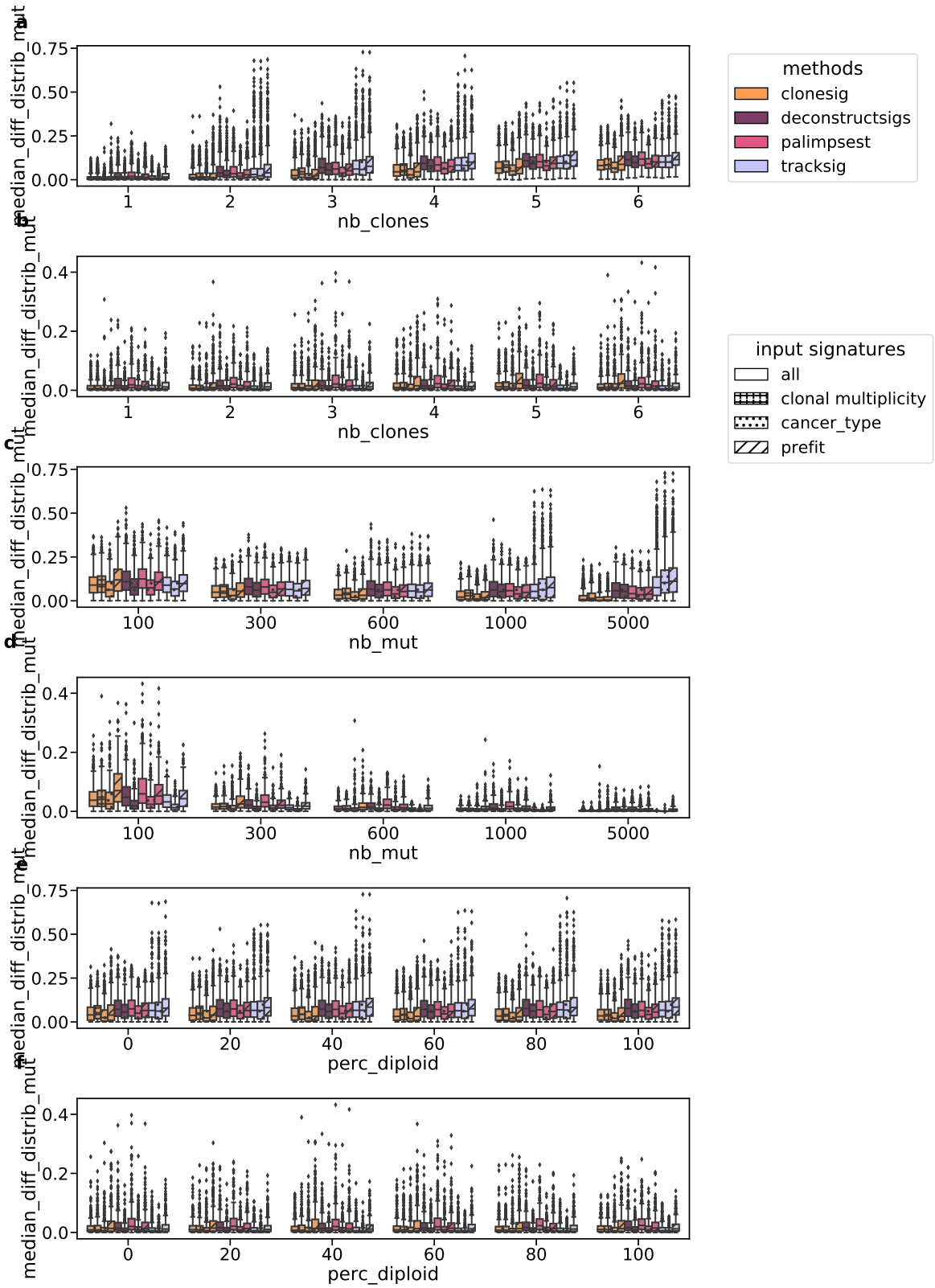

Figure B.22 – Median cosine distance between the true and estimated mutation type profile for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

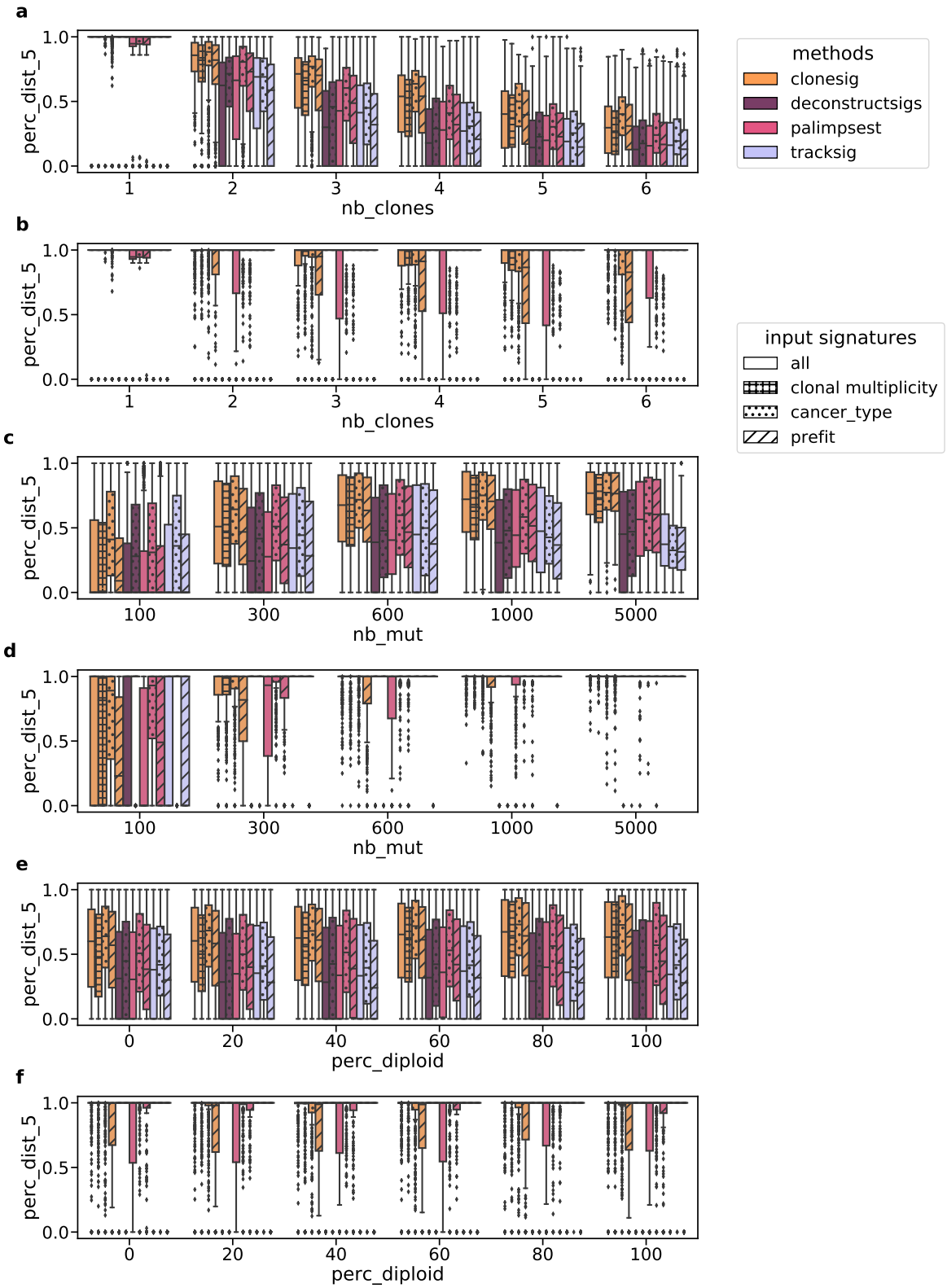

Figure B.23 – Proportion of SNVs with cosine distance between the true and estimated mutation type profile under 0.05 for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

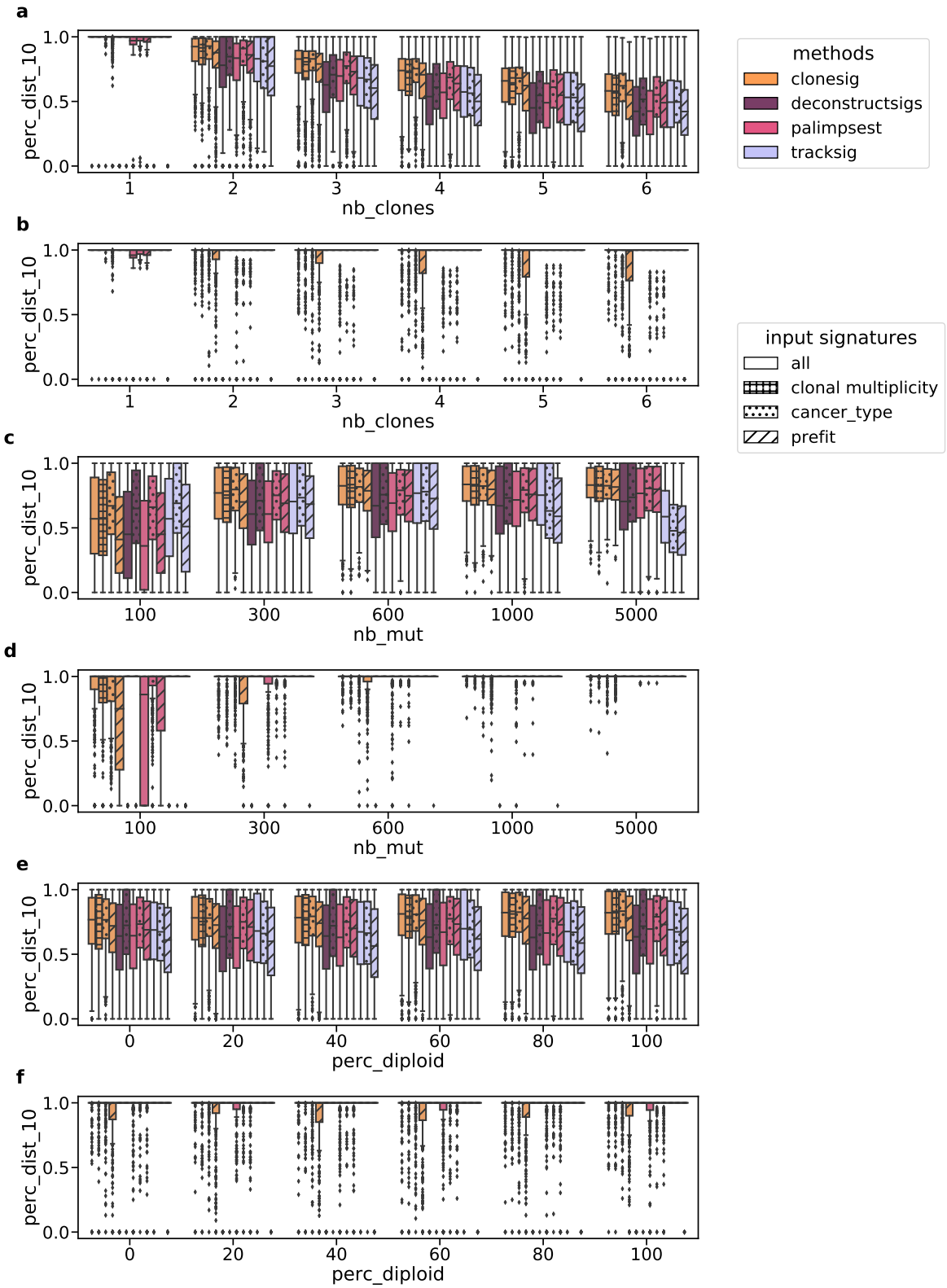

Figure B.24 – Proportion of SNVs with cosine distance between the true and estimated mutation type profile under 0.10 for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

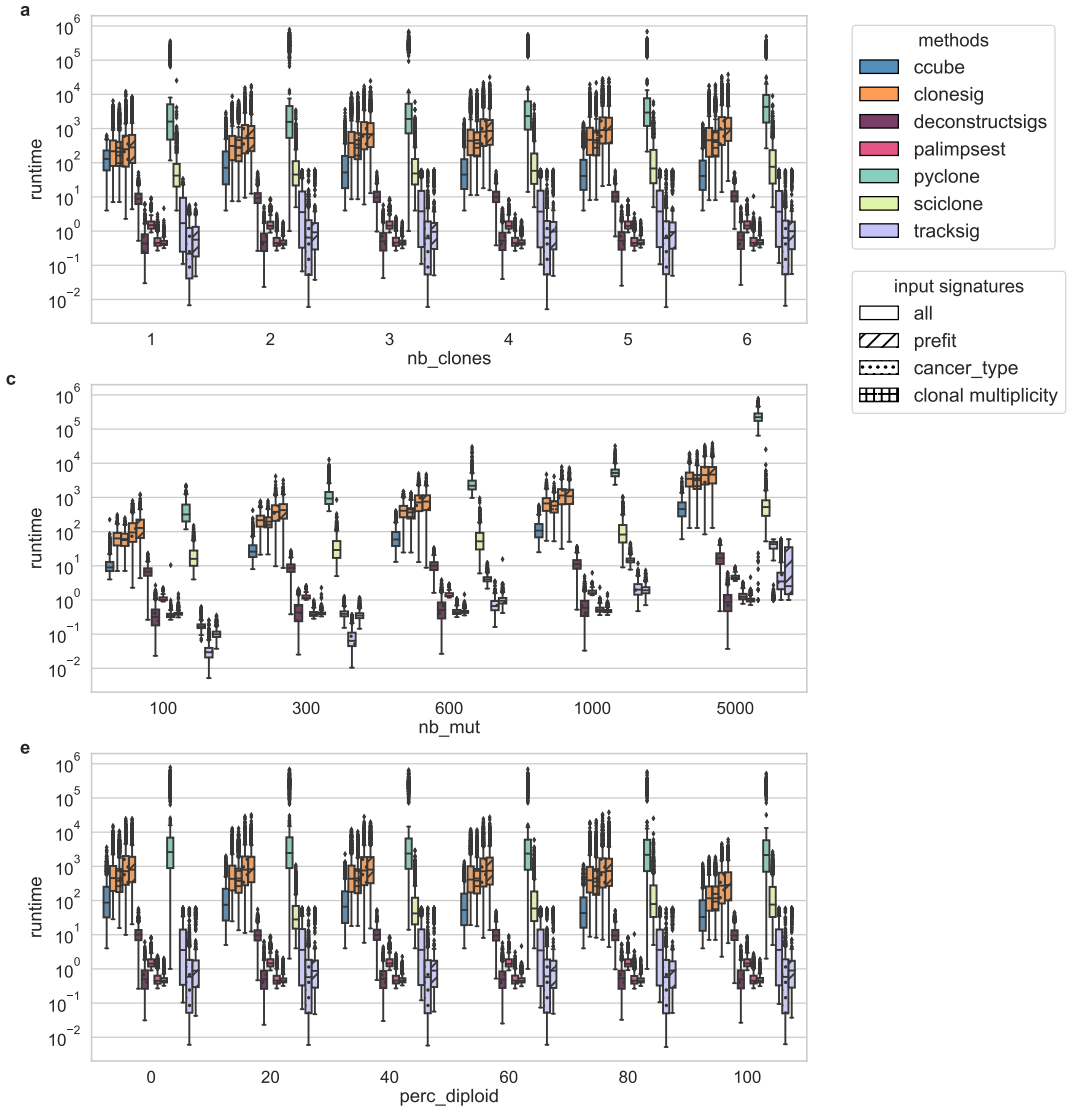

Figure B.25 – Runtime for ITH reconstruction and signature deconvolution methods on simulated data, with varying number of clones (a), number of observed mutations (b) and diploid percent of the genome (c). Results with varying signature between clones only are shown but similar results were obtained on simulations with constant signatures.

B.3 Complete overview of TCGA results

To complete the analysis of the TCGA, we present here heatmaps to delineate an overview of each cancer type in Figures B.26 to B.56. For each type, the first panel represents the difference between subclonal and clonal signature activities (in case of a significant change in activity), and the bottom panel represents the absolute values of each signature activity for clonal SNVs (belonging to the clone of largest CCF estimated by CloneSig), and in the main subclone (in terms of number of SNVs). This allows researchers to fully explore CloneSig's results on the TCGA, and further compare their results in future studies. For each panel, we have added several clinical variables, in particular, the patient's age at diagnosis, the stage of the tumor, the size class of the primary tumor, and the patient's sex. Overall, we found no trend of association between signature activities or change in activities and those clinical characteristics, as previously observed in the particular case of prostate cancer [Espiritu et al., 2018].

In most types, like CESC (Figure B.29), HNSC (Figure B.35) and others, we observe groups of patients with different patterns of signature activity. The clinical significance of such groups remains to be further explored.

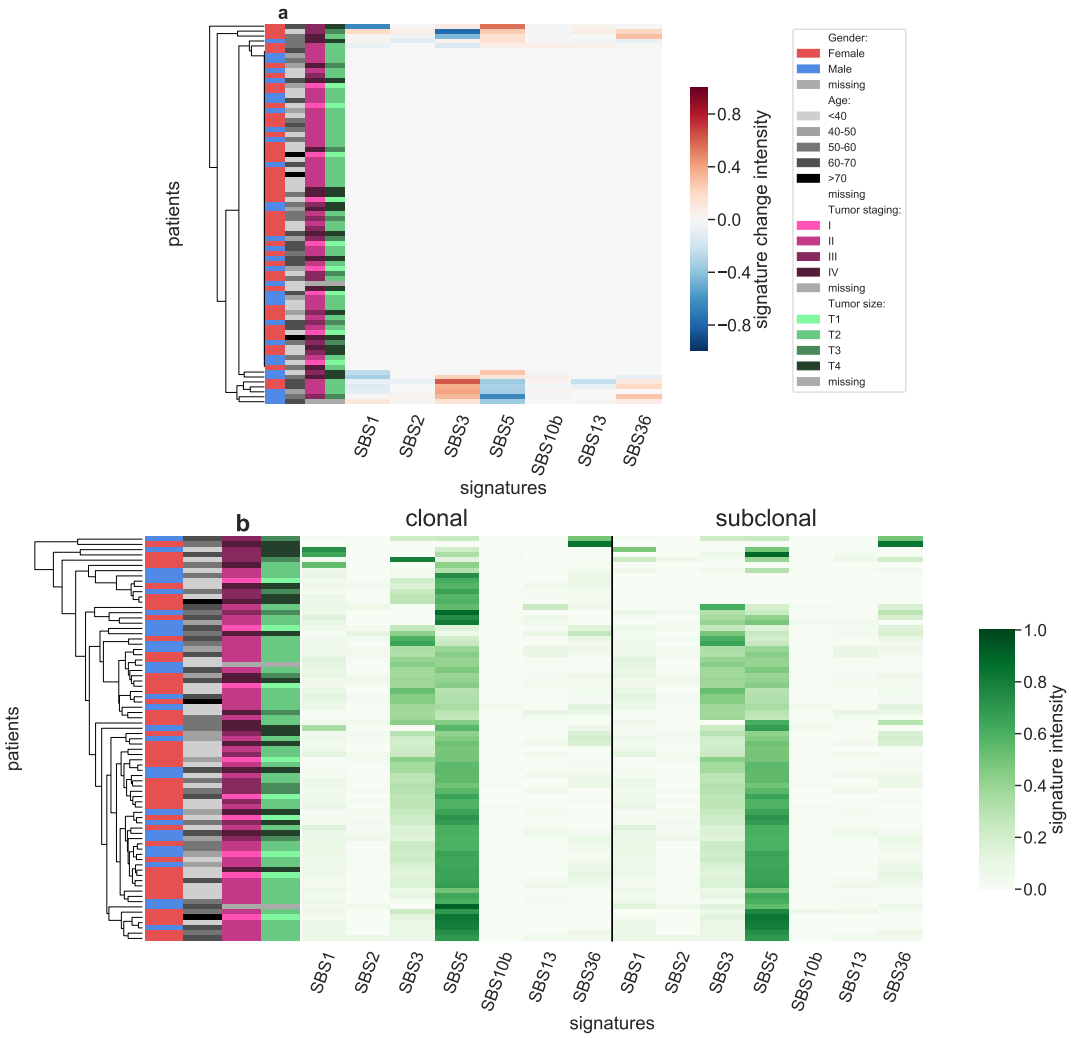

Figure B.26 – Panel a: Stratification of patients depending on their pattern of signature change for ACC patients (77 patients, including 12 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

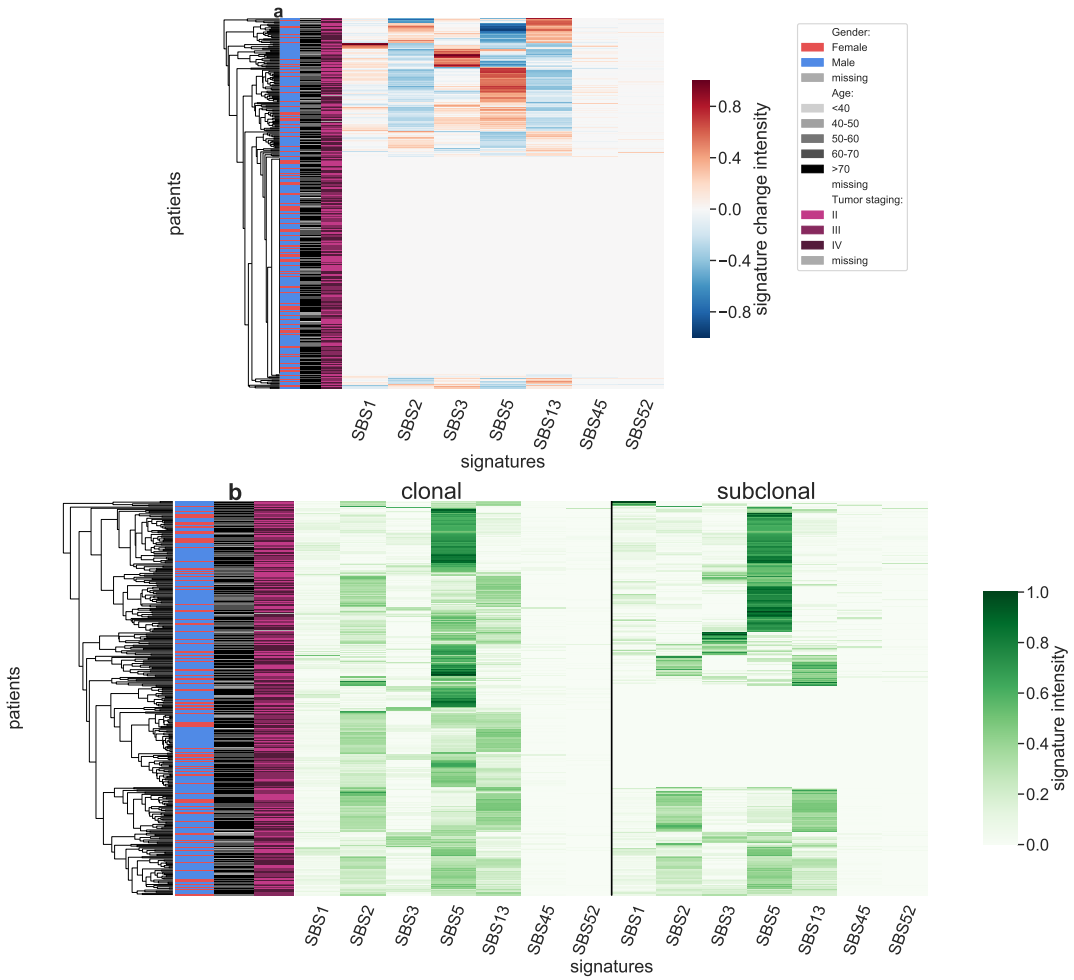

Figure B.27 – Panel a: Stratification of patients depending on their pattern of signature change for BLCA patients (354 patients, including 147 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

