Analyses of genomic and transcriptomic profiles of metastatic tumors from precision medicine clinical trials

Yoann Pradat

université
PARIS-SACLAY

Analyses of genomic and transcriptomic profiles of metastatic tumors from precision medicine clinical trials

Analyses des profils génomiques et transcriptomiques de tumeurs métastatiques issus d'essais cliniques de médecine de précision

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 582: cancérologie : biologie - médecine - santé (CBMS),
Specialite de doctorat : Sciences du cancer
Graduate School : Life Sciences and Health. Référent : Faculté de médecine

Thèse préparée dans les unités de recherche **MICS EA 4037 Mathématiques et Informatique pour la Complexité et les Systèmes** (Université Paris-Saclay, CentraleSupélec), et **PMNCO U 981 Prédicteurs moléculaires et nouvelles cibles en oncologie** (Université Paris-Saclay, Inserm, Gustave Roussy)

sous la direction de **Paul-Henry COURNÈDE**, Professeur,
la co-direction de **Fabrice ANDRÉ**, Professeur,
le co-encadrement de **Sergey I. NIKOLAEV**, Docteur

Thèse soutenue à Paris-Saclay, le 14 février 2024, par

Yoann PRADAT

Composition du Jury

Membres du jury avec voix délibérative

Edwin Cuppen Professeur, UMC Utrecht, Pays-Bas	Président
Tatiana Popova Chargée de recherche (INSERM), Université PSL, France	Rapporteur & Examinatrice
Eric Letouzé Directeur de recherche (INSERM), Université de Nantes, France	Rapporteur & Examinateur
Macha Nikolski Directrice de recherche (CNRS), Université de Bordeaux, France	Examinatrice
Elli Papaemmanuil Docteur, MSKCC, États-Unis	Examinatrice
Isidro Cortés-Ciriano Docteur, EMBL-EBI, Royaume-Uni	Examinateur

Thèse de doctorat

NNT: 2024UPASL010

Acknowledgments

Although the decision to engage in a PhD is individual, the journey to its successful completion is collective. I, therefore, cannot overstate how thankful I am to all the people who have surrounded me during this long, challenging, but also mind-blowing experience that has allowed me to pursue my insatiable quest for scientific knowledge.

This journey would never have begun without the support of my PhD advisor, Paul-Henry Cournède. From the outset and throughout the years, you have shown genuine enthusiasm for my research ideas and progress. I am immensely grateful to you for your mentorship during my PhD program. I also want to extend my sincere thanks to Fabrice André, my co-advisor, for his enthusiastic support and insightful comments on my various research projects. The trust you place in me is invaluable. Lastly, I owe a great deal to Sergey Nikolaev, whose day-to-day mentorship has been incredibly valuable. I have learned a lot from you, and without your guidance, the core study of my PhD would not have been possible.

I would like to express my heartfelt gratitude to everyone at the MICS laboratory in CentraleSupélec who have made my time there so memorable. Special thanks go to Gautier, whose enthusiasm for table football added a lot of fun to our interactions, even though he statistically significantly lost more times to me than I did to him. I also want to acknowledge Mahmoud, Brice, Antonin, Sylvain, Marin, Léo, Ludovic, and Romain for their friendship and their willingness to engage in table football matches, despite the odds being against them. Lastly, I am thankful to all the other members of the MICS laboratory with whom I interacted, whether it was to invite them to play or to discuss research-related or unrelated topics.

I would like to extend my sincere gratitude to all the individuals from Gustave Roussy with whom I had the privilege to collaborate. Special thanks to Marc and Gérôme for their consistent support and responsiveness to my numerous requests. I am also grateful to Semih, Fernanda, Julien Viot, Julien Vibert, Luigi, Guillaume, and Loic for their invaluable contributions and the knowledge I gained from them throughout these years.

I would like to extend heartfelt thanks to all the individuals who have been present at various stages of my life. I am particularly grateful to my parents and siblings for their unwavering support and patience. Special appreciation goes to Elli for welcoming me into her lab and igniting my passion for cancer research, and to Elsa for her continuous support, invaluable feedback, and for having me in her team for the next phase of my research journey.

I am also very grateful to all my jury members for accepting to examine my work and particularly to Eric Letouzé and Tatiana Popova for their very encouraging feedback.

These acknowledgments would not be complete without all the patients who have consented to donate samples for research and without whom this work would not be possible.

Lastly, I want to express my profound gratitude to Charlène, my partner and constant source of support, whose unwavering encouragement sustains me through both joyful and challenging times and will continue to do so for the rest of our lives.

Contents

Abstract

In the era of massive data acquisition and analysis, our understanding of cancer onset and evolution has improved in light of the results derived from the analyses of the molecular portraits of tens of thousands of tumors across the globe. The advent of next-generation sequencing technologies in the 2000s has revolutionized how we investigate the tumor cells of patients with cancer. These technologies were first used to characterize specific genomic regions but have matured over time to allow for the systematic profiling of the whole exome, transcriptome, and even whole genome of patients enrolled in clinical trials. As sequencing technologies continued development will fuel new research areas and discoveries for many years, a complete understanding of the different aspects of sequencing data analysis is of utmost importance. Although high-throughput sequencing is not yet part of the standard of care for all cancer patients, extensive molecular profiling has been offered to significant numbers of patients participating in clinical trials and is now used in routine for an increasing number of indications. This large quantity of data is now available to support many research opportunities, ranging from drawing detailed molecular portraits of particular groups of patients to deciphering the links between tumors genotypes and patients clinical outcomes and, eventually, to advancing precision oncology.

This thesis covers many aspects involved in the retrospective analysis of a large cohort of cancer patients, along with detailed reviews of important concepts and tools from modern oncology. The first chapter introduces general concepts about cancer biology and classification, which are fundamental to the immediate treatment decisions but also to research wherein patients are organized according to clinically defined groups. The last section of this first chapter additionally introduces the reader to considerations about the growing place of molecular profiling and their impact on classifications and trial designs. The second chapter extensively reviews the computing tools and databases employed to analyze sequencing data and extract clinically meaningful information. The first two chapters serve as the laying bricks behind the analysis of a large pan-cancer cohort of patients presented in the third chapter. This cohort, META-PRISM, comprises 1,031 patients from two large precision medicine trials led at Gustave Roussy in the decade 2010-2020. About a third of the enrolled patients benefited from whole-exome and RNA sequencing technologies at trial entry and, for a subset of them, at resistance. Compared to other analyses of large pan-cancer cohorts, this study stands out for its focus on patients that were refractory to treatments and the derivation of highly detailed and curated clinical histories. Notably, all patients shared the common characteristic of being in the advanced stages of their disease, deemed incurable by a multidisciplinary board. The comparative analyses against an international cohort of primary untreated tumors have shed light on a few tumor-type-agnostic genetic differences

and multiple tumor-type-specific differences. Additionally, predictive modeling of patients survival using molecular biomarkers has shown that even late-stage patients can benefit from sequencing for important therapeutic decisions, particularly trial eligibility. The fourth and last chapter focuses on the analysis of known and emerging genomic markers of treatment resistance in the META-PRISM cohort, but also in two other cohorts from recent clinical studies led by Gustave Roussy, one interested in a recently approved antibody-drug conjugate for breast cancers and another in a particular class of inhibitors for bladder cancers. These two studies demonstrated that alterations in the expression or structure of the target or mutation-induced activation of alternative pathways are important contributors to drug resistance.

Preface

Context

Precision medicine is now becoming a reality in the era of extensive data collection and analysis. Massive data acquisition has been fueled by the widespread adoption of automatized protocols and computer systems across all services but, most importantly, by the development of high-throughput sequencing technologies at increasingly cheap costs per sequenced sample. Whereas the cost of establishing the first draft sequencing of the human genome during the Human Genome Project (HGP) stood at about 300 million dollars, the current cost of sequencing a complete genome is estimated at about 200 dollars using the latest sequencing machine produced by Illumina, NovaSeq X, which can produce up to 20,000 genomes per year. The dramatic improvements in sequencing technologies now permit the systematic profiling of selected gene panels in the clinical routine of many cancer centers but also more extended sequencing of whole exomes or genomes for patients enrolled in precision medicine trials. The archival of the sequencing files accumulated over the years and of the clinical reports of treated patients now holds the potential to retrospectively analyze in depth the molecular characteristics of swathes of tumors and correlate them with the observations made in the clinic. Many discoveries have already been delivered from the sequencing data analysis of large cohorts of cancer patients, as done by The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), Hartwig Medical Foundation (HMF), and many other teams worldwide investigating diverse tumor types. Continued analyses of the increasing databases of sequencing data in correlation with complete clinical histories are critical for further enhancing our understanding of cancer onset, development, and response to therapies.

The increased diversity of sequencing techniques, coupled with the quick expansion of the current databases, has ignited the development of many computational tools for analyzing vast amounts of data. The early 2000s saw many papers harnessing the power of gene expression microarrays to unravel the molecular specificities of specific cancer types, such as the studies by Golub *et al.* (1999) of acute leukemias, by Alizadeh *et al.* (2000) of diffuse large B-cell lymphomas, by Perou *et al.* (2000) of breast cancers, by Louis *et al.* (2001) of brain cancers. These years were also contemporary with the advent of next-generation sequencing (NGS) that allowed the start of in-depth studies of the molecular profiles of multiple patients within selected tumor types and the development of dedicated tools. In the late 2000s, numerous studies utilized NGS to describe the mutations in the genome of individual patients (Mardis *et al.* 2009; Shah *et al.* 2009; Pleasance *et al.* 2010) whereas a few studies already started to investigate and compare the mutated profiles from multiple patients as done in the landmark studies by Wood *et al.* (2007) of 22 breast and colorectal

cancers and by Parsons *et al.* (2008) of 22 glioblastomas. In the years 2010, numerous studies utilized NGS to analyze increasingly large cohorts of cancer patients, contributing to exploring the clinical implications of detected alterations (Mardis 2014; Shen *et al.* 2015; Gagan & Van Allen 2015) and describing the inter- and intra-tumoral genetic diversity (Gerlinger *et al.* 2012). These studies have been critical in identifying driver mutations, characterizing clonal evolution, and guiding the development of targeted therapies for various cancer types. However, achieving each of these aims depends on the computational tools used to process the raw sequencing files and the more or less complex statistical models and, now, artificial intelligence models used to analyze the processed data. It is common practice to organize all the steps involved in the processing of raw sequencing data into *pipelines* whose composition, often left to the appreciation of dedicated teams of bioinformaticians, has an influential role in the results it produces for the downstream statistical analyses and modeling. Chapter 2 presents in details modern bioinformatic and statistical tools employed to process sequencing files from deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) sequencing experiments and analyze the resulting tables of genomic and transcriptomic alterations. An accurate understanding of all the steps involved in processing raw data and an extensive knowledge of standard genomic analyses and modeling practices are key ingredients for any thorough analysis of tumors' molecular profiles.

In the past decade, extensive studies of patients representing many cancer types, referred to as *pan-cancer* studies, have been led by national or international consortiums (Hoadley *et al.* 2014; Rokita *et al.* 2019; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020; Nguyen *et al.* 2022; Martínez-Jiménez *et al.* 2023). The TCGA initiative was a pioneering initiative in this field from its inception in 2005 to its completion in 2018. The project resulted in the generation of an enormous amount - 2.5 petabytes - of genomic, epigenomic, transcriptomic, and proteomic data, providing valuable insights into the genetic basis of cancer. However, TCGA and most other currently published pan-cancer studies are either focused on samples biopsied from primary tumors or have only scarce clinical information available, thereby limiting the extent of the analyses of the links between the molecular portraits of tumors and the clinical outcomes of patients. A few pan-cancer studies of metastatic samples have recently been released and have shed light on genomic differences with primary tumors (Robinson *et al.* 2017; Priestley *et al.* 2019; Martínez-Jiménez *et al.* 2023). Other tumor-type specific studies have investigated the genomes of advanced tumors in correlation with exposure to specific treatments so as to decipher the genetic or epigenetic mechanisms behind treatment resistance (Pao *et al.* 2005; Shi *et al.* 2014; Woyach *et al.* 2014; Hyman *et al.* 2015; Chandrasekar *et al.* 2015). Putative or confirmed resistance mechanisms are, however, lacking for the vast majority of treatment resistances observed in the clinic, which are the main reason behind our failure to control the disease, eventually, behind patient death. It is consequently of the utmost importance to investigate in depth the molecular landscapes of the tumors from past and future refractory patients in correlation with their treatment history to pursue the quest for resistance mechanisms for all observed resistances. The integrative study presented in this thesis is a small step toward this goal and a more significant step toward the systematic delivery of high-quality data combining detailed clinical histories and detailed profiling of tumors.

Outline

The first chapter is a general introduction to cancer incidence, biology, and classification. The concepts presented in this chapter, particularly the current state and evolution of cancer classifications, are the bricks essential for understanding the data collected and analyzed and for guiding the analyses towards clinically relevant results. This chapter results from an extensive literature review and the substantial expertise gathered throughout the Ph.D. journey from the numerous interactions with biologists and clinicians while reviewing and organizing data for the comprehensive study outlined in Chapter 3. The review study titled Unified classification and risk-stratification in Acute Myeloid Leukemia, published in Nature Communications in 2022, is the result of a retrospective study initiated during my internship in Papaemmanuil's lab. It serves as a prominent illustration of how molecular characteristics are reshaping cancer classifications.

The second chapter delineates the various steps involved in analyzing high-throughput sequencing data. This process starts with the processing of unaligned raw reads and continues with the statistical analyses of the genomic alterations detected with a chosen level of confidence depending on the application. It, too, results from an extensive literature review but also from all the efforts invested in designing pipelines for processing raw sequencing data and conducting downstream analyses for various translational projects I took part in. The extensive pipelines I developed during my PhD journey are structured and documented on the GitHub repositories https://github.com/gustaveroussy/MetaPRISM_WES_Pipeline and https://github.com/gustaveroussy/MetaPRISM_RNAseq_Pipeline.

The third chapter of the thesis delves into the genomic and transcriptomic landscapes of tumor samples obtained from 1,031 patients with refractory or advanced cancers deemed incurable by a multidisciplinary board prior to trial inclusion. This chapters encompasses many different facets of such a comprehensive study, spanning from the creation, curation, and organization of a large database housing harmonized data sourced internally and externally, to the scrutiny of the tumors profiles and the modeling of patients survival. It is an extended presentation of the original study titled Integrative Pan-Cancer Genomic and Transcriptomic Analyses of Refractory Metastatic Cancer, published in Cancer Discovery in May 2023, to which I made substantial contributions across all stages, with the collaboration of colleagues from Gustave Roussy (Villejuif, France) and CentraleSupélec (Gif-Sur-Yvette, France).

The fourth and final chapter widens the scope of the analyses introduced in the preceding chapter, focusing on the specific issue of treatment resistance. It also incorporates analyses of potential resistance mechanisms observed in patients participating in two distinct trials assessing innovative drugs for breast and bladder cancers, respectively. The results from these trials were presented in the original articles Trastuzumab deruxtecan in metastatic breast cancer with variable HER2 expression: the phase 2 DAISY trial and Resistance to selective FGFR inhibitors in FGFR-driven urothelial cancer, published in 2023 in Nature Medicine and Cancer Discovery, respectively.

Bibliography

1. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. en. *Nature* **403,** 503–511. doi:10.1038/35000501 (Feb. 2000).
2. Chandrasekar, T., Yang, J. C., Gao, A. C. & Evans, C. P. Mechanisms of resistance in castration-resistant prostate cancer (CRPC). eng. *Translational Andrology and Urology* **4,** 365–380. doi:10.3978/j.issn.2223-4683.2015.05.02 (June 2015).
3. Gagan, J. & Van Allen, E. M. Next-generation sequencing to guide cancer therapy. en. *Genome Medicine* **7,** 80. doi:10.1186/s13073-015-0203-x (Dec. 2015).
4. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. en. *New England Journal of Medicine* **366,** 883–892. doi:10.1056/NEJMoa1113205 (Mar. 2012).
5. Golub, T. R. *et al.* Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. en. *Science* **286,** 531–537. doi:10.1126/science.286.5439.531 (Oct. 1999).
6. Hoadley, K. A. *et al.* Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. en. *Cell* **158,** 929–944. doi:10.1016/j.cell.2014.06.049 (Aug. 2014).
7. Hyman, D. M. *et al.* Vemurafenib in Multiple Nonmelanoma Cancers with *BRAF* V600 Mutations. en. *New England Journal of Medicine* **373,** 726–736. doi:10.1056/NEJMoa1502309 (Aug. 2015).
8. Louis, D. N., Holland, E. C. & Cairncross, J. G. Glioma Classification. en. *The American Journal of Pathology* **159,** 779–786. doi:10.1016/S0002-9440(10)61750-6 (Sept. 2001).
9. Mardis, E. R. The translation of cancer genomics: time for a revolution in clinical cancer care. en. *Genome Medicine* **6,** 22. doi:10.1186/gm539 (2014).
10. Mardis, E. R. *et al.* Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. en. *New England Journal of Medicine* **361,** 1058–1066. doi:10.1056/NEJMoa0903840 (Sept. 2009).
11. Martínez-Jiménez, F. *et al.* Pan-cancer whole-genome comparison of primary and metastatic solid tumours. en. *Nature* **618,** 333–341. doi:10.1038/s41586-023-06054-z (June 2023).
12. Nguyen, B. *et al.* Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. en. *Cell* **185,** 563–575.e11. doi:10.1016/j.cell.2022.01.003 (Feb. 2022).
13. Pao, W. *et al.* Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain. en. *PLoS Medicine* **2** (ed Liu, E. T.) e73. doi:10.1371/journal.pmed.0020073 (Feb. 2005).
14. Parsons, D. W. *et al.* An Integrated Genomic Analysis of Human Glioblastoma Multiforme. en. *Science* **321,** 1807–1812. doi:10.1126/science.1164382 (Sept. 2008).
15. Perou, C. M. *et al.* Molecular portraits of human breast tumours. en. *Nature* **406,** 747–752. doi:10.1038/35021093 (Aug. 2000).

16. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. en. *Nature* **463,** 184–190. doi:10.1038/nature08629 (Jan. 2010).
17. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. en. *Nature* **575,** 210–216. doi:10.1038/s41586-019-1689-y (Nov. 2019).
18. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. en. *Nature* **548,** 297–303. doi:10.1038/nature23306 (Aug. 2017).
19. Rokita, J. L. *et al.* Genomic Profiling of Childhood Tumor Patient-Derived Xenograft Models to Enable Rational Clinical Trial Design. en. *Cell Reports* **29,** 1675–1689.e9. doi:10.1016/j.celrep.2019.09.071 (Nov. 2019).
20. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. en. *Nature* **461,** 809–813. doi:10.1038/nature08489 (Oct. 2009).
21. Shen, T., Pajaro-Van De Stadt, S. H., Yeat, N. C. & Lin, J. C.-H. Clinical applications of next generation sequencing in cancer: from panels, to exomes, to genomes. *Frontiers in Genetics* **6.** doi:10.3389/fgene.2015.00215 (June 2015).
22. Shi, H. *et al.* Acquired Resistance and Clonal Evolution in Melanoma during BRAF Inhibitor Therapy. en. *Cancer Discovery* **4,** 80–93. doi:10.1158/2159-8290.CD-13-0642 (Jan. 2014).
23. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. en. *Nature* **578,** 82–93. doi:10.1038/s41586-020-1969-6 (Feb. 2020).
24. Wood, L. D. *et al.* The Genomic Landscapes of Human Breast and Colorectal Cancers. en. *Science* **318,** 1108–1113. doi:10.1126/science.1145720 (Nov. 2007).
25. Woyach, J. A. *et al.* Resistance Mechanisms for the Bruton's Tyrosine Kinase Inhibitor Ibrutinib. en. *New England Journal of Medicine* **370,** 2286–2294. doi:10.1056/NEJMoa1400029 (June 2014).

Contributions

In this section, I have outlined the scientific contributions of my Ph.D. First, I've listed the published papers, followed by oral presentations, then code contributions, and finally, contributions related to data sharing.

Published papers

1. 2023. Facchinetti F, Hollebecque A, Braye F, Vasseur D, **Pradat Y**, Bahleda R, Pobel C, Bigot L, Deas O, Florez Arango JD, Guaitoli G, Mizuta H, Combarel D, Tselikas L, Michiels S, Nikolaev SI, Scoazec JY, Ponce-Aix S, Besse B, Olaussen KA, Loriot Y, Friboulet L. Resistance to selective FGFR inhibitors in FGFR-driven urothelial cancer. **Cancer Discovery** 13, 1998-2011. *doi:10.1158/2159-8290.CD-22-1441*

2. 2023. Mosele F, Deluche E, Lusque A, Le Bescond L, Filleron T, **Pradat Y**, Ducoulombier A, Pistilli B, Bachelot T, Viret F, Levy C, Signolle N, Alfaro A, Tran DTN, Garberis IJ, Talbot H, Christodoulidis S, Vakalopoulou M, Droin N, Stourm A, Kobayashi M, Kakegawa T, Lacroix L, Saulnier P, Job B, Deloger M, Jimenez M, Mahier C, Baris V, Laplante P, Kannouche P, Marty V, Lacroix-Triki M, Diéras V, André F. Trastuzumab deruxtecan in metastatic breast cancer with variable HER2 expression: the phase 2 DAISY trial. **Nature Medicine** 29, 2110-2120. *doi:10.1038/s41591-023-02478-2*

3. 2023. **Pradat Y**, Viot J, Yurchenko AA, Gunbin K, Cerbone L, Deloger M, Grisay G, Verlingue L, Scott V, Padioleau I, Panunzi L, Michiels S, Hollebecque A, Jules-Clément G, Mezquita L, Lainé A, Loriot Y, Besse B, Friboulet L, André F, Cournède PH, Gautheret D, Nikolaev SI. Integrative Pan-Cancer Genomic and Transcriptomic Analyses of Refractory Metastatic Cancer. **Cancer Discovery** 13, 1116-1143. *doi:10.1158/2159-8290.CD-22-0966*

4. 2023. Benkirane H, **Pradat Y**, Michiels S, Cournède PH. CustOmics: A versatile deep-learning based strategy for multi-omics integration. **PLoS computational biology** 19, e1010921. *doi:10.1371/journal.pcbi.1010921*

5. 2022. Tazi Y, Arango-Ossa JE, Zhou Y, Bernard E, Thomas I, Gilkes A, Freeman S, **Pradat Y**, Johnson SJ, Hills R, Dillon R, Levine MF, Leongamornlert D, Butler A, Ganser A, Bullinger L, Döhner K, Ottmann O, Adams R, Döhner H, Campbell PJ, Burnett AK, Dennis M, Russell NH, Devlin SM, Huntly BJP, Papaemmanuil E. Unified classification

and risk-stratification in Acute Myeloid Leukemia. **Nature communications** 13, 4622. *doi:10.1038/s41467-022-32103-8*

Oral communications

- 2023. Integrative pan-cancer genomic and transcriptomic analyses of refractory metastatic cancer. *Gustave Roussy Research open days*, Gif-Sur-Yvette. **Selected talk (Recipient of Young Researcher Award)**
- 2023. Efficient processing of TCGA data on cloud services. *Institut Curie Seminars*, Paris. **Invited talk**
- 2022. Integrative genomic and transcriptomic analysis of refractory metastatic cancers. *ESHG 2022*, Vienna. **Selected talk**
- 2021. META-PRISM: A retrospective analysis of more than a 1,000 metastatic tumor. *CBMS doctoral school PhD students day*, Paris. **Selected talk**

Original code

The codes I have developed are all original and have been deposited on one of the following GitHub or Gitlab accounts

- https://github.com/ypradat
- https://gitlab-research.centralesupelec.fr/mics_biomathematics
- https://github.com/gustaveroussy

The list of repositories below is organized in three categories with the code supporting the published studies first, then the bioinformatic pipelines, and eventually miscellaneous graphic or analysis packages, some still under development.

Studies

- https://github.com/gustaveroussy/MetaPRISM_Public
  - Created, organized, and maintain the repository.
  - R and python packages developed specifically for the project are available in the code/functions folder.
  - All analyses and figures in the subfolders in code/scripts. Each subfolder is organized via a Snakemake pipeline for **reproducibility**.
  - Development of the Snakemake whole-exome sequencing (WES) pipeline starting from the work of a former bioinformatician. Cleaned it, completely reorganized

the rules and added many scripts and rules to make it an end-to-end workflow starting from FASTQ and ending with tables of filtered and annotated somatic alterations. Available in code/pipelines/wes.
  - Deposited the code on a Code Ocean capsule https://codeocean.com/capsule/2014781/tree/v1
- https://github.com/gustaveroussy/DAISY_Public
  - Created, organized, and maintain the repository.
  - All analyses and figures in the subfolders of code/scripts except for the code/scripts/slides folder which is the work of another PhD student. Each subfolder is organized via a Snakemake pipeline for **reproducibility**.
- https://github.com/ypradat/TCGA_Facets
  - Developed Bash scripts to interact programmatically with the Google Cloud Engine (GCE). Starting from a list of binary alignment map (BAM) URIs on the GDC-controlled google bucket hosting TCGA data, the Bash scripts allow automatic spawning of VM (one for each predefined batch) with fine-tuned resources, monitoring of the VM and handling of preemption (allowing up to 91% reduction in computing costs), and copying of the Snakemake pipeline results to a centralized google bucket.
  - Reduced the Snakemake pipeline developed for META-PRISM to only the part required for calling copy-number alterations (CNAs) starting from BAM files and annotating them with OncoKB and CIViC databases.
  - Received technical support from ISB-CGC team and financial support from National Cancer Institute (NCI).

Pipelines

- https://github.com/ypradat/CivicAnnotator
  - Manually curated and completed the tables of evidence from CIViC January 2022 release to allow for automatic annotation. This tool was used for all three studies presented in this thesis.
  - Developed a user-friendly python script allowing to annotate tables of mutations, CNAs, gene fusions, and combinations of these using the CIViC database. It mirrors the tool developed for annotating with OncoKB database at https://github.com/oncokb/oncokb-annotator.
- https://github.com/gustaveroussy/MetaPRISM_WES_Pipeline
  - Development of the Snakemake WES pipeline for the META-PRISM project starting from the work of a former bioinformatician. Cleaned it, completely reorganized the rules and added many scripts and rules to make it an end-to-end workflow starting from FASTQs and ending with tables of filtered and annotated somatic alterations.
  - This pipeline was used for the analysis of DNA sequencing experiments of all three studies presented in this thesis.
- https://github.com/gustaveroussy/MetaPRISM_RNAseq_Pipeline

  - Development of a Snakemake RNA sequencing (RNA-seq) pipeline to allow for the detection of gene fusion events from RNA sequencing experiments. This pipeline will in the long-term also include the code used for quantifying gene expression.
  - This pipeline was used for the analysis of RNA sequencing experiments in META-PRISM study.

Tools

- https://github.com/ypradat/UltiSig (ongoing work)
  - Developed an R package for running different methods of mutational signatures deconvolution (see the founding paper https://doi.org/10.1016/j.celrep.2012.12.008).
- https://github.com/ypradat/VariantNMF (ongoing work)
  - Developed a Julia package implementing user-friendly functions for running the non-negative matrix factorisation (NMF) algorithm with full control over the objective function (via the alpha and beta divergences) to be minimized and the methods used to perform the minimization (multiplicative updates, projection pursuit, ALS, etc.)
- https://github.com/ypradat/tableExtra
  - R package implementing a tool for producing heatmaps with double information encoded in the shapes colors and sizes.
  - Implemented a Gitlab continuous integration running on the Gitlab servers of CentraleSupélec and making use of a docker image that I created.
  - **Released the package on CRAN** https://cran.r-project.org/web/packages/tableExtra/index.html.
- https://github.com/ypradat/PrettyPy
  - Python package implementing useful functions for drawing high-quality figures I often use.

Data sharing

I also invested time and effort in developing meticulously organized databases to support our analyses, and when feasible, share them with collaborators and the broader scientific community. Notably, standardized data tables were uploaded onto a local instance of cBioPortal managed by Gustave Roussy using the open-source code developed by a dedicated team at Memorial Sloan Kettering (MSK)[1].

- https://cbioportal.gustaveroussy.fr/study/summary?id=metaprism_2023 (public)
- https://mappyacts-portal.gustaveroussy.fr/study/summary?id=brca_daisy_2023 (controlled-access)

[1] https://www.cbioportal.org/

1. Cancer characterization and classification

Contents

Abstract Chapter 1

In this chapter, we will first provide more specific details about how cancer affects the general population and how the research and medical communities have organized themselves to draw detailed portraits of the disease. In the second part, we shall look more specifically at the molecular mechanisms currently described for their presumed or confirmed role in cancer onset and growth. Eventually, we will provide a brief overview of the existing cancer classifications and how research is currently reshaping these classifications.

CANCER is a highly complex medical condition that remains largely elusive to our understanding due to its heterogeneous origins, characteristics, and consequences. The disease is caused by the uncontrolled proliferation of cells, i.e the development of a tumor, which acquire malignant properties that enable them to invade nearby tissues and organs and, in some instances, spread to other parts of the body through a process called metastasis. Conversely, benign tumors are non-cancerous cell masses that do not spread to other parts of the body and have less impact on health. Tumors can arise in any part of the body of mammalian species with varying degrees of frequency depending on the body site. They can remain unnoticed for the lifetime of the host or, on the contrary, disrupt the normal functioning of other body parts and become life-threatening conditions. Though we continue to lack a complete or even satisfying understanding of the biological mechanisms underlying tumorigenesis, medical doctors, biologists, statisticians, and, more generally, researchers from all backgrounds have conducted countless observational, experimental, and theoretical studies to characterize and classify cancer, with the ultimate aim to treat patients with cancer better and, if feasible, cure them. The accumulation of decades of work has led to radical changes in how we approach the disease and remarkable improvements in the life expectancy and quality of affected individuals. This progress enables present and future medical doctors to propose therapeutic paths leading to stable remission in an increasing proportion of patients, thereby prolonging their lifetimes.