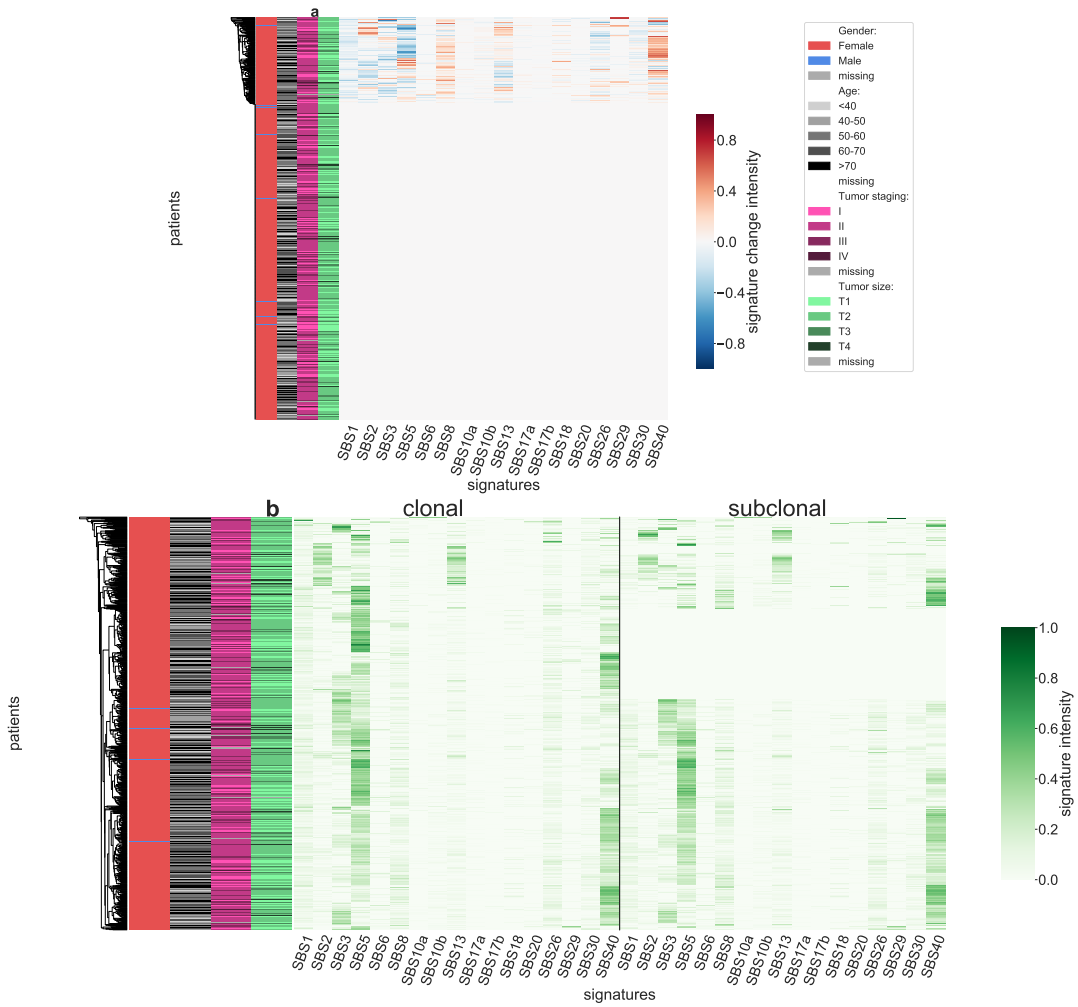

Figure B.28 – Panel a: Stratification of patients depending on their pattern of signature change for BRCA patients (931 patients, including 200 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

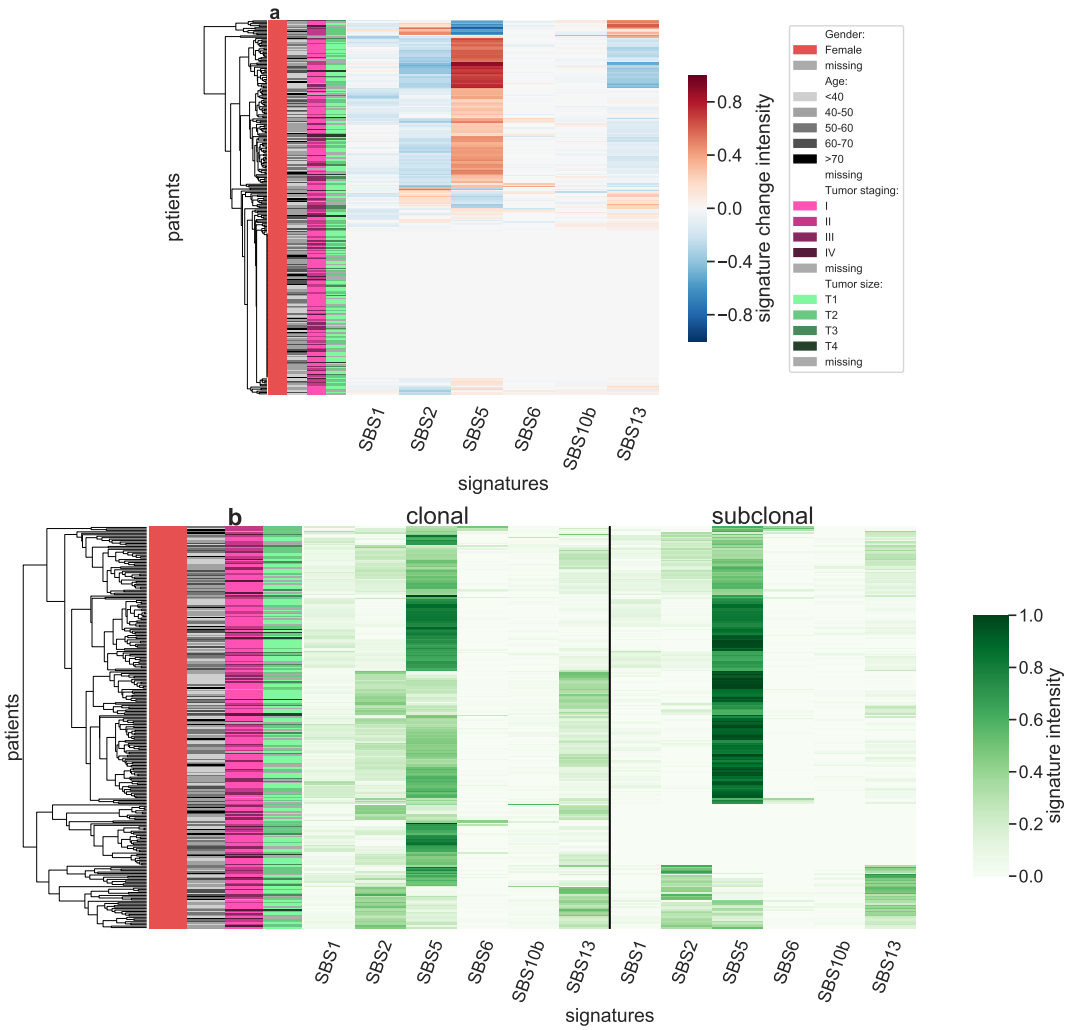

Figure B.29 – Panel a: Stratification of patients depending on their pattern of signature change for CESC patients (275 patients, including 168 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

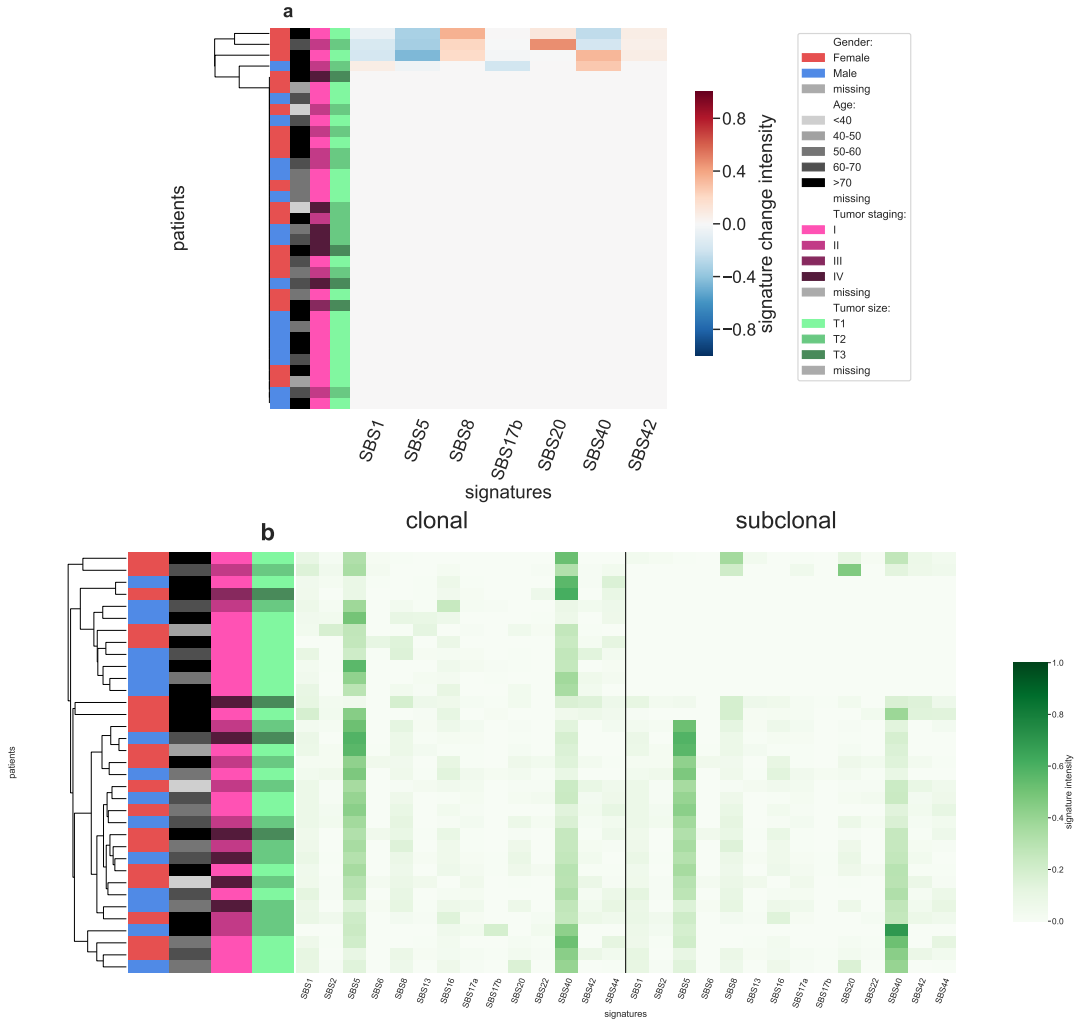

Figure B.30 – Panel a: Stratification of patients depending on their pattern of signature change for CHOL patients (35 patients, including 4 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

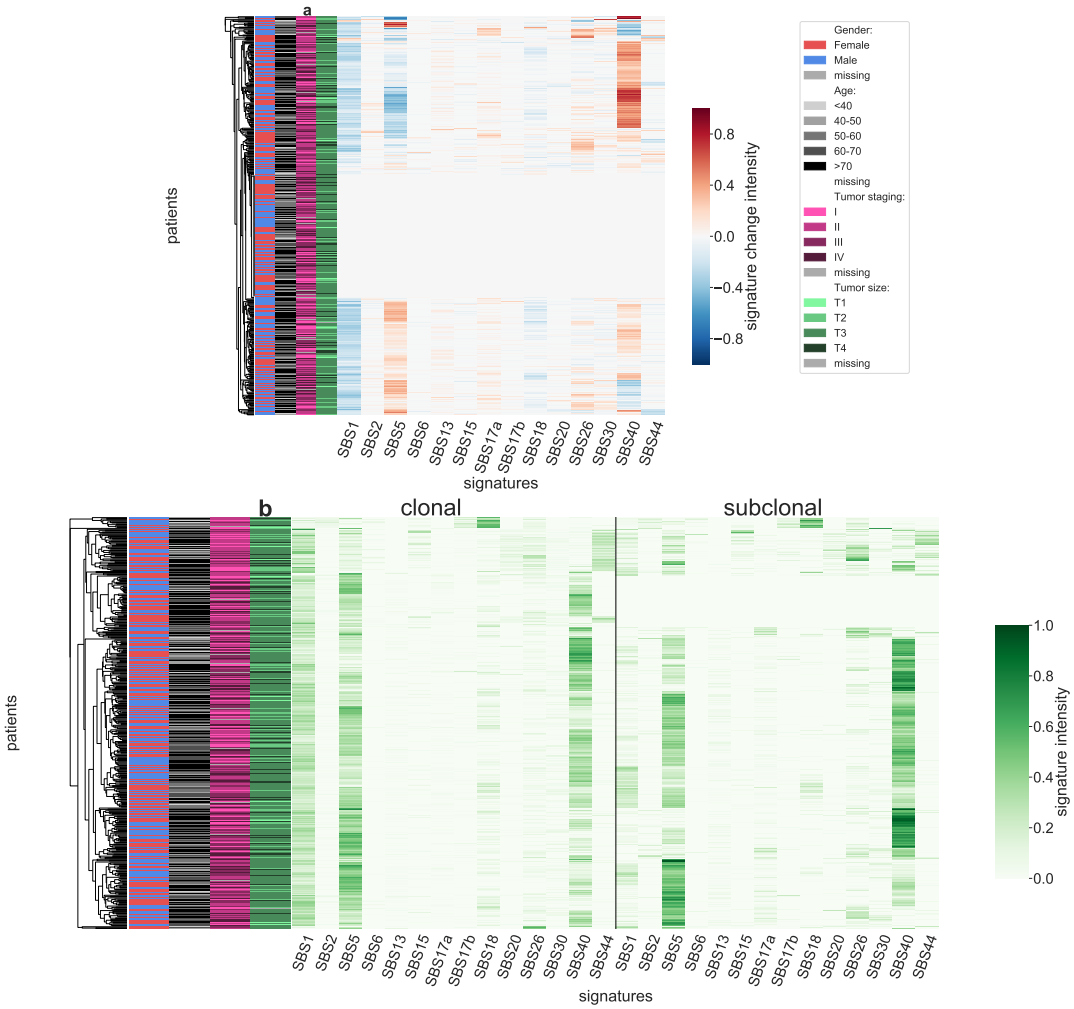

Figure B.31 – Panel a: Stratification of patients depending on their pattern of signature change for COADREAD patients (458 patients, including 318 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

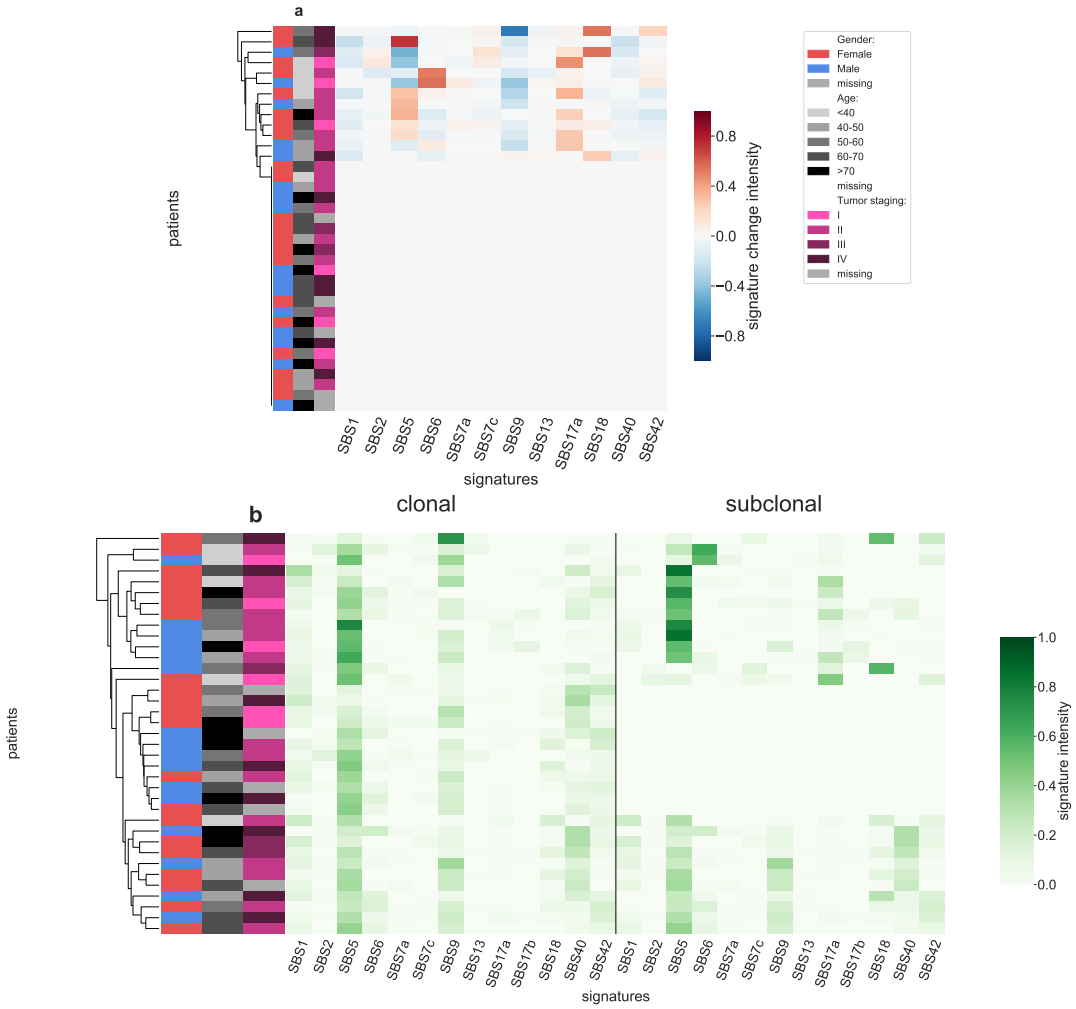

Figure B.32 – Panel a: Stratification of patients depending on their pattern of signature change for DLBC patients (37 patients, including 13 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

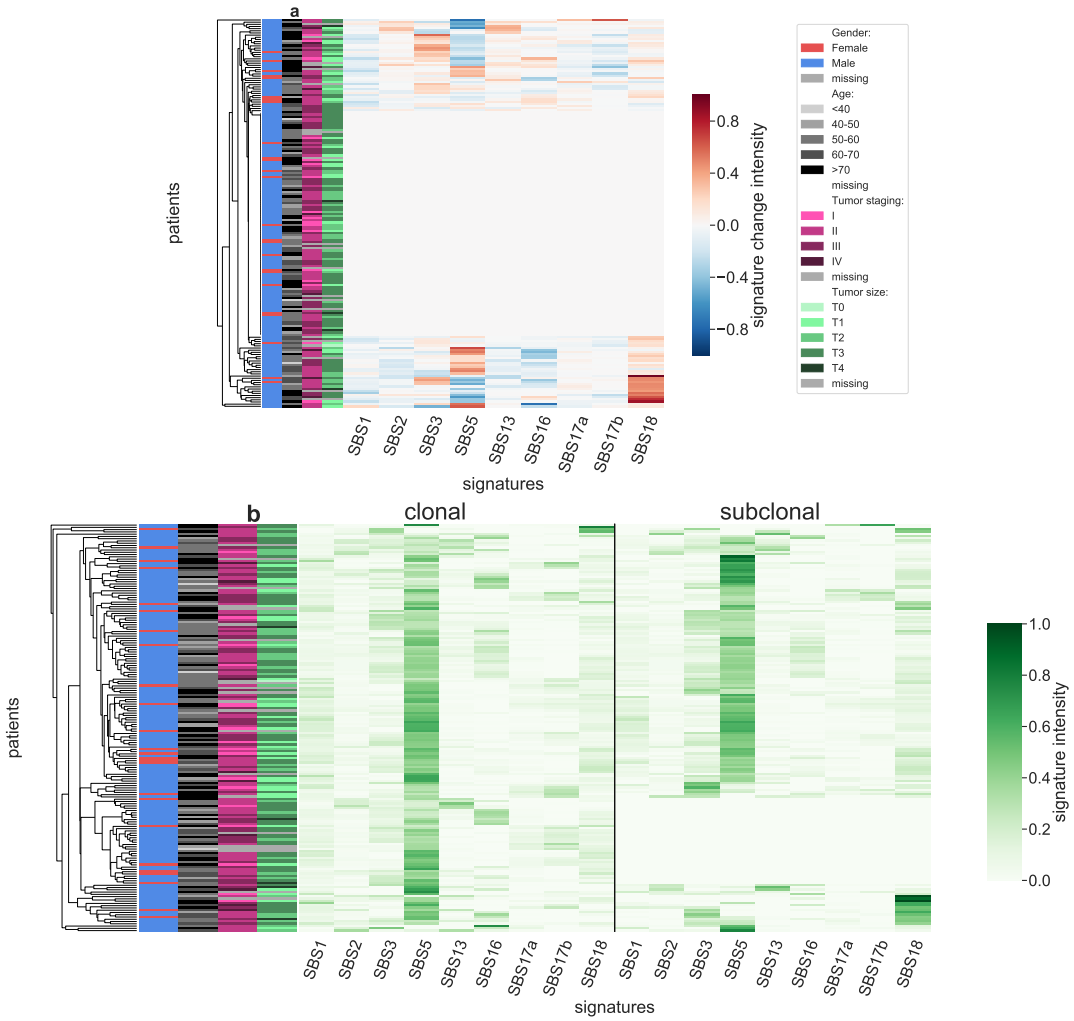

Figure B.33 – Panel a: Stratification of patients depending on their pattern of signature change for ESCA patients (180 patients, including 76 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

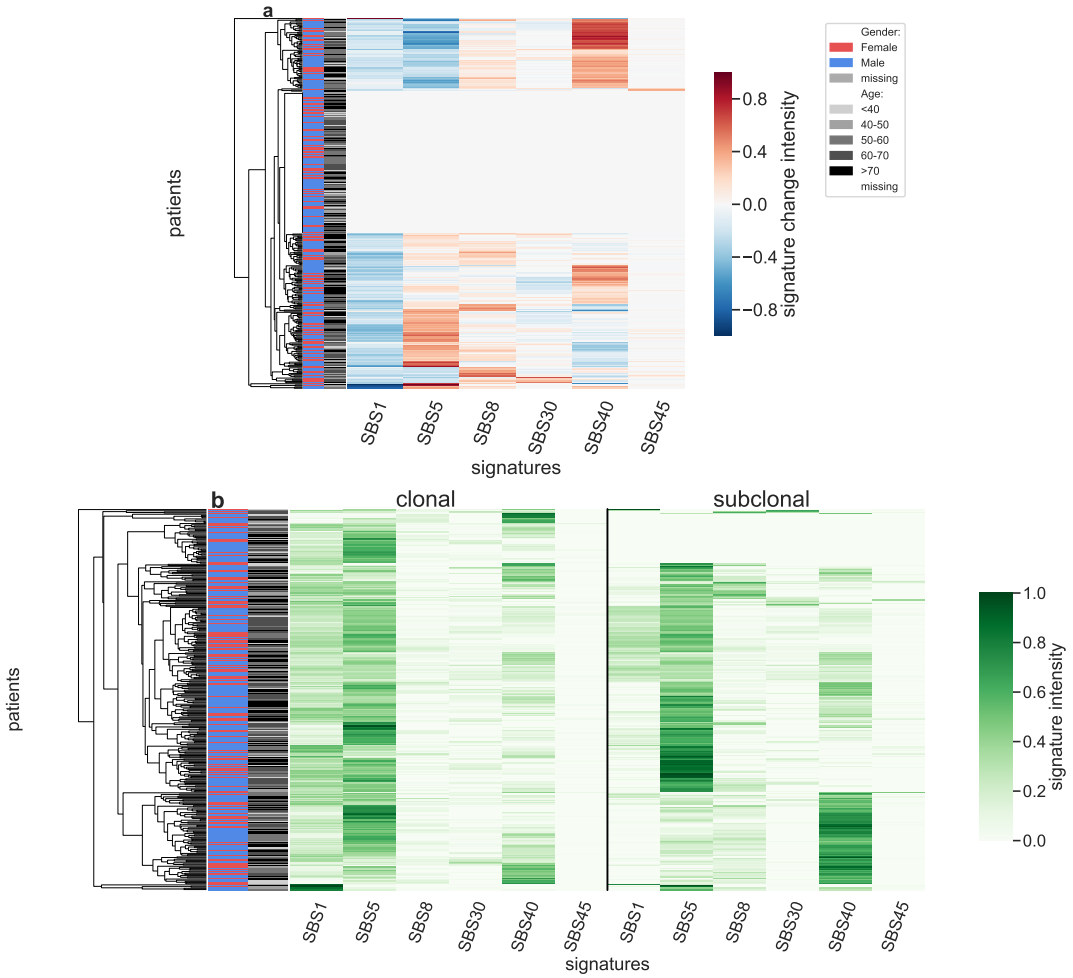

Figure B.34 – Panel a: Stratification of patients depending on their pattern of signature change for GBM patients (327 patients, including 202 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

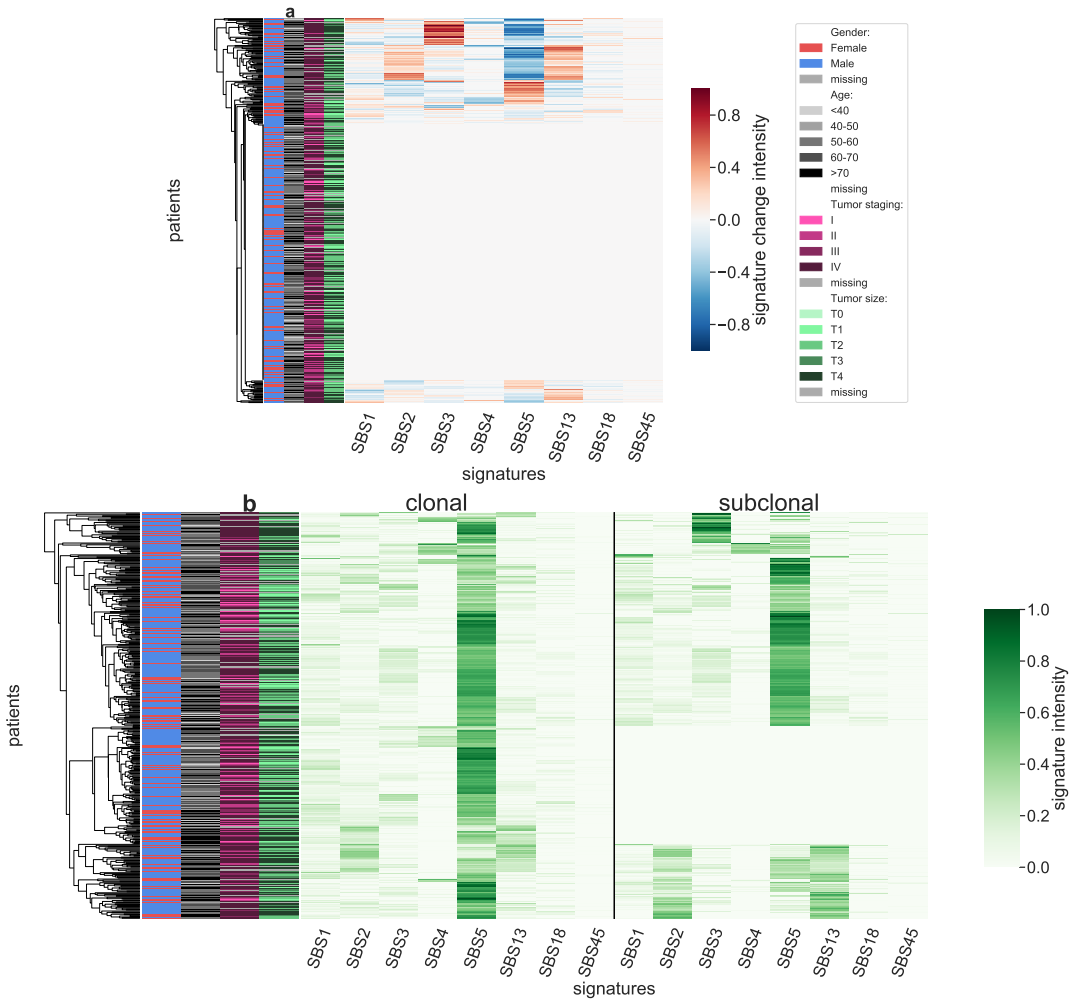

Figure B.35 – Panel a: Stratification of patients depending on their pattern of signature change for HNSC patients (445 patients, including 148 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

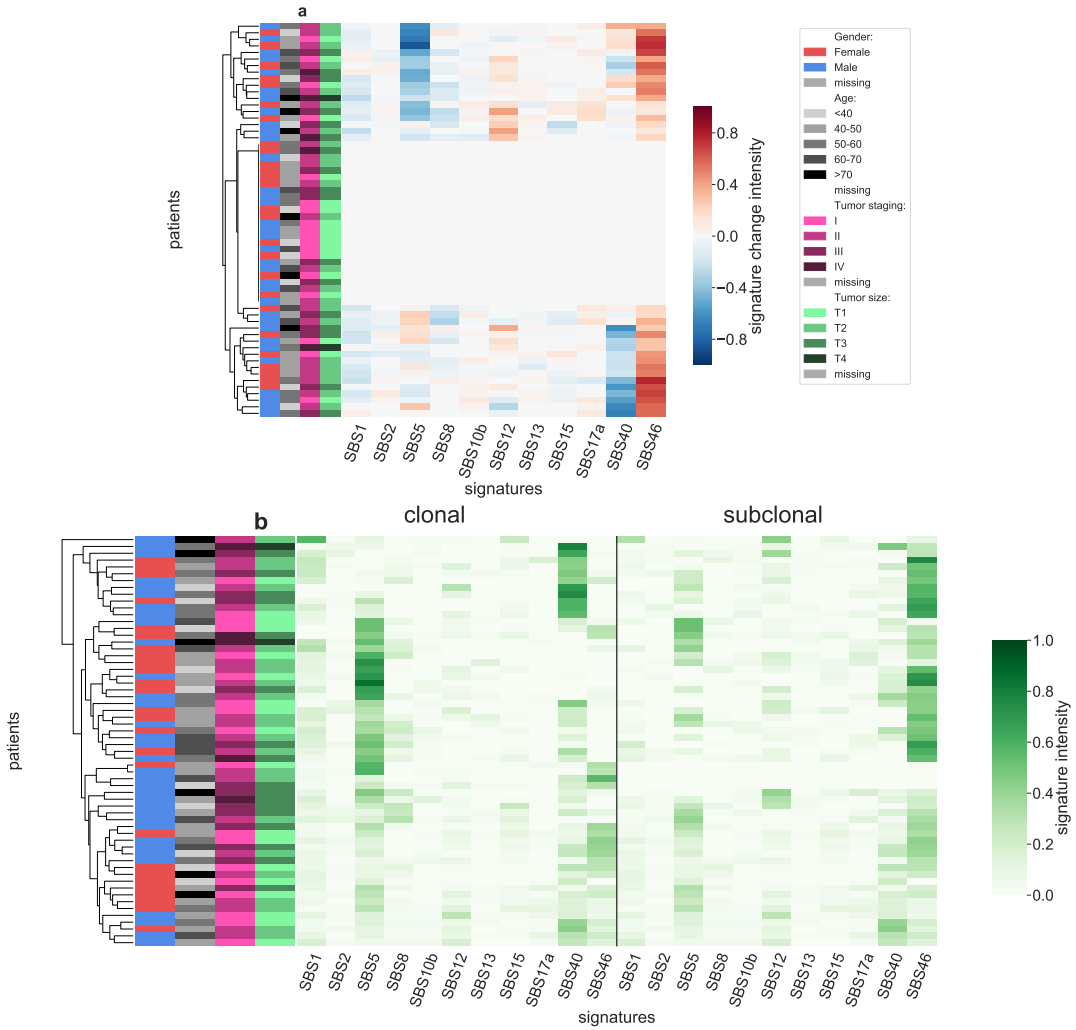

Figure B.36 – Panel a: Stratification of patients depending on their pattern of signature change for KICH patients (60 patients, including 35 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

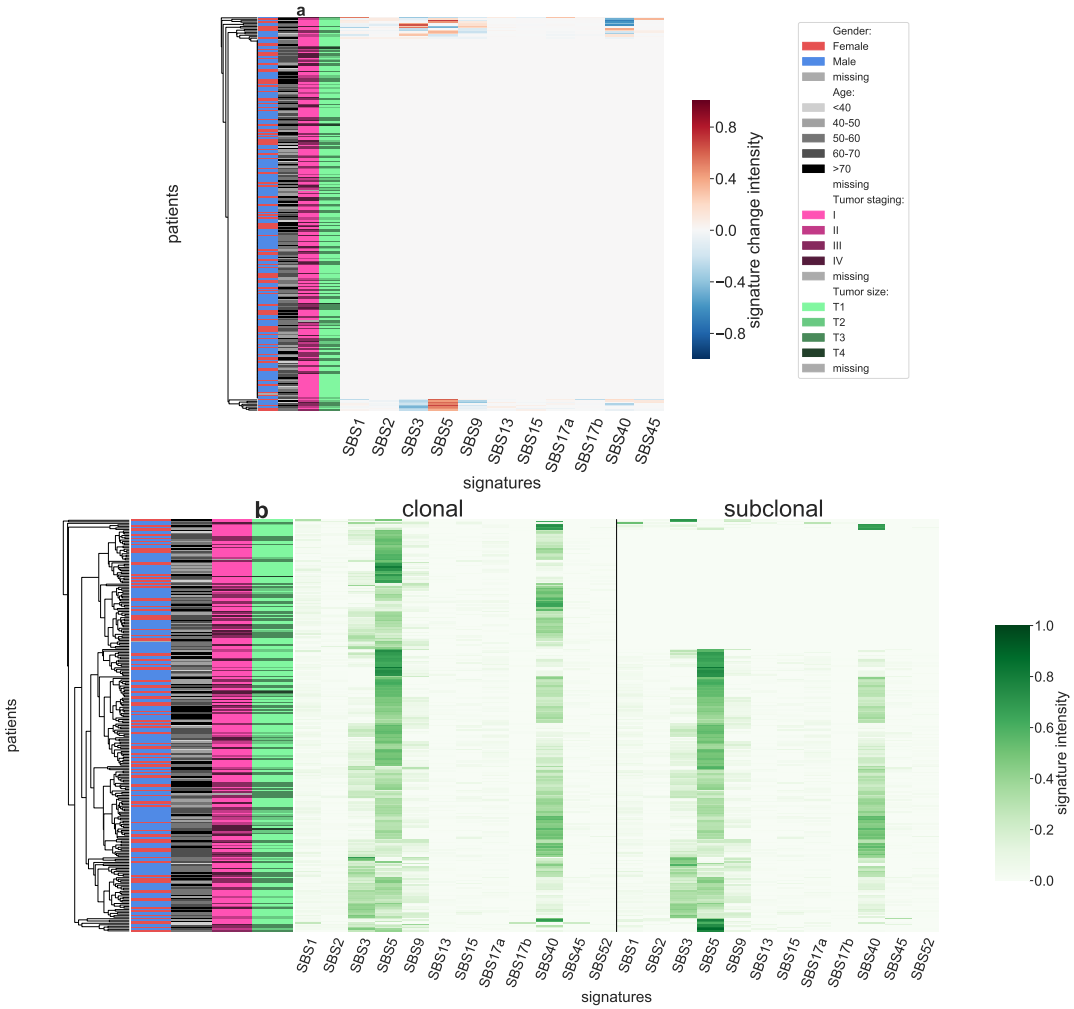

Figure B.37 – Panel a: Stratification of patients depending on their pattern of signature change for KIRC patients (271 patients, including 23 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

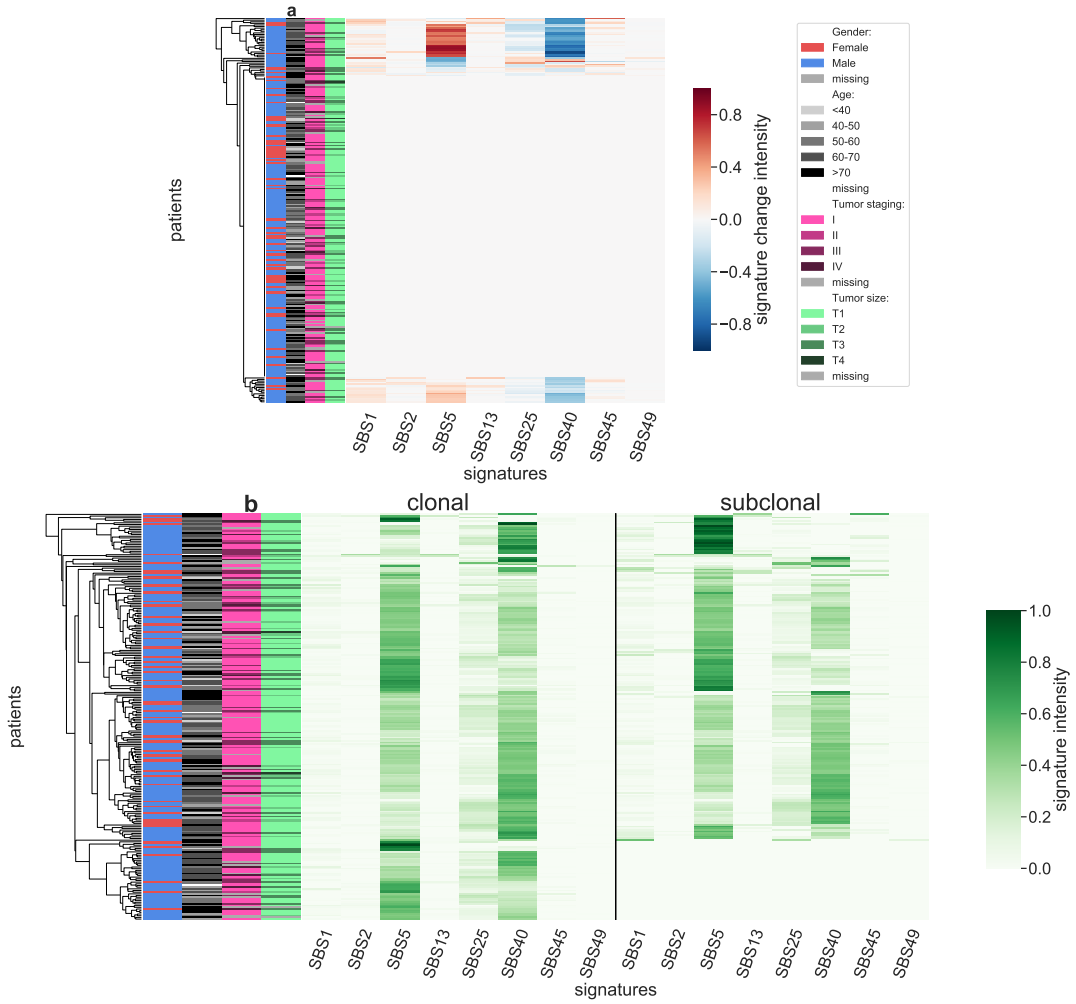

Figure B.38 – Panel a: Stratification of patients depending on their pattern of signature change for KIRP patients (242 patients, including 53 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

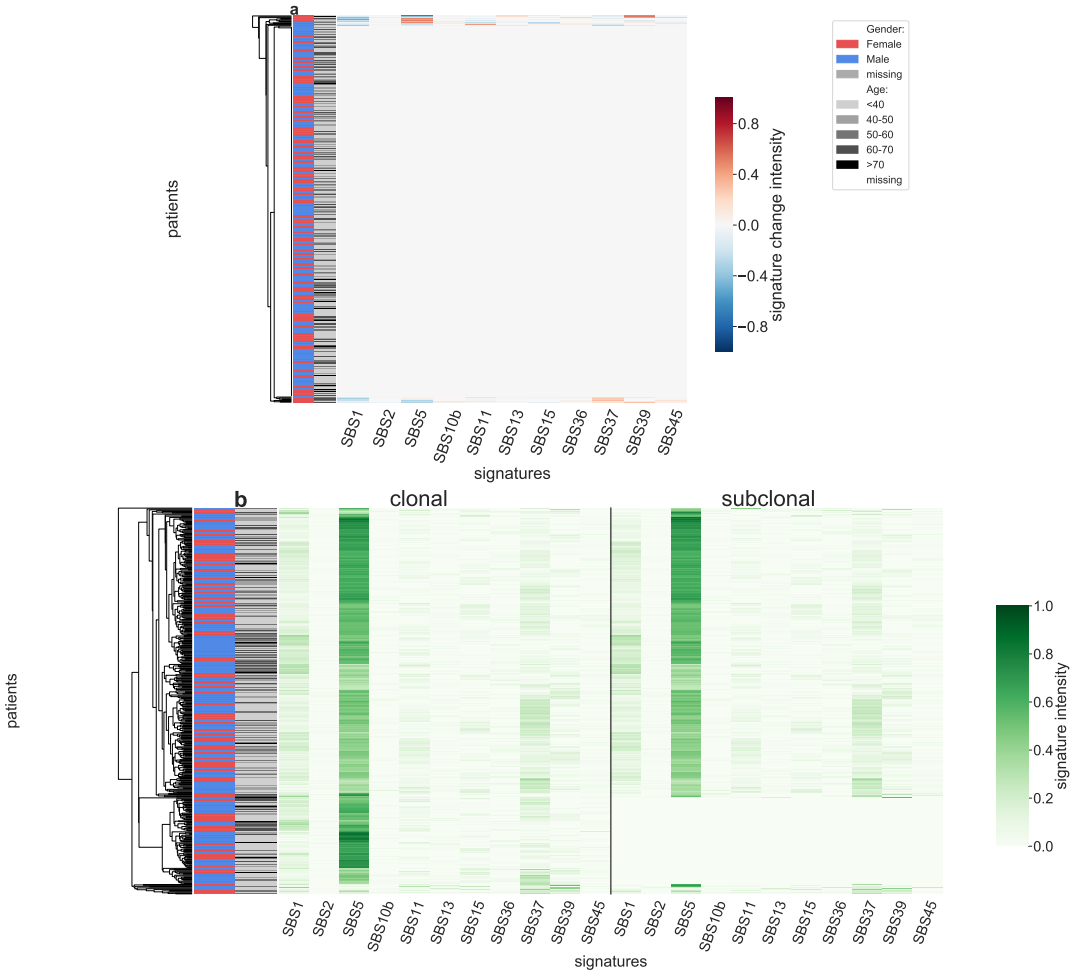

Figure B.39 – Panel a: Stratification of patients depending on their pattern of signature change for LGG patients (455 patients, including 20 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

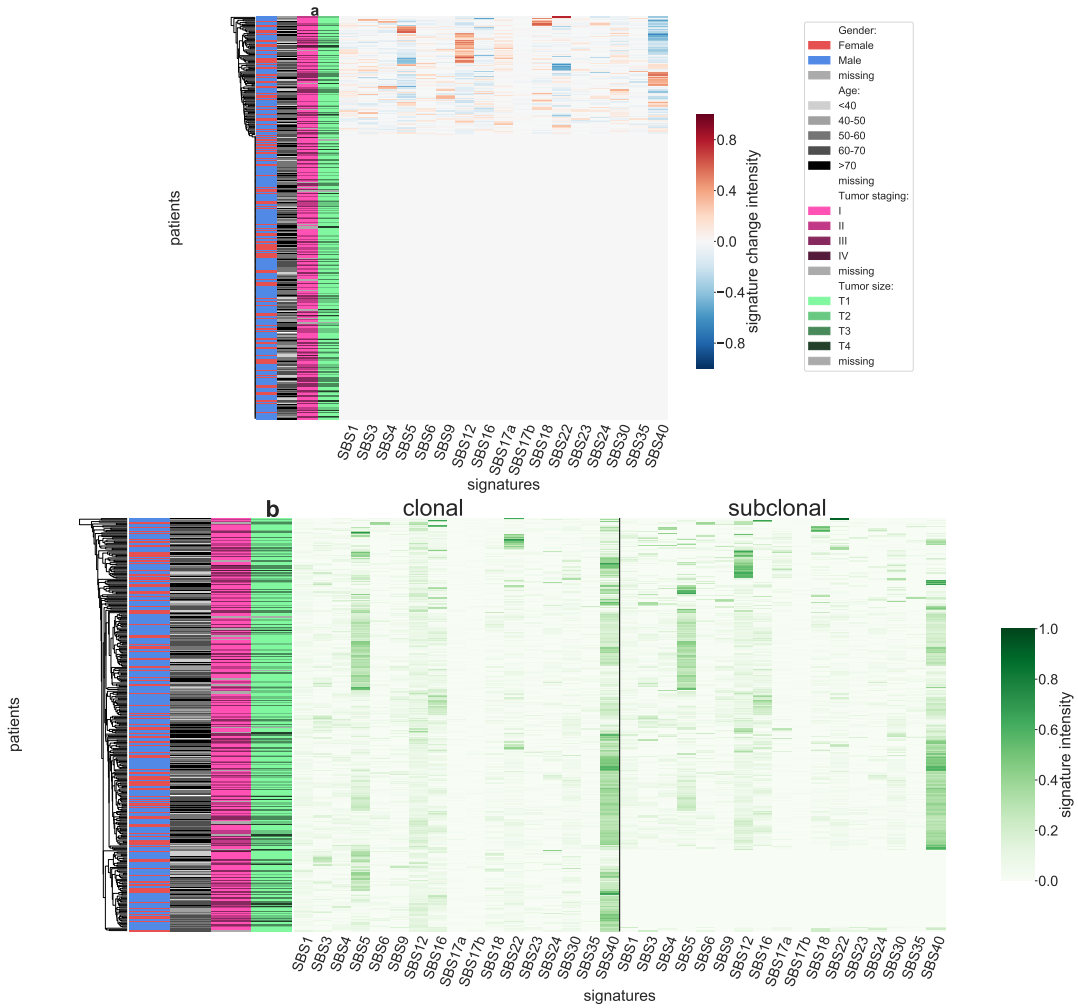

Figure B.40 – Panel a: Stratification of patients depending on their pattern of signature change for LIHC patients (347 patients, including 102 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