1.1. General introduction

1.1.1. About cancer incidence

The past and future trends of cancer statistics all point towards an increased incidence of cancer in the general population in the coming years. In its latest global incidence survey from 2020 (Sung *et al.* 2021) in 185 countries, the Global Cancer Observatory (GLOBOCAN) estimated that 19.3 million new cancer cases occurred globally and that nearly 10 million cancer deaths were registered during the elapsed year. In the United States only, the 2023 yearly report of cancer statistics (Siegel *et al.* 2023) anticipated 1.95 million new cases and 610,000 cancer-related deaths and estimates that *any individual has a 41% chance of developing invasive cancer throughout their life*. Incidence greatly varies according to the

primary site, with breast cancer now the most frequent (11.7% worldwide, 15.3% US), followed by lung (11.4%, 12.1%), colorectal (10.0%, 8.9%), prostate (7.3%, 14.7%), and stomach (7.7%, 1.4%) cancers (Sung *et al.* 2021; Siegel *et al.* 2023). In France, the French National Cancer Institute (INCA) estimated that 382 000 new cancer cases and 157,000 cancer deaths occurred in 2018, making cancer the leading cause of premature deaths. The GLOBOCAN 2020 report *projects a 47% increase in cancer incidence over the next 20 years* compared to a projected general population increase of 18%, highlighting the incoming pressure of the disease on our societies and healthcare systems.

However, trends are not uniform across primary sites, countries, ages, or gender. In the United States, the incidence of breast, skin, and liver cancers increases steadily, but that of lung and colorectal cancers is slowly decreasing. In young US women, cervical cancer has registered a dramatic 65% drop in the last decade due to widespread papillomavirus vaccine coverage (Lei *et al.* 2020; Falcaro *et al.* 2021). The most important increase in cancer incidence is anticipated in older adults, as a consequence of increased lifetimes, improved screening, but also of increased numbers of people that have been exposed to one or several risk factors throughout their life (Smith *et al.* 2009). Worldwide, the incidence rate was 19% higher in men than in women in 2020. Interestingly, incidence rates correlate positively with the human development index, with rates of about 100 per 100,000 individuals in low-index compared to rates of 300 per 100,000 in high-index countries, and are projected to increase substantially across all development levels but more markedly in emerging countries. Of note also, the distribution of cancer types in emerging countries is transitioning towards the distribution currently observed in high-income countries, reflecting lifestyle changes and increased exposure to risk factors usually seen in the latter countries (Sung *et al.* 2021).

Cancer is now a worldwide leading cause of death, partly due to the increased incidence of the disease and partly to the decreased mortality from other causes, in particular cardiovascular diseases. Across primary sites, *lung cancer is the leading cause of cancer death in 93 countries*, accounting for nearly 20% of cancer-related deaths worldwide, with country trends reflecting the maturity of the tobacco epidemic but also increasingly the exposure to air pollution. Colorectal and liver cancers rank second and third by the absolute number of deaths, respectively, though, if discriminating between men and women, breast cancer ranks first in women (Sung *et al.* 2021). In the United States, the 5-year survival rate for all cancers has increased from 49% in the 70's to 68% in the last decade, but in some cancer types, this rate remains very low, particularly in pancreas (12%), liver (21%), and esophagus (21%) cancers (Siegel *et al.* 2023). Metastatic dissemination is by far the most frequent cause of cancer death, accounting for 90% of these (Chaffer & Weinberg 2011).

1.1.2. About cancer studies and consortiums

According to the World Health Organisation (WHO), the number of registered clinical studies investigating malignant neoplasms has grown dramatically from 738 studies in 1999

to 104,491 in 2021[1]. This sharp rise in numbers reflects the incredible technological and pharmaceutical advancements made in the last two decades, as well as the increasing public interest in cancer research. In 2021, the United States registered one-third of all cancer studies, while China accounted for 15%, Japan for 11%, and France and Germany for 9% each. Each clinical study is categorized into one of four phases based on the drug development stage, with phase II trials being the most common. Panel 1.1 presents more in detail the categorization of cancer clinical trials.

Complementary to clinical trials, which are focused on assessing the efficacy and side effects of one or multiple drugs, alone or in combination, scientific consortiums have emerged over the years to gather data about individuals harboring specific medical conditions at an unprecedented scale. In cancer research, the most prominent of these efforts was first achieved by TCGA consortium, which started in 2005 as a joint effort between the NCI and National Human Genome Research Institute (NHGRI) US institutes in order to explore the entire spectrum of genomic alterations found in exonic regions of human cancer genomes. From 2006 to 2014, diagnostic samples from 11,315 patients representing 33 different cancer types were collected across 22 countries and sequenced. A first interim analysis of glioblastomas was published by The Cancer Genome Atlas Research Network (2008), opening the path for a long series of tumor type-specific and pan-cancer landmark studies that have each defined or redefined the molecular landscape and classification of cancer and pushed research far beyond its limits. TCGA reached its peak in 2018 through its Pan-Cancer Atlas, a collection of 27 papers simultaneously published in Cell (Sanchez-Vega *et al.* 2018; Hoadley *et al.* 2018; Ding *et al.* 2018), each presenting in-depth studies of the cancer genome from different aspects and sequencing technologies. Nowadays, TCGA continues to represent the largest database ever assembled of multimodal sequencing of human cancer, with more than 2.5 petabytes of data made publicly available for anyone to use in the research community through the Genomic Data Commons (GDC) data portal[2].

Multiple other large-scale international consortiums ave improved our understanding of the molecular mechanisms underlying tumorigenesis. Building upon the work of the TCGA Pan-Cancer Atlas project, the PanCancer Analysis of Whole Genomes (PCAWG) project jointly led by TCGA and ICGC aimed at extending the study of the nature and consequences of genomic variations in both coding and non-coding regions through the sequencing of more than 2,600 whole cancer genomes. The output from this project was presented to the public through the simultaneous publication of 23 papers in the February 2020 issues of Nature journals (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). Mirroring the TCGA effort for the specific question of childhood cancer, NCI's Therapeutically Applicable Research to Generate Effective Treatments (TARGET) is yet another program led by the NCI that aims at exploring in detail the genomic characteristics of more than 6,500 childhood cancers and is also made publicly available through the GDC data portal. The American Association for Cancer Research (AACR)'s Genomics Evidence Neoplasia

[1] https://www.who.int/observatories/global-observatory-on-health-research-and-development/monitoring/number-of-trial-registrations-by-year-location-disease-and-phase-of-development

[2] https://portal.gdc.cancer.gov/

Panel 1.1: Cancer clinical trials

Clinical trials are classified into **phases** according to the drug development stage. There are three main phases of clinical trials: phases I to III. Phase IV trials, often referred to as post-market, are not as systematic as other phases and tend to be more observational. The following definitions describe precisely each phase.

- **I** Assesses the drug's safety and dosage and in its first clinical use on up to 100 individuals.
- **II** Assesses the drug side effects and efficacy on up to hundreds of individuals. Only one drug in three moves to the next phase.
- **III** More comprehensively assesses the drug efficacy and monitors adverse reactions over a longer period and a larger cohort of up to 3,000 individuals. Only one drug in four passes this phase and meets the conditions for approval.
- **IV** Further extends the objectives of previous phases over a longer time but takes place after the drug has been approved and is on the market.

Randomization is another important characteristic of clinical trials. A randomized controlled trial (RCT) is a specific type of randomized trial in which there is a control group that serves as a comparison for the group receiving the intervention or treatment under investigation. In an RCT, participants are randomly assigned to one of two or more groups: an experimental group (receiving the intervention being studied) and a control group (receiving a placebo or standard treatment). The purpose of the control group is to provide a baseline for comparison, allowing researchers to assess the true impact of the experimental intervention by minimizing the influence of confounding variables. RCTs are considered the gold standard in clinical research and are widely used in medical and scientific studies to evaluate the efficacy of treatments or interventions compared to standard treatments.

The **blinding** of the randomization is another important characteristic of RCT to mitigate various biases. While random allocation eliminates the primary source of bias, namely selection bias, other potential biases may occcur and may be mitigated through simple, double, or triple-blinding. Simple blinding, i.e the concealment of treatment allocation from trial participants, allows to control for cognitive biases. On the other hand, double and triple blinding, corresponding to the additional concealment of information from experimenters and evaluators, respectively, serves to control for confirmations biases.

Nowadays, with the increasing use of biomarkers to guide treatment decisions, a novel glossary has emerged to describe novel designs of clinical trials (Yates *et al.* 2018; Ravi & Kesari 2022). Cancer clinical trials have been traditionally categorized by cancer type. However, there is a discernible shift towards trials that include two or more cancer types within their design, with the aim of evaluating whether biomarker-driven strategies can transcend the conventional paradigm of cancer type-centric treatment decisions. Main examples include

- **Basket trial** a biomarker-based trial investigating a therapeutic intervention targeting a single molecular alteration across different cancer types.
- **Umbrella trial** a biomarker-based trial investigating different therapeutic interventions targeting different molecular alterations in a single cancer type.
- **Platform trial** multi-arm multi-stage trial allowing the flexible addition of new treatment arms or subgroups to compare multiple interventions to a control arm. This type of trial may run for a long time.

Information Exchange (GENIE) project is another ongoing large-scale effort aiming at building an international cancer registry of clinical and genomic data assembled through the data sharing of 19 leading cancer centers spread worldwid. It projects to make its data accessible through both GDC's data portal and MSK's cbioportal.

Some national initiatives have also successfully screened large cohorts of individuals and established databases that are fueling research throughout the world. In the United-States, the MSK cancer center has established a clinical assay targeting a couple hundreds of known cancer genes, named MSK-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT), used in the pilot phase on 284 tumor samples (Cheng *et al.* 2015) and later on extended to more than 10,000 patients (Zehir *et al.* 2017), leading to the first Food and Drug Administration (FDA) approval of a tumor-profiling assay in November 2017. Even more recently, the assay was extended to cover more genomic regions, and the database of all profiled samples served as the basis for a landmark comparative study of 25,000 primary and metastatic samples (Nguyen *et al.* 2022). All generated data, except for raw sequencing files, has been made publicly available through MSK's cbioportal[3]. In the United Kingdom, the nationwide *100,000 Genomes Project* led by Genomics England sequenced the complete genomes of 85,000 patients with cancer or rare diseases and of their relatives to investigate the genetic components of these conditions and, more generally, elucidate "the role of genes in health and disease"[4]. The project closed for recruitment in 2018, and the data collected is now being actively analyzed to support many research questions across diverse topics[5]. Importantly, the initiative is evolving to embrace the potential of multi-modal data and long-read sequencing technologies. Concurrently, the United Kingdom hosts the *UK Biobank*, a long-term health study that has amassed data from approximately 500,000 middle-aged individuals, primarily of European ancestry. This study has progressed to the sequencing phase, with 200,000 genetic profiles having been made accessible to the research community by the end of 2021 (Kaiser 2021), and the remaining data anticipated to follow soon. A noteworthy feature of the database is the identification of incidental cancers, occurring subsequently to enrollment, in 46,021 individuals as of June 2022, rendering the study an invaluable resource for cancer research. The HMF serves as another prominent example of a large-scale national initiative, operating in the Netherlands, that has effectively established an extensive repository featuring paired sequencing data and clinical information from patients afflicted with metastatic cancer. The foundation's establishment in 2015 was motivated by the goal of making cutting-edge sequencing technologies accessible to patients while simultaneously aggregating a comprehensive database to support research. As of November 2023, the HMF's repository encompasses sequencing and clinical data from 5,891 metastatic patients[6], representing a diverse spectrum of cancer types. This valuable resource has already contributed to landmark pan-cancer studies (Priestley *et al.* 2019; Martínez-Jiménez *et al.* 2023) and has been featured in over 60 additional publications in esteemed scientific journals [7].

[3] https://www.cbioportal.org/
[4] https://www.genomicsengland.co.uk/initiatives/100000-genomes-project
[5] https://www.genomicsengland.co.uk/research/publications
[6] https://database.hartwigmedicalfoundation.nl/
[7] https://www.hartwigmedicalfoundation.nl/en/research-and-science/scientific-publications/

Importantly, cancer study is also dependent on large-scale sequencing projects run on general populations or populations selected for diseases unrelated to cancer. Indeed, the output data of these projects are now commonly used in the interpretation of genomic variation in nearly all genotype-phenotype association studies and particularly in cancer studies. In 2009, the National Heart, Lung, and Blood Institute (NHLBI) initiated the Exome Sequencing Project (ESP) to identify rare and putatively functional protein-coding variants associated with heart-, lung-, and blood-related diseases through the sequencing of the exome of 6,700 individuals of either European or African ancestry. In 2014, the Exome Aggregation Consortium (ExAC) collected and harmonized exome sequencing data from large-scale sequencing projects around the world. The 2016 ExAC release included over 60,000 exomes (Exome Aggregation Consortium *et al.* 2016), and subsequent releases doubled this number to 125,748 exomes from any ancestry or medical condition. ExAC has now been merged with additional harmonized genome sequencing data resulting in the Genome Aggregation Database (gnomAD) which, in its v3.1.2 release, spanned all 125,748 exomes from ExAC and an additional 76,156 complete genomes. In the United Kingdom, the 1000 Genomes Project (1000G), which began in 2012, combined whole-genome and targeted exome sequencing techniques to create a detailed catalog of human genetic variation (The 1000 Genomes Project Consortium 2012). Upon completion in 2015, the project successfully analyzed the genomes of 2,504 individuals from 26 populations (The 1000 Genomes Project Consortium 2015) and serves today as a reference catalog of human genetic variation alongside ESP and gnomAD. The database of Single Nucleotide Polymorphisms (dbSNP) database is a comprehensive resource that catalogs various types of genetic variations, including single-nucleotide polymorphisms (SNPs), insertions, deletions (i.e indels), and more, across the human genome. All variants identified by ESP and 1000G were submitted to dbSNP, while for gnomAD not all but the majority of the variants are also listed in dbSNP. As of March 2023, dbSNP v155 contained variants from more than 1 billion locations, 15 million of which have a minor allele frequency (MAF) greater than 1% in the 1000G phase 3 dataset[8]. It is now common practice to report the variant frequencies in the three aforementioned large-scale wild-type populations, namely ESP, gnomAD, and 1000G, and in the cancer database Catalogue of Somatic Mutations In Cancer (COSMIC), in the tables of mutations identified in samples of cancer patients. The dbSNP identifier of known variants is also now systematically reported.

1.1.3. About the cancer community

Nowadays, the field of cancer clinical care and research is bringing together a diverse community of passionate individuals from various backgrounds and areas of expertise, including physicians, biologists, physicists, engineers, statisticians, and data scientists. This collaboration is crucial for understanding cancer and developing effective treatments for all cancer patients. Modern cancer centers provide striking examples of this collaboration between people with different areas of expertise. Upon evaluating a patient, an oncologist establishes a clinical profile based on history and current condition, and further investigations are conducted to

[8] https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg38&g=dbSnp155Composite

determine the diagnosis and best therapeutic options. An interventional radiologist performs minimally invasive procedures to obtain a biopsy, which is then sent to hospital laboratories for assessment. A pathologist examines tissue slides to establish the precise diagnosis, and laboratory tests may be conducted to identify specific markers or genomic alterations. The tests serve as a concrete example of how research has advanced our understanding of cancer and has translated into new clinical care. Depending on the context, tumor sequencing may then be performed on one or multiple slides that were first carefully reviewed and selected based on their tumor cellularity content. Prior to sequencing, wet laboratory technicians perform the extraction of cells or nucleic acids and, more generally, all the library preparation protocols required for the specific machine to be used (Nangalia & Campbell 2019). Bioinformaticians then process the large files produced by the sequencer, applying quality-control, variant detection, and variant annotation algorithms to produce reports that are intelligible to the clinician and low-size data files meant to be analyzed by research scientists. In summary, multiple individuals from various departments work hand-in-hand to conduct all necessary analyses required to guide current clinical practices and to generate data that will support future research endeavors.

The success of research heavily relies on the existence, size, and quality of databases aggregated over the years. These databases are created and maintained by data managers with expertise in computer science, medicine, or both, and effective coordination is crucial for the efficient delivery and development of sound databases due to the large number and variety of experts involved in every step of the process. The data thus generated occupies swathes of researchers from all backgrounds who aim to decipher the molecular mechanisms of cancer onset and growth, as well as to help clinicians identify biomarkers that predict clinical course, particularly the response to therapies. Additionally, many other professions are involved in providing material and support to adequately collect, prepare, and process the ever-increasing amount of data collected from each cancer patient.

Furthermore, the community is very active through its numerous seminars, conferences, and congresses that play a critical role in disseminating research results to experts but also to the general public. The use of social networks have recently been shown to be very effective at that purpose (Morgan *et al.* 2022). With the constant influx of new research and emerging technologies, networks of renowned experts, such as the *OncoAlert Network*[9], have emerged to provide weekly summaries of essential news in oncology. Regular summaries are also provided by long-established associations, such as the European Society of Human Genetics (ESHG), American Society of Clinical Oncology (ASCO), European Society of Medical Oncology (ESMO), AACR, European Association for Cancer Research (EACR), among others.

[9] http://www.oncoalert360.com/

1.2. Biological aspects

1.2.1. Central dogma of molecular biology

The cell is the building block of all multicellullar organisms. The human body is composed of trillions of cells, and each of these cells behaves according to its intrinsic components but also extrinsic signals it receives from neighboring and distant cells. In eukaryotic organisms, which include humans, each cell has in its inner part a double membrane-bound organelle called the *nucleus* which holds the genetic material encoding the instructions required for organizing life. Most eukaryotic cells also carry another a class of double membrane-bound organelles called mitochondria, which enclose complementary instructions inherited from engulfed bacteria throughout evolution. This genetic material is stored in large DNA molecules that each consists in two long chains of covalently-linked nucleotides, called *strands*, intertwined in a double-helix structure and held together by hydrogen bonds. Each DNA molecule is folded into a structure that we call chromosome that consists in an assemblage of DNA and proteins that fold and pack the polynucleotide chain, thereby forming the *chromatin*, but also proteins involved in the nuclear processes of the cell, particularly DNA replication, repair, and expression (Figure 1.1). All human cells contain two copies of twenty-three chromosomes except for gametes that each contain only a single copy and some highly specialized cells. Chromosomes 1 to 22 are called *autosomes* and the last pair of chromosomes, called the sex chromosomes, differentiates females who have two X chromosomes from males who have one X and one Y.

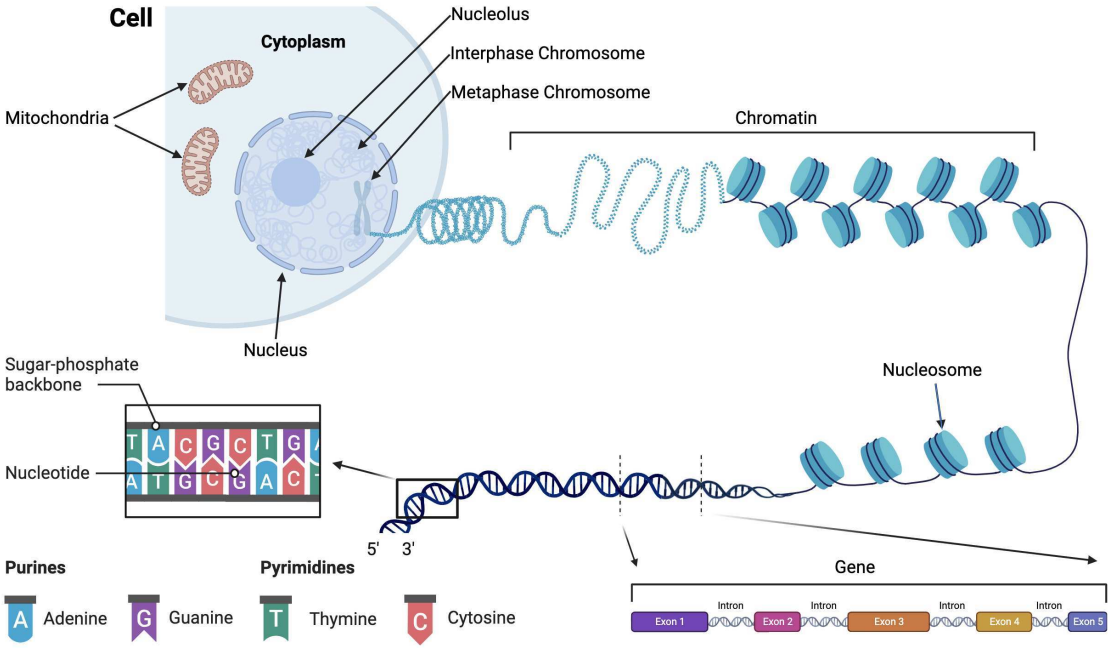


Fig. 1.1.: Structure of the DNA molecule in eukaryotic cells. Created with BioRender.com

Each human cell's complete set of chromosomes contains about 3.2 billion nucleotide pairs constituting the human genome. Specific DNA sequences called genes are scattered throughout the genome and code for the different organic molecules found in cells. The primary assembly of the latest version of the human reference genome[10] lists 41,762 genes, 19,895 of which encode for proteins (*protein-coding* genes) while others encode for non-coding RNA molecules (*non-coding* genes). While the number of protein-coding genes has been relatively stable over time, the actual number of non-coding genes varies significnatly according to estimates, largely due to a class of genes encoding a specific type of molecules known as long non-coding RNA (lncRNA) (Ponting & Haerty 2022). The average gene length is approximately 42 kilobases, but some genes can span very extensive regions, such as *RBFOX*, the largest human gene currently annotated spanning 2.47 million bases.

The information required for building complex proteins is encoded in the sequence of nucleotides of genes. More specifically, in eukaryotes only small fractions of the genes sequences called *exons* actually code for the amino acids constituting proteins. The other regions of genes, referred to as introns (Figure 1.1), have historically been described as "junk DNA", although contemporary insights increasingly acknowledge the multifaceted roles they play in our cells (Jo & Choi 2015). The transformation of nucleotides to amino acids is governed by the *genetic code*, which maps every possible sequence of three nucleotides, or codon, to one of the 20 standard amino acids or the stop codon. The genetic code is redundant as 61 codons map to the 20 standard amino acids, and three codons encode specific sequences which signals the end of the protein. In the process of gene expression, the introns are removed through a process called *splicing*. Interestingly, in eukaryotic organisms, a natural phenomenon called *alternative splicing* allows genes to be transcribed into different transcripts by selectively keeping or removing exons during splicing. On average, human genes have the capacity to generate approximately 4.24 different transcripts. Within the human genome, introns exhibit considerably greater lengths than exons. Although initial estimates from the first assembly of the reference genome indicated that introns accounted for 25% of the genome and exons comprised merely 1.1% (Venter *et al.* 2001), these figures have since been revised upwards due to the extensive genome annotation efforts that have followed.

In contrast to the chemical simplicity of DNA, composed of only four different subunits (A - adenine, C - cytosine, G - guanine, T - thymine), proteins are much more complex molecules of very heterogeneous sizes, ranging from a couple of dozen of amino acids to more than 35,000 for some isoforms of the largest known human protein (titin). At any given time, the composition of proteins and other biomolecules in a cell reflects the *expression* of its genes, both past and present. Cells express their genes by transcribing DNA into another nucleic acid called RNA, followed by the *translation* of the RNA into proteins for protein-coding genes. *Transcription* is a tightly regulated process triggered by proteins called transcription factors that signal RNA polymerase proteins to start RNA synthesis. As described in the previous paragraph, multiple RNA transcripts may be produced from one gene through alternative splicing, and each of these transcripts may direct the synthesis of multiple identical copies of a protein. Gene expression can be quantified in different ways, either by counting

[10] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/GCF_000001405.40-RS_2023_10/

transcripts or by counting proteins, with different quantification results. In some cases, the final products of gene expression are the RNA transcripts themselves, serving as regulatory, structural, and catalytic components of the cell. Examples include micro-RNAs (miRNAs) involved in gene regulation, ribosomal RNAs (rRNAs) that are key ribosome constituents, or transfer RNAs (tRNAs) which serve as adaptors during protein synthesis. The interested reader is greatly encouraged to read the excellent book by Alberts *et al.* (2019) for clear and detailed presentations of the molecular constituents and inner mechanisms of cells, particularly chapters 5 and 7 for questions related to DNA, RNA, genes, and gene expression.

1.2.2. Common molecular aberrations

Cancer is, by definition, the result of abnormal cellular behavior, which itself results from abnormalities in the instructions regulating intra- and extracellular processes. As the normal functioning of cells is reliant on a complex balance of all their components, genetic variations can disrupt this equilibrium by modifying directly the cells' gene expression or the regulators of this expression. Variations may occur at different scales, ranging from large rearrangements of one or multiple chromosomes to single-nucleotide variants (SNVs). Genetic alterations can have diverse consequences, particularly protein underexpression, overexpression, or disruption which can subsequently affect other proteins and eventually provide a selective advantage over neighboring cells. The consequences of these alterations mainly materialize either as the altered expression of unchanged proteins or the expression of altered proteins, some of which are known as *oncoproteins* due to their initiating or driving roles in tumorigenesis. Epigenetic changes represent another class of alterations that do not modify the sequence of the DNA itself but instead modify the state of the DNA molecule. Methylation patterns and chromatin conformation play key roles in gene expression regulations, and modifications of these are known to be directly involved in the transformation toward the cancerous state (P. A. Jones & Baylin 2002).

1.2.2.1. Chromosomal alterations

Chromosomes were first described in the nineteenth century, but their role as the root cause of medical disorders was only first evidenced by Lejeune *et al.* (1959) who, for the first time, described supernumerary chromosomes 21 in individuals with Down syndrome. Nowadays, changes in the number or structure of chromosomes are clinically recognized as the genetic cause of many medical disorders, including cancers that are characterized by chromosomal alterations. Structural and segmental rearrangements of chromosomes have long been implicated for their crucial role in tumor initiation and evolution, particularly in blood cancers where they serve as the basis of many hematological classifications. These chromosomal changes impact tumor growth by activating oncogenes or inactivating tumor suppressor genes (Mitelman, Johansson & Mertens 2007). Changes in chromosome structures are either balanced if the complete set of chromosomes is present but rearranged, or imbalanced. They are commonly classified into four categories (deletions, duplications, inversions, and translocations), all of which have distinct consequences on the genome. The most well-known

example is the "Philadelphia chromosome", resulting from the translocation of chromosomes 9 and 22 and was first reported by Rowley (1973). This specific translocation leads to the formation of the *BCR-ABL1* oncoprotein, a product of the gene fusion between *BCR* and *ABL1* genes and the root cause of chronic myeloid leukemia (CML). Other common rearrangements found in hematological malignancies include translocations t(8,21)(*RUNX1-RUNX1T1* genes), t(15,17) (*PML-RARA*), t(9,11) (*MLLT3-KMT2A*), t(6,9) (*DEK-NUP214*), t(1,22) (*OTT-MAL*), inversions on chromosomes 16 (*CBFB-MYH11*) and 3 (*GATA2-MECOM*). Multiple gene fusions have also been identified as recurrent in solid tumors, most notably *TMPRSS2* and *ETS* gene fusion resulting from interstitial deletion or translocation of chromosome 21 in prostate cancers (Tomlins *et al.* 2005), *MYB-NFIB* gene fusion consequence of a t(6,9) translocation in adenoid cystic carcinomas (Persson *et al.* 2009), and *EML4-ALK* gene fusions caused by small inversion within chromosome 2p in about 7% of lung cancers (Soda *et al.* 2007). As the role of cytogenetics in carcinogenesis becomes better understood, therapeutic approaches targeting these events are increasingly being developed. Notable examples of this are the FDA and European Medical Agency (EMA) approval of imatinib in 2001, which revolutionized the treatment of CML and transformed a fatal cancer into a manageable condition (Deininger *et al.* 2005), or larotrectinib in November 2018 for the tissue-agnostic treatment of pediatric and adult tumors with *NTRK* fusions[11] - a good example of how new therapies are being designed to target specific alterations.

In contrast to structural rearrangements, chromosomal numerical changes are not well understood in relation to cancer. Alterations in the number of copies of very localized regions, particularly copy gains in regions harboring oncogenes or copy losses in regions harboring tumor suppressor genes have long been implicated in cancer. Prime examples include the focal amplification of *ERBB2* on chromosome 17 in breast cancer which can reach *hundreds* of copies (Révillion *et al.* 1998) or the deletion of *CDKN2A* on chromosome 9 across multiple cancer types (Kamb *et al.* 1994). On a larger scale, *aneuploidy*, which refers to cells with either too many or too few chromosomes compared with normal cells, is the most common chromosomal abnormality, affecting up to 90% of solid tumors and 50% of blood tumors (Mitelman, Johansson & Fredrik 2012; Taylor *et al.* 2018). Although it is frequent in tumor cells, some consider aneuploidy to be a byproduct of their increased genome instability, while others argue that it plays a significant role in tumor growth, development, and adaptability (Gordon *et al.* 2012; Holland & Cleveland 2012). Whole-chromosome aneuploidies are diverse across chromosomes and tumor types, contributing to the overall heterogeneity of cancers (Taylor *et al.* 2018). Some aneuploidies are common across cancers, particularly gains of chromosome 8q and losses of chromosomes 8p and 17p, while others are only recurrently found in some tumor types, such as losses of chromosomes 5q or 7q in myeloid disorders (Jerez *et al.* 2012) or losses of chromosome 3p identified recurrently in squamous cancers (Taylor *et al.* 2018). In recent years, studies have focused more on chromosomal instability (CIN), a type of genomic instability that causes chromosomes to acquire gains, losses, and structural changes at a high rate over time and is known to originate mainly from missegregation errors during mitosis (Bakhoum, Silkworth, *et al.* 2014). Defects in DNA repair pathways,

[11] https://www.fda.gov/drugs/fda-approves-larotrectinib-solid-tumors-ntrk-gene-fusions-0

such as mutations in *BRCA1* or *BRCA2* genes which impair homologous recombination (HR) (Stewart *et al.* 2022) are presumed causes of CIN. Investigations of CIN in cell lines from distinct tumor types or large cancer cohorts, particularly TCGA, have demonstrated its association with poor prognosis, tumor heterogeneity, and drug resistance (A. J. Lee *et al.* 2011; McGranahan *et al.* 2012; Bakhoum & Cantley 2018; Watkins *et al.* 2020; Lukow *et al.* 2021; Drews *et al.* 2022). These studies unanimously recommend evaluating CIN for risk stratification and drug assessment in clinical trials or even propose new drug targets from correlative analyses with CIN patterns (Drews *et al.* 2022). However, these recommendations have yet to be implemented in clinical practice.