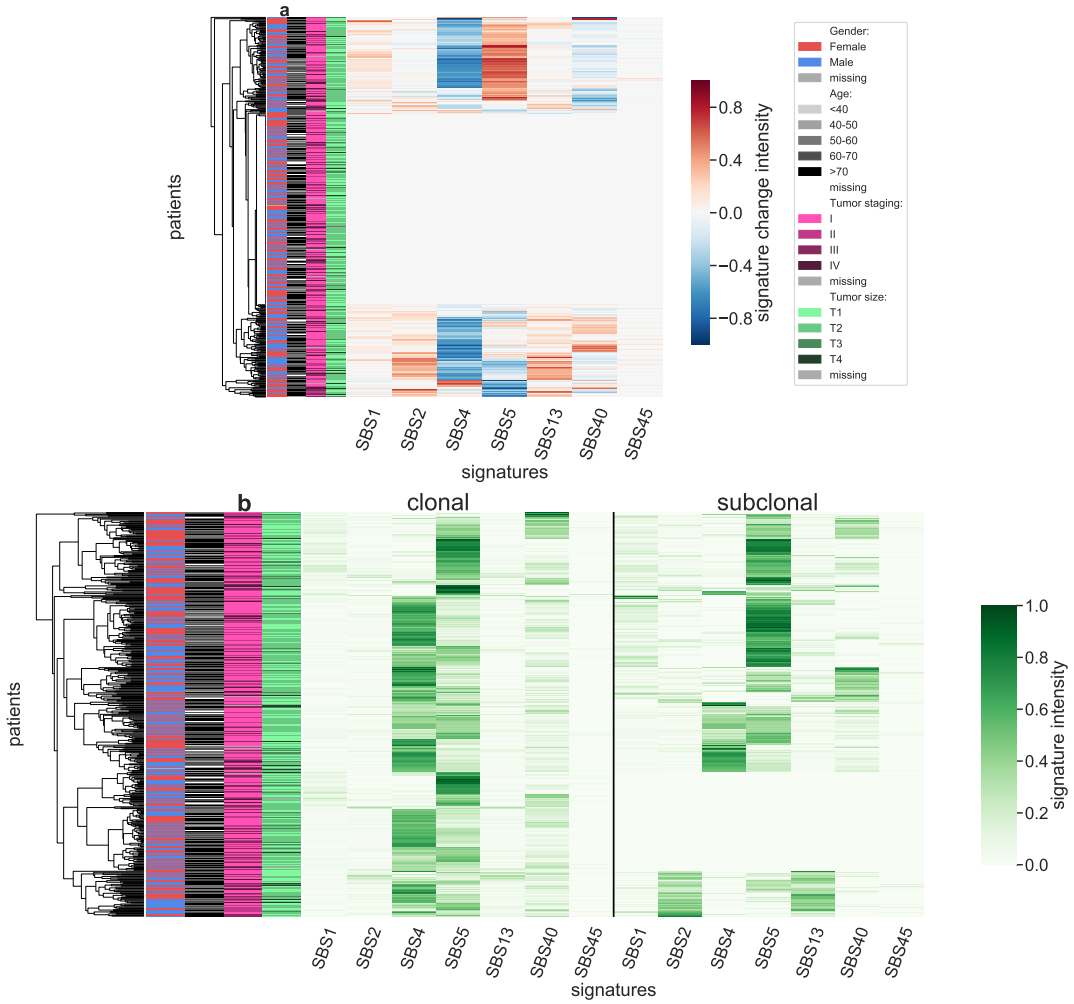

Figure B.41 – Panel a: Stratification of patients depending on their pattern of signature change for LUAD patients (433 patients, including 217 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

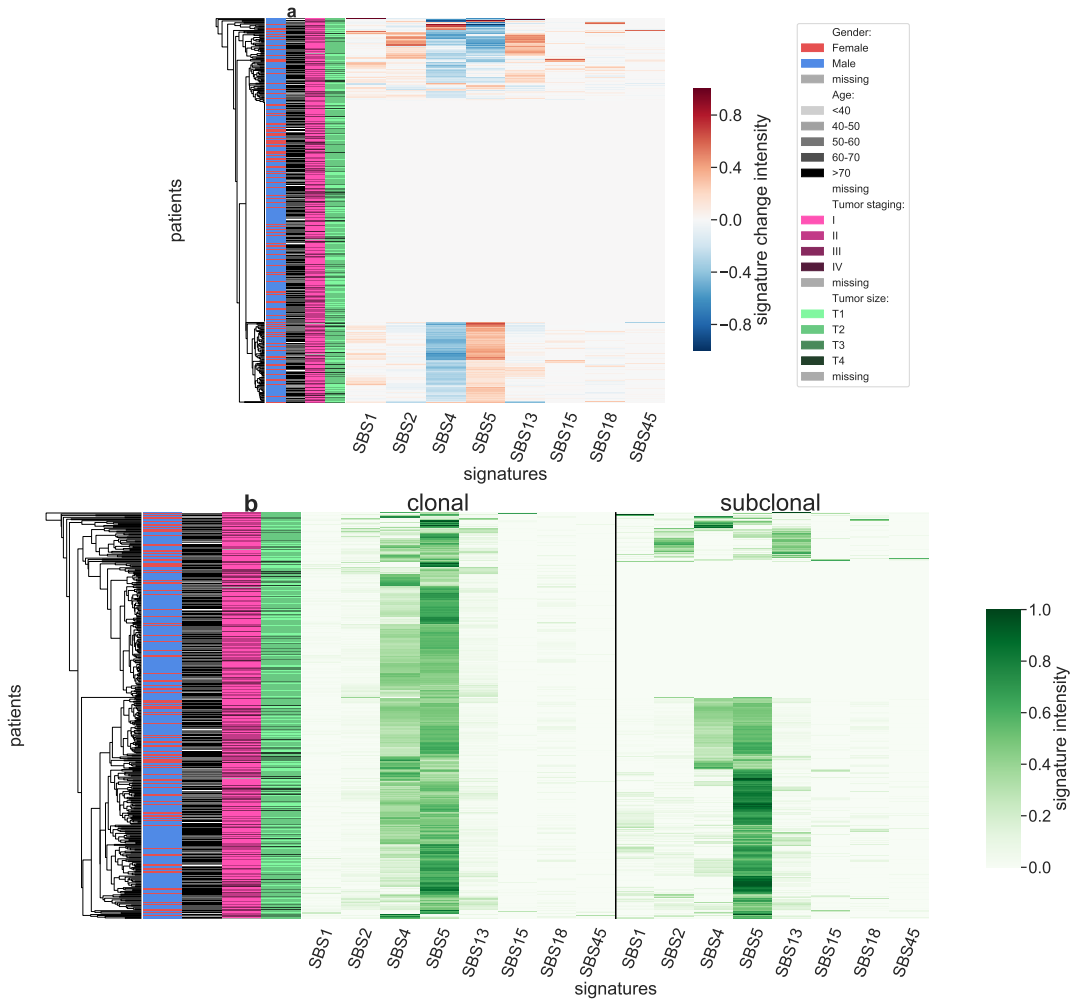

Figure B.42 – Panel a: Stratification of patients depending on their pattern of signature change for LUSC patients (423 patients, including 180 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

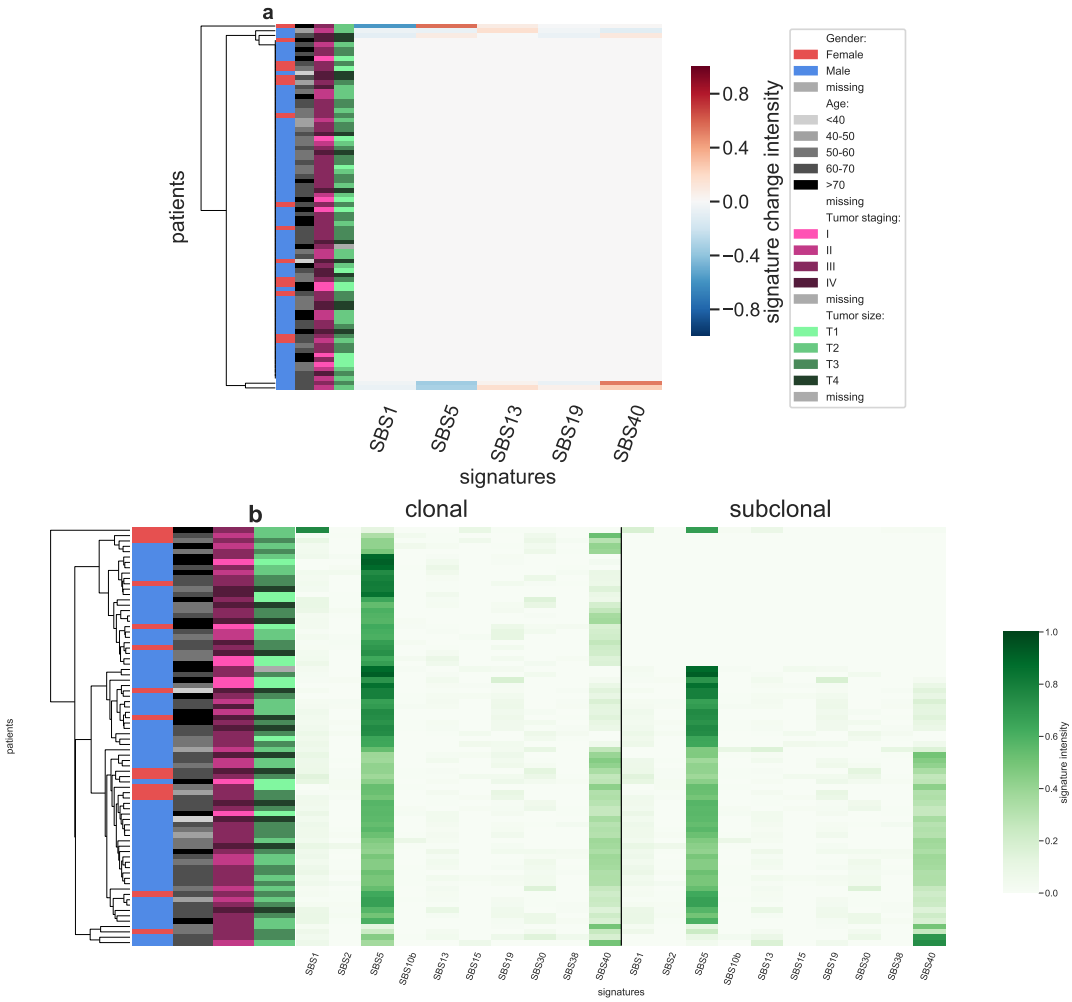

Figure B.43 – Panel a: Stratification of patients depending on their pattern of signature change for MESO patients (78 patients, including 5 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

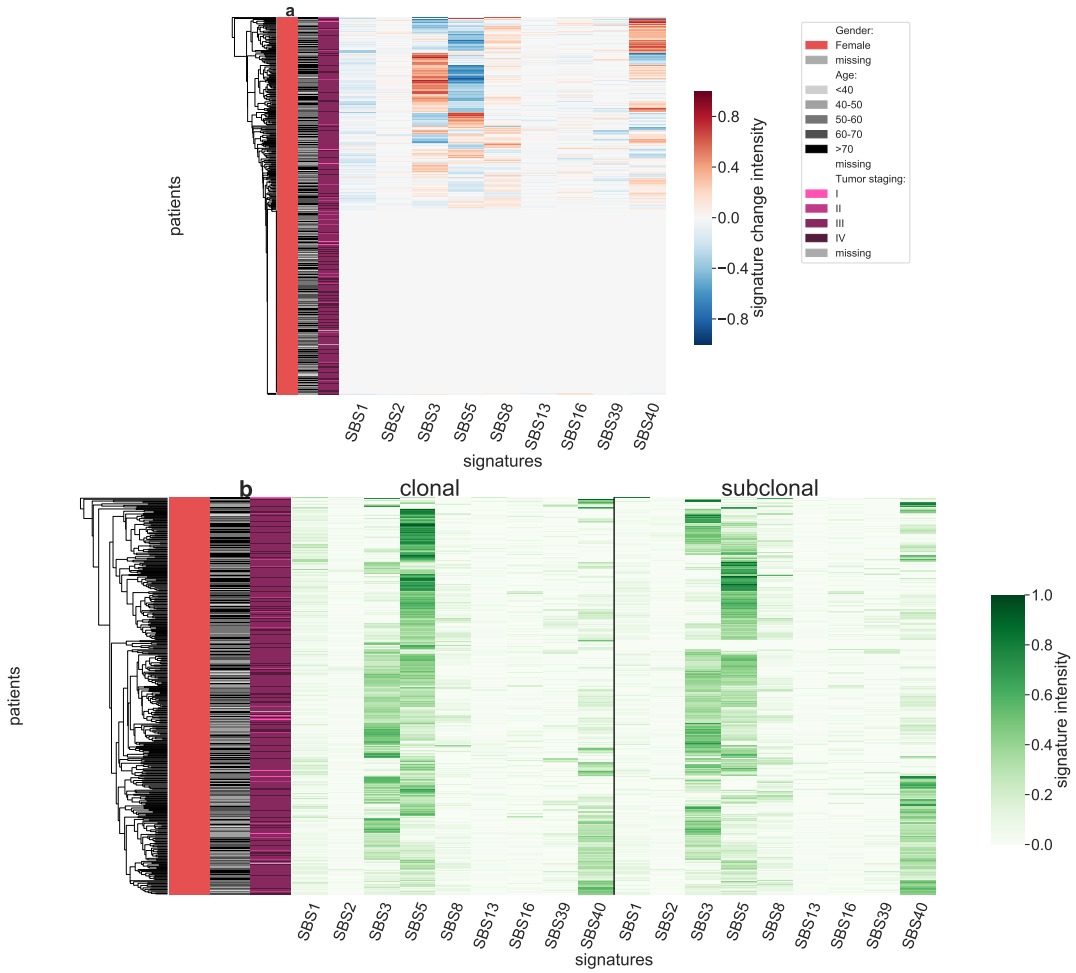

Figure B.44 – Panel a: Stratification of patients depending on their pattern of signature change for OV patients (390 patients, including 201 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

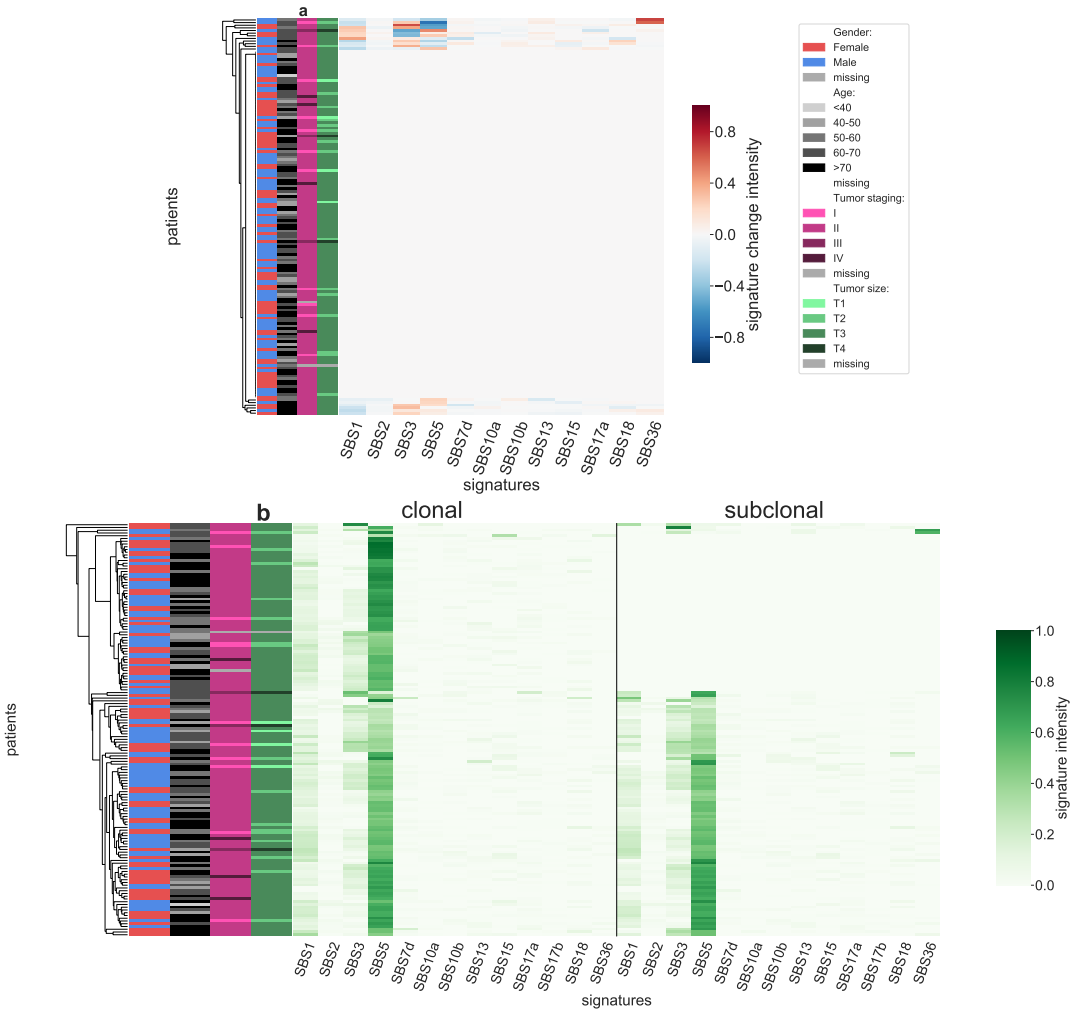

Figure B.45 – Panel a: Stratification of patients depending on their pattern of signature change for PAAD patients (150 patients, including 18 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

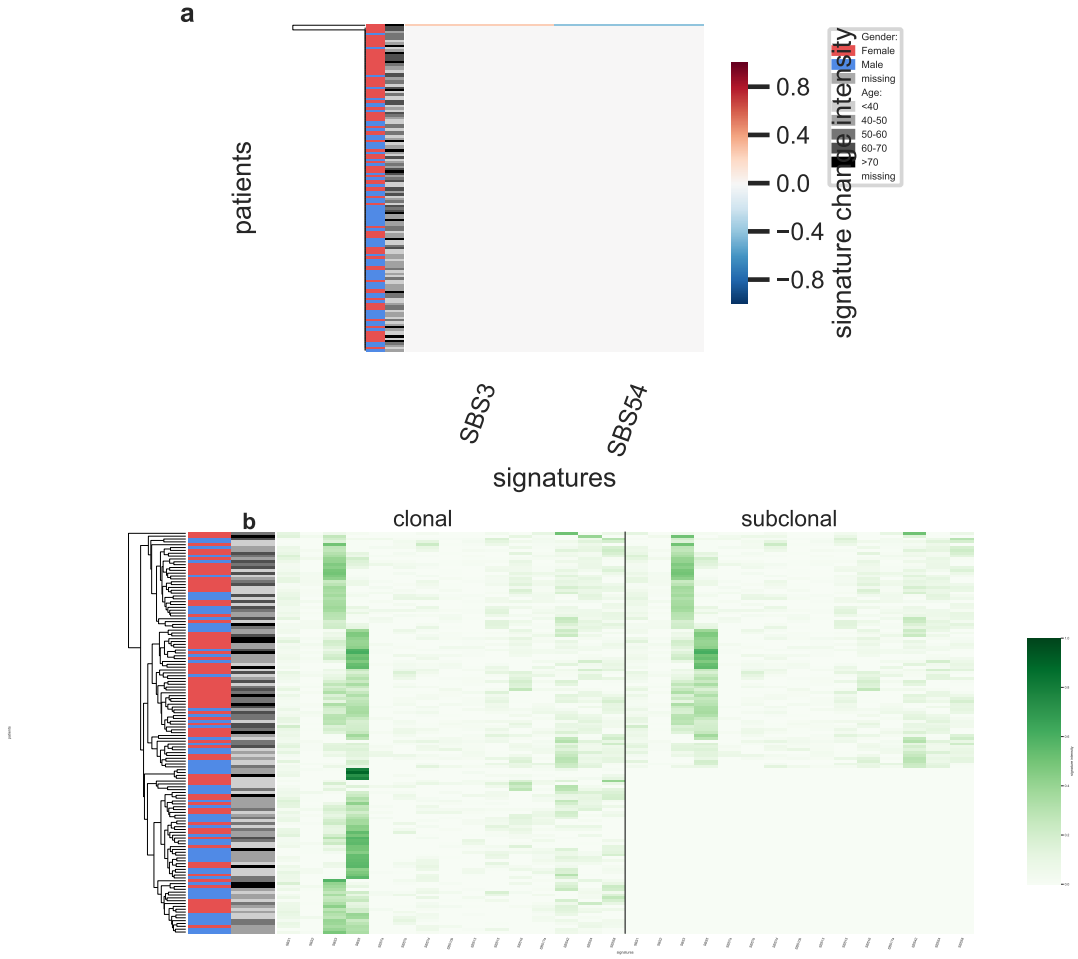

Figure B.46 – Panel a: Stratification of patients depending on their pattern of signature change for PCPG patients (141 patients, including 1 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

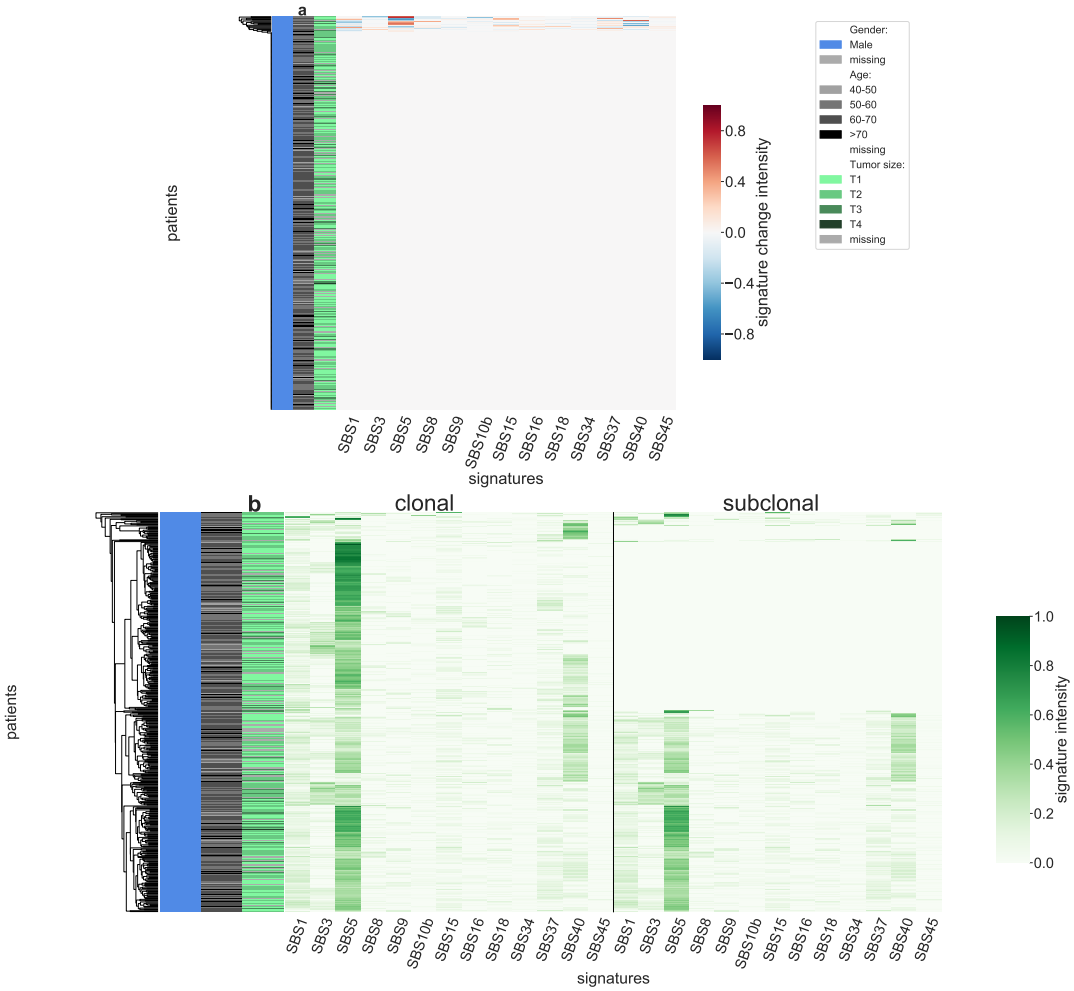

Figure B.47 – Panel a: Stratification of patients depending on their pattern of signature change for PRAD patients (458 patients, including 18 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

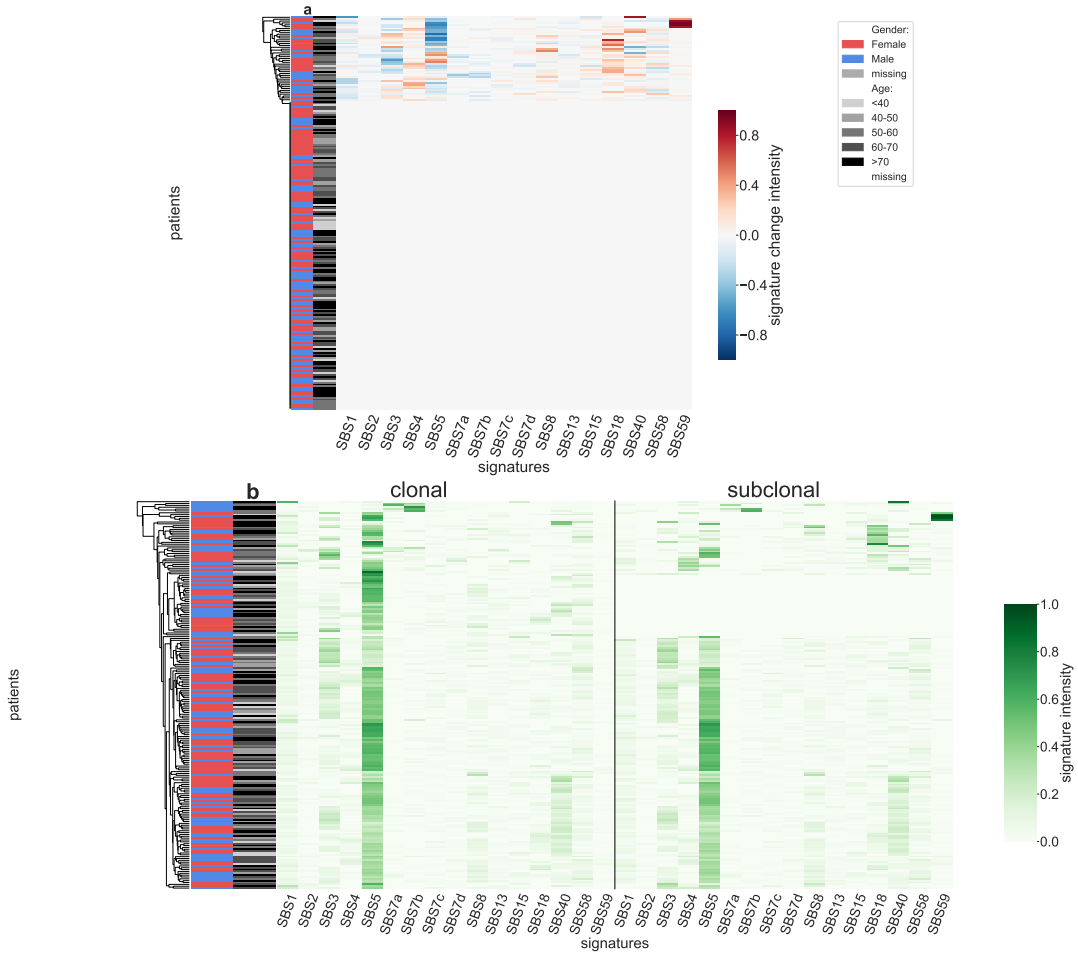

Figure B.48 – Panel a: Stratification of patients depending on their pattern of signature change for SARC patients (210 patients, including 45 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

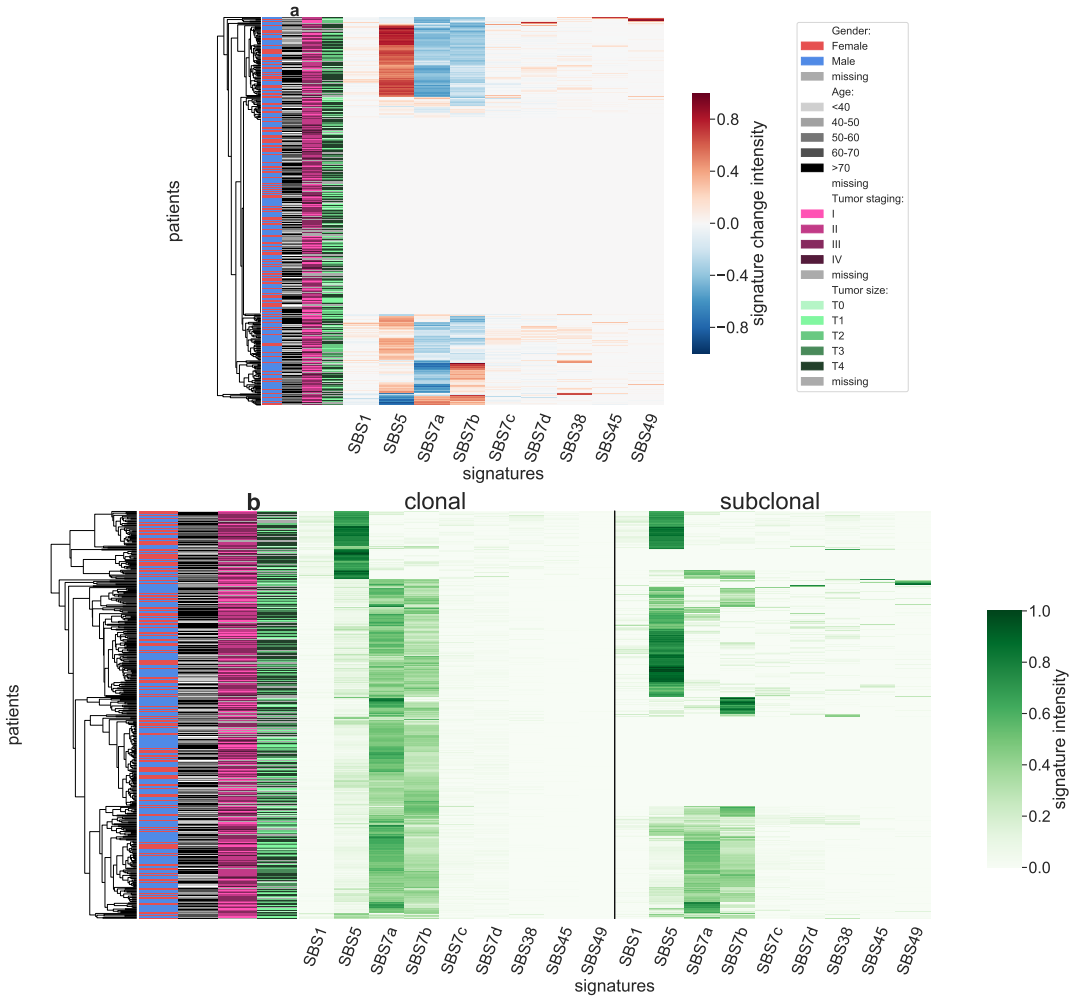

Figure B.49 – Panel a: Stratification of patients depending on their pattern of signature change for SKCM patients (423 patients, including 210 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

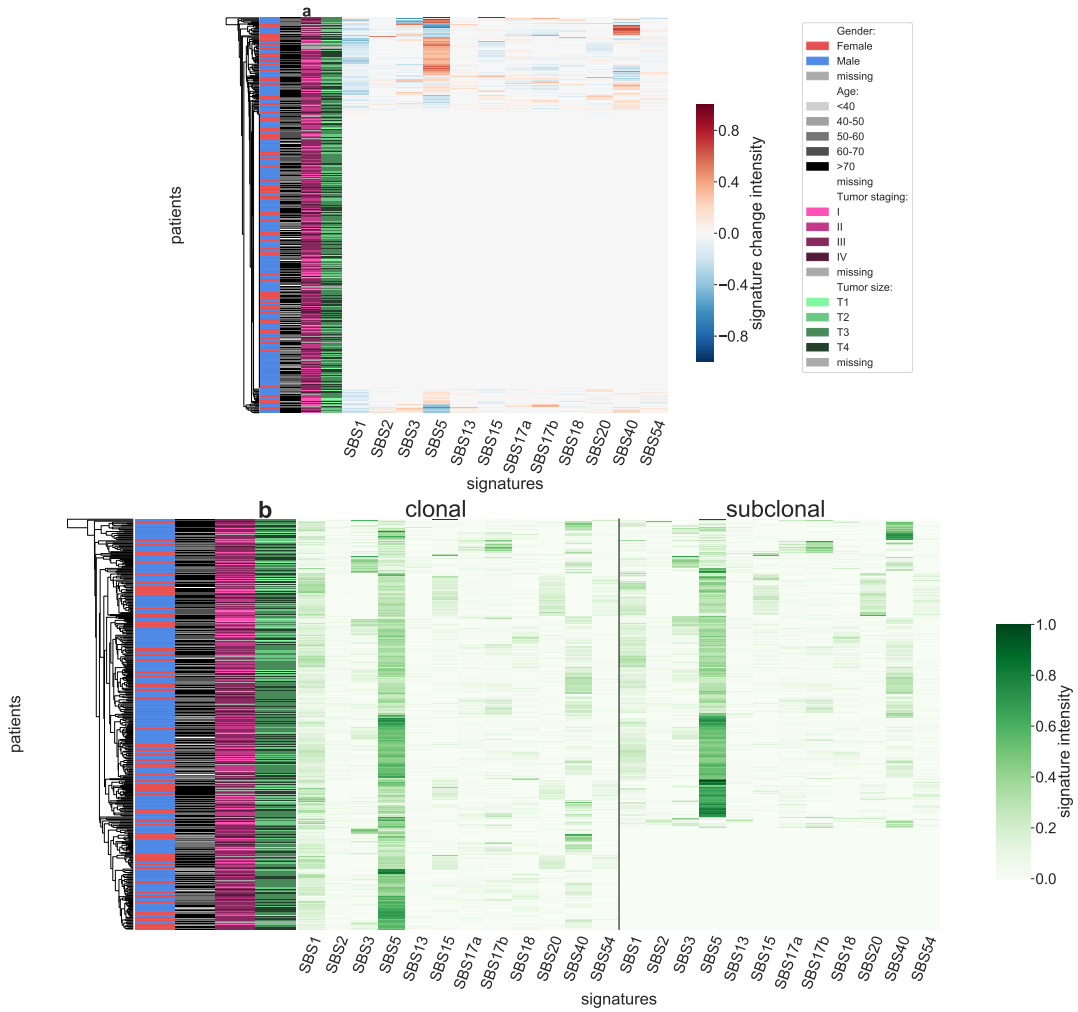

Figure B.50 – Panel a: Stratification of patients depending on their pattern of signature change for STAD patients (418 patients, including 127 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

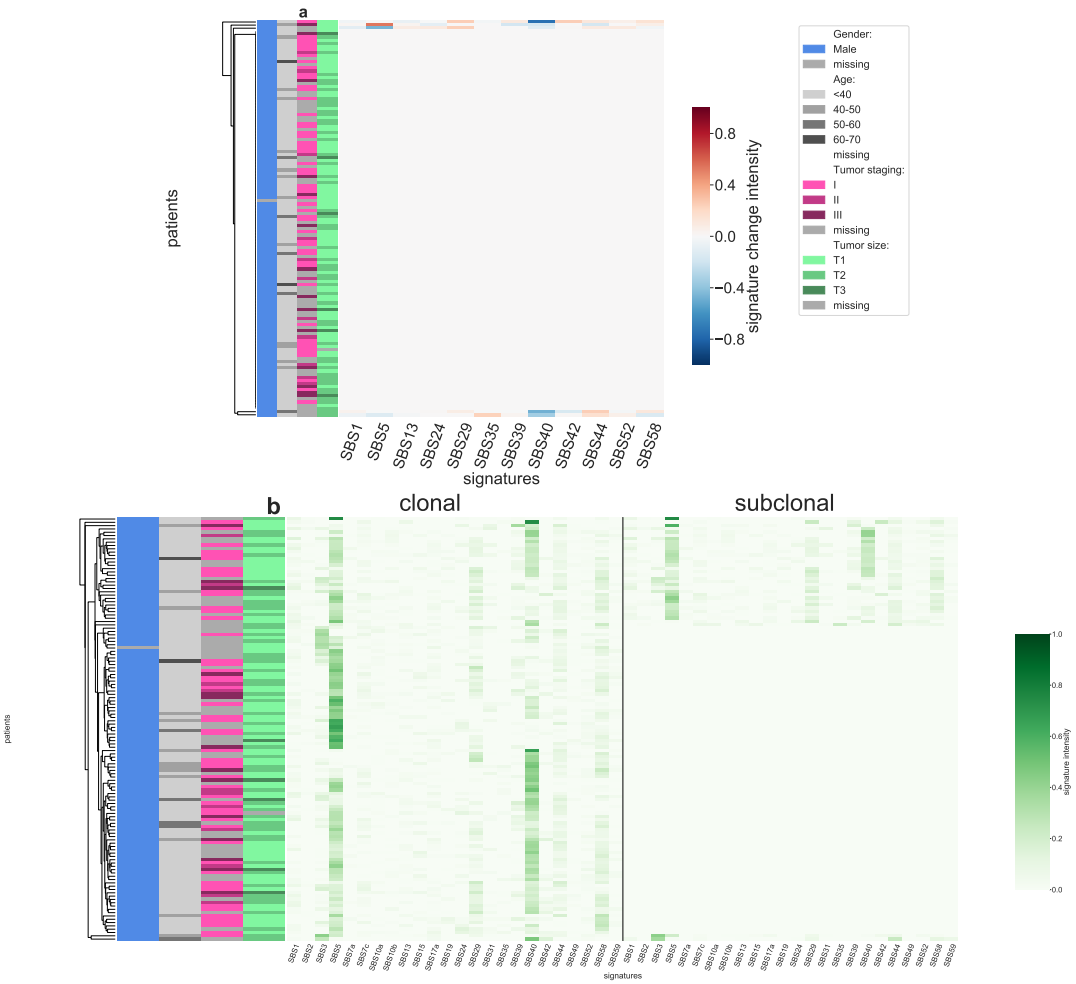

Figure B.51 – Panel a: Stratification of patients depending on their pattern of signature change for TGCT patients (128 patients, including 5 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

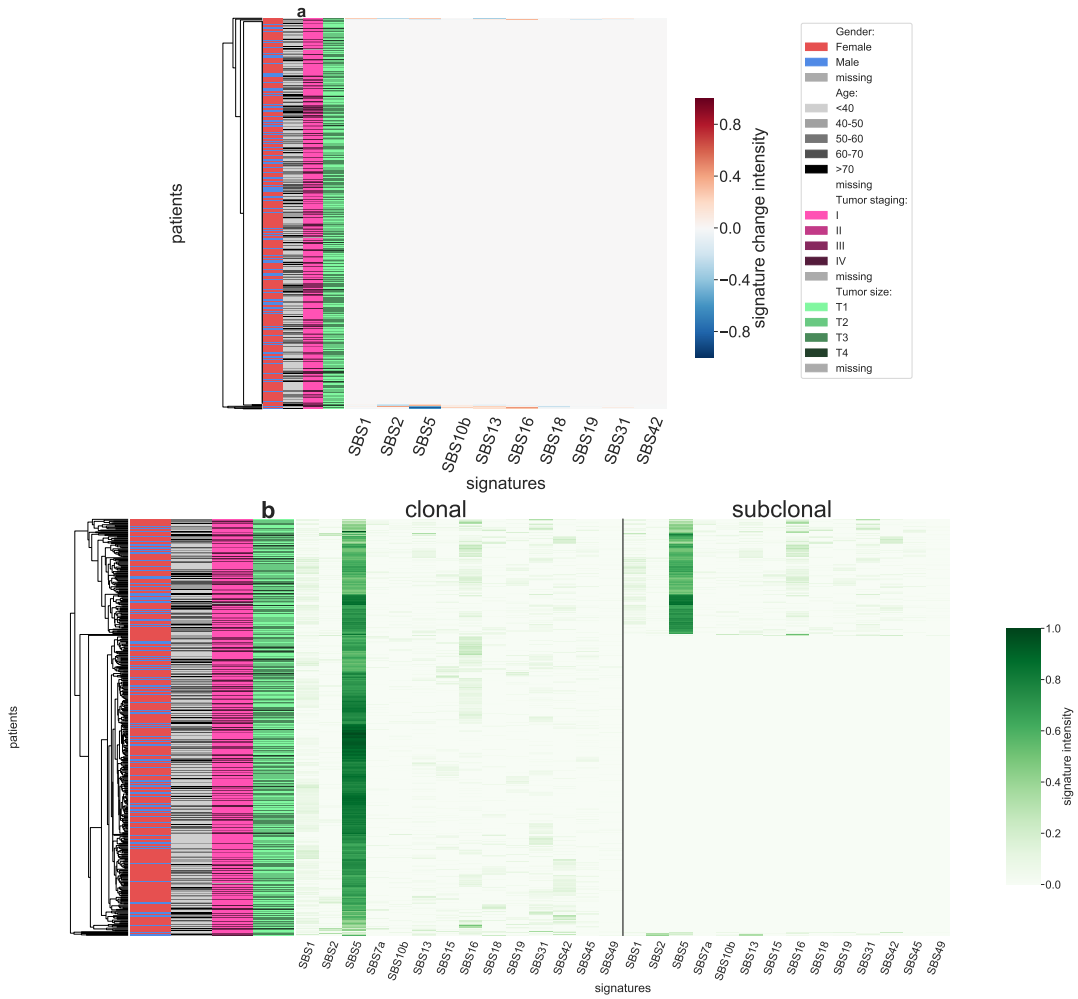

Figure B.52 – Panel a: Stratification of patients depending on their pattern of signature change for THCA patients (467 patients, including 9 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

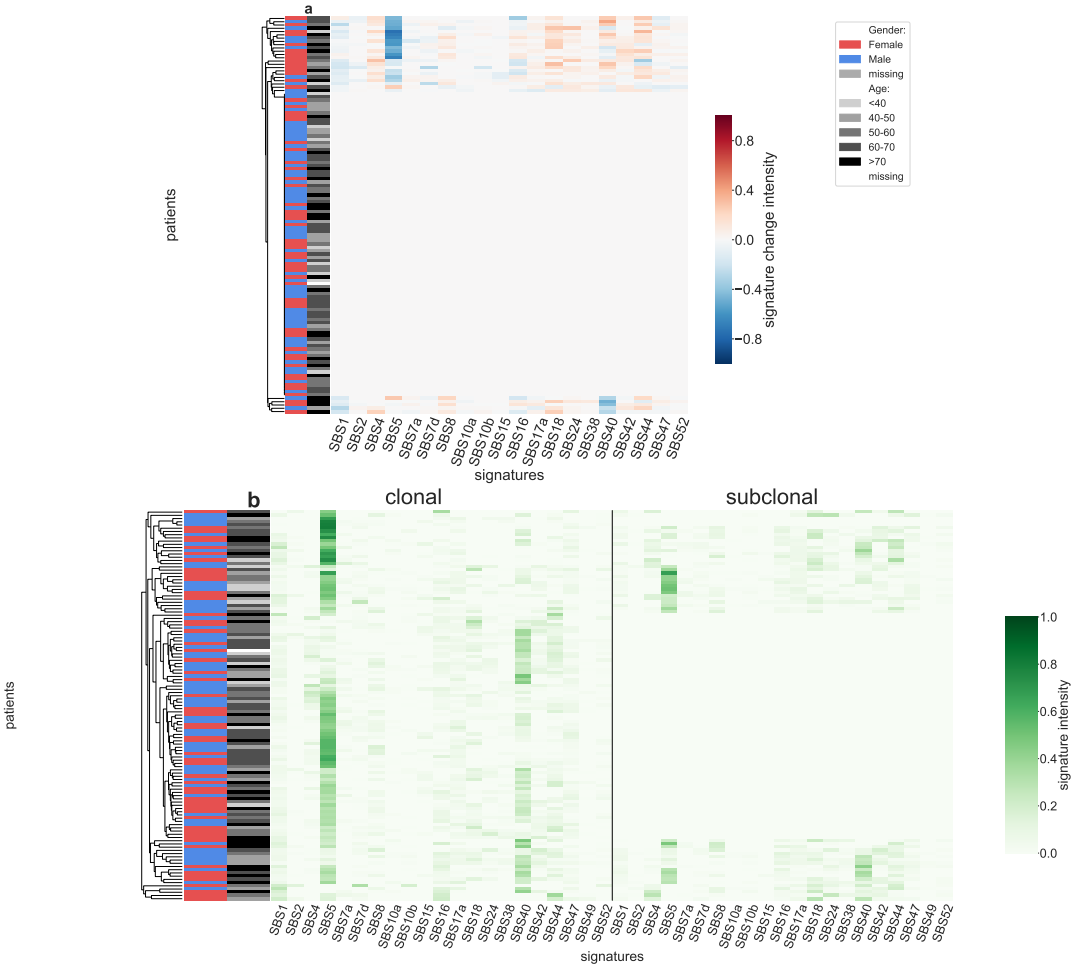

Figure B.53 – Panel a: Stratification of patients depending on their pattern of signature change for THYM patients (121 patients, including 28 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