1.2.2.2. DNA sequence mutations

The completion of the human genome project in the early 2000s paved the way for establishing the reference human genome but also for producing cancer maps drawn from the sequencing of dozens, hundreds, and, nowadays, tens of thousands of human tumor genomes. In one of the first exome-wise descriptions of breast and colorectal tumors, at the cost of about $100,000 per case, the exome sequencing of tumor DNA collected from 22 patients showed how only a small number of genes were frequently mutated compared to a large number of genes infrequently mutated (Wood *et al.* 2007). Simultaneous studies on glioblastomas (Parsons *et al.* 2008) and pancreatic cancers (S. Jones *et al.* 2008) made similar observations on the distribution of mutations but also highlighted the heterogeneity of the mutational landscape within and across tumor types. It also emerged from these studies that among all mutations found in tumors' genomes, only a few contribute to tumor initiation and growth and are drowned in an ocean of *passenger* mutations with no causal relationship with the tumor. These mutations fall in genes playing key roles in cancer onset and development and classified into either oncogenes or tumor suppressor genes depending on their role. Oncogenes are involved in tumor circuits through mutations or amplifications that allow the protein to gain a tumor-promoting function, whereas tumor suppressor genes which normally protect against cancer are commonly inactivated through double-hit alterations affecting both alleles (double mutations, mutation combined with a copy loss, or loss of the two copies) in cancer cells.

Mutations in the DNA sequence are conveniently categorized according to the number of nucleotides affected, the nature of the changes, but also according to their predicted consequences. Changes in the nucleotide sequence are broadly classified either as *substitutions* if the number of nucleotides remains unchanged or as *insertions* and *deletions* if nucleotides are inserted or deleted, respectively. Substitutions on a single base, known as SNV, are by far the most frequent type of substitution, while substitutions on multiple consecutive bases are known as multi-nucleotide variants (MNVs). As there are only four different nucleotides, there are 12 possible base substitutions which are further subdivided between transversions if the substituted base is from a different family (a purine - adenine or guanine - is replaced by a pyrimidine - cytosine or thymine, or vice versa) and transitions otherwise. The transition/transversion ratio is generally around two along the genome but tends to be more elevated in coding regions because transversions are more likely to result in amino acid

changes. Insertions and deletions, commonly called indels, may be *frameshift* if the number of affected bases is not a multiple of 3, thereby changing the reading frame, or *inframe* otherwise.

Though genomic alterations may occur in any region of the genome, they have been mostly studied in the protein-coding sequences (about 1.5% of the overall genome) because of the intuitive hypothesis that changes in these regions would have the most deleterious consequences and, therefore, the highest probability of playing crucial roles in tumorigenesis. The effect of alterations in protein-coding regions can conveniently be predicted by the consequences they have on the amino acid sequence of the proteins they code for. Of all base substitutions identified in coding regions of cancer genomes, about 23% are synonymous mutations (Y. Sharma *et al.* 2019) (also known as *silent* mutations) that do not alter the amino acid encoded by the affected codon due to the redundancy of the genetic code. About 90% of all other non-synonymous substitutions are missense changes, 8% nonsense changes, and 2% affect splice sites or untranslated regions adjacent to the start codon - 5' region - or stop codon - 3' region (Vogelstein *et al.* 2013). Many missense mutations in genes that we now refer to as oncogenes have been implicated in neoplastic processes. Mutations of *KRAS*, one of the most prominent oncogenes, are present in approximately 25% of cancers and assumed to drive 32%, 40%, and up to 90% of lung, colorectal, and pancreatic cancers, respectively, mainly through missense mutations of its twelfth amino acid residue, glycine, as first evidenced by the seminal study of Reddy *et al.* (1982). Other mutations favor tumor onset or growth by inhibiting tumor suppressor genes, as is the case for *PTEN*, which is mutated in about 13% of all cancers and has to date more than 1,993 unique mutations listed in COSMIC database (Y.-R. Lee *et al.* 2018). *TP53*, which is the most mutated gene in cancer with a pan-cancer mutation frequency of about 50%, has a special status because it is classified as a tumor suppressor gene although 90% of encountered mutations are missense and 28% of all mutations are localized in a handful of hotspot codons (Baugh *et al.* 2018), a pattern that is usually characteristic of oncogenes. Frameshift indels, which have deleterious consequences on the protein as they, with a high probability, introduce a stop coding shortly after the site of the alteration, are also commonly encountered mechanisms through which tumor cells inactivate genes. By contrast, inframe indels may have opposite consequences and play an activating or facilitating role for the tumor, as is the case for the inframe deletions in exon 19 of the epidermal growth factor receptor (*EGFR*) gene, which are the most common activating mutations in non-small cell lung cancer (NSCLC) and associate with sensitivity to tyrosine kinase inhibitor (TKI) treatments.

With the spread of whole-genome sequencing (WGS), many studies have investigated mutations in non-coding sequences of the DNA and made it increasingly clear that mutations falling outside exons play critical roles in the disruption of normal cell behavior (Rheinbay *et al.* 2020; Gutman *et al.* 2021). Examples include mutations in regulatory regions, such as promoter and enhancer regions like the well-studied *TERT* promoter (Huang *et al.* 2013), which directly influence the transcription rate of downstream genes through the destruction or formation of new transcription factor-binding sites (Melton *et al.* 2015); mutations in untranslated regions or introns of genes which can affect splicing, gene expression, messenger

RNA (mRNA) stability, protein folding, among other consequences (Kimchi-Sarfaty *et al.* 2007; Supek *et al.* 2014). Similarly, though many of the past and current cancer genomics studies discard all synonymous variants from their analyses, a growing number of studies argue that even if they do not change the amino acid sequence of the protein, synonymous mutations matter (X. Shen *et al.* 2022) and may, in fact, account for 6-8% of all driver mutations occurring due to single-base substitutions (Supek *et al.* 2014). Some of the latter studies are, however, controversial, particularly (X. Shen *et al.* 2022) for which a fiercely critical paper was very recently published in Nature (Kruglyak *et al.* 2023).

1.2.2.3. Epigenetic alterations

Epigenetics, originally defined by C.H. Waddington as the "causal interactions between genes and their products, which bring phenotype into being", now more generally encompasses all *heritable* changes in cells that allow the regulation of their gene expression without alterations in their DNA sequence. As epigenetic processes play a fundamental role in gene expression modulation, dysregulations of these processes can alter the gene expression and, in some instances, direct cells toward a neoplastic state. It has indeed been demonstrated in many studies that DNA methylation alterations, histone modifications, nucleosome repositioning, and modified expression of non-coding RNAs, particularly miRNAs, are all enabling characteristics that facilitate the acquisition of cancer hallmark capabilities (P. A. Jones & Baylin 2002; S. Sharma *et al.* 2010).

DNA methylation is a widely studied epigenetic mechanism that plays a crucial role in the modification of gene expression in tumor cells. Methylation, or the addition of a methyl group to a molecule, is a frequent event in the nucleus of cells where DNA methylation occurs naturally through the activity of specific enzymes called DNA methyltransferases. These enzymes covalently attach methyl groups to the fifth carbon of cytosine residues within the pyrimidine ring. This process allows cells to modulate the expression of DNA segments without modifying their underlying sequences, particularly in gene promoter regions, where it can function to repress gene transcription. It also is a source of mutations since methylated cytosines followed by a guanine, called CpG dinucleotides, tend to spontaneously deaminate to thymines resulting in C>T transitions. Interestingly, CpG dinucleotides are underrepresented in the human genome except for regions known as CpG islands (CGIs) that are very rich in CpGs and are normally unmethylated so that spontaneous transition from cytosine to uracil (the consequence of the deamination of the unmethylated cytosine) gets quickly repaired by the cell. The approximately 25,000 CGIs of the human genome are found in about half of all gene promoters. In contrast, the other half of genes are mostly methylated and silent and tend to be tissue-specifically expressed.

Hypermethylation of CGI, which inversely correlates with gene expression, is the secondary mechanism by which tumor suppressor genes are inactivated and is found to be mutually exclusive with genetic mutations (P. A. Jones & Baylin 2007). Silencing of DNA repair genes via promoter hypermethylation is a common feature of many cancers, such as silencing of *BRCA1* in breast and ovarian cancer, *VHL* in clear cell renal cell carcinoma, or *MGMT* in

gliomas and colorectal cancers. *MLH1* mismatch repair gene is another example of a gene for which methylation-induced decrease of its expression enables genomic instability and, in particular, microsatellite instability (MSI) by strongly associating with hypermethylation of other CGIs. Aside from canonical genes, methylation is also an essential regulator of non-coding RNAs, particularly miRNAs which are critical regulators of gene expression. Examples include epigenetic silencing of miRNA miR-127 in bladder cancers, which normally regulates *BCL6* oncogene and can successfully be reactivated by treatment with chromatin-modifying drugs such as 5-aza-2-deoxycytidine (Saito *et al.* 2006); silencing of miRNA miR-124a that activates the *CDK6-RB1* oncogenic pathway in acute lymphocytic leukemia (ALL) (Agirre *et al.* 2009); or the silencing of vault RNA vtRNA1-3 which is associated with decreased survival in myelodysplastic syndrome (MDS) patients. Although much less frequent, hypomethylation of gene promoters is also a known mechanism through which some tumor cells upregulate oncogenes, as is the case for *ELMO3* in NSCLC patients (Su *et al.* 2014) or *IRX1* in osteosarcoma cell lines (Lu *et al.* 2015).

Over the past two decades, extensive research has delved into the role of epigenetic alterations and epigenetic regulator changes in cancer initiation, progression, and response to treatment. Pioneering work by Feinberg, Ohlsson, *et al.* (2006) has shed light on the crucial early role of epigenetic disruption in stem or progenitor cells, mediated by "tumor-progenitor genes", as a significant contributor to both tumor heterogeneity and progression alongside more commonly known genetic lesions. Building upon this work and ensuing research, the same authors expanded their framework for cancer epigenetics by introducing two additional gene categories, termed "epigenetic modifiers" and "epigenetic modulators" (Feinberg, Koldobskiy, *et al.* 2016), alongside the "tumor-progenitor genes" which they relabelled "epigenetic mediators". As outlined by the authors, the contribution of epigenetics in cancer is well exemplified by pediatric cancers, which often harbor little to no mutations, suggesting that critical tumor-driving events likely arise from epigenetic modifications. The identification of biallelic loss in the chromatin remodeler gene *SMARCB1* in malignant pediatric rhabdoid tumor stands as a prominent example of a cancer driver mechanism rooted in epigenetics. Large-scale sequencing initiatives, as previously described, have underscored the prevalence of mutations in epigenetic modifiers in various cancer types, impacting all levels of the epigenetic machinery from DNA methylation to histone modification and chromatin remodeling. Conversely, epigenetic mediators are seldom mutated but are recurrently the target of epigenetic modifications. The final category, denoted as epigenetic modulators, is presented as genes positioned upstream in signaling pathways serving as ways through which environmental factors exert stress on cells, pushing them toward a neoplastic state.

The reversible nature of epigenetic changes make them attractive therapeutic targets and have been the basis of promising therapies in multiple tumor types. Notable drugs targeting epigenetic alterations include DNA methyltransferase inhibitors (DNMTis) azacitidine, the first FDA and EMA-approved drug for treating high-risk MDS, and decitabine which was approved two years later in 2006 also for treating MDS. They may also increase the sensitivity to chemotherapies, such as for the alkylating agent temozolomide which is mostly indicated

in the treatment of brain tumors and was found to be particularly effective in cells where the repair gene *MGMT* has been silenced through methylation (Esteller *et al.* 2000).

1.3. Cancer classifications

1.3.1. Clinical descriptions

1.3.1.1. The primary site

It is generally observed that cancer originates from one specific location, called the primary site, before invading nearby tissues and spreading to distant parts of the organism through a mechanism called metastasis in the most severe cases. Though rare, some patients are diagnosed with multiple primary sites, either synchronous or metachronous. The coexistence of multiple primary sites should not be confounded with the metastatic spread of a single-primary cancer and is characterized by the observation of cell masses that are histologically different and located in distinct parts of the body. Depending on the definition used, the frequency of multiple primary sites is estimated in the range of 2-17% (Vogt *et al.* 2017), with concomitant metachronous sites three times more frequent than synchronous (Testori *et al.* 2015).

The primary site generally serves as the main descriptor of cancers. There is not a single test that can diagnose cancer and the localization of its primary site. Medical doctors assess possible tumor sites by analyzing the patient's symptoms, clinical and family histories, and any other information that may be relevant for guiding the tests that will help in establishing the diagnosis. There are various imaging tests available for diagnosing cancer, including bone scans, computerized tomography (CT) scans, magnetic resonance imaging (MRI), positron emission tomography (PET) scans, ultrasound, and X-rays. Mammograms are an example of X-ray imaging tests commonly used for routine surveillance in women and which can diagnose a breast tumor at the earliest stages of its development. Any of the 78 organs that make up the human body can be the site of a malignant tumor, and specific organs or groups of organs usually make up distinct cancer specialties organized in separate medical departments. The most commonly found oncology departments are hematology which deals with all hematological malignancies; neuro-oncology, which takes charge of brain and nervous system tumors; dermatologic oncology, which is interested in all skin cancers; head and neck oncology, which treats all tumors arising from the oral cavity, sinuses, salivary glands, and pharynx; thoracic oncology, which gathers all tumors from the chest (breast excluded) such as lung, esophageal, and thymic tumors; breast oncology which considers breast tumors exclusively; gastrointestinal oncology which considers all tumors from the digestive system (stomach, colon, rectum, liver, gallbladder, pancreas); genitourinary oncology which encompasses all tumors starting from the kidney, urinary tract, bladder, and male reproductive system (prostate, testes); gynecologic oncology which considers all tumors arising from organs of the female reproductive system (ovarian, uterine, cervical cancers). Of note, cancer specialties are not necessarily exclusive from each other, and the anatomical site of the neoplasm is not the determinant of some specialties, such as for pediatric cancers or sarcomas, which consider cancer patients according to their age and the nature of the

neoplastic tissue, respectively, rather than the specific site of their tumor.

The type of tissue from which the neoplasm grows is another essential factor for discriminating between cancers and is the basis for their histological classification. From a histological point of view, there are hundreds of cancer types defined in national or international nomenclatures such as the French Association pour le Développment de l'Informatique en Cytologie et Anatomie Pathologique (ADICAP)[12], the WHO International Classification of Diseases for Oncology, third edition (ICD-O-3)[13], the WHO International Classification of Diseases, tenth revision (ICD-10) and now eleventh revision (ICD-11)[14], the "Clinical Modification" of ICD-10 used by the United States to classify mortality data (ICD-10-CM)[15], or the WHO classifications of tumors[16], fifth and earlier editions, which consist in reference books on specific organs or groups of organs, known as the WHO/International Agency for Research on Cancer (IARC) Blue Books, mirroring the current spectrum of cancer specialties. Though tables exist for converting between the nomenclatures, many cancer types in one classification map to none specifically or on the contrary to multiple in other classifications. This makes the harmonization of data from different studies often difficult, although the accurate classification of cancer patients is crucial for clinical care and research, where the tumor type is almost invariably the primary factor for grouping patients. Broadly speaking, cancers may be grouped into four categories according to their tissue of origin, namely epithelial tissue, hematopoietic and lymphoid tissues, muscle and other connective tissues, and nervous tissues.

1.3.1.2. Cancers of epithelial tissue

All cancers originating from epithelial cells are classified into the general category of *carcinomas*. Based on their shape, epithelial cells are divided between between squamous, cuboidal, and columnar. Depending on the number and organization of cell layers, epithelial tissue is also divided into simple, stratified, and pseudostratified. Eventually, epithelial cells are also categorized according to their specialization into keratinized, transitional, olfactory, or glandular. The histological classification of carcinomas closely follows the classifications of epithelial cells, and accordingly, some histological subtypes are only found in specific organs.

Glandular cells are modified epithelial cells specialized to secret products found in the glands or the surface of certain organs as scattered unicellular glands. *Adenocarcinomas* are cancers arising from the glandular epithelial tissue and are the most common type of cancer in humans. They account for almost all prostate, colorectal, pancreatic, stomach, and breast cancers (>90% of cancers for each organ) and a large proportion of lung cancers (40%), among others. Other types of cancers that are not technically classified as adenocarcinomas, although they also arise from glandular tissues, include adenoid cystic carcinomas, which are

[12] https://smt.esante.gouv.fr/catalogue-des-terminologies

[13] https://www.who.int/standards/classifications/other-classifications/international-classification-of-diseases-for-oncology

[14] https://icd.who.int/en

[15] https://www.cdc.gov/nchs/icd/icd-10-cm.htm

[16] https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-Tumours

most commonly encountered in salivary gland cancers but also in the breast, skin, prostate, and various other areas, and sebaceous carcinomas which are a rare type of skin cancer originating from sebaceous glands.

Squamous cells are thin and flat cells found in different body parts, either as a single layer in blood vessels, lungs, kidneys, and the heart or as stratified layers in the skin, oral cavity, respiratory tract, vagina, and anal canal. Cancers arising from squamous cells are classified as *squamous cell carcinomas* and, although less frequent than adenocarcinomas, are often the second most frequent cancer type of multiple organs such as lung squamous cell carcinoma (LUSC), which represent 20% of all lung cancers, cutaneous squamous cell carcinoma (20% of all skin cancers), or the most frequent cancer type for other areas such as head and neck squamous cell carcinoma (HNSC), cervical squamous cell carcinoma, vulvar squamous cell carcinoma, anal squamous cell carcinoma (>90% of cancers from each area).

Transitional epithelium is a type of epithelium found in organs able to distend. In the human body, they are found only in certain parts of the urinary tract, namely the urethra, bladder, ureters, and renal pelvis. Cancers arising from these cells are classified as transitional cell carcinoma, also known as urothelial cell carcinoma, and are by far the most common type of bladder cancer. Similarly, basal cell carcinomas is a type of cancer that starts from specific epithelial stem cells found in the bottom layer of the epidermis and is the most common type of skin cancer. Neuroendocrine cells are yet another specific type of epithelial cells which give rise to malignant tumors called neuroendocrine tumors (NETs) or neuroendocrine carcinomas (NECs) depending on their growth rate. Neuroendocrine cells are present in various organs, including the gastrointestinal tract, lungs, liver, and pancreas, with approximately 50% and 20% of NETs initiating in the gastrointestinal tract and lungs, respectively. NECs tend to be highly aggressive and have limited therapeutic options (Oronsky *et al.* 2017).

Besides the epithelium, two additional prominent types of epithelial cell layers include the mesothelium and endothelium, which form the lining of internal organs, blood vessels, and body cavities. These layers are also the potential origin sites for carcinomas. Specifically, mesotheliomas are cancers that emerge from the mesothelium tissue, such as the peritoneum, pleura, or pericardium. In contrast, endothelial cell cancers are rare neoplasms that are generally classified as sarcomas, such as angiosarcoma, or Kaposi's sarcoma. While many other types of carcinomas exist and contribute to the vast diversity of cancers, they are not described here for conciseness.

1.3.1.3. Cancers of hematopoietic and lymphoid tissues

Leukemias and lymphomas, also known as "blood cancers", are neoplastic diseases that affect the blood and the immune system. leukemia originates from hematopoietic stem and progenitor cells located in the bone marrow, the tissue that produces the different blood cells (red blood cells, platelets, and white blood cells). Some leukemias infrequently manifest themselves outside the bone marrow, such as in the lymph nodes or spleen, and are referred to as extramedullary or aleukemic leukemias, although these are generally associated with leukemias originating in the bone marrow. The chronic or acute nature of leukemia

is dependent on the maturation stage of the affected cells. Chronic leukemia develops gradually and may take a long time before causing symptoms, whereas acute leukemia is a life-threatening condition that necessitates prompt and aggressive intervention. Leukemias can be divided into two broad categories based on the type of stem cells from which they arise: myeloid (or myelogenous) if they develop from myeloid progenitors or lymphoid (or lymphoblastic, lymphocytic) if they arise from cells that will become lymphocytes. Overall, we typically distinguish four subtypes of leukemias based on the lineage and maturity of the affected cells: acute myeloid leukemias, chronic myeloid leukemias, acute lymphocytic leukemias, and chronic lymphocytic leukemias (CLLs).

In contrast, lymphoma initiates in the lymphatic system's infection-fighting cells, known as lymphocytes, which are located in the primary lymphoid organs (bone marrow, thymus) and secondary lymphoid organs, including lymph nodes, spleen, tonsils, or the mucosa-associated lymphoid tissue (MALT) situated in various submucosal membranes in the body, such as the gastrointestinal tract, nasopharynx, thyroid gland, breast, salivary glands, eyes, and skin. Lymphomas are classified as Hodgkin's lymphomas (which make up 11% of all lymphomas) or non-Hodgkin's lymphomas. Hodgkin's lymphomas are characterized by the presence of abnormally large and multi-nucleated lymphocytes known as Reed-Sternberg cells.

Multiple myeloma is a distinct form of cancer that falls outside the leukemias/lymphomas classification. It represents the most common lymphoid neoplasm involving plasma cells and is characterized by bone lesions, anemia, renal insufficiency, and hypercalcemia. Unlike leukemia, it lacks circulating tumor cells, and it is not a lymphoma since it does not originate from lymph node tissues. Instead, multiple myeloma arises from differentiated plasma cells in the bone marrow, which are a type of white blood cell responsible for antibody production. In this condition, cancerous plasma cells produce an excess of a specific type of immunoglobulins called monoclonal proteins. The presence of elevated immunoglobulin levels in the blood serves as one of the diagnostic indicators for myeloma.

1.3.1.4. Cancers of muscle and other connective tissues

Sarcomas are types of cancer that develop from connective tissues such as bones, cartilage, fat, and blood vessels, as well as from muscle and nerve tissues. Although they account for approximately 1% and 15% of all adult and childhood cancers, respectively, there are more than 175 different subtypes of sarcomas (mondiale de la santé & international de recherche sur le cancer 2020). They are classified based on the type of tissue from which they originate, with a broad classification into either bone or soft tissue sarcoma and according to their characteristics.

Bone sarcomas are a highly heterogeneous group of malignancies representing less than 0.2% of all cancers. They are thought to arise primarily from mesenchymal stem cells which can differentiate into different mesenchymal tissues, including bones and articular cartilage. Osteosarcoma, Ewing sarcoma, and chondrosarcoma are the three main types of bone tumors and differ in the affected populations, locations, and biological characteristics. Osteosarcoma is the most common malignant primary bone tumor and has a higher incidence in adolescents

and young adults. It arises from osteoblasts which form bone tissue and usually affects the long bones of the arm and leg. Ewing sarcoma accounts for 2% of childhood cancers and is more predominant in males. It also affects long bones but can occasionally originate from soft tissues. It is characterized by undifferentiated small round cells with a high nuclear-cytoplasmic ratio when viewed under the microscope. Chondrosarcoma is the third most common bone sarcoma and develops in cartilage cells, also known as chondrocytes.

Soft tissue sarcomas are also a diverse group of cancers that arise from the body's soft tissues. There are many subtypes of soft tissue sarcoma, but some of the main ones include leiomyosarcoma which arises from smooth muscle cells and mostly occurs in the uterus and gastrointestinal tract; rhabdomyosarcoma which arises from skeletal muscle cells that have not fully matured and is most common in young patients; liposarcoma which arises from fat cells and most commonly occurs in the limbs, but can also occur in the abdomen and other locations; synovial sarcoma that arises from cells in the synovial lining of joints in the limbs, head and neck, and other areas; malignant peripheral nerve sheath tumor (MPNST) which arises from cells that surround nerves and can occur anywhere in the body; gastrointestinal stromal tumor (GIST), a cancer that arises from cells in the wall of the gastrointestinal tract, most commonly in the stomach or small intestine; undifferentiated pleomorphic sarcoma (UPS) for which the cells have a spindle-shaped appearance and lack distinct morphological features of specific cell types.

1.3.1.5. Cancers of nervous tissues

The nervous system is composed of two main parts: the central nervous system (CNS), which includes the brain and the spinal cord, and the peripheral nervous system, which is made up of nerves that extend from the spinal cord to all parts of the body. Based on this division, nerve cell tumors are categorized as either CNS tumors or peripheral nerve tumors depending on their primary location. The classification of nervous system cancers primarily considers CNS tumors, as tumors arising from peripheral nerves are typically benign, such as acoustic neuroma, neurofibroma, and schwannoma, or classified as soft tissue sarcomas for the malignant types, particularly MPNSTs.

Malignant tumors of the brain and other parts of the CNS account for approximately 1% of all invasive cancers in the United States, but they contribute to a much higher proportion of cancer-related deaths due to their high fatality rate. In fact, these tumors have a low five-year survival rate of approximately 35%, with certain subtypes, such as glioblastomas, having even lower rates that do not exceed 7% at present (Miller *et al.* 2021). The epidemiology, treatment, and prognosis of malignant brain and other CNS tumors vary significantly between adults and children, which has resulted in different classification systems being used for adult and pediatric patients.

Malignant tumors of the CNS are classified based on their cell type of origin and location within the brain or spinal cord. CNS stem cells, known as neuroepithelial cells, differentiate into intermediate progenitor cells, such as radial glial cells, which can develop into neurons or non-neuronal cells, including glial cells that support and nourish neurons. CNS cancers are

classified as gliomas when they arise from glial cells, neuronal tumors when they originate from neuronal cells, and glioneuronal tumors when they consist of both glial and neuronal components. Most CNS tumors come from glial and other non-neuronal cells. Diffuse gliomas, which grow diffusively and invade functional tissue of the CNS, are the most common glial tumors in adults. Based on the similarity of tumor cells with non-neoplastic glial cells, most diffuse gliomas can be classified as astrocytomas (astrocytes), oligodendrogliomas (oligodendrocytes), and ependymomas (ependymal cells). Gliomas are also subdivided based on tumor grade into high-grade glioblastomas (grade IV), anaplastic gliomas (grade III), and low-grade gliomas (LGGs, grade II). Neuronal tumors are a rare group of brain and spinal tumors composed of abnormal neurons. These tumors may originate purely from neurons or have mixed neuronal and glial components, which comprise a subset of glioneuronal tumors. The 2021 WHO classification of CNS tumors includes 14 distinct tumors within this classification (mondiale de la santé & international de recherche sur le cancer 2021). In most patients, these tumors grow slowly, have well-defined borders, and are therefore considered benign.

Embryonal tumors are a specific type of malignant tumors that arise from embryonal cells left over from fetal development. While most embryonal tumors occur within the CNS, some forms can develop outside the CNS, such as neuroblastoma, a type of cancer that develops from immature nerve cells, called neuroblasts, and which most commonly occurs in the adrenal glands; retinoblastoma, which develops in the retina and typically occurs in young children; and hepatoblastoma, a type of liver cancer that arises from immature liver cells also in young children. CNS embryonal tumors are classified into medulloblastomas and other embryonal tumors. Medulloblastoma, the most common malignant pediatric tumor, starts from the lower part of the brain known as the cerebellum. It is histologically classified into classic, desmoplastic/nodular, and large cell/anaplastic based on the appearance of cells under a microscope. Other embryonal tumors include atypical teratoid/rhabdoid tumors, embryonal tumors with multilayered rosettes, both of which are almost exclusively seen in children aged three years or younger, and CNS neuroblastoma, which is rare and for which only isolated case reports are available in the literature. Other rare subtypes of CNS tumors include choroid plexus tumors (3% of pediatric brain tumors), pineal tumors (1% of all primary brain tumors), or pituitary tumors (very few of which are malignant), which grow in the choroid plexus, pineal gland, and pituitary gland regions of the brain, respectively. The tumor grade is used to distinguish additional subtypes with distinct clinical characteristics.

1.3.1.6. Other cancers and classifications

Cancers may be further subdivided according to the *morphological aspect* of tumor cells as assessed by trained histopathologists. Examples of morphological descriptions of cancer cells include tubular adenocarcinomas which describe adenocarcinomas where malignant cells form tubular shapes, a pattern mostly seen in breast and gastrointestinal tumors; lobular and ductal carcinomas distinguishing breast cancers with cells forming lobular or ductal structures, respectively; papillary and follicular carcinomas which are characterized by the presence of papillary or follicular growth patterns, respectively, and are used to describe the two most

common forms of thyroid cancer; small cell and non-small cell carcinomas which distinguish two major entities of lung cancer; spindle cell sarcomas which arise from a type of elongated cells that have a characteristic shape resembling a spindle; clear cell cancer which describe cancerous cells with a clear appearance under the microscope and are a distinct subtype in kidney cancer known as clear-cell renal cell carcinoma; cribiform carcinomas characterized by small holes when viewed under a microscope and representing $<$4% of breast cancers, 0.5% of papillary thyroid carcinomas, but also observed in lung, stomach and colon cancers; round cell tumors which describe a group of highly aggressive malignant tumors characterized by round cells with increased nuclear-cytoplasmic ratio and encompasses entities such as peripheral neuroectodermal tumor, rhabdomyosarcoma, synovial sarcoma, non-Hodgkin's lymphoma, neuroblastoma, hepatoblastoma, Wilms tumor, and desmoplastic small round cell tumor.