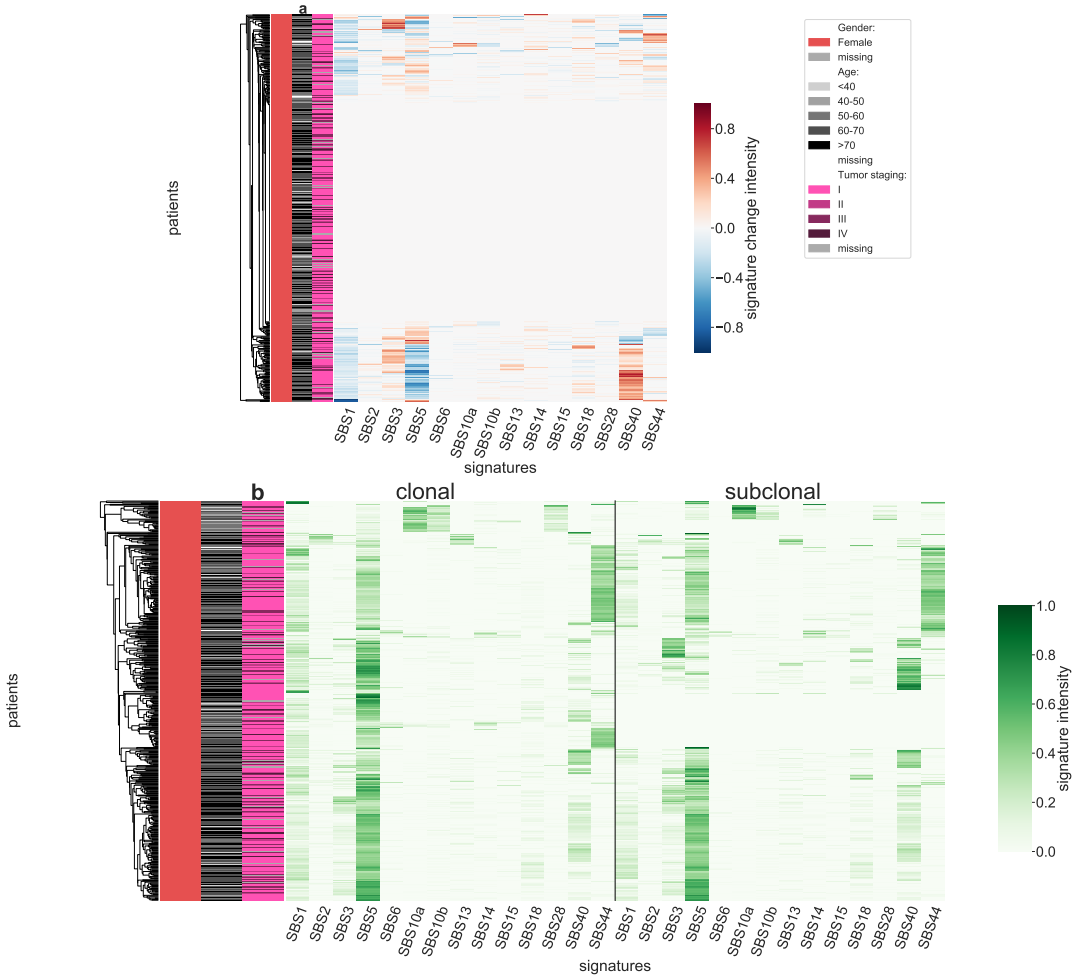

Figure B.54 – Panel a: Stratification of patients depending on their pattern of signature change for UCEC patients (487 patients, including 213 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

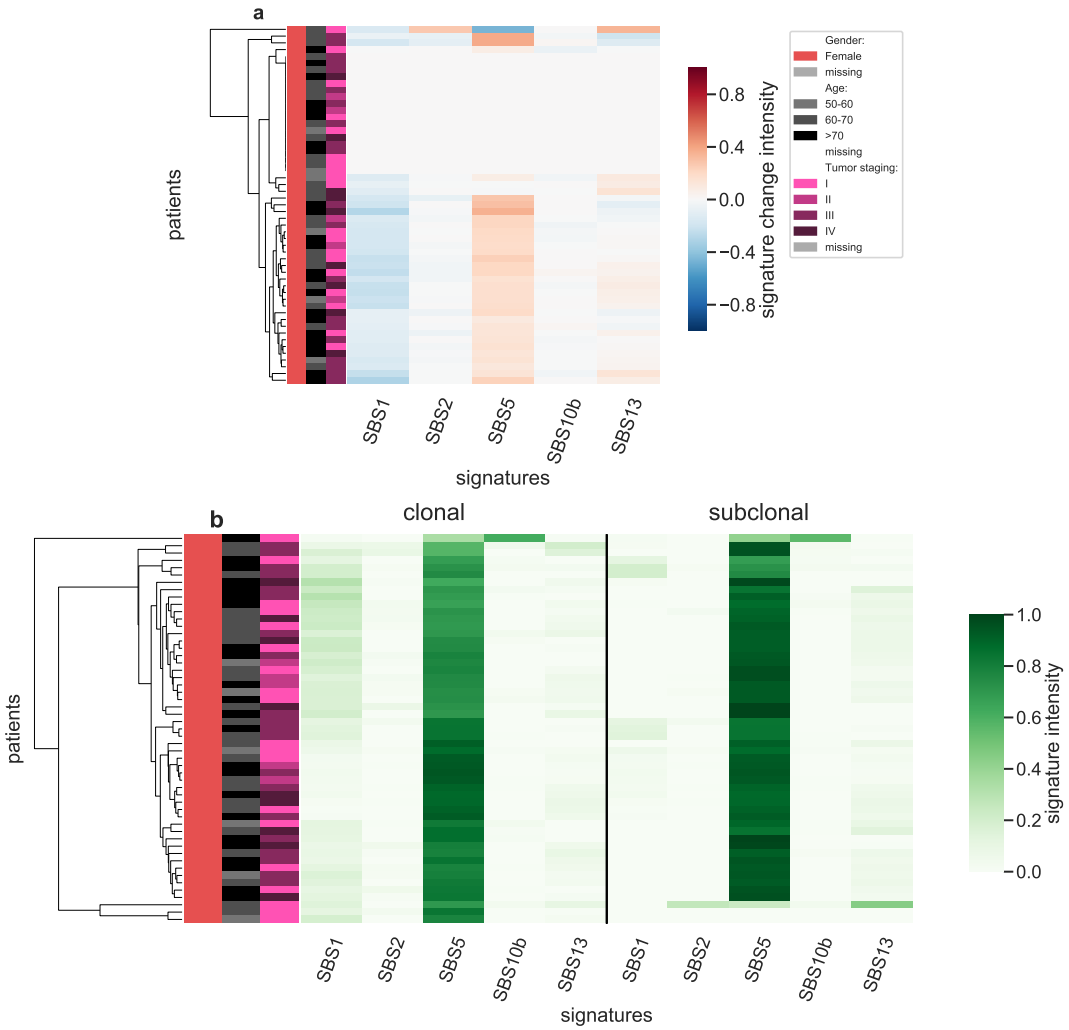

Figure B.55 – Panel a: Stratification of patients depending on their pattern of signature change for UCS patients (53 patients, including 35 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

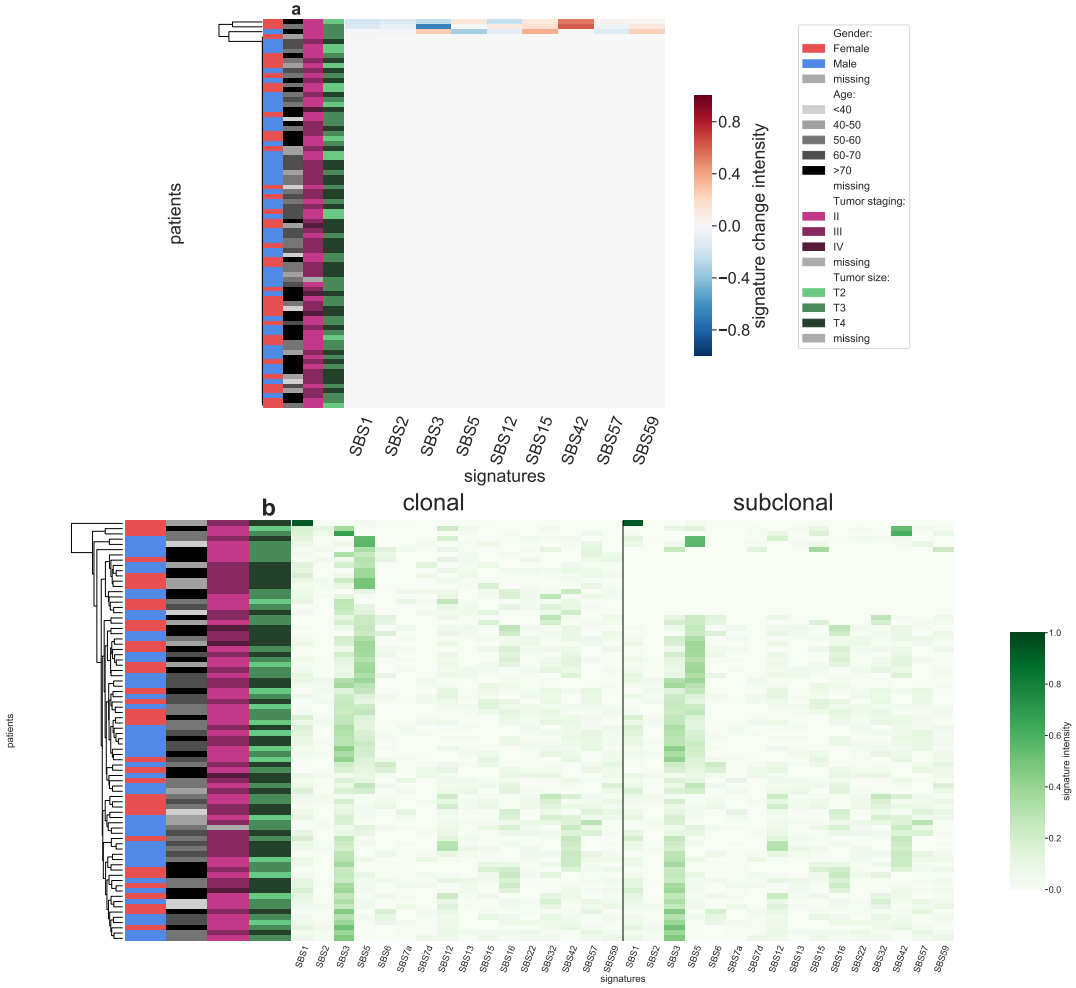

Figure B.56 – Panel a: Stratification of patients depending on their pattern of signature change for UVM patients (80 patients, including 3 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

Supplementary figures

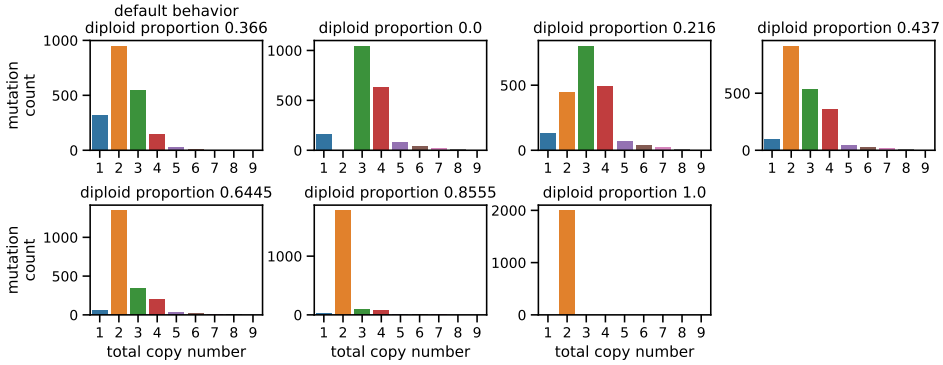

Figure B.57 – An example of empirical distribution of the total copy number for samples with 2000 mutations. In the first panel, labeled "default behavior", the user does not specify the percentage of genome that is diploid, and the total copy number values are drawn as specified in the Methods section. On the other panels, the user specifies a desired percentage of genome that is diploid (0, 0.2, 0.4, 0.6, 0.8, 1) respectively for the cases shown. The distribution is slightly different from the default behavior.

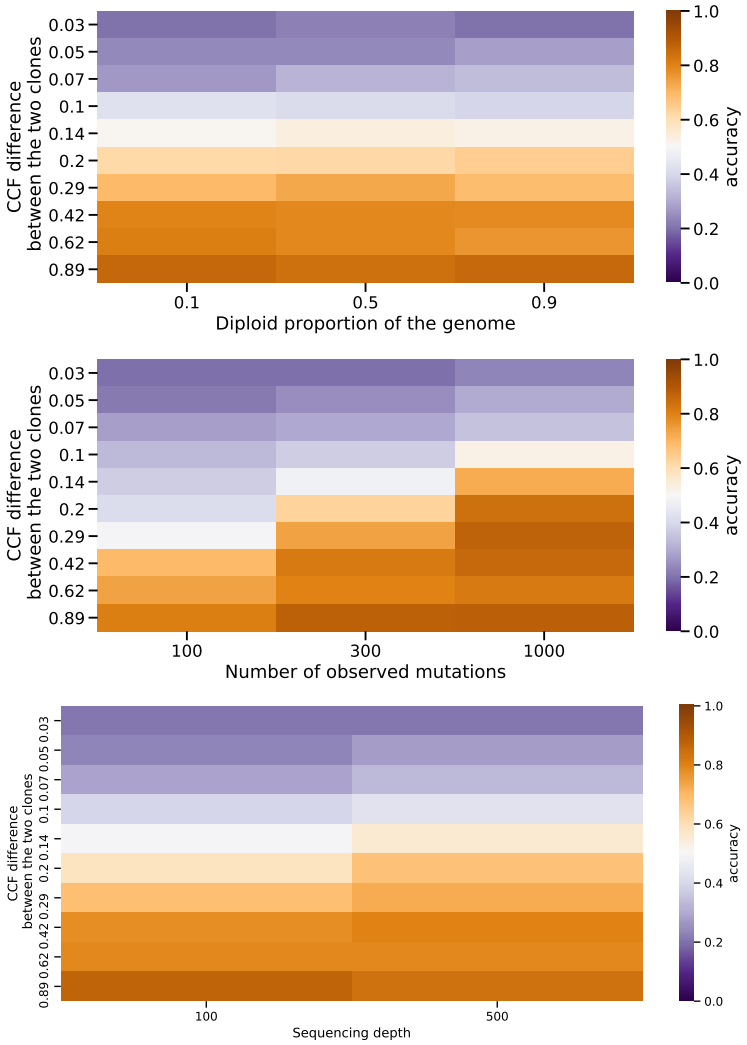

Figure B.58 – CloneSig's ability to distinguish 2 clones depending on the CCF distance between the two clones, and other relevant variables: number of mutations, and percentage of diploid genome, and sequencing depth

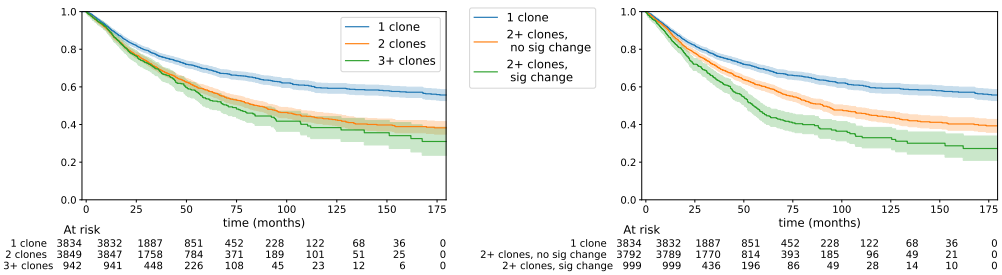

Figure B.59 – Kaplan-Meier curves for all TCGA samples (8625) distinguishing tumors only along the number of clones (left) or along the number of clones and the presence of a significant change in signatures along tumor evolution (right) using the public input mutation sets. A multivariate Cox model was fitted in both cases, and indicates for 2 clones, hazard ratio (HR) of 1.38 (95% confidence interval (CI): $[1.27, 1.49]$, $p = 2.62\text{e} - 15$), and 3 clones (HR= 1.54, CI= $[1.36, 1.74]$, $p = 4.53\text{e} - 12$) (left). For the distinction based on signature change, without signature change (HR= 1.32, CI= $[1.22, 1.43]$, $p = 7.55\text{e} - 12$), and with signature change (HR= 1.80, CI= $[1.60, 2.02]$, $p = 4.89\text{e} - 23$) (right)

Supplementary tables

Cancer type	**Mean nb mutations (protected)**	**Mean nb mutations (public)**	**Standard deviation nb mutations (protected)**	**Standard deviation nb mutations (public)**	**Number of samples**	**Median followup (months)**	**Number of events**
ACC	324.34	110.58	467.43	240.88	77	39.22	27
BLCA	732.18	350.38	886.65	424.23	354	17.21	153
BRCA	472.91	121.88	981.68	372.65	931	27.00	123
CESC	921.07	366.51	2869.13	1323.24	275	21.42	67
CHOL	358.97	100.97	505.33	225.34	35	12.65	15
COADREAD	2085.67	621.14	6247.55	1708.77	458	21.42	93
DLBC	568.30	203.68	276.72	123.15	37	24.67	5
ESCA	707.39	247.19	560.34	251.02	180	13.02	75
GBM	790.05	245.66	2583.80	1191.70	327	11.27	246
HNSC	454.21	201.73	543.69	271.16	445	20.96	185
KICH	209.32	50.12	264.68	142.42	60	85.66	8
KIRC	330.32	73.08	338.69	52.35	271	36.33	64
KIRP	280.86	82.57	130.50	37.90	242	25.13	37
LGG	212.73	76.89	1462.32	770.54	455	20.04	115
LIHC	511.53	157.00	445.93	174.27	347	19.25	117
LUAD	892.89	381.00	985.35	393.91	433	22.01	146
LUSC	909.66	382.83	686.09	289.58	423	22.27	169
MESO	203.23	47.08	114.19	48.74	78	NA	0
OV	593.79	160.15	562.30	152.24	390	31.34	227
PAAD	560.23	203.05	3780.28	1828.93	150	15.14	83
PCPG	78.70	14.09	16.71	7.43	141	NA	0
PRAD	171.30	61.78	824.76	467.50	458	30.80	8
SARC	424.27	130.72	723.66	309.04	210	30.96	81
SKCM	1876.97	886.84	2621.06	1204.90	423	35.28	184
STAD	989.96	455.07	1822.53	909.56	418	14.01	163
TGCT	148.88	22.16	33.90	11.64	128	43.05	3
THCA	120.98	16.22	95.88	14.59	467	31.01	12
THYM	253.69	36.40	175.18	83.19	121	39.17	8
UCEC	4647.22	1791.01	12341.73	4832.10	487	30.12	80
UCS	680.43	198.00	1743.77	738.28	53	NA	0
UVM	88.54	24.60	96.34	62.57	80	27.52	11

Table B.3 – Characteristics of the TCGA cohort used in this study.

tcga_name	ACC	BLCA	BRCA	CESC	CHOL	COADREAD	DLBC	ESCA	GBM	HNSC	KICH	KIRC	KIRP	LGG	LIHC	LUAD	LUSC	MESO	OV	PAAD	PCPG	PRAD	SARC	SKCM	STAD	TGCT	THCA	THYM	UCEC	UCS	UVM
SBS1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SBS2	1	1	1	1	1	1	1	1	0	1	1	1[a]	1	1	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1
SBS3	1	1	1	0	0	0	0	1	0	1[d]	0	1[a]	0	0	1	0	0	0	1	1[b]	1	1[b]	1	0	1[c]	1	0	0	1[b]	0	1
SBS4	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1[e]	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0
SBS5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SBS6	0	0	1[f]	1	1[g]	1[h]	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1[i]	0	1[j]
SBS7a	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	0	1
SBS7b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0
SBS7c	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0
SBS7d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	1	0	0	1
SBS8	0	0	1[f]	0	1[g]	0	0	0	1[b]	0	1	0	0	0	0	0	0	1[b]	0	0	1[k]	1	0	0	0	0	1	0	0	0
SBS9	0	0	0	0	0	0	1[l]	0	0	0	0	1[a]	0	0	1	0	0	0	0	0	1[k]	0	0	0	0	0	0	0	0	0
SBS10a	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	1	0	0
SBS10b	1	0	1	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	1	1	1	0	0	0	1	1	1	1	1	0
SBS11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS12	0	0	0	0	0	0	0	0	0	1	0	0	0	1[e]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1[j]
SBS13	1	1	1	1	1	1	1	1	0	1	1[b]	1[a]	1	1	0	1	1	1	1	1	1	0	1	0	1	1	1	0	1	1	1
SBS14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
SBS15	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1
SBS16	0	0	0	0	1[g]	0	0	1	0	0	0	0	0	1	0	0	0	1[m]	0	1	1[k]	0	0	0	0	1	1	0	0	1[j]
SBS17a	0	0	1[f]	0	1[g]	1	1	1	0	0	1	1[a]	0	0	1[e]	0	0	0	0	1[b]	1	0	0	0	1	1	0	1	0	0	0
SBS17b	0	0	1[f]	0	1[g]	1	1	1	0	0	0	1[a]	0	0	1[e]	0	0	0	0	1[b]	0	0	0	0	1	0	0	0	0	0	0
SBS18	0	0	1[b]	0	0	1[b]	1[l]	1	0	1[b]	0	0	0	0	1[b]	0	1[b]	0	0	1[b]	0	1[b]	1	0	1[b]	0	1[b]	1	1[b]	0	0
SBS19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0
SBS20	0	0	1[f]	0	1[g]	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
SBS21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS22	0	0	0	0	1[g]	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
SBS23	0	0	0	0	0	0	0	0	0	0	0	0	0	1[e]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS24	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
SBS25	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS26	0	0	1[f]	0	0	1[h]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
SBS29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
SBS30	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
SBS32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
SBS33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1[b]	0	0	0	0	0	0	0	0	0	0
SBS35	0	0	0	0	0	0	0	0	0	0	0	0	0	1[b]	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
SBS36	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1[b]	0	0	0	0	0	0	0	0	0	0	0	0
SBS37	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1[b]	0	0	0	0	0	0	0	0	0	0
SBS38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0
SBS39	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1[b]	0	0	0	0	0	0	1	0	0	0	0	0	0
SBS40	0	0	1[b]	0	1	1	1	0	1	0	1	1	1	0	1	1	0	1	1	0	0	1	1	0	1	1	0	1	1	0	0
SBS41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS42	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0	0	1
SBS43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS44	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
SBS45	0	1	0	0	0	0	0	1	1	0	1	1	1	0	1	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0
SBS46	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
SBS48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS49	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0
SBS50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS52	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
SBS53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
SBS55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
SBS58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0
SBS59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1
SBS60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table B.4 – Table of presence of signatures in the different cancer types. A 1 indicates the presence of the signature, and a 0 an absence. The background color indicates whether the presence of the signature comes from SigProfiler detection in the TCGA (lavender) or from the literature (green), in which case, the matching paper is referenced in the table cell.

[a] Warsow et al. [2018]
[b] Dentro et al. [2018]
[c] Liu et al. [2018b]
[d] Verhagen et al. [2018]
[e] Letouzé et al. [2017]
[f] Nik-Zainal et al. [2016]
[g] Shibata et al. [2018]
[h] Roerink et al. [2018]
[i] Ashley et al. [2019]
[j] Royer-Bertrand et al. [2016]
[k] Espiritu et al. [2018]
[l] Ren et al. [2018]
[m] Macintyre et al. [2018]

Appendix C

Participation to the Dream challenge for ensemble variant calling

Joint work with Paul Deveau.

During the early months of my PhD, I had the occasion through my own experience with whole exome sequencing (WES) data to verify the difficulty and instability of variant calling, leading to subsequent difficulties to obtain robust intra-tumor heterogeneity estimates. Paul, a fellow PhD student had the same feeling, and we were both interested when a DREAM Challenge on that exact topic opened. Our team's name is **BDD**.

The ICGC-TCGA DREAM Somatic Mutation Calling Meta-pipeline Challenge (SMC-DNA Meta) lasted from October 2015 to March 2016 and was launched as the sequel of the ICGC-TCGA DREAM Somatic Mutation Calling Challenge (SMC-DNA) [Ewing et al., 2015]. This first challenge organized from December 2013 to August 2016, included several rounds and both real and synthetic datasets. The goal of this initial challenge was to evaluate algorithms performing variant calling, and to provide a good benchmark tool to the community. Before the challenge started, an analysis of the results of the first rounds was performed and conclusions were published Ewing et al. [2015], highlighting combination of results from several variant callers as a potential improvement: the organizers found that, for each sub-challenge, the majority vote of the five best ranked submissions systematically outperformed the best individual submissions. Moreover, an expected benefit of this strategy was also to reduce the sensitivity of variant calling to hyper-parameters which are very hard to fine-tune in real-life application as no ground truth is available. The objective of this challenge was hence to set up more advanced combination strategy to improve variant calling performances.

C.1 Description of available data

The organizers of the challenge provided for 4 synthetic datasets, called `IS1` to `IS4`. For each dataset, the data consist in:

- All genomic positions called by at least one variant caller in SMC-DNA Challenge.
- For each position, the presence/absence binary status per variant calling pipeline submitted in the SMC-DNA Challenge. For each pipeline, only the submitting team was provided, but no further details on the exact content of the pipeline.
- 13 genomic features for each position, such as base quality, reference and variant read counts, mapping quality, and strand bias.
- The true status for each position.

Synthetic datasets 1 to 4 (noted IS1–4) are of increasing difficulty, with addition of contamination by non-mutated normal cells, structural variants, and subclonal variants. Main

characteristics of datasets IS1–4 are summarized in table C.1. Each of these datasets was generated for a different sub-challenge, at different time points, allowing participants to modify their variant calling pipelines between sub-challenges based on their latest results. Hence, there is no indication to assume that the different submissions from the same team, either in the same or in different sub-challenges are similar.

In addition, 10 real normal-tumor pairs are provided, 5 are prostate cancers, and 5 pancreatic cancers. The same information is provided, except for the ground truth. These real datasets are evaluated on a separate leaderboard. The organizers were planning to provide independent experimental verifications of mutational status at numerous positions to evaluate submissions.

dataset	Number of calls	number of true positives	number of submitted callers
IS1	214541	3535	119
IS2	51108	4303	69
IS3	22884	7709	67
IS4	129091	15163	223

Table C.1 – Characteristics of synthetic datasets from the DREAM Challenge

C.2 Materials and Methods

C.2.1 Selection of pipelines

Both for synthetic and real data, the DREAM Challenge limits the number of pipelines to either 5 or 50, although running 50 variant callers to get a result seems unrealistic in terms of computational time. We implemented two different strategies:

Maximize consensus , where we constructed first a simple majority vote consensus among all available methods, and then selected the pipelines closest to that consensus using the Jaccard distance.

Maximize recall , where we selected with a greedy procedure the pipelines in order to maximize the number of considered positions.

An intermediate approach , that consists in selecting some of the pipelines with the first strategy, and some with the second.

C.2.2 Feature engineering

We have designed additional variables to better account for every aspect of variant calling: the CG content of a 50bp window around the variant position, and the homopolymer rate, defined by the sum of squared homopolymer lengths normalized by the length of the sequence. For instance, the sequence "AAATTGAGG" has an homopolymer rate of $\frac{3^2+2^2+1^2+1^2+2^2}{9} = \frac{19}{9} \approx 2.11$. These features can help detect error-prone regions, due to a lower coverage or the occurrence of polymerase slippage in homopolymer regions.

C.2.3 Implemented algorithms

The main difficulty with the Challenge setting is that each dataset comes from a separate sub-challenge of the SMC-DNA Challenge, therefore, there is no intersection between the calling pipelines run on each dataset, so any model learnt using results from variant callers as features on one dataset cannot be applied to a different dataset.

Three main approaches were tested:

Aggregation: for each dataset aggregate predictions from available variant caller results with different strategies (varying threshold for vote, etc), and apply this prediction to the dataset itself

Autotrain: for each dataset use an aggregation of prediction as labels to train a supervised classifier with genomic features and variant callers results, and apply this prediction to the dataset itself.

deepLearning: use IS4 dataset as a training sample, with several supervised Random Forest Classifiers trained on different combinations of genomic and variant caller features, and finally combined by a last supervised Random Forest Classifier.

Hence, in the two first approaches, only the available calls are used to provide a new prediction, while in the third approach, the calls from the previous challenge are used to design a new feature, the proportion of pipelines calling each position, which is concatenated with the 13 genomic features provided, and the new features described in section C.2.2 and provided as input to the Random Forest Classifier. This latter approach was mainly designed by Paul Deveau.

We have designed an automated framework to evaluate new algorithms on the four test datasets using a fixed train and test split for the four synthetic datasets by ourselves, as the number of allowed submissions is limited.

C.3 Results

The results on this "local leaderboard" are shown in figure C.1. Overall, the "aggregation" models exhibit very stable performances across all attempts, and are close to the best "deepLearning" runs. It appears that **autotrain** baseline models perform very badly, so we have not explored them any further. We have tried a few strategies of aggregation (majority vote, more than 80 % of callers etc).

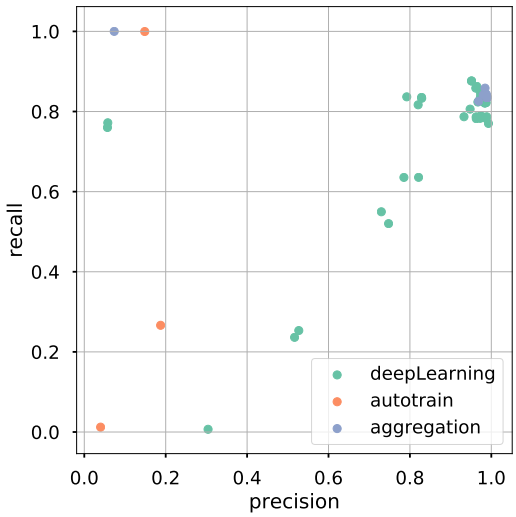

Figure C.1 – Precision and recall of different methods on synthetic datasets, divided in train and test sets, our "private leaderboard". Each point represents one of our trained model, and the color represents the category of algorithm.

The dream challenge presented four distinct leaderboards (synthetic and real tumors, with 5 or 50 callers), with each time a score averaged on all available datasets. In FigureC.2, results of the two leaderboards with 50 callers are presented, as the global behavior is the same with 5 pipelines. One important thing is that the performance metric used to evaluate submissions changes between the two settings. Altogether, all teams have obtained very close scores. Regarding our models, "deep learning" approaches are slightly better on the synthetic dataset, but "aggregation" runs exhibit better specificity on the real data.

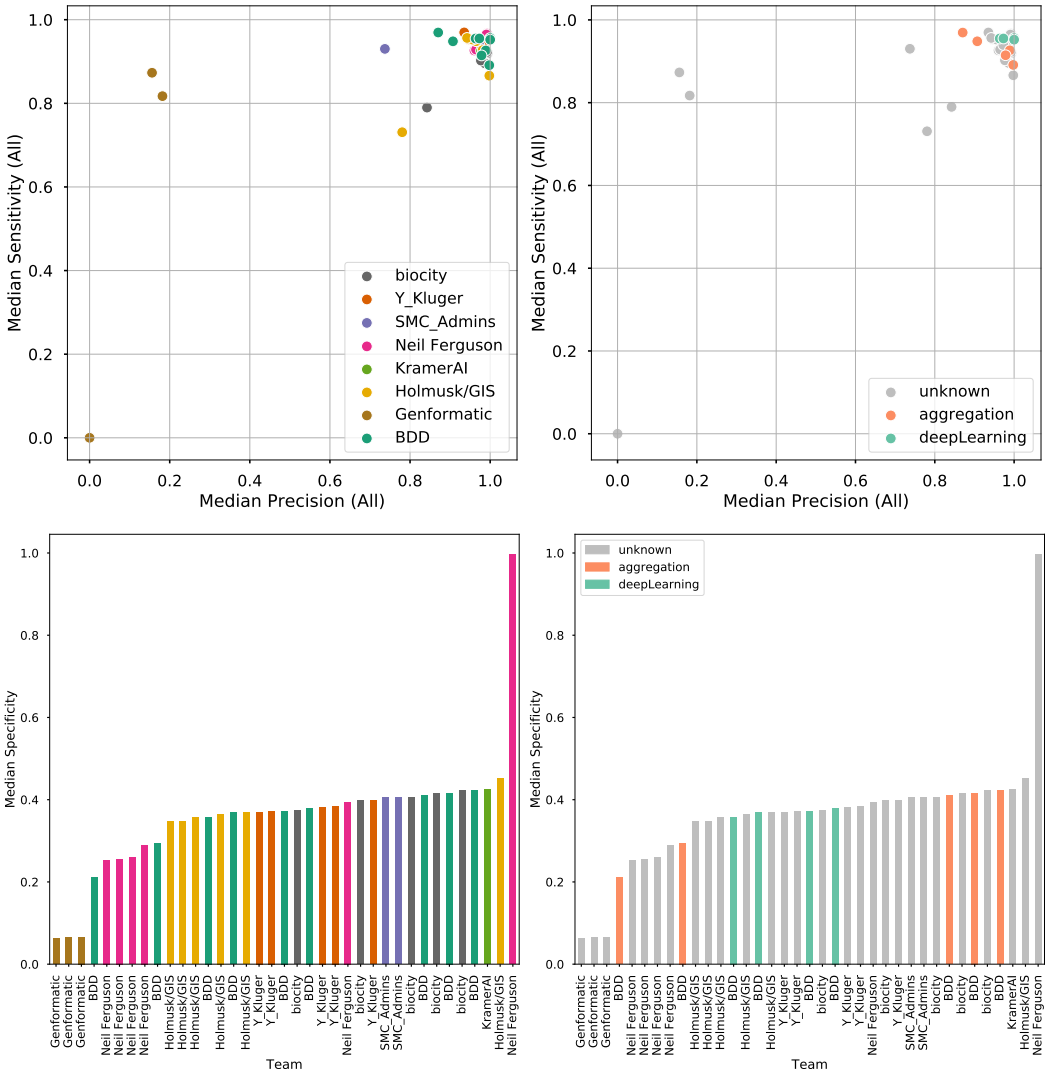

Figure C.2 – Results from the challenge leaderboard on simulated samples (top), and real tumors (bottom). On left panels, we see all submissions with one color per Team, and on right panels, we see all submissions in gray, except ours, separated by used approach. We see that "Deep Learning" strategies performed best on the synthetic dataset, and "aggregation" on the real tumors, but with very close scores that do not really allow us to conclude.

C.4 Discussion

Our team **BDD** has not won in any category, but we are quite close to the best scores. We see that "Deep Learning" strategies performed best on the synthetic dataset, and "aggregation" on the real tumors, but with very close scores that do not really allow us to conclude. We note however that the gap in performance of the "Deep Learning" strategies is small compared to the additional complexity of the models. The difference of ordering of the methods performances between real and simulated data can be explained in two ways: the difference in metrics in the two cases, or the (in)ability of the simulated datasets to recapitulate the patterns of real data variant calling.

The official results of the challenge have not been published (either in a journal or on the dream challenge website). Unfortunately, the experimentally-verified positions on real data, which could have been used to further test models have not been published either. They could have been very interesting to further develop methodologies on the topic, as was done in Kim et al. [2014], where ground truth was available on a part of a real dataset, and features were homogeneous (i.e. the same set of callers is run on all data).

Appendix D

Supplementary for intra-tumor heterogeneity methods review

Table D.1 – Characteristics of ITH methods. The methods are sorted by publication date. This table completes Figure 2.3, where more binary features were extracted, whereas here, more detailed descriptions of the methods and underlying algorithms are provided.

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
TuMult	2010/7/22	Letouzé et al. [2010]	no	yes (breakpoints and segmented copy number)			http://bioserv.rpbs.univ-paris-diderot.fr/services/TuMult/ALGORITHM.html	TuMult reconstructs the evolution of the cancer genomes between different samples from the same patient. The method relies on shared breakpoints across samples. This method treats each sample as homogeneous and uses standard parsimony phylogenetic reconstruction				greedy bottom-up agglomerative clutering
GRAFT	2011/10/12	Greenman et al. [2012]	no	yes			https://www.sanger.ac.uk/science/tools/graft	GRAFT analyses and reconstitutes the succession of CNAs using copy number information, and breakpoints with a graph-based approach. A second step resorts to MCMC to infer the timing of mutations through probabilistic modeling.				construction of a graph, then traversing of the graph, then combinatorial ordering, then MCMC for time estimation
DPclust	2012/5/25	Nik-Zainal et al. [2012]; Dentro et al. [2017]	yes	no			https://github.com/Wedge-Oxford/dpclust	DPClust models estimated CCF of SNVs, corrected for copy number using a DP, the posterior is estimated through MCMC, and peak detection is applied to get the final clustering. A post-processing step removes small clones.				MCMC
Battenberg	2012/5/25	Nik-Zainal et al. [2012]	no	yes			https://github.com/cancerit/cgpBattenberg	Battenberg uses BAFs at heterozygous SNPs, pahsed by resorting to known haplotypes, for instance from the 1000 genomes project, and refines purity and ploidy estimates for each segment, thus being able to call subclonal segments				segmentation, then resegmentation, then copy number fitting
MATH	2012/10/15	Mroz and Rocco [2013]	yes	no			NaN	MATH is a single value score measuring the "diversity" of the SNV VAF distribution by MATH = 100 * MAD/median (median absolute deviation)				NaN
PurBayes	2013/6/6	Larson and Fridley [2013]	yes	no			https://CRAN.R-project.org/package=PurBayes	PurBayes is a multinomial mixture of binomial distributions accounting for read counts of SNVs in copy-number neutral regions, and sample purity. Estimation is conducted with MCMC and the number of components is set via a penalized expected deviance criterion.				MCMC
TrAp	2013/7/27	Strino et al. [2013]	yes	no			https://sourceforge.net/projects/klugerlab/files/TrAp/	TrAp is an algorithm to infer a tree structure for pre-clustered SNVs from a tumor sample. Tree inference relies on constraints such as parsimony and shallowness, with a first step identifying "easy" relationships, and then a brute force step enumerating all possibilities to find the tree best explaining the estimated CCFs of each cluster.				NaN

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
THetA	2013/7/29	Oesper et al. [2013]	no	yes (segmented read counts)			https://github.com/raphael-group/THetA	THetA enumerates the potential copy number profiles for each clone, and estimates the mixtures in each case (convex optimization). The overall complexity is $O(m^k)$, with m the number of segments, and k the number of populations. The model is chosen by BIC criterion, with respect to the likelihood of raw read counts per segments (with correction for segment legnth and mappabiity).				NaN
cancerTiming	2013/9/23	Purdom et al. [2013]	yes	yes			https://cran.r-project.org/web/packages/cancerTiming/index.html	CancerTiming relies on SNVs in CNA regions to provide a relative timing of the CNA occurrence. However, each segment is analysed separately, and no population estimation is provided				EM
EXPANDS	2013/10/30	Andor et al. [2014]	yes	yes			https://cran.r-project.org/package=expands	ExPANdS characterizes subclonal populations by estimating for each SNV a probability distribution of its CCF, accounting for copy number alterations, and then clusters those CCFs using hierarchical clustering. A statistical test is then used to filter non-significant clusters.				hierarchical clustering
PhyloSub	2014/2/1	Jiao et al. [2014]	yes	no			https://github.com/morrislab/phylosub/	PhyloSub relies on a Tree-structured stick-breaking process (tree-based mixture model) to model raw SNV counts. The parameters of the model are inferred using MCMC, and the sampling process respects evolutionary constraints of the tree.				MCMC
PyClone	2014/3/16	Roth et al. [2014]	yes	yes			https://shahlab.ca/projects/pyclone/	PyClone is a Dirichlet Process mixture model that maximizes the likelihood of observed SNV read counts, accounting for copy number (only one variant genotype is allowed). Inference is performed by MCMC.				MCMC
SciClone	2014/4/12	Miller et al. [2014]	yes	no			http://github.com/genome/sciclone	SciClone models the subclonal structure of one or several tumor samples as a mixture of beta distributions over SNVs CCFs for SNV in copy number unaltered regions. The parameters are inferred through a variational inference strategy.				variational inference
MEDICC	2014/4/17	Schwarz et al. [2014]	no	no			https://bitbucket.org/rfs/medicc	This method uses several tumor samples from the same patient; each sample is considered homogeneous. The algorithm is based on automata (finite state transducer) to find the shortest path of transformations (amplifications or deletions) to transform an integer copy number profile into another. Classical algorithms for phylogenetics reconstruction from distance (in this case the Fitch-Margoliash method) can be applied to recover the tree.				phylogenetic reconstruction from distance

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
OncoSNP-SEQ2	2014/4/30	Yau [2014]	no	yes (snp position read counts)			https://sites.google.com/site/oncosnpseq/	OncoSNP-SEQ2 models read counts covering SNPs, with a factorial HMM. Inference is made using Viterbi algorithm. Careful exploration of the copy-number detection sensitivity association with false positive clones is conducted.				viterbi algorithm
cloneHD	2014/5/29	Fischer et al. [2014]	yes	yes			https://github.com/andrej-fischer/cloneHD	CloneHD is an ensemble of three coupled HMM models for copy number, BAF and SNVs, that jointly represent the subclonal structure of tumor samples. Inference is achieved using a forward-backward algorithm and BIC is used for model selection.				forward-backward algorithm
Rec-BTP	2014/6/11	Hajirasouliha et al. [2014]	yes	no			http://compbio.cs.brown.edu/software/	Rec-BTP proposes a combinatorial formulation of the subclone phylogeny problem. Rec-BTP takes as input SNV clusters along with their estimated CCF and adds nodes to provide a conflict-free tree, respecting the assumption that the CCF of a clone is equal to the sum of its children CCFs				NaN
Clomial	2014/7/10	Zare et al. [2014]	yes	no			https://bioconductor.org/packages/Clomial/	Clomial formulates the subclonal as a probabilistic matrix factorization problem, with one matrix representing the subclone (binary) genotypes, and another the mixture of clones in the different avaialable tumor samples. Estimation is done using an EM algorithm with a quasi-Newton method BFGS-B. The number of clones can be set by the user or chosen using BIC.				EM
TITAN	2014/7/24	Ha et al. [2014]	no	no			https://shahlab.ca/projects/TitanCNA/	TITAN jointly models subclonal populations and their associated copy number states as a factorial HMM model. The parameters are inferred by an EM algorithm, with observed data the counts of total and minor allele reads at SNP positions, normalized for CG content and mappability. The number of clones is chosen with the sdbw index.				EM
CLONET	2014/8/15	Prandi et al. [2014]	no	yes			https://bitbucket.org/deid00/clonet	CLONET is based on a local optimization which estimates the purity and ploidy of each CNV separately, to identify a few clonal events and provide heterogeneity-robust estimates.				CCF estimation of each CNV
SubcloneSeeker	2014/8/26	Qiao et al. [2014]	yes	yes			https://github.com/yiq/SubcloneSeeker	SucloneSeeker proceeds in 4 main steps: (i) estimate a CCF fraction for all alterations (SNVs and CNVs), (ii) (multidimensional) clustering of alterations (if several samples), (iii) construction of a tree for each sample, through enumeration of all possibilities and (iv) merging of the resulting trees.				multi-dimensional clustering, and enumeration of possible trees

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
BreakDown	2014/9/8	Fan et al. [2014]	no	yes			https://bioinformatics.mdanderson.org/public-software/breakdown/	BreakDown relies on three data types to infer the CCF of each CNV from WGS data: coverage by normal reads, discordant reads (for paired-end sequencing, read pairs not in the expected order or orientation), and soft-clipped reads (reads overlapping a non-reference jonction).				closed-form MLE
CHAT	2014/9/25	Li and Li [2014]	yes	yes			https://sourceforge.net/projects/clonalhetanalysistool/	CHAT first estimates subclonal CNAs, and then uses those to model the observed SNV VAFs using a Dirichlet process Gaussian mixture model inferred with MCMC.				MCMC
THetA2	2014/10/8	Oesper et al. [2014]	no	yes			https://github.com/raphael-group/THetA	THetA2 builds on THetA, but introduces further constraints on the possible subclonal profiles (matrix C), thus allowing the algorithm to accomodate more segments (and hence WES data), and more subclonal populations.				NaN
BayClone	2015/1/4	Sengupta et al. [2015]	yes	no			http://health.bsd.uchicago.edu/yji/soft.html	BayClone relies on a categorical Indian Buffet Process to model the SNV read counts based on three possible latent states for each SNV in each clone haplotype: non mutated, heterozygous or homozygous. Inference is performed using MCMC.				MCMC
CITUP	2015/1/6	Malikic et al. [2015]	yes	no			https://sourceforge.net/projects/citup/	CITUP infers clonal tree phylogeny from multiple samples using a combinatorial method associated with quadratic integer programmation or an iterative heuristic, to fit observed SNV VAFs. BIC criterion is used to choose a minimal tree structure.				quadratic integer programmation or an iterative heuristic
MixClone	2015/1/21	Li and Xie [2015]	no	yes			https://github.com/uci-cbcl/MixClone	MixClone models subclonal populations based on segment coverage (CNVs) and heterozygous SNP read counts, via a generative probability mixture model. Inference is performed by EM and the number of clones is chosen heuristically.				EM
BitPhylogeny	2015/2/13	Yuan et al. [2015]	yes	yes			https://bitbucket.org/ke_yuan/bitphylogeny	BitPhylogeny attempts at reconstructing a tree and infers the number and genotypes of its nodes from various types of data, using a tree-structured stick-breaking process, and adapting the "emission probability distribution" depending on the data. Inference is achieved with MCMC sampling.				MCMC
PhyloWGS	2015/2/13	Deshwar et al. [2015]	yes	yes			https://github.com/morrislab/phylowgs	PhyloWGS builds up on PhyloSub model, and takes as input CNVs in addition to SNVs. The model hence accounts and corrects the parameter estimation for SNVs affected bu CNVs.				MCMC

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
LICHeE	2015/5/6	Popic et al. [2015]	yes	no			http://viq854.github.io/lichee/	LICHeE infers lineage trees for multiple bulk tumor samples, with a first clustering step, then construction of a graph based on the ISA and the pigeonhole principle, and finally finds the best tree from all possible spanning trees.				evaluation of all possible spanning trees
AncesTree	2015/6/15	El-Kebir et al. [2015]	yes	no			https://github.com/raphael-group/AncesTree	AncesTree's goal is to reconstruct the evolutionary history of a tumor from one or several samples. First, a clustering step is performed to group SNV depending on VAF and presence patterns across samples, before appliying a MILP algorithm find the best tree				MILP
MAD Bayes	2015/7/16	Xu et al. [2015]	yes	no			https://web.ma.utexas.edu/users/yxu/software.html	MAD Bayes models SNV read counts as a mixture of binomial distributions, and present an extension of the beta process means algorithm to infer the parameters, which is faster than MCMC and parallelizable.				MCMC
SCHISM	2015/10/5	Niknafs et al. [2015]	yes	yes			https://karchinlab.org/apps/appSchism.html	SCHISM implements a genetic algorithm to reconstruct the phylogenetic history from previously clustered SNV CCFs (ideally corrected for CNV), by minimizing violoations to the ISA rule, and to violations to the independantly determined ancestor-descendant relations of each SNV pair, using a genetic algorithm.				genetic algorithm
BubbleTree	2015/11/17	Zhu et al. [2016]	no	no			https://www.bioconductor.org/packages/release/bioc/html/BubbleTree.html	BubbleTree graphically matches logRatio score and BAF to an allele-specific genotype and a cellular prevalence after adjustment for tumor purity and ploidy. Further post-processing can lead to the reconstruction of a tree				closed-form formulas
BayClone2	2016/1/12	Lee et al. [2016]	yes	no			https://cran.r-project.org/web/packages/BayClone2/index.html	BayClone2 extends the cIBP model of BayClone to model CNVs overlapping SNVs, not accounting however for input CNVs.				MCMC
bayesian feature allocation	2016/2/17	Lee et al. [2015]	yes	no			NaN	Bayesian feature allocation model to infer haplotypes and their proportion in tumor samples from SNV read counts, using an Indian Buffet Process mixture of binomial distributions, with MCMC to infer parameters.				MCMC
OncoPhase	2016/3/31	Chedom-Fotso et al. [2016]	yes	yes			https://github.com/chedonat/OncoPhase	OncoPhase estimates the CCF of each SNV separately, by considering the ratio of variant reads at that position compared to the ones of phased germlines SNPs, and correcting for copy number. However, the SNVs are not grouped into clones or genotypes.				closed-form formulas