The *level of differentiation* of tumor cells also serves as an important cancer descriptor. In cancer histopathology, a four-level grading system is used to describe how abnormal cancer cells look under the microscope. Low-grade or grade I and II tumors are well-differentiated, which means that the tumor cells are organized and look more like normal tissue, while high-grade or grade III and IV tumor cells are poorly differentiated. Poorly differentiated cancers are usually aggressive in nature and have an unfavorable prognosis compared to other cancers. In prostate cancers, the *Gleason score*, which dates back to 1966 (Gleason 1966), is a more elaborate grading system also used to describe the microscopic aspect of tumor cells. It sums the four-level grade values of the two most common grades in the tissue sample and has a value between three and seven. The Gleason score has been recently improved to describe five different microscopic patterns and is now used to define five Gleason grades according to the value of the updated Gleason score (from six or less for grade I to ten for grade V; Epstein *et al.* (2016)).

On rarer occasions, cancers may also display characteristics of two or more usual subtypes and are therefore considered as *mixed subtypes*. These malignancies are rare and typically associated with a poorer prognosis. Examples include basaloid squamous cell carcinomas found in the oral cavity, respiratory tract, and lungs, which display features of both squamous and basal cell carcinomas; adenosquamous carcinomas, containing both squamous and glandular cells, found in various body areas; mixed hepatocellular and cholangiocarcinomas; carcinosarcomas affecting the uterus and ovaries, which are a combination of carcinoma and sarcoma; mucoepidermoid carcinomas that can arise in the salivary or thyroid glands and other locations and contain both squamous cells and mucin-secreting cells (whether they are distinct from adenosquamous carcinomas or not is a topic of debate; White *et al.* (2022)); mixed lobular and ductal breast cancers. It is worth noting that cancer classifications are frequently updated as our comprehension of their genomic, histopathological, and phenotypic profiles improve, as well as as cases and pieces of evidence accumulate, particularly for ultra-rare cancers which are defined by incidence rates below one per million (Loskutov *et al.* 2022). However, in spite of all the above considerations, some cancers still remain unclassifiable today and are labelled cancers of unknown primary (CUPs). They represent 2% of all cancers diagnoses in our societies[17].

[17] https://www.cancer.org/cancer/types/cancer-unknown-primary/about/key-statistics.html

1.3.2. Molecular descriptions

Advancements in technologies are rapidly improving our ability to profile the genome, epigenome, transcriptome, and proteome of cancer cells, leading to a greater understanding of the molecular specificities of cancer and refinements of the classifications used to stratify patients in the clinic. Myeloid neoplasms serve as a prime example of how tumor classifications have been reshaped by the inclusion of recurrent molecular aberrations in the classification criteria. In 2001, after nearly three decades of categorization based on the French-American-British (FAB) morphological classification proposed by Bennett *et al.* (1976), the 3rd edition of the WHO/IARC Blue Books proposed a new stratification of myeloid neoplasms (Vardiman, Harris, *et al.* 2002). In this new classification, the blast threshold for diagnosing acute myeloid leukemia (AML) was reduced from 30% to 20% but, more importantly, it was advised to classify as AML myeloid malignancies harboring clonal recurrent cytogenetic abnormalities t(8,21), inv(16) or t(16,16), and t(15,17) *regardless* of the blast percentage. In the group of AML with recurrent abnormalities, encompassing about 30% of all de novo AMLs, four distinct entities were proposed based on four distinct genomic abnormalities. This substratification was motivated by the strong correlation between the molecular aberrations and morphological aspects of the cells, as well as the distinctive clinical features and more favorable response to therapy observed in these groups, rendering them "truly distinct clinicopathologicgenetic entities" (Vardiman, Harris, *et al.* 2002).

As new evidence emerged, the WHO has gradually revised its classification of myeloid neoplasms with notable changes including the definition of new subentities of AML defined by translocations t(9,11), t(6,9), t(3,3) (or inv(3)), or t(1,22), as well as provisional entities with *NPM1* and *CEBPA* mutations in the 2008 edition (Vardiman, Thiele, *et al.* 2009) which became definitive entities in the 2016 revision (Arber, Orazi, R. Hasserjian, *et al.* 2016) alongside two other provisional entities defined by the *BCR-ABL1* gene fusion (confirmed in 2022) or mutations in *RUNX1* (eliminated in 2022). These changes were informed by large-scale sequencing studies, such as the work of Papaemmanuil *et al.* (2016), which proposed 11 classes of AML based on the co-mutation patterns observed in more than 1,500 patients. Some of the proposed classes overlapped established entities, particularly the classes defined by gene fusions (inv(16), t(15,17), t(8,21), t(6,9)), but also new classes emerged from specific alterations in driver genes, such as the *CEBPA* biallelic inactivation, *IDH2* R172 hotspot mutation, *NPM1* mutations, or *TP53* aneuploidy. This molecularly-defined classification was recently refined in the work of Tazi *et al.* (2022), to which I contributed and which aimed at providing a unified classification of AML. The European Leukemia Net (ELN) international consortium also established classifications of AML in 2010 (Döhner, E. H. Estey, *et al.* 2010), 2017 (Döhner, E. Estey, *et al.* 2017), and 2022 (Döhner, Wei, *et al.* 2022). Unlike the WHO, the ELN classifications are based *solely* on genetic alterations and divide patients into originally four, and now three, prognostic groups. Single or combined genetic abnormalities define risk group criteria, such *NMP1*-mutations, where concomitant absence or presence of *FLT3*-ITD determines favorable or intermediate risk, respectively (Döhner, Wei, *et al.* 2022). In contrast to the WHO and ELN classifications, the International Consensus Classification (ICC) of myeloid neoplasms, established by a consortium separate from the

WHO for lack of consensus, utilizes molecular features even more extensively as distinctive criteria. For example, they advise considering *TP53*-mutated neoplasms as a separate entity due to their unique genetic profile, poor-risk cytogenetics, and overall dismal outcome and argue that single or multi-hit mutations of *TP53* should override morphological variants and other classification criteria (Arber, Orazi, R. P. Hasserjian, *et al.* 2022).

Breast cancer serves as another example of how studies of the molecular landscape of cancer specimens have been used to define clinical subentities. In the study of Perou *et al.* (2000), variations in the expression of approximately 8,000 genes in 65 breast cancer specimens were examined, leading to the identification of four subtypes: luminal, HER2-overexpressing, basal-like, and normal-like. The luminal subtype expressed estrogen receptor (ER) target genes, while epidermal growth factor 2 (HER2)-overexpressing was associated with *ERBB2* gene amplification. The basal-like subtype did not express HER2 nor hormone receptors, and normal-like specimens were similar to normal tissue and thought to reflect low-purity samples rather than a distinct entity. Further investigation by the same authors on 85 biospecimens revealed that the luminal subtype could be subdivided into luminal A and luminal B, and potentially luminal C, based on the expression patterns of a short list of 427 genes (Sorlie *et al.* 2001). The luminal A subtype was characterized by ER-positive/Ki67-negative tumors, while luminal B encompassed ER-positive/Ki-67-positive tumors (Prat & Perou 2011). A fifth intrinsic breast cancer subtype known as *Claudin-low* was established in the study of human and mouse tumors Herschkowitz *et al.* (2007). This subtype was associated with poor prognosis, absence of epidermal growth factor (HER2) or hormone receptors (ER, PR), and high frequency of metaplastic or medullary differentiation. As different classifications emerged from different studies of gene expression patterns in breast cancer samples, demonstration of the robustness of the predicted subtypes became a crucial step for clinical use. Kapp *et al.* (2006) showed that three main intrinsic subtypes of breast cancer could accurately be predicted using only the expression of the genes encoding the HER2 (*ERBB2*) and ER proteins (*ESR1*), resulting in the three-group classification ER-/HER2- (basal-like or triple-negative), HER2+ (HER2-overexpressing) and ER+/HER2- (luminal A and B combined). Although various assays have been developed to classify single samples into one of the five main intrinsic subtypes, such as the 50-gene PAM50 predictor (Parker *et al.* 2009), or to risk-stratify patients using limited gene panels (Paik *et al.* 2004), the three-group classification based on the presence or absence HER2 and hormone receptors is the only molecular classification widely used in the breast cancer clinics.

Characterizing thousands or even tens of thousands of cancer samples in large-sequencing projects has revealed the molecular portraits of the different tumor types defined by anatomical and histological considerations. In particular, the TCGA research network has conducted extensive molecular analyses of each of the 33 tumor types they have studied and have defined for many of them molecular substratifications using joint clustering methods (Vaske *et al.* 2010; The Cancer Genome Atlas Network 2012; R. Shen *et al.* 2012; Qianxing Mo 2017). Table A.1 lists the different subtypes used or defined in the analyses of the TCGA, along with references to the corresponding studies. These large-scale sequencing efforts have also proven useful to help classify patients for whom the primary site of the tumor cannot be determined

by a trained histopathologist, a condition referred to as CUPs. Such cases still represent about 2% of all cancer diagnoses today. Recent studies leveraging the vast amount of data collected by the TCGA or PCAWG have demonstrated that highly accurate prediction of the primary site could be achieved via the profiling of the transcriptome (Vibert *et al.* 2021), genome (Marquard *et al.* 2015; Jiao *et al.* 2020; Moon *et al.* 2023), or epigenome (Moran *et al.* 2016) of cancer cells. It is also worth noting that some studies, such as the work of Capper *et al.* (2018), have demonstrated how tumor sequencing could help clinicians establish precise cancer diagnoses, particularly for neoplasms with many histological subtypes, such as brain tumors or soft tissue sarcomas.

Bibliography

1. Agirre, X. *et al.* Epigenetic Silencing of the Tumor Suppressor MicroRNA *Hsa-miR-124a* Regulates CDK6 Expression and Confers a Poor Prognosis in Acute Lymphoblastic Leukemia. en. *Cancer Research* **69,** 4443–4453. doi:10.1158/0008-5472.CAN-08-4025 (May 2009).
2. Alberts, B. *et al. Essential cell biology* Fifth edition. eng (W. W. Norton & Company, New York London, 2019).
3. Arber, D. A., Orazi, A., Hasserjian, R. P., *et al.* International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. en. *Blood* **140,** 1200–1228. doi:10.1182/blood.2022015850 (Sept. 2022).
4. Arber, D. A., Orazi, A., Hasserjian, R., *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. en. *Blood* **127,** 2391–2405. doi:10.1182/blood-2016-03-643544 (May 2016).
5. Bakhoum, S. F. & Cantley, L. C. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. en. *Cell* **174,** 1347–1360. doi:10.1016/j.cell.2018.08.027 (Sept. 2018).
6. Bakhoum, S. F., Silkworth, W. T., *et al.* The mitotic origin of chromosomal instability. en. *Current Biology* **24,** R148–R149. doi:10.1016/j.cub.2014.01.019 (Feb. 2014).
7. Baugh, E. H., Ke, H., Levine, A. J., Bonneau, R. A. & Chan, C. S. Why are there hotspot mutations in the TP53 gene in human cancers? en. *Cell Death & Differentiation* **25,** 154–160. doi:10.1038/cdd.2017.180 (Jan. 2018).
8. Bennett, J. M. *et al.* Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. en. *British Journal of Haematology* **33,** 451–458. doi:10.1111/j.1365-2141.1976.tb03563.x (Aug. 1976).
9. Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. en. *Nature* **555,** 469–474. doi:10.1038/nature26000 (Mar. 2018).
10. Chaffer, C. L. & Weinberg, R. A. A Perspective on Cancer Cell Metastasis. en. *Science* **331,** 1559–1564. doi:10.1126/science.1203543 (Mar. 2011).
11. Cheng, D. T. *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). en. *The Journal of Molecular Diagnostics* **17,** 251–264. doi:10.1016/j.jmoldx.2014.12.006 (May 2015).
12. Deininger, M., Buchdunger, E. & Druker, B. J. The development of imatinib as a therapeutic agent for chronic myeloid leukemia. en. *Blood* **105,** 2640–2653. doi:10.1182/blood-2004-08-3097 (Apr. 2005).
13. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. en. *Cell* **173,** 305–320.e10. doi:10.1016/j.cell.2018.03.033 (Apr. 2018).
14. Döhner, H., Estey, E. H., *et al.* Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. en. *Blood* **115,** 453–474. doi:10.1182/blood-2009-07-235358 (Jan. 2010).

15. Döhner, H., Estey, E., *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. en. *Blood* **129,** 424–447. doi:10.1182/blood-2016-08-733196 (Jan. 2017).
16. Döhner, H., Wei, A. H., *et al.* Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. en. *Blood* **140,** 1345–1377. doi:10.1182/blood.2022016867 (Sept. 2022).
17. Drews, R. M., Barbara, H. & Markowetz, F. A pan-cancer compendium of chromosomal instability. en. *Nature,* 24 (June 2022).
18. Epstein, J. I. *et al.* The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. en. *American Journal of Surgical Pathology* **40,** 244–252. doi:10.1097/PAS.0000000000000530 (Feb. 2016).
19. Esteller, M. *et al.* Inactivation of the DNA-Repair Gene *MGMT* and the Clinical Response of Gliomas to Alkylating Agents. en. *New England Journal of Medicine* **343,** 1350–1354. doi:10.1056/NEJM200011093431901 (Nov. 2000).
20. Exome Aggregation Consortium *et al.* Analysis of protein-coding genetic variation in 60,706 humans. en. *Nature* **536,** 285–291. doi:10.1038/nature19057 (Aug. 2016).
21. Falcaro, M. *et al.* The effects of the national HPV vaccination programme in England, UK, on cervical cancer and grade 3 cervical intraepithelial neoplasia incidence: a register-based observational study. en. *The Lancet* **398,** 2084–2092. doi:10.1016/S0140-6736(21)02178-4 (Dec. 2021).
22. Feinberg, A. P., Koldobskiy, M. A. & Göndör, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. en. *Nature Reviews Genetics* **17,** 284–299. doi:10.1038/nrg.2016.13 (May 2016).
23. Feinberg, A. P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. en. *Nature Reviews Genetics* **7,** 21–33. doi:10.1038/nrg1748 (Jan. 2006).
24. Gleason, D. F. Classification of prostatic carcinomas. eng. *Cancer Chemotherapy Reports* **50,** 125–128 (Mar. 1966).
25. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. en. *Nature Reviews Genetics* **13,** 189–203. doi:10.1038/nrg3123 (Mar. 2012).
26. Gutman, T., Goren, G., Efroni, O. & Tuller, T. Estimating the predictive power of silent mutations on cancer classification and prognosis. en. *npj Genomic Medicine* **6,** 67. doi:10.1038/s41525-021-00229-1 (Aug. 2021).
27. Herschkowitz, J. I. *et al.* Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology* **8,** R76. doi:10.1186/gb-2007-8-5-r76 (2007).
28. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. en. *Cell* **173,** 291–304.e6. doi:10.1016/j.cell.2018.03.022 (Apr. 2018).
29. Holland, A. J. & Cleveland, D. W. Losing balance: the origin and impact of aneuploidy in cancer: Exploring aneuploidy: the significance of chromosomal imbalance review series. en. *EMBO reports* **13,** 501–514. doi:10.1038/embor.2012.55 (June 2012).

30. Huang, F. W. *et al.* Highly Recurrent *TERT* Promoter Mutations in Human Melanoma. en. *Science* **339,** 957–959. doi:10.1126/science.1229259 (Feb. 2013).
31. Jerez, A. *et al.* Loss of heterozygosity in 7q myeloid disorders: clinical associations and genomic pathogenesis. en. *Blood* **119,** 6109–6117. doi:10.1182/blood-2011-12-397620 (June 2012).
32. Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. en. *Nature Communications* **11,** 728. doi:10.1038/s41467-019-13825-8 (Feb. 2020).
33. Jo, B.-S. & Choi, S. S. Introns: The Functional Benefits of Introns in Genomes. en. *Genomics & Informatics* **13,** 112. doi:10.5808/GI.2015.13.4.112 (2015).
34. Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. en. *Nature Reviews Genetics* **3,** 415–428. doi:10.1038/nrg816 (June 2002).
35. Jones, P. A. & Baylin, S. B. The Epigenomics of Cancer. en. *Cell* **128,** 683–692. doi:10.1016/j.cell.2007.01.029 (Feb. 2007).
36. Jones, S. *et al.* Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. en. *Science* **321,** 1801–1806. doi:10.1126/science.1164368 (Sept. 2008).
37. Kaiser, J. 200,000 whole genomes made available for biomedical studies. en. *Science* **374,** 1036–1036. doi:10.1126/science.acx9689 (Nov. 2021).
38. Kamb, A. *et al.* A Cell Cycle Regulator Potentially Involved in Genesis of Many Tumor Types. en. *Science* **264,** 436–440. doi:10.1126/science.8153634 (Apr. 1994).
39. Kapp, A. V. *et al.* Discovery and validation of breast cancer subtypes. en. *BMC Genomics* **7,** 231. doi:10.1186/1471-2164-7-231 (Dec. 2006).
40. Kimchi-Sarfaty, C. *et al.* A "Silent" Polymorphism in the *MDR* 1 Gene Changes Substrate Specificity. en. *Science* **315,** 525–528. doi:10.1126/science.1135308 (Jan. 2007).
41. Kruglyak, L. *et al.* Insufficient evidence for non-neutrality of synonymous mutations. en. *Nature* **616,** E8–E9. doi:10.1038/s41586-023-05865-4 (Apr. 2023).
42. Lee, A. J. *et al.* Chromosomal Instability Confers Intrinsic Multidrug Resistance. en. *Cancer Research* **71,** 1858–1870. doi:10.1158/0008-5472.CAN-10-3604 (Mar. 2011).
43. Lee, Y.-R., Chen, M. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor: new modes and prospects. en. *Nature Reviews Molecular Cell Biology* **19,** 547–562. doi:10.1038/s41580-018-0015-0 (Sept. 2018).
44. Lei, J. *et al.* HPV Vaccination and the Risk of Invasive Cervical Cancer. en. *New England Journal of Medicine* **383,** 1340–1348. doi:10.1056/NEJMoa1917338 (Oct. 2020).
45. Lejeune, J., Gautier, M. & Turpin, R. Etude des chromosomes somatiques de neuf enfants mongoliens. fr. *Comptes rendus hebdomadaires des séances de l'Académie des sciences* **248,** 1721–1722 (Mar. 1959).
46. Loskutov, J. *et al.* Abstract 6224: The bad, the ugly and the ultra-rare: All cancers are equal in the face of personalized medicine. en. *Cancer Research* **82,** 6224–6224. doi:10.1158/1538-7445.AM2022-6224 (June 2022).

47. Lu, J. *et al.* IRX1 hypomethylation promotes osteosarcoma metastasis via induction of CXCL14/NF-B signaling. en. *Journal of Clinical Investigation* **125,** 1839–1856. doi:10.1172/JCI78437 (May 2015).
48. Lukow, D. A. *et al.* Chromosomal instability accelerates the evolution of resistance to anti-cancer therapies. en. *Developmental Cell* **56,** 2427–2439.e4. doi:10.1016/j.devcel.2021.07.009 (Sept. 2021).
49. Marquard, A. M. *et al.* TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. en. *BMC Medical Genomics* **8,** 58. doi:10.1186/s12920-015-0130-0 (Dec. 2015).
50. Martínez-Jiménez, F. *et al.* Pan-cancer whole-genome comparison of primary and metastatic solid tumours. en. *Nature* **618,** 333–341. doi:10.1038/s41586-023-06054-z (June 2023).
51. McGranahan, N., Burrell, R. A., Endesfelder, D., Novelli, M. R. & Swanton, C. Cancer chromosomal instability: therapeutic and diagnostic challenges: Exploring aneuploidy: the significance of chromosomal imbalance review series. en. *EMBO reports* **13,** 528–538. doi:10.1038/embor.2012.61 (June 2012).
52. Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. en. *Nature Genetics* **47,** 710–716. doi:10.1038/ng.3332 (July 2015).
53. Miller, K. D. *et al.* Brain and other central nervous system tumor statistics, 2021. en. *CA: A Cancer Journal for Clinicians* **71,** 381–406. doi:10.3322/caac.21693 (Sept. 2021).
54. Mitelman, F., Johansson, B. & Fredrik, M. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. doi:https://mitelmandatabase.isb-cgc.org/ (2012).
55. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. en. *Nature Reviews Cancer* **7,** 233–245. doi:10.1038/nrc2091 (Apr. 2007).
56. *Soft tissue and bone tumours* 5th ed. eng (eds mondiale de la santé, O. & international de recherche sur le cancer, C.) *World health organization classification of tumours* **Vol. 3** (OMS, Geneva, 2020).
57. *Central nervous system tumours* 5th ed. eng (eds mondiale de la santé, O. & international de recherche sur le cancer, C.) *World health organization classification of tumours* **6** (International agency for research on cancer, Lyon, 2021).
58. Moon, I. *et al.* Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary. en. *Nature Medicine* **29,** 2057–2067. doi:10.1038/s41591-023-02482-6 (Aug. 2023).
59. Moran, S. *et al.* Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. en. *The Lancet Oncology* **17,** 1386–1395. doi:10.1016/S1470-2045(16)30297-2 (Oct. 2016).
60. Morgan, G. *et al.* The (R)evolution of Social Media in Oncology: Engage, Enlighten, and Encourage. en. *Cancer Discovery* **12,** 1620–1624. doi:10.1158/2159-8290.CD-22-0346 (July 2022).

61. Nangalia, J. & Campbell, P. J. Genome Sequencing during a Patients Journey through Cancer. en. *New England Journal of Medicine* **381,** 2145–2156. doi:10.1056/NEJMra1910138 (Nov. 2019).
62. Nguyen, B. *et al.* Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. en. *Cell* **185,** 563–575.e11. doi:10.1016/j.cell.2022.01.003 (Feb. 2022).
63. Oronsky, B., Ma, P. C., Morgensztern, D. & Carter, C. A. Nothing But NET: A Review of Neuroendocrine Tumors and Carcinomas. en. *Neoplasia* **19,** 991–1002. doi:10.1016/j.neo.2017.09.002 (Dec. 2017).
64. Paik, S. *et al.* A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. en. *New England Journal of Medicine* **351,** 2817–2826. doi:10.1056/NEJMoa041588 (Dec. 2004).
65. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. en. *New England Journal of Medicine* **374,** 2209–2221. doi:10.1056/NEJMoa1516192 (June 2016).
66. Parker, J. S. *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. en. *Journal of Clinical Oncology* **27,** 1160–1167. doi:10.1200/JCO.2008.18.1370 (Mar. 2009).
67. Parsons, D. W. *et al.* An Integrated Genomic Analysis of Human Glioblastoma Multiforme. en. *Science* **321,** 1807–1812. doi:10.1126/science.1164382 (Sept. 2008).
68. Perou, C. M. *et al.* Molecular portraits of human breast tumours. en. *Nature* **406,** 747–752. doi:10.1038/35021093 (Aug. 2000).
69. Persson, M. *et al.* Recurrent fusion of *MYB* and *NFIB* transcription factor genes in carcinomas of the breast and head and neck. en. *Proceedings of the National Academy of Sciences* **106,** 18740–18744. doi:10.1073/pnas.0909114106 (Nov. 2009).
70. Ponting, C. P. & Haerty, W. Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review. en. *Annual Review of Genomics and Human Genetics* **23,** 153–172. doi:10.1146/annurev-genom-112921-123710 (Aug. 2022).
71. Prat, A. & Perou, C. M. Deconstructing the molecular portraits of breast cancer. en. *Molecular Oncology* **5,** 5–23. doi:10.1016/j.molonc.2010.11.003 (Feb. 2011).
72. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. en. *Nature* **575,** 210–216. doi:10.1038/s41586-019-1689-y (Nov. 2019).
73. Qianxing Mo, R. S. *iClusterPlus* 2017. doi:10.18129/B9.BIOC.ICLUSTERPLUS.
74. Ravi, R. & Kesari, H. V. Novel Study Designs in Precision Medicine Basket, Umbrella and PlatformTrials. en. *Current Reviews in Clinical and Experimental Pharmacology* **17,** 114–121. doi:10.2174/1574884716666210316114157 (July 2022).
75. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. en. *Nature* **300,** 149–152. doi:10.1038/300149a0 (Nov. 1982).
76. Révillion, F., Bonneterre, J. & Peyrat, J. ERBB2 oncogene in human breast cancer and its clinical significance. en. *European Journal of Cancer* **34,** 791–808. doi:10.1016/S0959-8049(97)10157-5 (May 1998).

77. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. en. *Nature* **578,** 102–111. doi:10.1038/s41586-020-1965-x (Feb. 2020).
78. Rowley, J. D. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. en. *Nature* **243,** 290–293. doi:10.1038/243290a0 (June 1973).
79. Saito, Y. *et al.* Specific activation of microRNA-127 with downregulation of the proto-oncogene BCL6 by chromatin-modifying drugs in human cancer cells. en. *Cancer Cell* **9,** 435–443. doi:10.1016/j.ccr.2006.04.020 (June 2006).
80. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. en. *Cell* **173,** 321–337.e10. doi:10.1016/j.cell.2018.03.035 (Apr. 2018).
81. Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. en. *Carcinogenesis* **31,** 27–36. doi:10.1093/carcin/bgp220 (Jan. 2010).
82. Sharma, Y. *et al.* A pan-cancer analysis of synonymous mutations. en. *Nature Communications* **10,** 2569. doi:10.1038/s41467-019-10489-2 (June 2019).
83. Shen, R. *et al.* Integrative Subtype Discovery in Glioblastoma Using iCluster. en. *PLoS ONE* **7** (ed Brusic, V.) e35236. doi:10.1371/journal.pone.0035236 (Apr. 2012).
84. Shen, X., Song, S., Li, C. & Zhang, J. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. en. *Nature* **606,** 725–731. doi:10.1038/s41586-022-04823-w (June 2022).
85. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. en. *CA: A Cancer Journal for Clinicians* **73,** 17–48. doi:10.3322/caac.21763 (Jan. 2023).
86. Smith, B. D., Smith, G. L., Hurria, A., Hortobagyi, G. N. & Buchholz, T. A. Future of Cancer Incidence in the United States: Burdens Upon an Aging, Changing Nation. en. *Journal of Clinical Oncology* **27,** 2758–2765. doi:10.1200/JCO.2008.20.8983 (June 2009).
87. Soda, M. *et al.* Identification of the transforming EML4ALK fusion gene in non-small-cell lung cancer. en. *Nature* **448,** 561–566. doi:10.1038/nature05945 (Aug. 2007).
88. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. en. *Proceedings of the National Academy of Sciences* **98,** 10869–10874. doi:10.1073/pnas.191367098 (Sept. 2001).
89. Stewart, M. D. *et al.* Homologous Recombination Deficiency: Concepts, Definitions, and Assays. en. *The Oncologist* **27,** 167–174. doi:10.1093/oncolo/oyab053 (Mar. 2022).
90. Su, S.-F. *et al.* A Panel of Three Markers Hyper- and Hypomethylated in Urine Sediments Accurately Predicts Bladder Cancer Recurrence. en. *Clinical Cancer Research* **20,** 1978–1989. doi:10.1158/1078-0432.CCR-13-2637 (Apr. 2014).
91. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. en. *CA: A Cancer Journal for Clinicians* **71,** 209–249. doi:10.3322/caac.21660 (May 2021).
92. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. en. *Cell* **156,** 1324–1335. doi:10.1016/j.cell.2014.01.051 (Mar. 2014).

93. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. en. *Cancer Cell* **33,** 676–689.e3. doi:10.1016/j.ccell.2018.03.007 (Apr. 2018).
94. Tazi, Y. *et al.* Unified classification and risk-stratification in Acute Myeloid Leukemia. en. *Nature Communications* **13,** 4622. doi:10.1038/s41467-022-32103-8 (Aug. 2022).
95. Testori, A. *et al.* Multiple primary synchronous malignant tumors. en. *BMC Research Notes* **8,** 730. doi:10.1186/s13104-015-1724-5 (Dec. 2015).
96. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. en. *Nature* **491,** 56–65. doi:10.1038/nature11632 (Nov. 2012).
97. The 1000 Genomes Project Consortium. A global reference for human genetic variation. en. *Nature* **526,** 68–74. doi:10.1038/nature15393 (Oct. 2015).
98. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. en. *Nature* **490,** 61–70. doi:10.1038/nature11412 (Oct. 2012).
99. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. en. *Nature* **455,** 1061–1068. doi:10.1038/nature07385 (Oct. 2008).
100. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. en. *Nature* **578,** 82–93. doi:10.1038/s41586-020-1969-6 (Feb. 2020).
101. Tomlins, S. A. *et al.* Recurrent Fusion of *TMPRSS2* and ETS Transcription Factor Genes in Prostate Cancer. en. *Science* **310,** 644–648. doi:10.1126/science.1117679 (Oct. 2005).
102. Vardiman, J. W., Harris, N. L. & Brunning, R. D. The World Health Organization (WHO) classification of the myeloid neoplasms. en. *Blood* **100,** 2292–2302. doi:10.1182/blood-2002-04-1199 (Oct. 2002).
103. Vardiman, J. W., Thiele, J., *et al.* The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. en. *Blood* **114,** 937–951. doi:10.1182/blood-2009-03-209262 (July 2009).
104. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. en. *Bioinformatics* **26,** i237–i245. doi:10.1093/bioinformatics/btq182 (June 2010).
105. Venter, J. C. *et al.* The Sequence of the Human Genome. en. *Science* **291,** 1304–1351. doi:10.1126/science.1058040 (Feb. 2001).
106. Vibert, J. *et al.* Identification of Tissue of Origin and Guided Therapeutic Applications in Cancers of Unknown Primary Using Deep Learning and RNA Sequencing (TransCUPtomics). en. *The Journal of Molecular Diagnostics* **23,** 1380–1392. doi:10.1016/j.jmoldx.2021.07.009 (Oct. 2021).
107. Vogelstein, B. *et al.* Cancer Genome Landscapes. en. *Science* **339,** 1546–1558. doi:10.1126/science.1235122 (Mar. 2013).
108. Vogt, A. *et al.* Multiple primary tumours: challenges and approaches, a review. en. *ESMO Open* **2,** e000172. doi:10.1136/esmoopen-2017-000172 (2017).

109. Watkins, T. B. K. *et al.* Pervasive chromosomal instability and karyotype order in tumour evolution. en. *Nature* **587,** 126–132. doi:10.1038/s41586-020-2698-6 (Nov. 2020).
110. White, V. A. *et al.* Mucoepidermoid carcinoma (MEC) and adenosquamous carcinoma (ASC), the same or different entities? en. *Modern Pathology* **35,** 1484–1493. doi:10.1038/s41379-022-01100-z (Oct. 2022).
111. Wood, L. D. *et al.* The Genomic Landscapes of Human Breast and Colorectal Cancers. en. *Science* **318,** 1108–1113. doi:10.1126/science.1145720 (Nov. 2007).
112. Yates, L. *et al.* The European Society for Medical Oncology (ESMO) Precision Medicine Glossary. en. *Annals of Oncology* **29,** 30–35. doi:10.1093/annonc/mdx707 (Jan. 2018).
113. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. en. *Nature Medicine* **23,** 703–713. doi:10.1038/nm.4333 (June 2017).

2. Analysis of high-throughput sequencing

Contents

Abstract Chapter 2

In this chapter, we will first cover the methods that have been developed to sequence DNA, process the sequencing files, and extract variations that exist in the analyzed genome when compared to the normal genome and reference genome. In a second section, we will examine how the genomic contexts surrounding localized somatic events in cancer genomes can reveal information about the mutagenic processes at play. Eventually, we will take a closer look at how cancer cells acquire and maintain malignant capabilities through specific genetic alterations known as driver events. All the concepts and data analyses described in this chapter lay the bricks for the genomic analyses presented in Chapters 3 and 4 analyzing cancer genomes from metastatic patients.

THE sequencing of DNA, i.e., the determination of the exact order of nucleotides that constitute the DNA chains, is considered one of the greatest accomplishments of modern science that has completely revolutionized life sciences and has spun the era of *computational biology*. As the first *essentially complete* sequence of the human genome was released in 2001 ten years into the publicly funded HGP (Lander *et al.* 2001; Venter *et al.* 2001), the achievements of this tremendous, collaborative, and international effort were beyond what scientists thought possible in 1988. The draft releases of the human genome revealed three critical features of human genomes: firstly, a human genome contains only 30,000 to 40,000 genes, which was considerably lower than expected; secondly, exons, which code for the amino acids making up all the proteins found in our cells, represent a mere 1.1% of our genome; lastly there are millions of SNPs locations and the genomes of any two individuals differ at a rate of 1 base per 1,250 on average (Venter *et al.* 2001). All of these initial estimates have been revised upwards over time as more complete genomes were assembled, many individual genomes were sequenced, and as many more genome annotations were described, but orders of magnitude have not changed. The HGP project cost about 3 billion dollars as a whole to obtain a finished assembled genome, including $300 million dollars for generating the first draft genome from April 1999 to June 2000, and lasted for 13 years[1]. Approximately 50% of the project overall cost was supported by the NCI and the other half by organizations from six countries.

The first rough draft of the human genome was released in June 2000 by the University of California, Santa Cruz (UCSC). A more complete version, known as the National Center for Biotechnology (NCBI) Build 33/hg15 (NCBI/UCSC versions), was released in April 2003. In May 2006, the sequencing of human chromosome 1, the largest of all our chromosomes, was completed (Gregory *et al.* 2006). Though the HGP came to an end in 2003, it is essential to note that its aim was not to sequence *all* the DNA found in human cells but only the euchromatic regions the nuclear DNA, constituting 92% of the genome. The 8% part of the genome not sequenced by the HGP consists of scattered heterochromatic regions found primarily in centromeres and telomeres and difficult to assemble due to their repetitive nature.

[1] https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

The Genome Reference Consortium (GRC) was founded in 2007 to improve the reference genome assemblies of human, mouse, and zebrafish. In February 2009, the GRC released the GRCh37/hg19 reference genome. The primary assembly of GRCh37 contained about 234 megabases (Mb) of unknown sequences distributed throughout the genome in 271 sequence gaps. This assembly was improved over time through patches until the 13th and final patch, GRCh37.p13, which was released in June 2013. In December 2013, the GRC released the GRCh38/hg38 version, which incorporates alternate contigs to represent common complex variation, most notably alternate haplotypes such as the HLA loci. The current patch of the latest genome assembly, known as GRCh38.p14, was released in February 2022 and now has only about 151 Mb of unknown sequence in 349 sequence gaps[2]. In 2022 also, the Telomere-to-Telomere Consortium published the first wholly assembled reference genome, known as T2T-CHM13, without any gaps in the assembly (Nurk *et al.* 2022).

The journey for assembling the human reference genome and building annotations as complete as possible has been a long one that has not yet reached its conclusion and is likely to continue for many years. In all the years since the first draft of the human genome was released, many individual genomes from healthy individuals or afflicted with a wide variety of Mendelian disorders have been sequenced and analyzed. Progressively more subtle and complex genetic variations have been described since the first human genome sequence as increasingly cheap technologies delivering increasingly high amounts of data were developed. Technological improvements have completely changed the amount and diversity of information that can be measured about DNA molecules and other molecules downstream in the biological workflow, namely RNAs and proteins. Exponential decreases in sequencing costs have allowed us to run sequencing experiments on a large collection of biological samples collected from humans and other species. The analyses of all the generated sequences have spun the building of many databases cataloging countless variants detected in healthy and unhealthy individuals. The low cost of sequencing, coupled with the availability of many algorithmic tools and extensive databases, now allows us to analyze sequencing data through many different techniques that provide quantitative and qualitative information about nucleic acid sequences. In this chapter, we will describe all the steps involved in the generation of sequencing data up to the extraction of biologically meaningful low-dimensional data that humans can comprehend and use to better understand the genomic underpinnings of disorders such as cancer.

2.1. From sequencing to variant detection

2.1.1. Sequencing techniques

Sequencing techniques are conventionally categorized into two overarching groups: first-generation sequencing techniques, which emerged in the late 1970s and remained the primary sequencing method for over three decades, and second-generation or NGS methods, which have been commercially accessible since 2005. Sections 2.1.1.1 and 2.1.1.2 offer brief presentations of these two broad types of sequencing. For conciseness and because they were

[2] https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh38.p14

not used in any of the project presented in this manuscript, single-cell sequencing technologies and the more recent spatial sequencing technologies are not discussed at all.

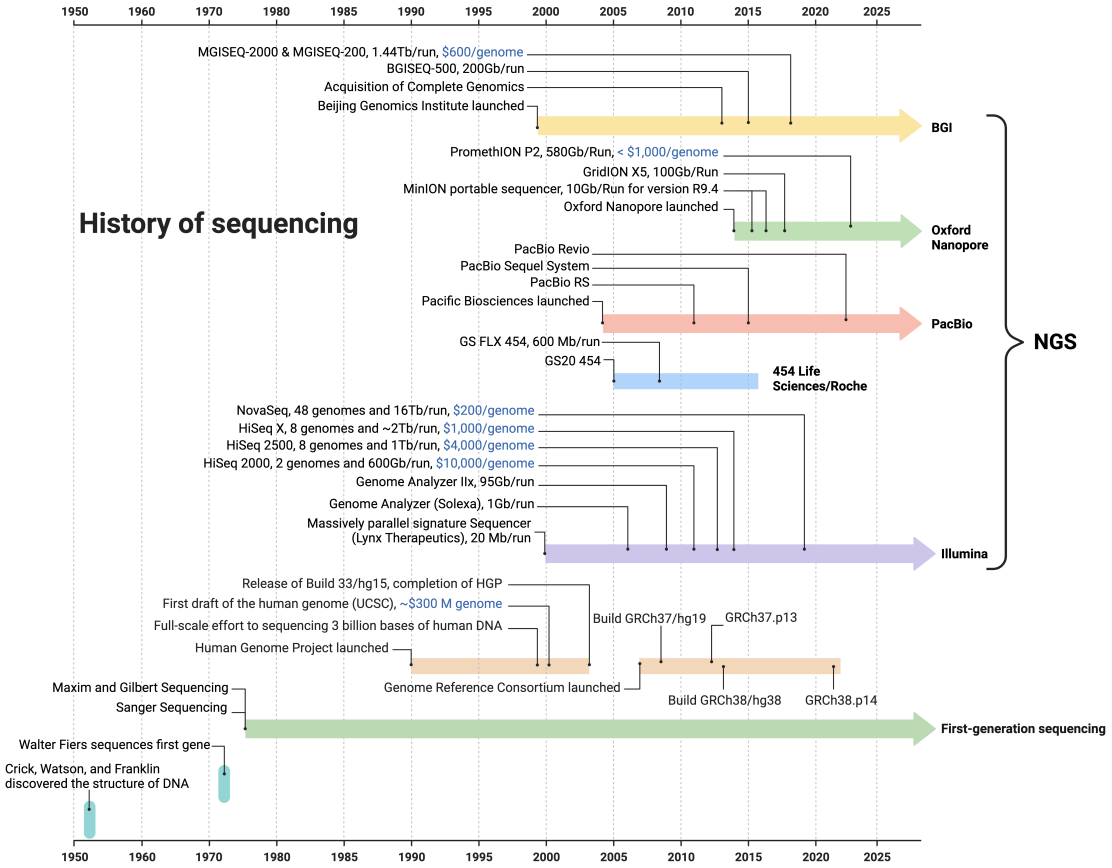


Fig. 2.1.: History of sequencing with a non-exhaustive list of the main next-generation sequencing products that have been developed. Created with BioRender.com

2.1.1.1. First-generation sequencing

The initial DNA sequences of specific organisms, including the preliminary blueprint of the human genome, were determined utilizing sequencing techniques commonly referred to as *first-generation sequencing* methods. In 1977, two distinct sequencing methodologies were introduced in the same year: *Maxam-Gilbert sequencing* (Maxam & Gilbert 1977) and *Sanger sequencing* (Sanger *et al.* 1977) (Figure 2.1). Both approaches rely on the fundamental idea that by generating incomplete copies of single-stranded DNA molecules with varying sizes starting from the first base, one can infer the nucleotide identity at each position of the template by examining the distribution of fragments of identical lengths.

In their pioneering technique, Maxam and Gilbert formulated four experiments employing diverse chemical treatments to modify specific bases and cleave subsequent nucleotides of

copies of a DNA template. In each experiment and for each copy of the DNA template, a cleavage point is designated, either selecting an A or G nucleotide (A+G), a G nucleotide exclusively (G), a C nucleotide exclusively (C), or either a C or G nucleotide (C+G). This generates fragment sizes that are unique to each experiment or overlapping with exactly one other experiment. Subsequently, X-ray file autoradiography or gel electrophoresis is employed to separate the fragments of each experiment based on their sizes. By scrutinizing the size distribution of fragments across all four experiments, the nucleotide sequence of the original DNA fragment can be readily determined. Due to its technical intricacy, challenges associated with scalability, and utilization of hazardous chemicals, the Maxam-Gilbert method gradually fell out of favor in favor of Sanger sequencing, which is presented in the next paragraph.

The method presented by Frederik Sanger from the MRC Center in Cambridge (UK) bears some similarities to the Maxam-Gilbert method, albeit with notable distinctions. *Sanger sequencing* begins with DNA extraction from a biospecimen comprising one or more cells, each housing its own DNA molecules. Depending on the sequencing objective and context, a DNA region of interest is chosen through the use of primers, either in close proximity (less than 700 base pairs apart) to hybridize with denatured DNA or at random following a process of natural or artificial DNA fragmentation to generate smaller, manageable fragments. Once a DNA fragment is selected, amplification becomes necessary to obtain an adequate quantity of material for sequencing. In the past, molecular cloning with bacteria served to generate numerous replicas of a specific DNA sequence. Presently, amplification is achieved through polymerase chain reaction (PCR). Typically, approximately 25 to 30 cycles of PCR are performed in Sanger sequencing to acquire sufficient material. Each PCR cycle entails three steps:

1. denaturation of the double-stranded DNA at $94°$C to separate it into single strands
2. primer binding (or annealing) at $54°$
3. DNA replication at $72°$

Subsequent cycles utilize the material from the previous cycle. As each cycle doubles the number of DNA chains, the theoretical number of PCR copies after n cycles is 2^n. The amplified DNA libraries are subsequently purified to isolate the PCR products. Each isolated PCR product is extended using a process similar to PCR, but with the incorporation of chain-terminating nucleotides or dideoxynucleotides (ddNTPs), alongside other PCR reactants. Each ddNTP is labeled with one of four radioactive or fluorescent tags differentiating the four types of nucleotides. The chain-terminating nature of ddTNPs allows the creation of incomplete and labeled copies of the original DNA fragment, encompassing all possible sizes. After approximately 20 cycles of DNA elongation, the partially elongated PCR products are denatured and separated by size using gel electrophoresis. Determination of each nucleotide within the DNA template of interest becomes a matter of examining the labels of copies sharing the same size.

2.1.1.2. Next-generation sequencing

In contrast to first-generation sequencing, the second-generation massively parallel sequencing techniques, also known as *next-generation sequencing (NGS)*, offer remarkable scalability. These methods have enabled the sequencing of extensive DNA regions at significantly low costs per sequenced base while maintaining high confidence levels. Targeted sequencing of DNA allows to sequence specific gene lists, representing genomic regions ranging from a couple hundreds kilobases (kb) to about 10 Mb. WES is used to sequence all or nearly all of the exome, covering between 30 and 50 Mb with capture kits. On the other hand, WGS encompasses the entire genome, i.e approximately 3 gigabases (Gb) of DNA, but at lower coverage depths than targeted sequencing or WES. While originally designed for sequencing DNA molecules, NGS also enables the sequencing of large libraries of RNA molecules, which are converted back into copy DNA (cDNA) through a process known as *reverse transcription*. Various library preparation protocols and capture techniques allow to focus on different types of RNAs: small RNA capture, for instance, facilitates the quantification of miRNAs and other small non-coding RNAs; polyA-enrichment selectively targets polyadenylated RNA species, primarily comprising mature mRNAs; probe-based depletion is employed to discard RNAs of little interest, primarily rRNAs, which can constitute between 80% and 98% of all RNA molecules in biological samples.

Several high-throughput sequencing techniques have been developed to enable rapid and cost-effective large-scale DNA and RNA sequencing. In 2000, Lynx Therapeutics introduced the massively parallel signature sequencer, considered the first NGS machine with a sequencing capability of 20 million base pairs per run. However, it wasn't until 2005 that the GS20 sequencer by 454 Life Sciences, later acquired by Roche, became commercially available for independent institutions and laboratories as depicted in Figure 2.1. In 2008, the Genome Sequencer FLX from 454 Life Sciences was used to sequence James Watson's complete genome (Wheeler *et al.* 2008). This and other individual genome sequencing projects revealed millions of genetic variations in any human genome compared to the human reference genome. Illumina played a significant role in advancing NGS with machines like the Genome Analyzer and HiSeq series, improving sequencing capabilities, data output, runtime, and cost. The NovaSeq series, launched in 2017, offers impressive sequencing capabilities, allowing parallel sequencing of up to 48 samples and sequencing up to 16 terabytes in less than 48 hours. While Illumina dominates the DNA sequencing market, other companies like Life Technologies, PacBio, QIAGEN, Thermo Fisher Scientific, Hoffman-La Roche, Oxford Nanopore Technologies, and BGI have developed comparable or complementary NGS techniques, some with potentially lower costs, such as BGI's claim of an $800 cost and a 10-day runtime for sequencing a complete genome[3].

The most prominent sequencing technique is the *sequencing by synthesis* method. In essence, sequencing by synthesis determines the DNA sequences of interest through real-time monitoring of inserted sequences along single-stranded DNA templates. During each cycle of the sequencing experiment, a new base is inserted and read using various techniques.

[3] https://www.bgi.com/global/service/whole-genome-sequencing-rapid

The most widely used such platforms are Illumina sequencers, which employ a reversible terminator-based method and monitor bases as they are inserted using fluorescent signals. Illumina devices are known for their high accuracy and throughput and are, in practice, the most commonly used. Ion Torrent sequencing or Roche-454 pyrosequencing are other notable methods that monitor bases inserted during DNA synthesis by detecting the release of hydrogen ions or pyrophosphatases, respectively.

Some technologies, such as Illumina sequencing or Ion Torrent sequencing, can only produce reads of limited sizes due to the decrease in the confidence of base calling as the number of cycles increases. These sequencers produce reads of sizes typically not exceeding a few hundred base pairs, known as *short reads*. Compared to Sanger sequencing, short-read NGS exhibits slightly lower accuracy, generates shorter read lengths, and has longer overall runtimes but offers significantly reduced costs and runtimes per sequenced base and much higher output. However, other technologies, such as PacBio single molecule real-time sequencing or Oxford Nanopore sequencing (Figure 2.1), allow the generation of reads of very long sizes. Nanopore sequencing has historically been limited by its high error rate, reaching up to 40% (Laver *et al.* 2015), and low throughput per experiment but has the advantage of being available in portable devices and producing sequences in real time. Additionally, recent improvements in chemistry and technology have pushed the accuracy of nanopore sequencing up to 99.9% using duplex sequencing (still lower than the 99.999% of Illumina sequencing) at data yields as high as 140 Gb for a single *Promethion* flow cell[4].

With the emergence of first NGS techniques, the question of how to quantify sequencing quality became a focal point. In 1998, Green and Ewing from the University of Washington introduced the *phred quality score* as a means to quantify sequencing quality (Ewing & Green 1998), and this score has been employed ever since in all NGS experiments. The phred score is calculated as ten times the negative logarithm of the probability of the sequenced base being incorrect ($Q = -10 \log_{10} P$) which is determined according to the peak shape and resolution of the base signal. Consequently, a phred score of 30 or higher indicates bases with a chance lower than one in a thousand of being incorrect. Notably, the nucleic acid amplification method developed by Kawashima, Farinelli, and Mayer from the Geneva Biomedical Research Institute stands as a significant milestone in the advancement of NGS[5].

Thanks to dramatic technological advances, exome, genome, and RNA sequencing costs decreased 100-fold in between 2008 and 2014[6] and were reduced by an additional 20-fold between 2014 and today. Today's cost of WES and WGS are estimated at below $1,000 per run including the library preparation costs and sequencing reagents, with WGS poised to become cheaper than WES and therefore likely become the new standard in incoming studies.

[4] https://nanoporetech.com/q20plus-chemistry
[5] https://patents.google.com/patent/WO1998044151A1/en
[6] https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

2.1.2. Libraries preparation and target enrichment

The library preparation and capture protocols employed differ according to the sequencing machine that will be used, the genomic regions or molecules species that investigators aim for, and the nature of the starting material, essentially either DNA or RNA for the data analyzed in this thesis. Although multiple NGS techniques exist as discussed in Section 2.1.1.2, this discussion focuses on Illumina dye sequencing and its associated library preparation protocols due to their widespread usage both in the clinical and research settings.

2.1.2.1. DNA sequencing

In targeted sequencing, the library preparation protocol shares similarities with Sanger sequencing. However, instead of using a single pair of primers, multiple pairs of primers are designed, spaced approximately 150 to 300 base pairs apart, according to the length of the sequencing reads. These primers are utilized in PCR cycles to amplify the specific regions of interest. Each primer is composed of two parts: a common sequence shared by all primers, serving as the starting point for the sequencing experiment, and a unique sequence specific to the region of interest. The number of PCR cycles is optimized to provide sufficient material for sequencing, considering that the number of PCR-induced errors increases with the number of cycles.

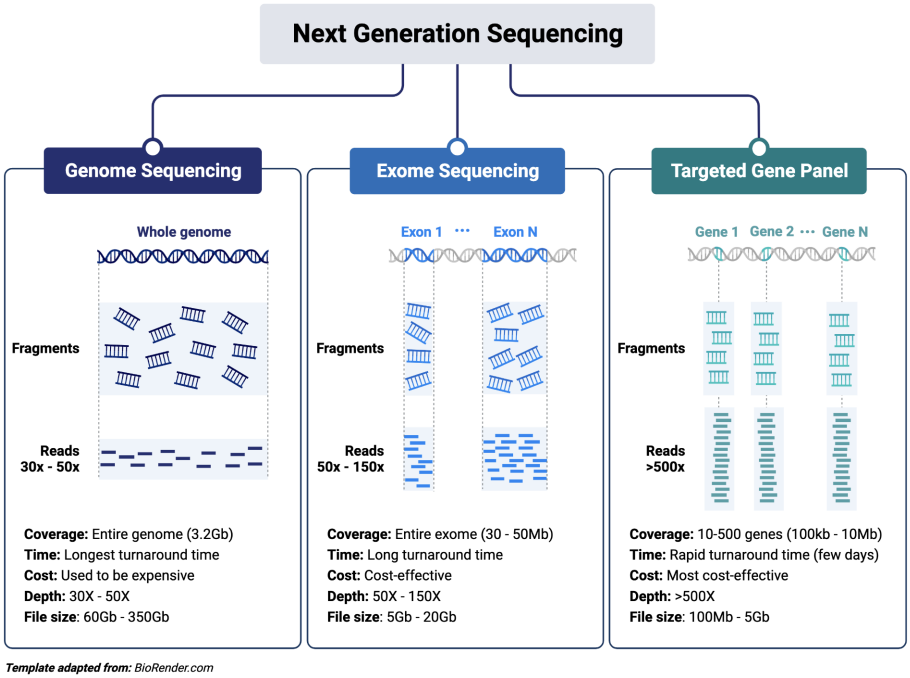


Fig. 2.2.: The three main applications of next-generation sequencing for DNA sequencing.

For WGS and WES, the genomic DNA is first randomly fragmented either mechanically by ultrasonication methods, or biologically by enzymatic digestion. Subsequently, adapters are

ligated to both ends of the DNA fragments, serving as anchoring points for the sequencing step. In the case of WGS, no capture is required, whereas WES necessitates a target enrichment method.

A comprehensive review article by Seaby *et al.* (2016), focusing on the applications of WES, offers a comparative table of major exome capture kits. Noteworthy kits, such as Agilents SureSelect Human All Exon, Roche-Nimblegen SeqCap EZ exome library, and Illumina TruSeq Exome enrichment, employ hybridization with hundreds of thousands or even millions of complementary baits to select exonic regions. In contrast, amplicon-based methods like Thermo Fisher AmpliSeq or Agilent HaloPlex directly amplify exonic regions. Hybridization allows the enrichment and sequencing of a larger number of targets per panel, exhibiting superior performance in terms of uniformity and complexity. On the other hand, amplicon sequencing is faster with higher on-target rates but at the expense of non-uniformity and higher error rates. Presently, WES is predominantly conducted using hybridization-based capture kits.

The retained DNA fragments undergo subsequent amplification through a minimal number of PCR cycles. The amplified DNA is then denatured and deposited in single strands onto a *flow cell*, essentially a glass slide featuring numerous attached probes called *oligonucleotides*. These oligonucleotides function as anchors for the single-stranded DNA fragments. Local replication of these fragments occurs via nearby probes in a process known as bridge amplification, generating numerous identical copies of each fragment. This amplification organizes replicated DNA fragments into clusters, each cluster containing approximately 1,000 copies of the original DNA fragment from the library. Each DNA cluster contributes to the generation of one read, either single- or paired-end. Real-time monitoring of complementary bases introduced by engineered polymerases enables the determination of reads. More specifically, during each cycle of the sequencing experiment, nucleotides with reversible dye-terminators are introduced and hybridized to the incompletely elongated DNA molecules on the flow cell. Fluorescence signals are used to read the inserted bases, after which the dye terminators are washed away to allow DNA elongation to continue, initiating a new sequencing cycle. For more details about Illumina dye sequencing, see their excellent tutorial video[7].

2.1.2.2. RNA sequencing

RNA sequencing (RNA-seq) is a technique employed to determine the sequences of RNA molecules within cells at a given time point. It provides a snapshot of gene activities in a given sample and comes in various modalities. Bulk RNA-seq sequences RNA molecules across a diverse array of cells, whereas single-cell RNA-seq achieves a resolution at the individual cell level. Spatial transcriptomics further refines this resolution to subcellular levels while offering information about the spatial distribution of RNAs (Stark *et al.* 2019). This section exclusively deals with library preparation protocols for bulk RNA-seq, as no data from single-cell RNA-seq or spatial transcriptomics experiments will be presented in this manuscript. Similarly, only protocols intended for short-read sequencing, the prevalent technique in clinical settings,

[7] https://www.youtube.com/watch?v=fCd6B5HRaZ8&ab_channel=Illumina

will be addressed. Long-read RNA sequencing is primarily pertinent for characterizing or improving transcriptomes of poor characterized species, as well as for discovering novel splice junctions or isoforms.

There is a specific algorithm assessing the results of library preparation through a score known as RNA integrity score. RNA is extracted and right away quality is assessed through RIN score. RIN 10, good; $>$ 7 is good; 7-4 is mediocre quality; below 4 is not usable. Peaks from some spectograms.

In short-read RNA sequencing, RNA nucleic acids are extracted from the sample, and contaminating DNA is eliminated using DNAse enzymes. Subsequently, the RNA undergoes pre-treatment to isolate the desired RNA molecules and remove unwanted ones, particularly rRNA, which can constitute a substantial portion of cellular RNA. RNA selection typically involves polyA-enrichment or rRNA depletion, each with its respective advantages and drawbacks. PolyA-enrichment selects polyadenylated RNA, including mature mRNA, but fails to capture non-polyA transcripts or partially degraded mRNAs. Conversely, rRNA-depletion captures both polyA+ and polyA- RNAs, encompassing mature and pre-mature mRNAs with intronic sequences as well as tRNAs (Zhao *et al.* 2018). 3' mRNA sequencing is a subtle variant of polyA-enrichment-based RNA-seq which attempts to minimize biases introduced in quantification results by the fact that longer transcripts are sheared into more fragments than shorter transcripts (Oshlack & Wakefield 2009). In this variant, mRNAs are not fragmented prior to reverse transcription so that only one cDNA is generated for each transcript.

Selected RNAs are subsequently fragmented and converted back into cDNA using reverse transcriptases. In most protocols, the RNA template used for cDNA synthesis is replaced with a proper DNA strand synthesized by DNA polymerases, known for their lower error rates compared to reverse transcriptases. Retaining information about the strandedness of the original RNA molecules can be achieved by incorporating uracil bases during the synthesis of the second DNA strand. The resulting double-stranded cDNA undergoes end-repair, and adapters are ligated as for DNA sequencing preparation. The cDNA is then denatured for single-stranded molecules ready for PCR amplification. To preserve strandedness information, the strand synthesized with uracils can be washed away at this stage. Following amplification, the library is prepared for sequencing using standard DNA sequencing methods, such as Illumina dye sequencing outlined in the previous section.

2.1.3. Processing of sequencing data

The data generated by sequencing machines are the entry point of a long series of data processing and transformation steps to generate tabular data of genetic variants or gene expression that humans can comprehend. The different steps for organizing, controlling, cleaning, and analyzing large sequencing data files are typically organized in *bioinformatic workflows* executed on high-performance computing *clusters*. A typical NGS workflow is depicted in Figure 2.3 and is described in more detail in the following sections.

Quality control of the sequencing data generated by sequencers is a major challenge

in NGS analyses. Checking that the data meets all the quality requirements is crucial for successful downstream analyses. Sometimes, bad-quality samples will fail on one or multiple bioinformatic tools, allowing them to be identified and removed easily. However, in many instances, samples will be successfully processed, albeit with some warnings from some tools. The results from processing such samples will, however, contain little to no meaningful information. Failing to identify such cases can mislead analyses and must, therefore, be properly identified.

Therefore, the analysis of sequencing data necessitates quality controls at all steps along the workflow to flag and discard low-quality samples or, when possible, remove the problematic sequences and flag the variants likely to be of artefactual rather than biological origin (Patel & Jain 2012; Conesa *et al.* 2016). Major quality control steps occur directly on the data produced by sequencing machines. Subsequently, in alignment-based workflows, post-alignment quality control using dedicated bioinformatic tools can help further identify low-quality or artefactual sequences. Quality control is also present at the variant-calling step to minimize the number of false positives ending up in the results tables while maintaining good sensitivity. Filtering procedures of variants are pretty diverse and elaborate and will be detailed in Section 2.1.4.4. All quality control steps of NGS workflows are essential, and none should be overlooked.