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
SPRUCE	2016/7/27	El-Kebir et al. [2016]	yes	yes			http://compbio.cs.brown.edu/projects/spruce/	SPRUCE infers an evolutionary tree by jointly model different possible combinations for SNVs and overlapping CNAs, and enumerate all compatible trees.				enumeration of all compatible trees
CloneCNA	2016/8/19	Yu et al. [2016]	no	no			http://bioinformatics.ustc.edu.cn/clonecna/	factorial HMM model (one chaine for the cluster, and one for the copy-number state) with BIC for the number of clones. Parameters estimated by EM. A limiting assumption is that only 1 aberrant phenotype exists per exon.				EM
Canopy	2016/9/13	Jiang et al. [2016]	yes	yes			https://cran.r-project.org/web/packages/Canopy/	Canopy infers the evolutionary phylogeny of a tumor using the SNV and CNA information from bulk sequencing tumor samples, using a probabilistic framework with MCMC inference				MCMC
MixPhy	2016/10/16	Hujdurovic et al. [2018]	yes	no			https://github.com/alexandrutomescu/MixedPerfectPhylogeny	The authors of MixPhy propose a faster heuristic to solve the problem as stated in Rec-BTP				heuristic
Cloe	2017/1/5	Marass et al. [2016]	yes	no			https://bitbucket.org/fm361/cloe/src/master/	Cloe is a bayesian hierarchical probabilistic model that represents observed read counts from one or several tumor samples as a mixture of clones from a phylogenetic tree. The tree probability model allows but penalizes violations to the ISA or to a single tumor origin assumption. Inference is performed using a MCMCMC algorithm, and selection of the number of clones is based on MAP.				MCMCMC
Treeomics	2017/1/31	Reiter et al. [2017]	yes	no			https://github.com/johannesreiter/treeomics	Treeomics takes as input multiple samples (tumor and metastases) from the same patient, and infers a tree compatible with potential migrations and sample locations, by assuming that all samples are homogeneous. A bayesian model allows to assess which SNVs are present in which samples, accounting for sequencing errors, and then a MILP algorithm is used to fit an envolutionary tree				MILP
PairClone	2017/2/24	Zhou et al. [2018]	yes	yes			http://www.compgenome.org/pairclone/ (broken...)	PairClone models pairs of phased SNVs to infer clone genotypes and abundances from one or several tumor samples, hence leveraging paired end sequencing data structure. PairClone was extended to jointly incorporates all SNVs and SNV pairs, and corrects for copy number. Inference is done by MCMC with parallel tempering				MCMC
TreeClone	2017/3/10	Zhou et al. [2019]	yes	no			http://www.compgenome.org/treeclone/	TreeClone extends PairClone to model a phylogenetic tree, with similar assumptions about the rest of the model. Though only copy-number neutral SNV pairs are considered				MCMC

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
CLImAT-HET	2017/3/15	Yu et al. [2017]	no	yes (snp position read counts)			https://github.com/USTC-HILAB/CLImATHET	Climat-HET models the known SNPs major and minor read counts as a factorial HMM to uncover intra-tumor heterogeneity. The counts can be adjusted for GC and mappability bias, hence allowing not to have a matched normal sample. The number of clones is selected using a custom-regularized BIC criterion				EM
CNTMD	2017/4/12	Zaccaria et al. [2017]	no	no			https://github.com/raphael-group/CNT-MD	CNTMD is a heuristic algorithm based on coordinate-descent paradigm that alternates LP and ILP steps to optimize the clone mixture matrix and the tree topology (genotype of each clone) alternatively				heuristic
CNT-ILP	2017/5/16	El-Kebir et al. [2017]	no	no			https://github.com/raphael-group/CNT-ILP	CNT-ILP considers the same setting and model as MEDICC and proposes a faster linear implementation				ILP
CTPsingle	2017/6/1	Donmez et al. [2017]	yes	no			https://github.com/nlgndnmz/CTPsingle	CTPsingle clusters SNVs in copy-number neutral regions using a DP, estimated through MCMC. The topology of the k clusters found in the first step is then inferred using a mixt integer linear programming algorithm				MILP (after probabilistic first step with MCMC for clustering)
PASTRI	2017/7/12	Satas and Raphael [2017]	yes	no			https://github.com/raphael-group/pastri	PASTRI proposes an efficient algorithm to reconstruct trees from the clonal structure deconvolution. PASTRI takes as input a posterior distribution of clone cancer cell fractions, and resamples to obtain a clustering compatible with a tree				importance sampling and MILP
GLClone	2017/7/20	Geng et al. [2017]	yes	no			not found	GLClone models read counts of SNVs (total and variant), and genotype/copy number at each SNV loci from average segment total read count (poisson distribution) as a hierarchical mixture model, inferred using variational inference. The strategy to choose the number of clones is not specified.				variational inference
ReMixT	2017/7/27	McPherson et al. [2017]	no	yes			https://bitbucket.org/dranew/remixt	ReMixT considers both copy number profile, and changes around a breakpoint to reconstruct subclonal copy number profiles, and their relative quantities. ReMixT is also able to reconstruct partial assemblies of rearranged subclonal genomes. Observed data are modeled as a generative hierarchical probabilistic and inference is carried out by a variational EM algorithm.				EM
SVclone	2017/8/4	Cmero et al. [2017]	yes + SV calls	no			https://github.com/mcmero/SVclone/tree/pymc2	SVclone can simultaneously cluster and determine the CCF of SNVs and SVs, using a dirichlet process mixture model inferred by MCMC. A major novelty consists in obtaining VAF estimates for SVs, and design of a filter to avoir false positive SVs				MCMC

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
ClonEvol	2017/9/11	Dang et al. [2017]	yes	no			https://github.com/ChrisMaherLab/ClonEvol	ClonEvol relies on clustered SNV CCFs, and estimates CCF confidence interval using bootstrapping, and then performs a phylogeny reconstruction, that accounts for noise in CCF estimates by enumerating possible architectures.				enumeration of all compatible trees
WSCUnmix	2017/10/23	Roman et al. [2017]	no	no			https://github.com/tedroman/WSCUnmix	WSCUnmix performs two steps by first grouping similar samples together (PCA + clustering), and then mixture deconvolution on each subgroup with a minimum spanning tree cost to get a parsimonious phylogeny, and finally groups all subtrees together by identifying similar clones in each.				minimum spanning tree
QuantumClone	2018/1/12	Deveau et al. [2018]	yes	yes			https://github.com/DeveauP/QuantumClone	QuantumClone is a finite mixture of binomial distributions to model observed SNV read counts, with a correction for the observed copy number. The number of mutated copies is also inferred, and a weight is attributed to each SNV to overrepresent SNVs in low-copy-number regions. Inference is done by an EM algorithm, with BIC to select the number of clones				EM
HetFHMM	2018/2/1	Rahman et al. [2018]	yes	yes			NaN	HetFHMM is a factorial HMM model, where each chain represents a clone genome, and the observations of each genomic location consist in SNV read counts and log ratio. The model infers the proportions of each clone, and the genotype states that best explain observations, thus modelling the genotype dependency between close SNVs. Inference alternates with an exponentiated gradient descent for proportions, and a MCMC with Gibbs sampling for genotypes.				MCMC
SIFA	2018/3/16	Zeng et al. [2019]	yes	yes			https://github.com/zengliX/SIFA	SIFA is a bayesian hierarchical model that deciphers the clone genotypes (in SNV and CNAs) and fractions, and phylogenetic relations from multiple WGS samples. The model is inferred via MCMC with parallel tempering. Model selection is achieved with a criterion based on bayes free energy				MCMC
MACHINA	2018/4/26	El-Kebir et al. [2018]	yes	no			https://github.com/raphael-group/machina	factorial HMM model (one chaine for the cluster, and one for the copy-number state) with BIC for the number of clones. Parameters estimated by EM. A limiting assumption is that only 1 aberrant phenotype exists per exon.				Sankoff algorithm

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
BayClone-C	2018/5/5	Dentro et al. [2018]	yes	yes			https://github.com/compgenome365/bayclonec	BayClone-C proposes a mixture of Gaussians to cluster the SNV CCFs corrected by copy number. The number fo clones is chosen by BIC, and a post-processing step merges close clusters using a ridgeline unimodal method				MCMC
palimpsest	2018/5/16	Shinde et al. [2018]	yes	yes			https://github.com/FunGeST/Palimpsest	Palimpsest relies on a binomial modelisation of variant read counts corrected for copy number to classify them as clonal or subclonal. Further options allow to characterize each subgroup in terms of mutational signature and structural variant timing				confidence interval computation
Sclust	2018/5/24	Cun et al. [2018]	yes	yes			http://www.uni-koeln.de/med-fak/sclust/Sclust.tgz	Sclust takes as input a pair of normal/tumor BAM files and a vcf, performs segmentation, and calling of copy number, allows for subclonal copy number, ie non-integer copy number for SNVs VAF normalization to CCF before clustering, but does not attribute CNAs to clones defined by SNVs. The clustering relies on a variational inference peak calling of the CCF histogram				variational inference
tusv	2018/6/27	Eaton et al. [2018]	no	no			https://github.com/jaebird123/tusv	tusv associates breakpoint calling to CNA profile to reconstruct a tree based on CNA-only data. Optimization of the clone mixture is done by simple constrained optimization, and alternates with the optimization of tree-compatible copy number profiles for each clone using integer linear programming.				linear programming
SuperFreq	2018/7/30	Flensburg et al. [2018]	yes	yes (BAM input)			https://github.com/ChristofferFlensburg/superFreq	SuperFreq calls CNAs and filters input SNVs from multiple tumor samples with or without matched normal sample. The clones are then inferred from a subset of high confidence SNVs using hierarchical clustering, in a way compatible with a tree structure (through post-processing of obtained clones), and then incorporates the low confidence SNVs				hierarchical clustering
CliP	2018/7/31	Yu et al. [2018]	yes	yes			https://github.com/wwylab/CliP	CliP sets the clustering of SNVs CCFs corrected for copy number as an optimization problem, where the objective function minimizes the difference between the estimated and the true CCF, and clustering is done by penalizing (Lasso, MCP, SCAD) the differences between the estimated CCFs for each SNV.				penalized cost function minimization
MIPUP	2018/8/8	Husić et al. [2019]	yes, binary	no			https://github.com/zhero9/MIPUP	MIPUP considers the problem of finding the minimum number of clones to be compatible with a perfect phylogeny, and stipulate that this is equivalent to finding an optimal branching in a direct acyclic graph (DAG), which is solvable by an ILP				ILP

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
GenoClone	2018/9/18	Zou et al. [2018a]	yes	no			http://augroup.org/GenoClone/GenoClone/GenoClone	GenoClone relies only on SNVs phased with germline SNPs to first infer the genotype of each SNP (1 or 2 mutated haplotypes), without CNV correction. This information is then used in a mixture setting to determines the composition and CCF of each subclone. The number of subclones is chosen through a fixed criterion.				monte carlo optimization
TargetClone	2018/11/29	Nieboer et al. [2018]	yes	yes			https://github.com/UMCUGenetics/targetclone	TargetClone takes as input the frequency of both SNPs and SNVs as input to infer copy number and genotype at all those positions from deep targeted sequencing from multiple samples. Alternate steps can modify the tree structure and the genotypes while accounting for phylogenetic and horizontal dependencies along the genome. A major limitation is the assumption that each sample corresponds to one clone/node in the tree				alternation of MLE and minimum spanning tree
CloneFinder	2018/12/1	Miura et al. [2018]	yes	yes			https://github.com/gstecher/CloneFinderAPI	CloneFinder starts with each tumor sample binary absence/presence representing a potential clone. The algorithm then alternates phases of estimation of clone proportions in each sample, and addition of extra-clone if needed, through phylogenetic inference of decomposition, until a stopping criterion of difference of estimated and true VAFs is reached.				NaN
ccube	2018/12/2	Yuan et al. [2018]	yes	no			https://github.com/keyuan/ccube	Ccube is a finite mixture of binomial distributions to model the observed read counts, and estimate the clones, their proportions and the multiplicity of each SNV. The inference is done via a variational EM algorithm, and the model with the best ELBO is chosen. Close clusters are then merged.				variational EM
p-SCNAClonal	2018/12/3	Chu et al. [2018]	no	no			https://github.com/Billy-Nie/pSCNAClonal	p-SCNAClonal first clusters segments with similar read depth after correction for the GC content, then further clusters according to BAF, and finally defines the copy number and clonal belonging for each cluster. In practice, 2 clusters are considered				EM
CloneDeMix	2018/12/12	Tai et al. [2018]	no	no			https://github.com/AshTai/CloneDeMix	mixture Poisson model which models the read count depending on inferred subclonal total copy number, CCF. vague untested extension to snvs in the supplementary, withtout further evaluation, or comment in the package doc				EM

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
HATCHet	2018/12/17	Zaccaria and Raphael [2018]	no	no			https://github.com/raphael-group/hatchet	HATCHet performs simultaneous matrix factorization to infer allele-specific copy number of multiple clones over multiple samples, and additionnally explicitely models whole genome duplication events. A final model selection step is implemented to select the number of clones and the WGD occurrence				Simultaneous matrix factorization
PhylogicNDT	2018/12/31	Leshchiner et al. [2019]	yes	yes			https://github.com/broadinstitute/PhylogicNDT	PhylogicNDT is a suite of tools to explore ITH, with a first step of clustering (Dirichlet process estimated via MCMC), then building of a tree, also via MCMC, with the possibility to alter the cluster. A number of other analyses including modeling tumor growth or neoantigen load or mutational signatures from the obtained clones are also implemented. Online methods are not available				MCMC
SeqClone	2019/1/5	Ogundijo and Wang [2019]	yes	no			https://github.com/moyanre/seqclone	SeqClone is a reimplementation of BayClone using sequencial monte carlo instead of MCMC.				sequential monte carlo
CALDER	2019/1/22	Myers et al. [2019]	yes	no			https://github.com/raphael-group/calder	CALDER leverages the pattern of co-occurrence of mutations along several longitudinal tumor samples to constraint the clustering of SNVs, and solves jointly the clustering and the tree structure inference using a mixt linear integer programming algorithm				MILP
CloneSeeker	2019/1/24	Zucker et al. [2019]	no	no			https://r-forge.r-project.org/projects/clonefinder/	cloneseeker proposes a bayesian generative model of allele-specific copy number and the multiplicity of mutations. It can work solely with SNV or CNA inputs, and optimizes parameters with a MAP approach				iterative search and selection of the MAP
tumor_clones	2019/1/25	Ogundijo et al. [2019]	yes	no			https://github.com/moyanre/tumor_clones	tumor_clones is a reimplementation of BayClone using sequencial monte carlo instead of MCMC, with the difference that 3 genotypes are allowed.				sequential monte carlo
RCK	2019/2/25	Aganezov and Raphael [2019]	no	yes			https://github.com/raphael-group/RCK	RCK solves the Cancer Genome Karyotype Reconstruction Problem using a mixed-integer linear program, from clonal and allele-specific CNVs from a third-party program				MILP
MOBSTER	2019/3/26	Caravagna et al. [2019]	yes	no			https://github.com/caravagn/MOBSTER (private for now)	MOBSTER is a probabilistic graphical model that accounts for the neutral accumulation of SNVs additionnaly to clonal expansion steps. The potential of confounding by those neutral mutations might be event larger with multiple samples. The parameters are estimated using EM, with the number of mixture components chosen with ICL criterion.				EM

...continued

id			input				description		method			
method	**date published**	**reference**	**SNV**	**CNV**	**one sample**	**mult. samples**	**tool availability**	**short description**	**graph or combinatorial**	**optimization**	**probabilistic**	**algorithm**
TRaIT	2019/4/25	Ramazzotti et al. [2019]	no	no			https://github.com/BIMIB-DISCo/TRaIT	TRaIT takes as input any genomic variation (SNVs, CNAs...) as binary vectors of presence/absence from multiple bulk or single cell tumor samples, and then applies diverse tree inference algorithms (Edmonds, Chow-Liu, Gabow, Prim) to infer a tree of all events.				diverse tree inference algorithms (Edmonds, Chow-Liu, Gabow, Prim)
BAMSE	2019/6/6	Toosi et al. [2019]	yes	no			https://github.com/HoseinT/BAMSE	BAMSE takes SNV read counts as input, performs a clustering of VAFs with k-means, and then uses a bayesian model, optimized via convex optimization to obtain the best tree architecture, with a ad-hoc filtering of potential tree in cases k>=6. The best trees are returned, potentially with varying k.				convex optimization, and heuristic to search the tree space
CLONETv2	2019/6/21	Prandi and Demichelis [2019]	no	yes			https://cran.r-project.org/package=CLONETv2	CLONETv2 extends CLONET to manage input data from all sequencing platforms, and report estimates of the CCF for all CNVs				CCF estimation of each CNV
lbdp	2019/8/1	Dinh et al. [2019]	yes	no			NaN	This method relies on evolutionary population genetics modeling to fit the observed SNV VAFs to observed read counts, as a mixture of binomials.				sampling
EXACT	2019/8/22	Ray et al. [2019]	yes	no			https://github.com/surjray-repos/EXACT	EXACT takes pre-clustered SNV CCFs as input, and relies on GPU parallelization of tree cost estimates to compute the score of each tree, and hence return an exact solutionof the best tree for the perfect phylogeny problem, within hours.				evaluation of all possible trees
AMTHet	2019/9/7	Consul and Vikalo [2019]	no	no			NaN	AMTHet formulates the problem as a matrix factorization with a vector of the proportions of clones in the sample, and a integer-valued matrix L representing the copy number profiles of each clone. Optimization alternates the two matrices, with a branch and bound to optimize L				matrix factorization and branch and bound
MAGOS	2019/10/2	Ahmadinejad et al. [2019]	yes	no			https://github.com/liliulab/magos	MAGOS provides a framework to better model SNV CCF variance in association with sequencing depth when attempting to distinguish subclones. The first step consists in building a hierarchical clustering of SNVs with a custom-designed distance, and then to perform a statistical test to partition the SNVs into clusters.				hierarchical clustering + statistical test
Meltos	2019/10/4	Ricketts et al. [2019]	yes	yes			https://github.com/ih-lab/Meltos	Meltos uses a tree built from reliable SNVs to refine SV calls, and then places them in the same tree.				EM and heuristic

RÉSUMÉ

L'obtention du répertoire des gènes de cancer mutés a été déterminant pour notre compréhension de la tumorigénèse. Cependant, les efforts menés pour caractériser les cancers au niveau génétique ne sont pas suffisants pour prédire la survie des patients, ou leur réponse aux traitements, ce qui est essentiel pour améliorer leur prise en charge. Cet échec est en partie attribué au caractère évolutif des cancers. En effet, comme toute population biologique capable d'acquérir des changements héréditaires, les cellules tumorales sont soumises à la sélection naturelle et la dérive génétique, résultant en une structure mosaique, dans laquelle coexistent plusieurs sous-clones ayant des génomes et des propriétés différentes. Cela a d'importantes conséquences sur les traitements anti-cancéreux, puisque ces sous-populations peuvent être sensibles ou résistantes à différentes thérapies, et de nouveaux phénotypes résistants peuvent continuer d'apparaître alors que la maladie continue à progresser.

Un nombre importants de méthodes mathématiques ou statistiques a été développé pour détecter et mesurer l'hétérogénéité intra-tumorale (ITH), mais aucune évaluation systématique de leurs performances et de leur application clinique potentielle n'a été effectué. Notre première contribution a donc été de réaliser une étude des approches existantes pour détecter l'hétérogénéité intra-tumorale, pour permettre de naviguer plus facilement entre les idées sous-tendant ces approches. Nous avons aussi proposé un cadre pour analyser la robustesse de ces approches, et leur usage potentiel pour la stratification des patients.

Cette enquête approfondie nous a aussi permis d'identifier un type de données encore non exploité pour la reconstruction de l'hétérogénéité intra-tumorale, et notre seconde contribution vise à combler ce manque. En effet, au-delà de la fréquence observée d'une mutation somatique dans un échantillon tumoral, qui permet de distinguer plusieurs clones, le contexte nucléotidique d'une mutation révèle les processus mutationnels causaux et non observables. Nous montrons, à la fois avec des données simulées et réelles la possibilité de modéliser ces deux aspects de l'évolution tumorale conjointement.

En conclusion, nous mettons en évidence le besoin de renforcer l'intégration de données de nature ou d'origine multiples pour exploiter pleinement le potentiel de l'évolution tumorale dans la prise en charge clinique du cancer.

MOTS CLÉS

Inférence bayésienne - Evaluation des performances - Génomique des cancers - Séquençage à haut-débit

ABSTRACT

Accessing the repertoire of cancer somatic alterations has been instrumental in our current understanding of carcinogenesis. However, efforts in genomic characterization of cancers are not sufficient to predict a patient's outcome or response to therapy, which is key to inform their clinical management. This failure is partly attributed to the evolutionary aspect of cancers. Indeed, as any biological population able to acquire heritable transformations, tumor cells are shaped by natural selection and genetic drift, resulting in a mosaic structure, where several subclones with distinct genomes and properties coexist. This has important implications for cancer treatment as those subpopulations can be sensitive or resistant to different therapies, and new resistant phenotypes can keep emerging as the diseases progresses further.

An important number of mathematical or statisical methods have been developed to detect and quantify the intra-tumor heterogeneity (ITH), but no systematic evaluation of their performances and potential for clinical application has been performed. Our first contribution consists in a survey of existing approaches to decipher ITH, that allows to navigate the different underlying ideas easily. We have also proposed a framework to assess the robustness of those approaches, and their potential for use in patient stratification.

This survey has allowed us to identify an unexploited type of data in the process of ITH reconstruction, and our second contribution fills remedies to this shortfall. Indeed, besides observed prevalences of somatic mutations within a tumor sample that allow us to distinguish several clones, the nucleotidic context of those mutations reveals the unknown causative mutational processes. We illustrate on both simulated and real data the opportunity to jointly model those two aspects of tumor evolution.

In conclusion, we highlight the need to reinforce data integration from several sources or samples to harness the potential of tumor evolution for cancer clinical management.

KEYWORDS

Bayesian inference - Benchmarking - Cancer Genomics - High-througput sequencing