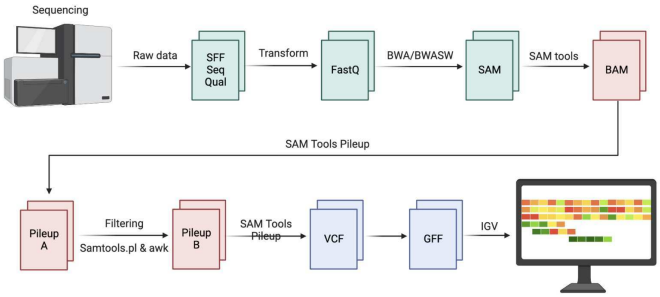


Fig. 2.3.: NGS workflow

2.1.3.1. The FASTQ file

Sequencing machines generate data in a specific file format known as the FASTQ format. Some devices, such as Illumina sequencing machines, go through the intermediate base call format (BCL) representation format that is then converted to the FASTQ format to make it easily usable by downstream processing tools. The FASTQ format is a text-based sequence file format that stores three pieces of information for every read: a text description identifying the read, the sequence of bases constituting the read, and the qualities of each base encoded using ASCII characters[8]. It was developed at the Wellcome Trust Sanger Institute (WTSI) but has become the de facto standard for storing the output of sequencing instruments.

After the sequencing has been performed, a series of steps is applied to control the quality of the generated FASTQ files before extracting meaningful information. Figure 2.3 represents a typical analysis workflow of NGS data. The most important parameters to check for in the raw sequencing data are the base qualities distribution, the nucleotide distribution, the GC content, and the duplication rate. Aside from these metrics that serve to flag bad samples, corrective actions are systematically applied to clean the sequencing data by removing, or if possible, correcting low-quality bases and removing low-quality reads. `FASTQC` (Andrews

[8]More info at https://en.wikipedia.org/wiki/FASTQ_format or https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html

2010) and `fastp` (S. Chen, Zhou, *et al.* 2018) are now the most commonly used bioinformatic tools for generating reports of quality control and applying corrective actions. Prior to the release of `fastp`, `Cutadapt` (Martin 2011) used to be the standard for read trimming, while `Trimmomatic` (Bolger *et al.* 2014) was the standard for read filtering and pruning.

Quality profiling Firstly, the *distribution of base qualities*, assessed using the *phred quality scores*, or Q scores, as presented in Section 2.1.1.2, should be high enough (usually more than 30 for Illumina sequencing) and relatively stable over cycles, i.e, along the reads. Visual inspection of plots showing base qualities against base positions in reads can reveal decaying quality at the end of reads for various reasons, such as adapter contamination or experimental issues in the last sequencing cycles. The problem of quality deterioration in the last cycles has been addressed through improved chemistry and, more significantly, by restricting sequencing experiments to short reads not surpassing 2x150 base pairs. It is worth noting that MiSeq Illumina sequencers, which are now phased out, were designed to generate reads of up to 2x300 base pairs.

Secondly, the *distribution of nucleotides* across sequencing cycles serves as a valuable quality control metric for whole genomes and exomes, though it is not as pertinent for amplicons or RNA-seq samples. In an ideal sequencing run, the distribution of the four nucleotides across all reads should remain relatively consistent, with minor fluctuations towards the start or end of the read. In practice, slightly uneven distribution is frequently observed in the first 12 bases of each run due to a biased selection by random primers, which are not so random. This problem cannot be fixed by processing and is not known to adversely affect the analyses. Nucleotide distribution is closely linked to base quality, and both metrics can be employed to assess the quality of raw data. Poor base quality is also often reflected in the nucleotide distribution plot.

The *GC content*, measured as the percentage of G or C nucleotides in a sequence, is another metric that is commonly looked at to assess quality. The GC content varies across species and genomic regions. In human whole genomes, it is about 38 to 39%, while in human exomes, it is around 49-51%. Deviation from the experiment-specific theoretical distribution typically indicates contamination by adapters (sharp peaks) or by other species (broad peaks).

The *level of duplication*, i.e, the proportion of reads in a sequencing file that have at least one duplicate, serves as another monitored metric for gauging the quality of a sample. Depending on the nature of the sequenced biological material, a variable level of duplication is acceptable. In RNA-seq libraries, different transcripts will be present at widely different levels in the sample used. If deep sequencing is used to capture lowly-expressed transcripts, it increases the likelihood that biologically highly expressed transcripts will generate large sets of duplicates.

Lastly, the number of reads per sequenced sample in a pooled experiment is also a metric that can be used to flag samples. It is indeed very common to pool multiple samples during one sequencing experiment to save costs. Demultiplexing of pooled biological fragments is made possible by the ligation of sample-specific indices prior to pooling and sequencing. A

high variation in the number of reads across pooled libraries with similar amounts of starting material can identify low-quality samples.

Corrective actions A certain number of corrective actions may be taken during the different stages of the bioinformatic processing to mitigate the effects of low-quality samples. In the initial phase, when dealing with raw sequencing data - before any subsequent processing for variant calling or quantification, a small number of corrective measures may be applied. Previously, these corrective actions were derived from disparate tools, resulting in inefficient data processing. Nowadays, the `fastp` tool (S. Chen, Zhou, *et al.* 2018) consolidates the key functionalities of various tools into a single and highly efficient program which is widely used.

Reads trimming, i.e, the removal of a certain number of consecutive bases at reads extremities, is the main corrective action applied at this stage. It serves to remove either contaminating adapters that should not be here, polyG tail in the reads of Illumina NextSeq or NovaSeq series, or bases with insufficient qualities located at read tails. On top of being versatile, `fastp` is quite elaborate as it implements different trimming strategies for different sequencing technologies and can learn adapter sequences automatically. *Base correction* is the second main corrective measure applicable here. In some specific scenarios, low-quality bases ($Q < 15$) may be corrected by replacing them with high-quality bases ($Q > 30$). This is only possible for paired-end sequencing data and only at positions overlapped by the two reads from the same pair.

Cleaned raw sequencing files are also FASTQ files, albeit with fewer or shorter reads. They are then ready to be further processed in different manners according to the experimental setting. In the case of DNA sequencing for variant detection in a species with a well-established reference genome, the next step in the processing is the alignment of reads against the reference. Aligned reads are then compared against the reference, and a wide number of statistics are compiled to identify deviations from the reference. In RNA sequencing, reads may or may not be aligned, against the reference genome or transcriptome, depending on the analysis aimed and the tool employed. Although RNA-seq may also be used to discover genetic variants fixed in DNA, the primary purpose of performing RNA-seq is to quantify the number of different transcripts to assess gene expression levels. Quantification of gene expression may or may not go through an intermediate alignment step since the development of pseudo-alignment methods, as will be described in Section 2.1.4.3.

2.1.3.2. The BAM file

The reads that stem from the sequencing of fragmented DNA molecules are then mapped to a reference genome to identify their origin, an information needed for many downstream analyses. This requirement also holds true for variant-discovery analyses conducted on RNA-seq data, mostly for structural variants but also possibly for short variants usually identified from DNA sequencing although this practice is not recommended. Alignment is also a possible but non-mandatory step prior to quantification analyses. Some quantification

algorithms now indeed have the capability to bypass alignment.

Alignment to the human reference genome of reads generated by DNA sequencing experiments is commonly done through the standard Burrows-Wheeler Aligner (H. Li & Durbin 2009) available in the program `BWA-MEM` since 2013 (H. Li 2013). The alignment tool outputs files in sequencing alignment map (SAM) format from the `SAMtools`[9] suite of tools. The specifications for this file format may be found on the hts-specs GitHub repository of the `SAMtools` project[10]. As aligned reads storage requires huge volumes of disk, the BAM format - binary equivalent of SAM - is used as a more efficient format for the long-term storage of sequencing reads. Likewise, specific tools have been developed to alignreads generated by RNA sequencing experiments either to the reference genome using splice-aware algorithms such as the popular STAR algorithm (Dobin *et al.* 2013), or the reference transcriptome. They also return aligned reads in SAM or BAM formats.

Aligned reads offer additional opportunities for overseeing the quality of sample preparation and sequencing processes. The parameters subject to control vary based on the nature of the sequencing experiment. In techniques such as targeted or capture-based sequencing, critical metrics include *capture efficiency* and *target coverage*. Capture efficiency gauges the proportion of reads mapping to target-included regions, while target coverages quantify the percentages of targeted regions covered by a minimum number of reads. In exome and targeted sequencing, it is not uncommon for capture efficiency to range between 40% and 70%. While exceedingly low efficiencies warrant caution, the critical metrics are predominantly the coverage of the target at various depths. The mean or median overall coverage serves as the primary metric when evaluating whether variant calling should proceed with a given sample. Coverage uniformity is another crucial quality metric. The proportions of regions covered at different read levels naturally depend on overall sequencing depth but should ideally fall within the range of 90% to 100% for at least one read and not exhibit significant drops for higher read coverage. Coverage metrics can be assessed using different standard tools such as `CollectHSmetrics` included in the Genome Analysis Tookit (GATK) suite [11] or `mosdepth` (Pedersen & Quinlan 2018). In the META-PRISM study, presented in Chapter 3, samples with an overall sequencing depth below 40 reads or a target coverage of at least 10 reads below 60% were excluded. This exclusion criterion should be contextualized with subsequent variant filtering rules, often necessitating a minimum number of reads for considering putative variants. Setting a minimum threshold of 10 reads for variant consideration in a sample with a 10X target coverage at 60% implies that 40% of targeted regions may go unassessed for variant discovery, potentially leading to a high rate of false negatives.

The identification and elimination of PCR duplicates represents a critical step consistently employed in all sequencing endeavors involving a capture step - comprising all previously mentioned sequencing experiments except for WGS which is minimally affected if at all. The technical duplication of reads through PCR amplification introduces sequence errors as well as

[9] http://samtools.sourceforge.net/
[10] https://github.com/samtools/hts-specs/blob/master/SAMv1.pdf
[11] https://gatk.broadinstitute.org/hc/en-us

potentially significant bias into variant-allele frequencies (VAFs) or copy numbers estimations, particularly when the amplification is uneven across targeted regions. To address this concern, it is common practice to designate reads sharing identical genomic coordinates PCR duplicates. Nevertheless, distinguishing technical duplicates from biological ones at this stage remains challenging. For RNA-seq, where the sequenced material inherently possesses high levels of redundancy (given that a single gene can generate thousands of identical transcripts), the question of whether to apply PCR duplicate removal remains unsettled. The standard tool for marking and removing duplicates is the `MarkDuplicates` tool within the GATK bundle. A faster implementation of this tool, known as `sambamba`[12], has recently become available.

For most cancer types, aligned DNA sequencing files usually come in pairs: one file contains DNA sequenced from healthy tissues, while the other file contains DNA sequenced from tumor tissues. In certain cancer types, particularly blood cancers, it may be overly complex to recover tissues not contaminated by the tumor (saliva or skin samples). In such instances, only one FASTQ or BAM file is usually available, and the identification of genetic variation compared to the reference genome becomes more involved and less accurate. This also holds true for the sequencing of cell-free circulating tumor DNA (ctDNA), an experiment that has gained a lot of attention in the recent years thanks to non invasiveness of the sample collection procedure. Aligned DNA sequences stored in BAM files are the starting point for many downstream analyses, most notably the detection of all kinds of genetic variants or diverse assessments of genome instability (MSI, mismatch repair deficiency (MMRd), homologous recombination deficiency (HRD)) as described in later Sections.

Nowadays, tremendous numbers of FASTQ and BAM files are available for download on dedicated platforms such as the database of Genotypes and Phenotypes (dbGaP) (217K WGS, 319K WES, 201K targeted sequencing as of December 2023), the European Genome Archive (EGA) (3.2M sequence files representing more than 13 petabytes of data as of December 2023), or the GDC data portal (34K WGS, 38K WES as of December 2023), for authorized users and projects.

2.1.3.3. The VCF file

The variant calling format (VCF) is the standard file format used for reporting genetic variants of all types. *Variant calling* algorithms analyze sequencing data from one or multiple samples compared to a reference sequence. The comparison outputs are stored in one or multiple VCF files which list all putative variants identified. There used to be different VCF formats depending on the variant-calling algorithm, the date on which it was produced, and the laboratory that produced the analysis. Even though there exist public specifications that try to make VCF a standard file, project, team, or technology-specific VCF files continue to exist in public databases and may not always be easily used in standard downstream analysis tools, limiting their utility. Fortunately, recent VCF files now tend to all abide by the reference specifications detailed on the `hts-spec` GitHub repository of the `SAMtools` project[13]. In this

[12] https://lomereiter.github.io/sambamba/
[13] https://github.com/samtools/hts-specs

repository, one may find documentation specifying how this file format should be structured for each type of variant. At the time of writing, the latest version released was VCF 4.4, but earlier versions, particularly 4.3, 4.2, and 4.1, are still commonly used.

In short, the VCF file contains *meta-information* lines, a *header* line, and then *records* containing information about a position in the genome where a putative variant was identified. Meta-information lines generally describe the specific commands of the different tools that have been run to generate the present file. They also provide technical information about the reference genome used and the meaning of the different abbreviations used in the `INFO` or `FORMAT` fields. The header line must contain the 8 following mandatory fields: `CHROM`, `POS`, `ID`, `REF`, `ALT`, `QUAL`, `FILTER`, `INFO` and one or multiple additional fields for every sample that was analyzed. Typical VCFs contain either a single sample field, for analyses in *tumor-only* mode, or two sample fields for the tumor and normal samples if sequencing data from both tissue types were analyzed. However, one may also find VCF files aggregating many samples, particularly the `GVCF` format recently described in the GATK documentation[14] and which aggregates all samples jointly genotyped and possibly all sites, whether there is a variant call or not. For each sample column, a column `FORMAT` indicates how the data is formatted. There can be multiple `FORMAT` columns if the format is not same from one sample to another. This happens for instance if the VCF gathers calls produced by multiple algorithms in which case one format field will be added for each variant-calling algorithm. Example VCF files can be found on the legacy GDC portal[15] for the TCGA breast invasive carcinoma (BRCA) study under Data Category > Simple nucleotide variation, Data Format > VCF and Platform > Illumina HiSeq or on the current GDC data portal where the latest somatic variant calling workflow incorporates four different algorithms[16]. Examples and detailed descriptions of the meaning of these fields can be found in the specifications of the latest VCF 4.4 format. The number of records in a single VCF file varies greatly according to the experiments and the number of samples that are aggregated. It can range from a dozen records to billions for public VCF files generated by large consortia as the ones originating from the gnomAD database with data volumes reaching hundreds of gigabytes[17].

2.1.4. Genetic variants and gene expression

Human *genetic variants* designate variations in the genomes of individuals compared to a reference genome and can take many forms and originate from many different sources. Of note, the terms *genomic variant* and *genetic variant* will be used interchangeably in this manuscript with a slight preference for using *genomic* in the context of variations implicating large DNA sequences.

To categorize genetic variants, a primary distinction is made between somatic and germline variants, discerning events acquired during an individual's lifetime, post-fertilization, from

[14] https://gatk.broadinstitute.org/hc/en-us/articles/360035531812-GVCF-Genomic-Variant-Call-Format
[15] https://portal.gdc.cancer.gov/legacy-archive
[16] https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline
[17] https://gnomad.broadinstitute.org/downloads#v3-hgdp-1kg

those naturally inherited from parental egg and sperm cells, respectively. In this manuscript, there is limited discussion of germline variants, and when addressed, their germline origin is explicitly stated. All other variants discussed are considered of somatic origin. Genetic variants are further classified into three broad categories, distinguishing various event types: SNVs or MNVs, which are localized changes without altering sequence size; indels, which are also localized but result in changes in the sequence size; and structural variants (SVs), encompassing diverse events involving one or multiple DNA segments of varying sizes, potentially up to entire chromosomes.

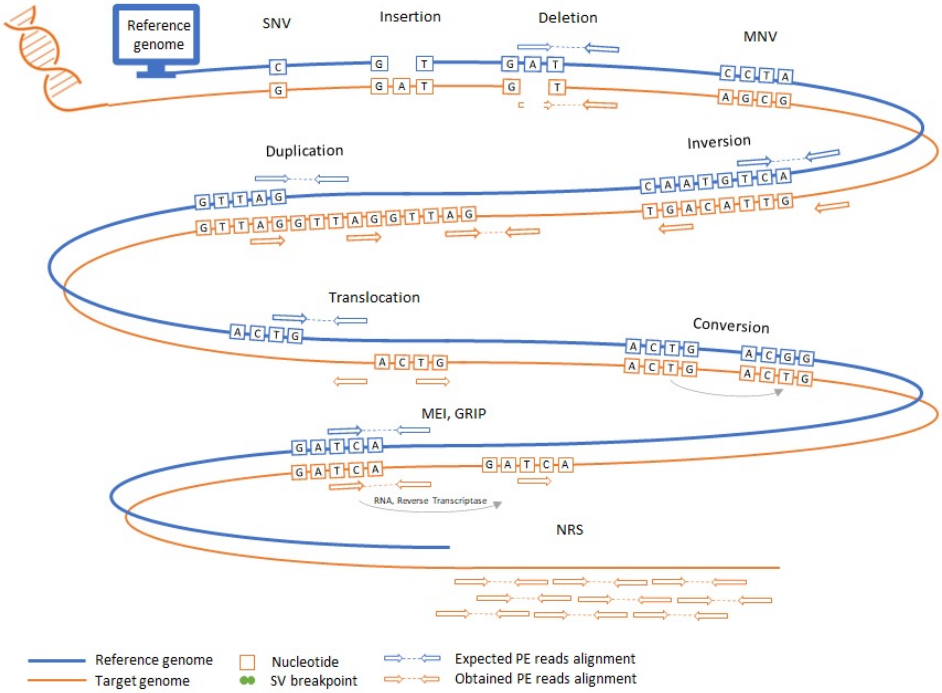


Fig. 2.4.: Diversity of DNA variant types and their consequences on paired-end reads. MEI, mobile element insertions; MNV, multi-nucleotide variants; NRS, non-reference sequences; SNV, single nucleotide variants; GRIP, gene retrotransposition insertion polymorphisms. Reproduced with permission from authors of Zverinova & Guryev (2022).

Figure 2.4 provides a visual representation of the different genetic variant types and their manifestations in the DNA molecule. The three aforementioned categories are depicted, with all events being subtypes of SVs, except for those at the top - namely, SNVs, MNVs, short insertions, and short deletions.

2.1.4.1. SNVs, MNVs, and indels

Changes in a single base are the simplest and most frequent form of genetic variant. Extensive genetic analyses conducted by the 1000G consortium, encompassing WGS data from over 2,500 individuals representing diverse ancestries, have elucidated that 99.9% of germline variants manifest as SNVs or indels, with SNVs occurring approximately ten times

more frequently than indels. Likewise, comprehensive investigations into cancer genomes have allowed us to precisely quantify the distribution of somatic genetic variants within cancer cells. Landmark studies led by the PCAWG consortium, involving more than 2,500 cancer patients, have delineated a comprehensive catalog of genetic variants spanning all three types of events. Analogous to germline variation, SNVs, MNVs, and indels - commonly referred to as mutations - constituted 99.3% of all somatic events. SNVs were nearly 20 times more prevalent than indels and approximately 100 times more frequent than MNVs (Rodriguez-Martin *et al.* 2020).

Further dissecting somatic events based on their occurrence in coding and non-coding regions revealed that a mere 0.8% of all localized genetic variants (i.e excluding SVs) are located in coding regions. On average, each cancer genome harbors approximately 140 coding somatic mutations and 18,000 non-coding somatic mutations. Despite this, the predominant focus of genetic studies on cancer genomes has historically centered on coding regions, partly due to the historically more economical nature of sequencing targeted regions compared to WGS, and partly to the intuitive hypothesis that genetic events influencing cellular behavior are more likely to be situated in protein-coding regions.

As succinctly outlined in Section 1.2.2.2, SNVs, MNVs, and indels can be classified based on their genomic localization relative to existing gene annotations. Further categorization is employed for those impacting coding regions based on their consequences on the gene product. Broadly speaking, mutations influencing regions beyond genes are designated as `intergenic`, while those affecting genes are either *intronic* or *exonic*. Intronic mutations are predominantly labeled as `intron variants`, which are generally silent except for mutations occurring in specific positions that play a role in splicing. Mutations located at the 5' and 3' ends of introns, within the 2-base splice donor and acceptor sites, are denoted as `splice donor` and `splice acceptor variants`, respectively. Mutations occurring a few bases before or after these splice sites are categorized as `splice region variants`.

Exonic mutations encompass a diverse spectrum of types based on their functional consequences. Indels in coding regions are conveniently classified as either `frameshift` if the affected base number is not a multiple of 3, thereby altering the reading frame, or `inframe` if otherwise. Substitutions (SNVs and MNVs) are categorized as `synonymous` if they do not modify the amino acid sequence, and `non-synonymous` otherwise. Non-synonymous substitutions are further subdivided into `missense` if they replace a codon coding for one amino acid with another, `nonsense` if they replace a codon coding for an amino acid with a stop codon, or `nonstop` if they replace a stop codon with a codon coding for an amino acid. Mutations in the 5' or 3' untranslated regions of genes are classified separately and their functional impact is less well understood. Additionally, mutations affecting the start or stop codons receive dedicated classes. For more detailed definitions of all potential functional consequences, the reader is directed to the Ensembl website page on variant consequences[18].

The detection of variants from sequencing data is performed by elaborate algorithms known as *variant callers* or *variant-calling* algorithms. Variant callers do not all have the same detection capabilities nor are they applicable to data generated from all sequencing

[18] https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html

experiments. The calling of indels and substitutions have historically been separated into different algorithms or different running modes of the same algorithm. Likewise, the calling of germline and somatic variants are performed using distinct callers or a single caller configured in distinct modes given the very different nature of these two types of mutations. Table A.2 lists some of the most commonly used variant callers for detecting SNVs, MNVs, and indels in germline or somatic settings. Some of the tools listed in this table have broader capabilities such as `VarScan2` which can detect somatic CNA, a frequently analyzed consequence of SVs, or `FreeBayes`, `Platypus`, and `VarDict` which can all additionally detect complex variants.

Most variant callers use joint-genotype inference methodologies derived from Bayesian or conventional statistical models, incorporating specific filters to deduce the most probable genotype from allelic counts on aligned reads. Notable examples encompass `MuSE`, `JointSNVMix`, `SomaticSniper`, `MuTect`, `LoFreq`, `Strelka`, `EBCall`, and `VarScan`. Conversely, alternative recent algorithms have started to transpose the transdisciplinary successes of neural networks to calling variants from NGS data. The `DeepVariant` CNN-based model, for instance, achieved the highest SNP-detection performance in the precisionFDA Truth Challenge run in 2016[19] and was shown to perform as good or marginally better than gold-standard germline variant callers `HaplotypeCaller` and `Strelka2`.

Alterations supported by low VAFs observed in tumor samples arise for various biological reasons, mostly *intra-tumor heterogeneity* wherein minor cellular clones with different genotypes coexist or local copy-number variations. They may also arise from technical sources, including tumor-normal cross-contaminations, DNA damage during sample preservation, base errors introduced during library preparation or sequencing, or mapping errors. Some variant callers have specialized in calling low-VAF variants, such as `EBCall`, which uses an empirical Bayesian framework for sorting somatic mutations from artifactual ones, or `LoFreq`, which claimed to be capable of calling variants with frequencies as low as 0.05% under optimal conditions of very high coverage (10,000x) and high-quality data (Q40).

However, the performance of specific variant callers exhibits variability across different datasets. Previous benchmark studies have underscored substantial divergence among the outputs of various callers for a given dataset (O'Rawe *et al.* 2013; Krøigård *et al.* 2016). In light of these challenges, *ensemble approaches* have arisen as a strategic solution, combining prediction results from multiple somatic variant callers to generate a consensus set of calls. This approach aims to enhance sensitivity without compromising specificity or vice versa. The efficacy of consensus methods depends on the individual performance of each caller and, crucially, the heterogeneity of statistical models employed. Different consensus methods have been developed which combine predictions from three to seven callers using simple majority voting rules (M. Wang *et al.* 2020) or more elaborate machine learning (S. Y. Kim *et al.* 2014; Fang *et al.* 2015; D. E. Wood *et al.* 2018; Anzar *et al.* 2019; W. Huang *et al.* 2019) or deep learning-models (Sahraeian *et al.* 2019) for generating consensus calls. Multi-caller approaches have been successfully applied by international consortia, as done for the latest update of somatic mutation calls on TCGA WES data during the MC3 project (Ellrott *et al.* 2018), or by

[19] https://precision.fda.gov/challenges/truth/results

the PCAWG consortium where a consensus script named `SNV-MERGE` implemented a logistic regression model to combine SNVs and indels calls from five pipelines (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020).

Of note, for the translational studies of WES data derived from the sequencing of tumor samples of patients with advanced cancer, presented in Chapter 3 and Chapter 4, SNVs, MNVs, and indels were detected using `Mutect2`. Specific filtering rules were subsequently applied for various purposes as will be detailed later.

2.1.4.2. Structural variants and their consequences

Structural variants

Structural variants (SVs) are conventionally characterized as genetic variations affecting DNA segments exceeding 50 base pairs in size. These variants encompass alterations such as amplifications, deletions, or rearrangements across the genome, presenting as junctions between genomic breakpoints. In essence, an SV denotes the contiguous positioning of two DNA segments normally separated by 50 bases or more. SVs are notoriously difficult to detect and interpret, making them an underappreciated class of genetic variants. Their diverse manifestations range from moderately localized events to entire chromosome rearrangements, displaying varying degrees of complexity based on their biological origins.

Historically, five canonical SV types have been identified and studied: *insertions*, *deletions*, *duplications*, *translocations*, and *inversions*. Nevertheless, more intricate SV types featuring localized or extended repeats of these canonical types have also been documented. The mechanistic origin of SV further serves to refine their classification; for instance, duplications are subdivided into tandem duplications (resulting from DNA replication errors), inverted duplications (stemming from sister-chromatid or telomere-telomere fusions), segmental duplications (arising from DNA repair errors), among others. Insertions also exhibit subcategories, including those caused by mobile elements (MEI, Figure 2.4), such as transposons, viral DNA insertions, or DNA replication errors, among various other mechanisms.

In a seminal pan-cancer study of SVs encompassing over 2,500 whole genomes, Y. Li *et al.* (2020) rely on a simple classification of SVs differentiating between *unbalanced* SVs, leading to changes in DNA copy numbers, and *reciprocal* or *balanced* SVs, which are nearly or entirely copy-neutral. Within this framework, insertions, deletions, and duplications fall under unbalanced SVs, while inversions and translocations are subtypes of balanced SVs. Y. Li *et al.* (2020) further delineated distinct SVs types characterized by complex copy-number and breakpoint patterns. Examples include *breakage-fusion-bridge* events (a few clustered inverted breakpoints with copy number changes), *chromoplexy* (sets of two or more SVs in which the chromosomal ends at either side of breakpoints are shuffled and rearranged), *chromotripsis* (a catastrophic one-time event characterized by the clustering of many SVs resulting in oscillating copy numbers and rearrangements junctions on one or multiple shattered chromosomes), and whole-genome duplication (WGD), each with unique features contributing to genomic

instability. A recent comprehensive review by Cosenza *et al.* (2022) offers a unified perspective on SVs, describing their biological origins as well as their role and prevalence in cancer. The authors proposed a mechanistic classification of known SVs into seven classes, with the five canonical SVs falling under the "simple events" class. Noteworthy distinctions are made for chromosomal aneuploidies and genome doublings, considered as separate classes due to their distinct mechanistic origins primarily rooted in mitotic segregation errors.

Explanations for the mechanistic underpinnings of SVs are notably limited and far less abundant than those available for base substitutions. Over the past decades, extensive characterization of both endogenous and exogenous mechanisms governing substitutions has been achieved, primarily through the elucidation of mutational signatures (Alexandrov, Nik-Zainal, *et al.* 2013; PCAWG Mutational Signatures Working Group *et al.* 2020). In contrast, mechanistic explanations and signatures associated with SVs have only recently begun to emerge (Nik-Zainal, Davies, *et al.* 2016; Y. Li *et al.* 2020). Section 2.2, in this context, offers an in-depth exploration of mutational signatures, elucidating how their foundational concept has been extended to derive signatures specific to somatic SVs or their CNAs consequences (Section 2.2.4). Additionally, a recent computational approach leveraging graph-based methodologies has surfaced and provided insights into numerous previously undisclosed complex SVs events (Hadi *et al.* 2020).

Pancancer studies of whole genomes have allowed to precisely quantify the prevalence and significance of somatic SVs in cancer. The PCAWG consortium has conducted a comprehensive analysis across 2,583 cancer genomes and revealed 288,416 somatic SVs, averaging 111 SVs per cancer genome (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). A parallel pan-cancer study on metastatic patients conducted by Priestley *et al.* (2019) described 653,452 somatic SVs in samples from 2,399 patients, averaging 272 somatic SVs per cancer genome. These somatic SVs coexist with germline SVs, which are known to number in the thousands (J. Wang *et al.* 2008; Ahn *et al.* 2009; The 1000 Genomes Project Consortium 2015; Collins *et al.* 2020) and to influence more genomic bases than the millions of germline SNVs and short indels (The 1000 Genomes Project Consortium 2015). These studies have also underscored the critical role of SVs in cancer. A remarkable 55% of cancer driver events identified by the PCAWG consortium are represented by SVs, surpassing the count of cancer driver mutations (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). This remarkable observation highlights the pivotal role played by SVs in cancer, particularly in amplifying oncogene expression through copy number increases or silencing tumor suppressor genes via deletions. Another significant consequence of SVs involves the generation of oncogenic gene fusions or the relocation of enhancers in proximity to oncogenes, a phenomenon referred to as *enhancer hijacking* (Northcott *et al.* 2014; Weischenfeldt *et al.* 2017). The consequences of SVs can be therapeutically targeted as exemplified by the targeted drugs ATRA, trastuzumab, and imatinib which inhibit the consequences of *PML-RARA* gene fusion in acute promyelocytic leukemia, *ERBB2* amplification in breast cancer, and *BCR-ABL1* gene fusion in CML, respectively.

The evolution of our capability to detect SVs is intrinsically tied to advancements in technology. In the early years of cancer genomics, only simple and macroscopic types of

SV were described. Cytogenetic studies examining karyotypes outlined deletions, inversions, duplications, and translocations with resolution limitations on the megabase scale. Notably, these techniques precluded the detection of *focal* SVs that are now recognized as playing crucial roles in cancer (Rheinbay *et al.* 2020; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). With the introduction of microarray technologies, notably CGH arrays and SNP arrays, assessments of copy numbers at significantly smaller resolutions became feasible. However, these methods faced limitations in detecting copy number-neutral SVs and were effective only in identifying copy number-altering SVs supported by consecutive probes. Consequently, SVs smaller than 50 kilobases in size remained undetected. WES has proven particularly valuable for identifying SVs within exonic regions and is well-suited for detecting both chromosome arm copy-number changes and focal alterations, provided they occur within coding regions. RNA-seq has also proved to be an effective means of detecting translocations expressed by the transcription machinery, a point that will be elaborated upon in a next paragraph. Pancancer studies analyzing WGS profiles have, however, revealed that a majority of SVs occur in intergenic regions, exerting oncogenic effects through alterations in regulatory regions and other mechanisms (Quigley *et al.* 2018; Rheinbay *et al.* 2020). While WGS does not have the same limitations as other technologies, short-read sequencing methods pose challenges in identifying SVs in repetitive regions. Long-read sequencing technologies, such as PacBio SMRT or Oxford Nanopore sequencing, offer new possibilities for precisely characterizing SVs but currently face limitations due to high error rates, limiting their practical utility compared to the gold-standard short-read sequencing techniques.

In the next two paragraphs, we shall describe tools for detecting two of the major consequences of SVs, namely CNAs and gene fusions.

Copy-number alterations

Copy-number alterations (CNAs) refer to changes in the number of copies of large sections of DNA, ranging from 1 kilobase to entire chromosomes. These alterations result from SVs, as previously described. We shall use the terms copy-number variation (CNV) and CNA interchangeably to denote changes in copy numbers of stretches of DNA larger than 1 kb. This one kilobase limit is slightly problematic for classifying all copy-number changes resulting from SVs affecting segments larger than 50 bases and smaller than one kilobase. Some consider them as long indels, but considering them as CNV may be more appropriate. How they are classified may not matter so long as these events are thoroughly considered. Several early studies of whole genomes have indeed demonstrated how small CNVs in the range 100b-1kb outnumber larger ones (J. Wang *et al.* 2008; Ahn *et al.* 2009), a difference that partly explains why WGS is much better equipped for detecting SVs and their associated CNAs than microarray or targeted sequencing techniques.

Until recent years, extensive analyses by international consortia primarily viewed SVs through the lens of their consequences, particularly focusing on CNAs and gene fusions. CNAs indeed stand out as the most readily identifiable consequence of SVs. Pancancer studies,

such as those by Beroukhim, Mermel, *et al.* (2010), S. L. Carter *et al.* (2012), and Zack *et al.* (2013) have underscored the ubiquity of somatic CNAs in cancer. Dedicated analysis tools like `GISTIC` (Beroukhim, Getz, *et al.* 2007; Mermel *et al.* 2011) have played a crucial role in singling out genomic regions where amplification or deletion played a pivotal role in cancer development. Notably, in their landmark pan-cancer study of somatic CNAs, Beroukhim, Mermel, *et al.* (2010) observed that the most prevalent somatic CNAs in cancer were either focal or encompassed the size of chromosome arms such p and q arms of chromosome 8 which are recurrently lost and amplified, respectively (Kou *et al.* 2020). Utilizing SNPs array profiles from over 3,000 solid cancer specimens, they noted that "in a typical tumor, 25% of the genome is affected by arm-level somatic CNAs and 10% by focal somatic CNAs, with 2% overlap". Employing the `GISTIC` tool, they detected over 150 regions subject to recurrent copy number alterations in their cohort, with 122 of these regions lacking explanation through the presence of a known cancer driver gene.

In the 2000s and early 2010s, prior to the widespread availability of NGS technologies, copy-number profiles of genomes were commonly analyzed using DNA microarrays. Most notably, CGH arrays, such as the Agilent microarrays 244K, 2x415K, or 1x1M, and SNP arrays, such as the SNP Affymetrix 6.0 array, were extensively utilized in genomic studies following the completion of the first human genome, including by the TCGA consortium, which made use of both technologies. CGH arrays involve labeling a test sample and a normal reference sample with distinct fluorescence markers, which are then hybridized onto a microarray containing tens of thousands or even millions of probes. The ability to incorporate two different DNAs on CGH arrays was a great advantage over the single-channel SNParrays as it allowed to control for many source of biases in copy number that are difficult to model and control for otherwise. However, it is unable to determine the genotype of specific segments. In contrast, SNP arrays can simultaneously identify copy-number changes and genotype information, but only at the positions of known SNPs. In contrast to microarray technologies, DNA sequencing of individual samples or paired tumor and normal samples presents numerous advantages for analyzing CNAs. These advantages include the ability to accurately estimate absolute copy numbers, compute total and allele-specific copy numbers, consider changes across the entire genome or exome, determine the precise location of SV breakpoints, particularly via WGS, and differentiate between somatic and germline alterations when a matched healthy tissue is available.

Table A.3 lists some of the *CNA callers* commonly used to analyze germline or somatic CNAs in samples sequenced through targeted sequencing, WES, or WGS. The table shows that some tools were developed to run specifically on data produced from one sequencing technology or only from paired sequencing files. FACETS (Shen & Seshan 2016) is, for example, a tool developed by a duo of statisticians from the MSK that was designed to analyze somatic CNAs from paired samples profiled on gene panels or the whole exome and which we will describe more extensively in Chapter 3.

Broadly speaking, tools for calling CNAs employ various strategies to identify and precisely characterize genomic variations of all sizes using NGS data. Different metrics and characteristics of sequencing reads serve as indicators for SVs, with notable features including variations

in *read density*, *distances between paired reads*, *reads orientations*, and *reads mapping qualities*. Among these, changes in read density play a crucial role in characterizing deletions (low-density) or amplifications (high-density), making it a prominent statistic in CNA-calling algorithms. Accordingly, many CNA calling algorithms estimate the boundaries of CNAs by applying *segmentation algorithms* to read density profiles, often utilizing hidden Markov models or circular binary segmentation (Olshen *et al.* 2004). In addition to read density changes, modified distances between paired-end reads provide valuable information regarding extra or missing bases in the analyzed genome compared to the reference. Unmapped or partially mapped reads, known as *split reads*, are useful for characterizing SVs, as the presence of one or multiple SV breakpoints within a read may be detected from alignment ambiguities. By performing local realignment of relevant segments within split reads, the precise location of breakpoints can be determined.

Calling CNAs from sequencing data, especially from targeted sequencing or WES, is, however, not without imperfections, as compared to WGS. The main reason for this disparity lies in the uneven coverage resulting from sequencing on gene panels or exomes, primarily due to the target capture step, which can critically mislead algorithms. Other confounding factors that hinder the accurate determination of CNAs include overall coverage depth, sample preservation methods (e.g., formalin-fixed, paraffin-embedded (FFPE) versus fresh frozen), tumor purity, and the GC-content or repeat density of the investigated regions. The work conducted by Chen and colleagues from the Somatic Mutation Working group of the SEQC-II consortium (Y.-C. Chen *et al.* 2021) has notably shed light on the influence of these confounding factors on CNA calls made by six different algorithms. The study revealed substantial variability across experiments, particularly in low-purity samples where purities below 50% (common in clinical settings) significantly reduce the number and concordance of calls across callers compared to high-purity samples. Another important finding from this study and other benchmarking investigations is the considerable heterogeneity observed in the calls made by different algorithms on the same data, particularly for focal and low-amplitude events which are more difficult to identify compared to chromosome arm-level CNA or amplifications of very high amplitude (R. Tan *et al.* 2014; Nam *et al.* 2016; Zare *et al.* 2017; Y.-C. Chen *et al.* 2021). These findings emphasize the urgent need for more robust algorithms and CNA calling strategies that involve multiple callers to address these challenges, similarly to the strategies that are currently emerging for calling SNVs or indels (Section 2.1.4.1).

As outlined in Chapter 3, significant efforts were directed towards standardizing the detection of somatic CNAs across the three cohorts we analyzed and compared. This endeavor aimed to mitigate potential sources of technical noise and enable a meaningful comparison of CNA profiles. Specifically, we employed the `FACETS` CNA caller (Shen & Seshan 2016) on the paired raw WES FASTQ or BAM files obtained from the three cohorts of interest.

Gene and RNA fusions

RNA fusions, also called *chimeric transcripts*, or *fusion transcripts*, or *chimeric RNAs*, are RNA molecules combining exons, and sometimes introns, from different parental genes. Upon translation, these molecules have the potential to generate *chimeric proteins* provided no event impeding proper translation, such as a frameshift event, results from the fusion. On the other hand, gene fusions refer to structural alterations at the genomic level where two genes are connected, leading to their transcription as a single chimeric transcript. It is worth noting that some SVs can fuse gene segments with intergenic regions or lncRNAs, which may still undergo partial transcription. Various computational tools have been developed to identify such events in addition to the more commonly acknowledged gene fusions but some do not consider them at all.

Gene fusions can result from different types of SVs causing genomic repositioning, most notably translocations, inversions, and interstitial deletions. RNA fusions, on the other hand, can result from gene fusions but also from alternative mechanisms at the RNA level. *Trans-splicing*, which involves the splicing of exons from different RNA molecules, and *read-through* events, which involve the extension of transcription beyond the typical termination signals, potentially contributing to the formation of extended transcripts, are two other important sources of RNA fusions that cannot be detected from DNA sequencing experiments. The PCAWG study of RNA alterations in cancer has notably revealed that 18% of fusions displayed no evidence of genomic rearrangement (Calabrese *et al.* 2020). Moreover, a comprehensive study of the RNA fusion landscape in expression data generated by TCGA and Genotype-Tissue Expression (GTEx) consortia has revealed that mRNA-mRNA fusions make up only 30.2% of RNA fusions in cancer. In contrast, the remaining 53.7% and 16.1% of fusion events were mRNA-lncRNA and lncRNA-lncRNA fusions, respectively (Guo *et al.* 2020). It is import to exercise caution in defining RNA fusions, considering the persisting incompleteness of genome and transcriptome annotations, despite the rapid pace of new annotations over the past two decades. Yuan *et al.* (2017), for instance, have argued in favor of the exclusion of read-through RNA transcripts from the definition of chimeric transcripts, acknowledging the possibility that these may indeed represent normal RNAs from unannotated genes.

The recurrence of gene and RNA fusions in cancer tissues has been extensively documented (Rabbitts 1994; Heim & Mitelman 2008; Hu *et al.* 2018; Balamurali *et al.* 2019). Chimeric transcripts associated with cancer typically arise from three principal subtypes of SVs: translocations, inversions, and deletions. Example of translocation-induced fusions include the *BCR-ABL1* protein, a product of the t(9,22) translocation between chromosomes 9 and 22 in CMLs, the *MYB-NFIB* chimeric protein resulting from the t(6,9) translocation in adenoid cystic carcinomas, and canonical chimeric transcripts frequently encountered in hematological malignancies, such as the *RUNX1-RUNX1T1* translocation t(8,21) or *PML-RARA* t(15,17) translocation (Section 1.2.2.1). Inversions represent a second significant mechanism for acquiring fusions in cancer, exemplified by the *EML4-ALK* fusion in lung cancers resulting from the inversion of the 2p chromosome arm, or the *CBFB-MYH11* fusion in leukemias stemming from an inversion of chromosome 16. Deletions constitute the third most prevalent

mechanism giving rise to gene fusions. A comprehensive review by Panagopoulos & Heim (2021) has reported cancer-associated gene fusions occurring on most chromosomes. Notable examples include the *TPMRSS2-ERG* gene fusion, identified in 40% of prostate cancers and attributed to an approximately 3 Mb interstitial deletion on chromosome 21, as well as the *DNAJB1-PRKACA* chimeric transcript observed in virtually all fibrolamellar hepatocellular carcinomas, originating from a 400kb deletion on chromosome 19.

Research teams have undertaken the establishment of databases containing information on fusion partners or specific fusion breakpoints, aiming to facilitate the detection, filtering, and interpretation of RNA fusions in both disease and healthy tissues. The foundational work by Mitelman and colleagues resulted in the creation of a reference database for cancer gene fusions, commonly referred to as the "Mitelman database"[20]. Other notable contributions include the TICdb collection, which catalogs translocation breakpoints in cancer (1,374 breakpoints in its v3.3 version) (Novo *et al.* 2007), the COSMIC teams' census of cancer gene fusions listing 305 cancer driver gene fusions as of December 2023[21], and the ChiTaRs database, which, in its fifth version, documented over 23,000 cancer breakpoints. However, it's essential to note that not all RNA fusions inherently possess oncogenic properties, as some have been observed in various healthy cells or tissues. Recurrent chimeric RNAs have indeed been documented in extensive studies of healthy tissues. For instance, Babiceanu *et al.* (2016) reported 291 recurrent fusions in an analysis of over 300 RNA-seq libraries. Similarly, the GTEx consortium's comprehensive analysis of more than 9,000 expression samples through fusion detection tools, identified over 14,000 chimeric RNAs detected five times or more (S. Singh *et al.* 2020). These findings now serve as fusion blacklists, aiding in the exclusion of fusions unlikely to be oncogenic in studies focusing on identifying chimeric RNAs involved in cancer, such as the comprehensive study presented in Chapter 3.

Table A.4 offers an almost exhaustive list of *fusion callers* utilized for detecting RNA fusions from short-read RNA-seq experiments. Notably, the initial tools developed were dedicated to detecting *splice junctions*, i.e., the exon-intron junctions where splicing occurs, to describe previously uncharacterized isoforms or chimeric transcripts. Examples include `TopHat`, `SpliceMap`, and `MapSplice`, all employing *splice-aware* mapping to the reference genome to identify the localization of splice junctions. The prevalent use of mapping to the reference genome in fusion callers can be attributed to the incomplete transcriptomes prevalent during the early years of RNA-seq, even in extensively studied species like humans and mice. While current reference transcriptomes are more comprehensive (Gencode v44, released in July 2023, includes 252,835 transcripts), capturing most splice junctions in healthy cells, the approach of mapping to the reference genome persists as a widely used technique for characterizing novel transcripts or detecting chimeric junctions.

In fusion calling, two core characteristics of reads that serve to identify chimeric events are the abnormal orientations of paired reads and the mapping ambiguities which can contribute evidence for breakpoints in DNA or RNA sequencing data. RNA fusion callers in bulk RNA-seq data heavily depend on two distinct categories of reads. The first category comprises *split*

[20] https://mitelmandatabase.isb-cgc.org/about
[21] https://cancer.sanger.ac.uk/cosmic/fusion

reads, also referred to as *chimeric reads*, which traverse the breakpoint within their sequence. These reads are logically identified by in reads with mapping ambiguities and most notably soft-clipped bases. The second category, exclusive to paired-end sequencing experiments that have largely supplanted single-end RNA-seq, consists of *spanning read pairs*, also known as *bridge read* pairs, or *discordant read pairs*. These designate a pair of reads that deviate from the expected mapping to contiguous genomic regions based on the average insert size between mate pairs. Instead, they map to distant regions of the same chromosome or even different chromosomes. The advent of paired-end sequencing has significantly enhanced the precision of RNA fusion callers.

Detecting genuine novel splice junctions and chimeric transcripts from standard RNA-seq experiments poses a formidable challenge due to factors such as the shortness of reads and the low coverage of genes with low transcription levels, markedly reducing the probability of detecting split or spanning reads. However, the primary challenge faced by RNA fusion discovery methods is distinguishing authentic RNA fusions from artefactual ones. This challenge is made evident by the disparities in the numbers of reported putative novel chimeric RNAs across different callers and their low concordance (Carrara *et al.* 2013; Liu *et al.* 2016; Haas *et al.* 2019).

Computational tools designed for detecting fusions in short-read RNA-seq data can broadly be categorized into *alignment-based* methods and *assembly-based* methods. Alignment-based methods utilize the output of mappers to identify discordant reads in paired-end sequencing data and perform local alignments on putative split reads to precisely determine the breakpoint of the junction, if present. On the other hand, assembly-based methods, exemplified by tools like `JAFFA`, `TrinityFusion`, or `novoRNAbreak`, assemble reads into longer transcripts before proceeding with fusion identification. The first category of methods is computationally efficient, highly sensitive to detecting known fusion partners present in reference databases, and well-suited for short-read sequencing data. However, they are less effective at identifying entirely novel fusions and struggle with fusions involving repetitive or homologous regions due to mapping ambiguities. In contrast, alignment-based methods excel at describing novel or complex fusions but are computationally intensive and more sensitive to data quality. In general, short-read RNA-seq is suitable for quantifying known fusions or describing novel fusions involving well-characterized regions that are neither repetitive nor homologous.

Variations between alignment-based and assembly-based tools depend on the choice of aligners or assembler algorithms and the diverse filtering criteria employed to eliminate low-confidence or specific types of chimeric transcripts. Commonly applied filters include the exclusion of read-through chimeric transcripts, the discarding of putative fusions lacking support from a minimum number of spanning and split reads, the removal of fusions involving homologous genes or regions, and the exclusion of putative junctions supported by split reads with insufficient anchoring lengths on either side of the breakpoint (Carrara *et al.* 2013). The reported number of putative fusions by callers heavily relies on the choices made by the authors to strike a balance between sensitivity and specificity. For example, the authors of `Arriba` acknowledge the use of stringent filters to minimize false positives at the cost "that occasionally driver gene fusions are discarded and events with subtle evidence in RNA-seq

data are lost entirely" (Uhrig *et al.* 2021). Conversely, `SQUID` authors prioritized sensitivity over specificity, aiming to detect transcriptomic SVs beyond gene fusions, i.e non-fusion gene events, such as events involving rearrangement of a tumor suppressor genes with an intergenic or lncRNA region, often resulting in truncated transcripts and potential loss of function. Such variants, not strictly classified as gene fusions, are typically overlooked by most fusion callers.

Similarly to the detection of CNAs, improvements can be achieved through technical and technological advancements. On the technical front, employing consensus strategies that integrate results from multiple fusion callers is a promising axis for enhancing specificity while maintaining high sensitivity. On the technological side, the renewed interest for long-read sequencing technologies like PacBio SMRT sequencing or Oxford Nanopore sequencing opens new opportunities for identifying unknown isoforms, precisely characterizing common fusions, and detecting events too intricate to be reliably delineated by short-read sequencing techniques (Weirather *et al.* 2015).

In the analysis of RNA fusions from RNA-seq profiles of metastatic patients presented in Chapter 3, we combined calls from four top-performing callers and fine-tuned our consensus criteria for the relatively straightforward task of confidently detecting fusions with known or putative roles in cancer. However, it is acknowledged that our consensus criteria may not be optimal in other research contexts where prioritizing sensitivity is crucial or when investigating unknown fusions, such as those implicating intergenic or lncRNA regions.

2.1.4.3. Gene expression quantification

RNA sequencing has significantly advanced transcriptome coverage and resolution in comparison to prior methodologies such as microarrays and first-generation sequencing. As detailed in Section 2.1.2.2, specific library preparation protocols enable the targeted analysis of distinct RNA species, including miRNA, lncRNA, and total RNA, alongside the conventional mRNA. Quantifying genes expression from NGS data involves assessing the abundance of mRNA transcripts in a biological sample. In the early years of RNA-seq, two options existed for gene expression quantification: relying on a reference and mapping to it, or assembling de novo transcripts before proceeding to quantification. Similarly to fusion detection, computationally intensive assembly-based methods are well-suited for constructing de novo a transcriptome of an uncharacterized species or detecting novel transcripts. However, the human reference genome and transcriptome, being extensively studied and well-annotated, are now quite comprehensive. Consequently, de novo assembly of human RNA-seq data is now used only for specific projects that do not aim to provide a comprehensive gene expression landscape but rather investigate specific questions regarding the reference transcriptome or the discovery of novel isoforms. Presently, translational research projects universally rely on the human reference genome or transcriptome to quantify the expression of known genes or transcripts.

A contemporary RNA-seq pipeline typically beings with quality control on FASTQ files, followed by *alignment*, *counting*, and *normalization*. Quality control for raw sequencing data, presented in Section 2.1.3.1, primarily involves read trimming and the removal of low-quality bases. The second step in the pipeline usually involves alignment to the reference genome

or transcriptome, utilizing specific aligners. Subsequently, mapped reads are assigned to either transcripts or genes in a process known as *counting* or *quantification*. However, a transformative class of algorithms known as *pseudo-alignment* or *pseudo-mapping* methods emerged a decade ago, fundamentally altering the practices of gene expression quantification due to their significantly faster execution time. Consequently, current gene expression quantification pipelines are broadly categorized as either alignment-based or pseudo-alignment-based. Alignment-based methods can be further classified based on whether they map to the reference genome, the reference transcriptome, or both. The two following paragraphs will describe these two types of gene expression quantification methods in more detail.

Alignment-based methods require mapping RNA-seq reads, which turned out to be more challenging than mapping reads from DNA sequencing experiments, given that many reads will span exon-exon splice junctions that are not contiguous in the reference genome. Traditional aligners like `Bowtie` and `BWA` are deemed unsuitable for this task, necessitating the use of splice-aware mapping tools such as `TopHat` (Trapnell, Pachter, *et al.* 2009), `MapSplice` (K. Wang *et al.* 2010), `BowTie2` (Langmead & Salzberg 2012) on which `TopHat2` (D. Kim, G. Pertea, *et al.* 2013) relies, `STAR` (Dobin *et al.* 2013), or `HiSat2` (D. Kim, Langmead, *et al.* 2015). Alternatively, unspliced mapping to a reference transcriptome has also become popular, employing splice-aware alignment tools like `TopHat2` and `STAR`. The `RUM` aligner is a notable hybrid tool that maps to both the genome and transcriptome for aligning RNA-seq reads (Grant *et al.* 2011). Numerous comparative studies have been conducted to evaluate the impact of employing different alignment tools on RNA-seq analysis results. These evaluations have uncovered significant differences in alignment-related metrics and exon junction discovery (Grant *et al.* 2011; Borozan *et al.* 2013; Engström *et al.* 2013). As for the mapping step, diverse tools have been developed to quantify gene expression, such as Cufflinks (Trapnell, Williams, *et al.* 2010), eXpress (Roberts & Pachter 2013), RSEM (B. Li & Dewey 2011), HTSeq (Anders *et al.* 2015), or StringTie (M. Pertea *et al.* 2015). Once again, benchmarking studies such as the one by Teng *et al.* (2016) have pointed out variable results across quantification methods, with `RSEM` standing out as a superior method in this benchmark.

Pseudo-alignment methods have completely changed how people quantify gene expression in practice due to their high-speed execution while retaining excellent quantification accuracy. The three main pseudo-alignment quantification methods are `Sailfish` (Patro, Mount, *et al.* 2014), `Kallisto` (Bray *et al.* 2016), and `Salmon` (Patro, Duggal, *et al.* 2017). Of note, the authors of `Salmon` have developed a novel mapping algorithm called *selective alignment* that can overcome the usual mapping errors of pseudo-alignment techniques while retaining computational efficiency (Srivastava *et al.* 2020). This method is now incorporated in `Salmon` tool as the recommended option to improve sensitivity and used internally for reads that cannot be confidently mapped using the pseudo-mapping method. Interestingly, all pseudo-alignment methods accept as input aligned reads and can therefore be used only for their quantification method, as done in the benchmark study of quantification methods mentioned earlier (Teng *et al.* 2016).

Quantification results can be reported either as the direct raw count of reads that align

to the specific feature of interest, namely a transcript or a gene, or using a transformation over raw counts. Many normalization methods have indeed been devised to counter common biases, such as the one introduced by the fact that in short-read sequencing, more reads will naturally align to longer transcripts or by the fact that library sizes are not uniform across samples. The reads per kilobase per million mapped reads (RPKM) metric has been one of the first introduced to account for library size and transcript size effects in single-end experiments. The metric is now superseded by the fragment per kilobase per million mapped reads (FPKM) metric which is nearly identical except that it is designed not to count twice paired reads from paired-end sequencing experiments. Transcripts per million mapped reads (TPM) normalization is a more recent metric that is very similar to RPKM and FPKM in that it controls for library and feature-length effects but uses a different order of mathematical operations to increase inter-sample comparability. Briefly, if we denote as R_c^i and L^i the read counts and length of feature i, RPKM and TPM of feature i are defined as

$$\mathrm{RPKM}_i = \frac{\mathrm{R}_c^i \times 10^9}{\left(\sum_j \mathrm{R}_c^j\right) \times \mathrm{L}^i}, \qquad \mathrm{TPM}_i = \frac{\frac{\mathrm{R}_c^i}{\mathrm{L}^i} \times 10^6}{\sum_j \frac{\mathrm{R}_c^j}{\mathrm{L}^j}} \tag{2.1}$$

Both RPKM/FPKM and TPM are normalization methods used within a sample, with TPM being designed for comparing expression levels across different samples by representing them relative to one million within each sample. The upper-quartile normalization method, introduced by Bullard *et al.* (2010), is another commonly used intra-sample normalization metric and has been notably utilized by TCGA in their initial RNA-seq data release. Inter-sample normalization methods have also been explored, including trimmed mean of m-values (TMM) introduced by Mortazavi *et al.* (2008), and the relative log expression normalized by sample size factors as developed in the `DESeq2` differential expression analysis tool by Love *et al.* (2014). Various benchmarking studies have explored the impact of these normalization methods on downstream analyses (Dillies *et al.* 2013; Maza *et al.* 2013; Lin *et al.* 2016; Quinn *et al.* 2018), with Maza *et al.* (2013) recommending against using FPKM or upper-quartile normalization for downstream differential expression analysis and instead suggesting TMM normalization.

Several benchmarking studies have explored the challenges that bioinformaticians face when selecting specific steps in quantification analysis or entire RNA-seq pipelines. These studies have revealed that the use of different tools can result in significant discrepancies in results (P. Li *et al.* 2015; Teng *et al.* 2016; Srivastava *et al.* 2020). Other benchmarking studies have compared entire RNA-seq pipelines to replicate real-world usage. For example, Arora *et al.* (2020) compared five top-performing RNA-seq pipelines for quantifying mRNA transcript abundances. Although 88% of protein-coding genes had comparable abundance estimates, 12% of genes showed up to 4-fold variation in abundance estimation despite using the exact same samples and sequencing files. Various types of discrepancies were observed in the estimations, suggesting that inter-pipeline differences play a significant role in the uncertainty of mRNA abundance estimates. A recent comprehensive benchmarking study by Corchete *et al.* (2020) compared the quantification results of 192 RNA-seq pipelines and the

impact of pipeline selection on downstream differential expression analysis. Using orthogonal qRT-PCR-based measures of expression of 107 housekeeping genes, a type of genes that are consistently expressed across tissues, non-parametric measures of accuracy and precision revealed that pseudo-alignment methods, notably `Salmon` with the TPM normalization, were the top precision performers. In contrast, the top accuracy performers were mostly alignment-based pipelines that incorporated `HTSeq-Union` counting and TMM normalization. The top ten pipelines that balanced accuracy and precision all featured `HTSeq-Union` or `HTSeq-INTER` counting and TMM normalization. The aligners in this short list were `RUM`, `STAR`, or `TopHat2` (Corchete *et al.* 2020).

The lack of consensus regarding the most appropriate algorithms and pipelines has led to a common practice of reprocessing raw RNA-seq sequencing files using the same pipeline to achieve datasets that are harmoniously processed. Notably, early harmonization efforts were accomplished by the TCGA and GTEx consortia. TCGA initially utilized `TopHat` aligned and `Cufflinks` for transcript assembly and expression quantification. However, with the emergence of new algorithms that offered improved speed and accuracy, many projects, including TCGA, transitioned towards the `STAR` and `RSEM` quantification pipeline. This transition was performed by GDC, who additionally reanalyzed the RNA-seq data of various other large-scale cancer studies to achieve harmonization. GTEx, on the other hand, has applied the `TOPMed` pipeline[22], which internally runs `STAR` and `RSEM`, on their 9,661 samples.

Several independent efforts have also reanalyzed data from large-scale projects, most notably TCGA, with newer pipelines as they emerged. For instance, Rahman *et al.* (2015) applied a pipeline utilizing `Subread` aligner and `featureCounts` quantification to generate integer gene-level read counts, unlike the publicly released TCGA Level 3 data. Similarly, Zheng *et al.* (2019) used `Kallisto` pseudo-aligner followed by `TxImport` (Soneson *et al.* 2016) to generate gene-level counts on all TCGA, which they made publicly available. The gene expression tables thus generated were utilized in our comparative analyses involving TCGA in Chapter 3. Another notable reprocessing effort is the recount2 project (Collado-Torres *et al.* 2017), wherein RNA-seq raw reads from >70,000 human samples deposited in TCGA, GTEx, and SRA were processed uniformly using the `Rail-RNA` aligner, and gene expression counts were obtained via a custom tool named `recountNNLS` (Fu *et al.* 2018). Participants in the recount2 project have further extended their harmonization work into the recount3 project (Wilks *et al.* 2021), which has uniformly processed over 700,000 human and mouse specimens through the `Monorail` pipeline that relies on `STAR` aligner and `Megadepth` quantifier to generate gene-level counts.

In our transcriptomic analyses of Chapter 3, we also opted to analyze RNA-seq sequencing files harmoniously using a single RNA-seq quantification pipeline. As quantification tables of TCGA RNA-seq data generated by `Kallisto` were made publicly available by Zheng *et al.* (2019), we used the exact same pipeline as the one released alongside the paper to analyze our RNA-seq samples from metastatic patients as well as the samples from the validation cohort.

[22] https://topmed.nhlbi.nih.gov/sites/default/files/TOPMed_RNAseq_pipeline_flowchart_COREyr3.pdf

2.1.4.4. The art of variant filtering

As mentioned previously, there exist a myriad of confounding factors that make the identification of variants (SNVs, indels, CNAs) from sequencing data a complex process for which decisions must be made to distinguish artifactual variants from biological ones. As there are no perfect rules for performing this task, any study analyzing molecular alterations from sequencing data must decide on a balance between sensitivity and specificity according to the study objectives so as to adjust the filtering criteria accordingly. Whenever a variant is detected in sequencing data, one must always keep in mind that it may have a technical origin. Errors introduced during the replication steps of the PCR amplification cycles or base reading by the sequencing machine, generation of chimeric sequences during adapter ligation, contamination by other samples, or degradation of the nucleic acids preserved in FFPE are example sources of artifacts during the library preparation and sequencing steps. Some studies have investigated in details specific sources of artifacts, such as the work of Costello *et al.* (2013) who described how oxidative DNA damage during sample preparation can cause C>A (G>T) *transversion* artifacts in targeted capture data. Similarly, Arbeithuber *et al.* (2016) demonstrated that amplifiable DNA lesions, such as 8-oxoguanine and deaminated 5-methylcytosine, can introduce error sources in ultrasensitive sequencing applications, leading to artifactual mutations that can be indistinguishable from true mutations or variants. Haile *et al.* (2019) have shown that FFPE storage can produce strand-split artifact reads due to damage to nucleic acids during treatment with formalin and extraction. Importantly, though FFPE preservation is known to cause fragmentation and chemical modification of the embedded nucleic acids and make downstream molecular analyses more difficult and prone to artifacts, comparative studies with fresh-frozen samples have demonstrated a remarkable concordance in the variants identified to allow for the application of NGS to FFPE samples (Schweiger *et al.* 2009; Oh *et al.* 2015). As the vast majority of tissues stored in medical biobanks are preserved via FFPE, these important studies have opened up vast amounts of data to NGS analysis.

On top of the putative errors introduced during the sample preparation and sequencing, the bioinformatic processing of the sequencing files is also a potential source of artifacts. Errors during the alignment step, one of the first steps of many bioinformatic pipelines, is a common source of artifacts as it affects all downstream steps, particularly the variant calling algorithms. As a consequence, quality controls steps are usually applied throughout pipelines to mitigate the potential sources of artifacts and retain only reads of good enough quality. The removal of all duplicate reads sharing identical boundaries after alignment is, for instance, an effective strategy for removing PCR duplicates[23], but it may remove real reads and some argue that it actually has minimal effect on the variant calling accuracy (Alzheimers Disease Neuroimaging Initiative *et al.* 2016). Careful consideration of the number of reads supporting the variant and their sequencing and alignment qualities are key metrics considered by most algorithms and manual curators to help identify and discard artifactual variants. There are many other metrics that algorithms consider to decide on the status of candidate

[23] See Picard MarkDuplicates https://gatk.broadinstitute.org/hc/en-us/articles/360037052812-MarkDuplicates-Picard-

variants, such as the orientation of the reads supporting the alternative allele, the presence of surrounding artifacts, the presence of common polymorphisms at low frequencies indicating a possible contamination, or the presence of the alteration in a matched healthy sample or in a panel of unmatched healthy samples indicating a variant of germline rather than somatic origin.

For all the reasons detailed above, variant calling algorithms have developed filtering strategies and introduced options that the user can tune in order to further adjust the filtering stringency. Given the difficulties of accurately calling somatic CNA from exome sequencing data, filtering on the tumor purity may be critical for dropping overly noisy samples. Ciania and colleagues have for instance applied a minimum threshold of 20% on the estimated purity from 8,183 primary exome samples from the TCGA in their analysis of CNA profiles from 27 tumor types, effectively reducing the number of analyzable samples by approximately 40% (Ciani *et al.* 2022).

2.2. Signatures of mutational processes

2.2.1. Origin

The concept of mutational signatures was initially introduced in 2012 by researchers from the WTSI. In their study, they conducted a comprehensive analysis of the entire genomes of 21 cases of breast cancer, considering the nucleotides flanking the 5'- and 3'- ends of each substituted nucleotide (Nik-Zainal, Alexandrov, *et al.* 2012). The researchers employed a blind-source separation method to empirically examine their hypothesis, which posited that any given mutated genome results from a relatively limited number of mutagenic processes, each affecting specific genomic loci. The fundamental premise of their approach involved summarizing the complete set of mutations observed within a particular genome into a 96-category *mutation profile*. This categorization was achieved by classifying mutations based on both the specific nucleotide alteration and the nucleotides flanking it. The formulation of this hypothesis was likely influenced by well-established facts, such as the prevalence of CC:GG > TT:AA double nucleotide substitution in mutations associated with UV light exposure, and the preponderance of C:G > A:T transversions in individuals with lung cancer who are smokers. The rest of this section and following sections will allow us to delve further into the computational process behind mutational signatures and their implications for our understanding of mutagenesis in human cancers.

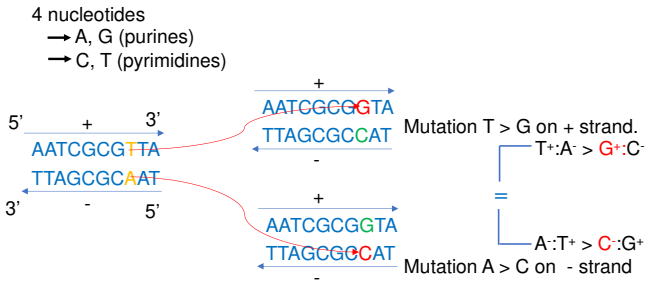


Fig. 2.5.: Occurrence of a single-base substitution (SBS)

To begin with, as there are precisely four different nucleotides constituting the DNA, and as each of them may be substituted by any of the three other bases, there are theoretically 12 possible SBSs. If, additionally, we consider the nucleotides immediately upstream (5') and

downstream (3') of the base under substitution in the DNA chain, a total of 16 potential contexts emerge, thereby engendering 192 feasible SBSs in trinucleotide contexts. However, as depicted in Figure 2.5, when we observe genomes exhibiting mutations, the original strand (forward or reverse) upon which the lesion prompting the mutation initially occurred remains indiscernible. This lack of distinction arises from the identical ultimate outcome, wherein one pair is exchanged for another. As a pair is always the association of a purine with a pyrimidine, any mutation can be codified in a way that the original base is the pyrimidine, such as T > G, or the purine, such as A > C (Figure 2.5). Opting to represent the 12 possible SBSs with the pyrimidine first results results into the six distinct types of SBS: C > A, C > G, C > T, T > A, T > C, and T > G. Considering that each of these six types of substitutions may occur across 16 trinucleotide contexts engenders the 96-categories classification. A concrete illustration of this system can be observed in the mutation G[T>G]T showcased in Figure 2.5. Figure 2.6 provides further illustration of a mutational profile wherein the bars heights depicts the mutation proportion of every type. In this example genome, the most frequently encountered mutation types manifest as C > T transitions in C[C>T]T and C[C>T]C contexts.

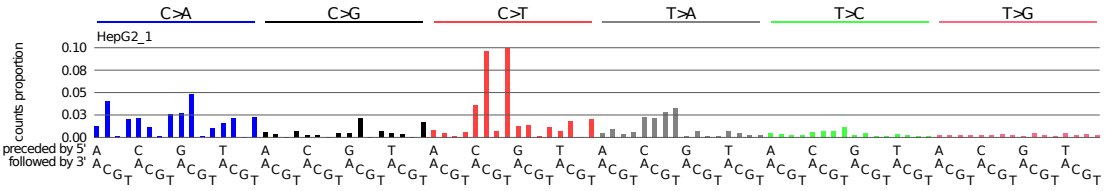


Fig. 2.6.: Example of a mutational profile derived from a mutated genome

To test the hypothesis that only a small number of common mutagenic processes contribute distinctly and additively to the mutational catalog observed in any tumor's genome, Nik-Zainal, Alexandrov, *et al.* (2012) developed a mathematical framework using a blind-source separation algorithm to characterize unknown mutagenic sources from any given set of mutated genomes. After developing the method on a small set of 21 whole genomes, the authors applied it to much larger datasets of cancer genomes aggregated by international consortia, in particular the set of all human cancer exomes profiled by TCGA (Australian Pancreatic Cancer Genome Initiative *et al.* 2013), and later on the set of more than 2,600 cancer genomes as described in PCAWG Mutational Signatures Working Group *et al.* (2020). The application of the method to such large datasets covering the most frequently observed tumor types has permitted the extraction of a large number of mutational signatures that are now maintained and updated in a reference database made publicly available on a dedicated page of the COSMIC portal[24]. The set of extracted signatures now serves as a reference against which any mutated genome can be analyzed individually to identify the signatures that contributed to the mutations observed using one of the multiple projection algorithms that have been developed over the years, as will be described in section 2.2.3.

In nowadays analyses, one may either factorize *de novo* a large matrix of mutational profiles

[24] https://cancer.sanger.ac.uk/signatures/

or use the set of existing signatures to compute the signature activities in every analyzed genome using projection methods. The former choice requires having a sufficiently large dataset in order to perform meaningful analyses. As this setting is very rarely encountered in clinical studies, researchers often resort to projection algorithms to uncover the activities of known mutagens in their samples. If we denote by F the number of mutation types (96 for SBS-based signatures), N the number of mutational profiles (i.e. the number of samples), K the *unknown* number of mutational processes, and $\mathbf{M} \in {\mathbb{R}_+}^{F \times N}$ the matrix of all mutational profiles. Each method, may it be a projection or de novo extraction, aims at decomposing the matrix $\mathbf{M}$ into a low-approximation product of a matrix $\mathbf{W} \in {\mathbb{R}_+}^{F \times K}$ of loadings or factors (or signatures) and a matrix $\mathbf{H} \in {\mathbb{R}_+}^{K \times N}$ of scores or weights, i.e.

$$\mathbf{M} = \mathbf{WH} = \sum_{k=1}^{K} \mathbf{W}_{,k}\mathbf{H}_{k,} \tag{2.2}$$

where each $\mathbf{W}_{,k}\mathbf{H}_{k,}$ is a rank-1 matrix resulting from the product between the k^{th} column of $\mathbf{W}$ and the k^{th} row of $\mathbf{H}$. In case one of the two matrices is already known, different projection algorithms such as non-negative least squares (NNLS) may be used. In other cases where both $\mathbf{W}$ and $\mathbf{H}$ need to be estimated (i.e for *de novo* extraction), the algorithm used for achieving the decomposition is known as NMF. Formally, NMF solves the following non-convex optimization problem

$$\begin{cases} \underset{\mathbf{W},\mathbf{H} \in \mathbb{R}^{F \times K} \times \mathbb{R}^{K \times N}}{\arg\min} & d(\mathbf{M}|\mathbf{WH}) \\ \text{s.t. } W_{fk} \geq 0, H_{kn} \geq 0, & \forall f, k, n \end{cases} \tag{2.3}$$

with d a divergence function applied and summed element-wise on the matrices. Details about the possible cost functions, optimization algorithms, and rules for selecting an optimal number of K are given in Annex A.2.2. A complete example is provided in the next section which presents the original method developed at the WTSI for the discovery of mutational signatures.

2.2.2. WTSI de novo extraction

The first algorithm developed for extracting mutational signatures was released in Matlab in 2012[25] before being translated into Python and R[26]. The authors now discourage the usage of the last Matlab version in favor of the Python tools (Islam *et al.* 2022). The original implementation of the procedure for extracting mutational signatures is presented in Algorithm 1, which relies on NMF to decompose the non-negative matrix $\mathbf{M}$ of all mutational profiles into two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ representing the mutational signatures (columns of $\mathbf{W}$) and the activities of these signatures in each genome (columns of $\mathbf{H}$). While NMF has been previously applied in the realm of biomedical sciences, the analysis

[25] http://www.mathworks.com/matlabcentral/fileexchange/38724
[26] https://github.com/AlexandrovLab

of mutational signatures analysis stands as one of its most successful applications. Prior to this application, NMF found utility in the early 2000s for identifying cancer subtypes from gene expression microarray experiments. Examples include the work of Brunet *et al.* (2004), as well as Gao & Church (2005), who applied this technique to AML/ALL and CNS tumors. Additionally, NMF was employed to extract biological features that remain invariant to technical variations, known as "metagenes", facilitating cross-platform and cross-species analyses (Tamayo *et al.* 2007). The reader is referred to (Devarajan 2008) for more examples of early applications of NMF in biomedical research.

As is customary for any application involving NMF, a set of key choices must be made, mainly three: the selection of the cost function, the optimization algorithm, and the criteria for determining an optimal value for K. A comprehensive exploration of potential selections for each of these fundamental components of NMF applications is presented in Annexes A.2.2.1, A.2.2.2, and A.2.2.3, respectively. In their original methodology for mutational signature extraction, Alexandrov and colleagues opted for the Kullback-Leibler divergence as the cost function, along with the straightforward multiplicative update rules introduced by Lee and Seung in their seminal work (D. D. Lee & Seung 2001). The selection of an optimal K value involved iteratively performing extractions for various candidate ranks and subsequently selecting the rank that best balances between two quantitative metrics, one assessing the stability of each factorization - S_K - and another its distance to the matrix $\mathbf{M}$ - E_K.

Algorithm 1 delineates the process of mutational signature extraction for a specific value of K. The algorithm may be dissected into three principal, consecutive parts: 1. the repeated application of NMF on subtly modified instances - referred to as *bootstrapped* - of the mutation count matrix; 2. the aggregation of all factorizations through a clustering algorithm, resulting in the derivation of mean factor matrices $\bar{\mathbf{W}}$ and $\bar{\mathbf{H}}$; 3. the application of a procedure that induces sparsity on the mean factor matrices. It is common practice to repeat NMF, often utilizing different initial factor matrices. This repetition is motivated by certain algorithms' lack of robust theoretical properties, such as the update rules presented by Lee and Seung (Chih-Jen Lin 2007), and by the non-convex nature of problem (2.3), which complicates the search for a global minimum. In their setting, the authors opted to repeat NMF on slightly modified versions of the original count matrix. These modifications were sampled from a multinomial distribution, with parameters reflective of observed mutation frequencies in each mutation type of matrix $\mathbf{M}$. Upon the conclusion of all NMF iterations - with the authors recommending a minimum of 1,000 iterations - a clustering algorithm summarizes all factor matrices into two final mean factor matrices $\bar{\mathbf{W}}$ and $\bar{\mathbf{H}}$. As the order of the signatures may vary from one iteration to another, the computation of these final mean factor matrices is more complex than a mere arithmetic averaging of all factor matrices. To achieve this, a variant of the K-means algorithm is employed, capitalizing on the distinctive structure of matrix $\mathbf{BW}$ to exclusively allocate each column of the constituent matrices $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(I)}$ to an individual cluster. This clustering procedure yields a partitioning denoted as $\mathscr{C}_1, \ldots, \mathscr{C}_K$ of the set $[\![1, KI]\!]$ with partitions of uniform size corresponding to the number of NMF iterations. Subsequently, the silhouette index of the resulting clustering

arrangement is calculated, serving as a metric to assess the stability of the NMF procedure for the specific value of K currently under exploration.

Algorithm 1: WTSI `SigProfiler` algorithm fixed K

Input: $\mathbf{M} = \left(m_f^n\right)_{f\in[\![1,F]\!],n\in[\![1,N]\!]}, K, I$

Preprocess Drop largest subset of count types accounting for less than 1% of all counts. Let $\mathbf{M}$ again be the possibly reduced matrix.

Bootstrap(*I*)
for $i = 1, \ldots, I$ *in parallel* **do**
  for $n = 1, \ldots, N$ **do**
    $\mathbf{M}^{(i)}_{:,n} \sim \mathrm{Multi}(\sum_f m_f^n, (m_1^n, \ldots, m_F^n))$
  $\mathbf{W^{(i)}}, \mathbf{H^{(i)}} \leftarrow$ **NMF**$(\mathbf{M}^{(i)}, K)$

Cluster
Compute $\mathbf{B^W} = \begin{bmatrix}\mathbf{W^{(1)}} & \ldots & \mathbf{W^{(I)}}\end{bmatrix}$ and $\mathbf{B^H} = \begin{bmatrix}\mathbf{H^{(1)}}^\top & \ldots & \mathbf{H^{(I)}}^\top\end{bmatrix}^\top$
Compute K-class clustering using a variant of K-means exploiting the structure of $\mathbf{B^W}$
Compute the clustering silhouette index S_K

Stability
for $k = 1, \ldots, K$ **do**
  $\bar{\mathbf{W}}_{:,k}, \mathbf{S^W}_{:,k} \leftarrow$ avg and std-dev of $\mathbf{B^W}_{:,\mathscr{C}_k}$
  $\bar{\mathbf{H}}_{k,:}, \mathbf{S^H}_{k,:} \leftarrow$ avg and std-dev of $\mathbf{B^H}_{\mathscr{C}_k,:}$

Enforce sparsity
for $n = 1, \ldots, N$ **do**
  Let $\mathscr{K}_0 = \{k \in [\![1,K]\!] | \bar{\mathbf{H}}_{k,n} > 0\}$ and $\mathscr{K}_{\mathrm{can}} = \mathscr{K}_0$
  Compute new weights $\mathbf{h^0}$ using NNLS
  $$\mathbf{h^0}_{\mathscr{K}_\mathbf{0}} = \underset{h\in[0,\mathbf{m_n^\top 1}]^{|\mathscr{K}_0|}}{\arg\min} \|\mathbf{M}_{:,\mathbf{n}} - \bar{\mathbf{W}}_{:,\mathscr{K}_0}\mathbf{h}\|_2; \qquad \mathbf{h^0}_{-\mathscr{K}_\mathbf{0}} = 0$$
  Let $\mathbf{m} = \bar{\mathbf{W}}\mathbf{h^0}$, $d_c^0 = 1 - \frac{\mathbf{M}_{:,\mathbf{n}}\mathbf{m}}{\|\mathbf{M}_{:,\mathbf{n}}\|_2\|\mathbf{m}\|_2}$, and $d_e^0 = \|\mathbf{M}_{:,\mathbf{n}} - \mathbf{m}\|_2$
  for $k = 1, \ldots, |\mathscr{K}_0| - 1$ **do**
    for $l \in \mathscr{K}_{\mathrm{can}}$ **do**
      Compute new weights $\mathbf{h^l}$ as above using NNLS on $\mathscr{K}_{\mathrm{can}} \setminus \{l\}$
      Compute the corresponding values for d_c^l and d_e^l
    Let $l^* = \underset{l\in\mathscr{K}_{\mathrm{can}}}{\arg\min}$ harmonic mean(d_c^l, d_e^l)
    if $d_c^{l^*} - d_c^0 < 1\%$ **then**
      $\bar{\mathbf{H}}_{:,n} \leftarrow \mathbf{h^{l^*}}$
      $\mathscr{K}_{\mathrm{can}} \leftarrow \mathscr{K}_{\mathrm{can}} \setminus \{l^*\}$
    else
      Break
Compute reconstruction quality $E_K = \|\mathbf{M} - \bar{\mathbf{W}}\bar{\mathbf{H}}\|_2$

Postprocess Reinstate dropped count types by setting the corresponding coefficients to 0.

Output: $\bar{\mathbf{W}}, \bar{\mathbf{H}}, \mathbf{S^W}, \mathbf{S^H}, S_K, E_K$

The last step of the algorithm, the "sparsity-enforcement" step, is actually crucial in determining the final set of weights and, therefore, the activity of each signature in each sample. For each profile under scrutiny, a set of candidate active signatures $\mathscr{K}_{\text{can}}$ is identified using the column of weights $\bar{\mathbf{H}}_{:,n}$ origination from the mean factor matrix. Subsequently, each signature undergoes a comprehensive evaluation to ascertain its possible removal. This assessment entails the reassignment of signature activities within the profile across all candidate signatures, with the exception of the one currently under examination. This is executed employing an NNLS procedure. The evaluation of the refined factorization's quality is undertaken by quantifying the cosine and Euclidean distances, abbreviated as d_c^l and d_e^l, respectively. The signature found to contribute least to the profile is identified by determining the minimum harmonic mean between d_c^l and d_e^l. The said signature is then eliminated from the profile, provided that the change in cosine distance exhibits a deviation of less than 1% relatively to the profile derived from the average matrix $\bar{\mathbf{H}}_{:,n}$ during the preceding step.

The procedure presented in 1 is repeated for multiple candidate values of K, yielding two metrics S_K and E_K for every candidate. These two metrics are subsequently plotted against K, and a visual inspection of the curves for the best trade-off between the factorization stability and the reconstruction fidelity chooses the best value for K. Figure 2A from the article (Alexandrov, Nik-Zainal, *et al.* 2013) provides example curves from simulated cancer genomes. The overall code for the mutational signature extraction, denoted as `SigProfiler`, is made available in diverse implementations including Matlab, Python, and R, as mentioned in the introductory remarks.

Building upon this foundational framework, several research groups have developed analogous tools aimed at performing *de novo* extraction, as reviewed in several recent publications (Omichessan *et al.* 2019; Y.-A. Kim *et al.* 2021; Islam *et al.* 2022). One method in particular, which relies on a Bayesian formulation of NMF allowing automatic selection of the optimal rank of the factorization (V. Y. F. Tan & Févotte 2009), was first employed for the extraction of mutational signatures by Kasar *et al.* (2015) in CLL samples. Shortly after this first application, Kim and colleagues used it also for characterizing the signature associated with nucleotide-excision repair (NER) disruption in urothelial cancers (J. Kim *et al.* 2016). This distinct version of NMF-based signature extraction was integrated into an algorithm termed `SignatureAnalyzer`[27], subsequently utilized in the PCAWG reference study involving 4,645 whole genomes PCAWG Mutational Signatures Working Group *et al.* 2020. The results of this novel extractor were benchmarked against outcomes produced by the original *SignatureProfiler* algorithm within the same study, culminating in overall concordant findings, albeit diverging in cases of hypermutated profiles, where `SignatureAnalyzer` discerned a greater number of signatures.

More recently, a novel study has emerged, addressing the intricacies of rank selection in the context of NMF for mutational signature extraction. A cross-validation-based approach was proposed as an easy and robust solution (D. Lee *et al.* 2022). In this methodology, a subset of mutation counts is intentionally removed from the observation matrix $\mathbf{M}$, thus

[27] https://github.com/getzlab/SignatureAnalyzer

generating missing values. These missing values are then concurrently imputed during the NMF factorization process, employing an expectation/conditional maximization algorithm. The sum of the prediction errors between observed and imputed values across the validation splits serves as a metric for evaluating the appropriateness of the NMF for a given candidate rank. The rank yielding the lowest prediction error is subsequently selected as the optimal number of signatures.

2.2.3. Reference catalog and its applicability

As previously indicated, the concept of mutational signature originated from the examination of the distributions of SBSs detected in WGS of 21 breast cancer (Nik-Zainal, Alexandrov, *et al.* 2012). This seminal work led to the identification of five distinct mutational processes, each delineated by its characteristic profile of 96 mutation types. However, only one of these signatures, namely Signature A, could be confidently attributed to a biological mechanism - specifically, arising from the deamination of 5-methyl-cytosine at NpCpG trinucleotides. In a subsequent publication by the same researchers introducing 1 (Alexandrov, Nik-Zainal, *et al.* 2013), it was revealed that only four signatures could confidently be identified from the 21 breast cancer samples highlighting some sensitivity of signature identification to methodological changes.

The framework for extracting mutational signatures was then extended to thousands of tumor samples using the high-quality data gathered by international consortia, mostly TCGA and ICGC, but also datasets from peer-reviewed papers. The first large-scale application of the method unveiled 21 distinct signatures through the analysis of 7,042 whole-exome-sequenced tumors representing over 30 cancer types (Australian Pancreatic Cancer Genome Initiative *et al.* 2013). Some of these signatures displayed ubiquitous presence across various cancer types, such as signatures 1A and 1B mediated by 5-methyl-cytosine deamination. Conversely, other signatures exhibited cancer type-specific prevalence, such as signature 3, exclusively detected in breast, ovarian, and pancreatic cancer samples. Remarkably, this signature was statistically associated with *BRCA1*/*BRCA2* mutations within each of these three distinct cancer types. However, it is important to acknowledge that while some newly identified signatures could be tentatively linked to putative etiologies, a comprehensive understanding of all these signatures necessitated further rigorous investigation. In response to this need, comprehensive analyses involving 10,250 cancer genomes (Alexandrov, Jones, *et al.* 2015), 560 whole genomes of breast cancers (Nik-Zainal, Davies, *et al.* 2016), 50 cases of oral squamous cell carcinomas (India Project Team of the International Cancer Genome Consortium *et al.* 2013), and cell lines from the COSMIC cell line project[28] were combined to provide a general landscape of mutational signatures. This comprehensive effort resulted in the development of a first reference set comprising 30 mutational signatures, publicly introduced in 2015 under version 2[29].

The application of `SigProfiler` and `SignatureAnalyzer` algorithms in a recent study

[28] https://cancer.sanger.ac.uk/cell_lines/
[29] https://cancer.sanger.ac.uk/signatures/signatures_v2/

encompassing 4,645 whole genomes and 19,184 exomes, which was part of the ICGC collection of 23 papers released in the February 2020 issue of Nature (PCAWG Mutational Signatures Working Group *et al.* 2020), yielded and expanded repertoire of 67 mutational signatures. This enhanced version, denoted as 3.0, has been incorporated into the reference database of signatures as part of COSMIC v89 release in May 2019. In addition to single-nucleotide signatures, this version introduced 11 double-nucleotide and 17 indel signatures. Further details regarding the characteristics of these signatures are provided in section 2.2.4. Among the new set of 67 signatures, 34 had a confirmed biological origin and 15 were of unknown etiology, while speculative attributions have been posited for some. The rest of the signatures were identified as potentially stemming from sequencing artifacts. At the time of writing, the most recent iteration of single-nucleotide signatures stands at version 3.3, encompassing a total of 79 signatures, with 41 of these signatures having confirmed biological origins.

The analysis of mutational signatures has now assumed a key role within cancer genome analysis pipelines, largely facilitated by the availability of diverse computational tools categorized as either *de novo* discovery, refitting, or hybrid discovery tools (Cortés-Ciriano *et al.* 2022). Refitting tools, relying on various mathematical approaches and heuristic principles, are designed to project individual sets of mutations onto a predefined reference set of signatures. This projection challenge is often formulated as an NNLS problem, although the suitability of utilizing the Euclidean distance metric for this purpose warrants further investigation. Numerous tools have been developed to address this challenge, with comprehensive evaluations and comparisons presented in recent literature (Omichessan *et al.* 2019; Y.-A. Kim *et al.* 2021; Islam *et al.* 2022). For instance, the R package `deconstructSigs`, introduced by Rosenthal *et al.* (2016), employs an iterative approach to the NNLS projection problem. Notably, this approach restricts false positives through pre-processing steps and imposes an empirical sparsity-enforcing threshold in post-processing, which ensures that signature contributions below a certain threshold - 6% - are disregarded. However, this approach may lead to false negatives for signatures with low contributions. Another illustrative example is `MutationalPatterns`, a tool released in 2018 (Blokzijl *et al.* 2018) and updated in 2022 (Manders *et al.* 2022). This tool incorporates a rapid implementation of an NNLS algorithm and enables the detection of signature activities in single samples. Comparisons with `deconstructSigs` showcased similar outcomes, with `MutationalPatterns` exhibiting significantly faster runtimes. In their comprehensive review, Omichessan and colleagues proposed an even faster algorithm by adopting a geometric perspective of the decomposition problem. They formulated the problem as a projection onto a cone, with the reference signatures defining the edges of this cone (Omichessan *et al.* 2019). It is important to note that many of these tools recommend a minimum threshold of mutations, typically ranging from 50 (Rosenthal *et al.* 2016) to 200 (Blokzijl *et al.* 2018), within the profile under scrutiny to ensure reliable identification of signature activities. However, the applicability of this criterion depends on the specific sequencing methodology employed (targeted, WES, or WGS) and the particular type of cancer under investigation, potentially leading to the exclusion of substantial numbers of samples from the analysis.

As a result of the community's growing interest in the concept of mutational signatures

and rapid development of refitting tools, many translational studies have included the analysis of mutational signatures as part of their pipeline. In 2016, Nik-Zainal and colleagues conducted extensive genomics analyses of the WGS of 560 breast cancers. They were able to extract twelve signatures, including two novels - SBS26 and SBS30 - that have later been incorporated in version 2 of the catalog of mutational signatures (Nik-Zainal, Davies, *et al.* 2016). In 2019, two genomic studies of metastatic breast cancers (Bertucci *et al.* 2019; Angus *et al.* 2019) both drawing comparison with the genomic landscape of primary breast cancers have observed shifts in the contributions of mutational signatures, most notably an increase in signatures SBS2 and SBS13 - known to be associated with APOBEC-depending mutagenesis - in the hormone receptor-positive/HER2-negative subtype among other differences. Other works have specifically addressed the problem of characterizing new signatures from known mutagens, validating putative biological etiologies, or describing new mechanisms for unexplained signatures. Most notably, in 2017, an experimental study of cancer-associated mutational signatures using CRISPR-Cas9 to inactive key genes involved in DNA repair has allowed us to lift the veil on the origin of SBS30 (Drost *et al.* 2017). Genetically-engineered experiments showed that disruption of base-excision repair (BER) by inactivation of *NTLH1* could reproduce this signature first observed in breast cancers (Nik-Zainal, Davies, *et al.* 2016). In 2019, Kucab *et al.* (2019) set out to characterize the mutational signatures of 79 suspected or known carcinogens using 324 WGS of human-induced pluripotent stem cells and were able to extract characteristic signatures for 41 of them. Following a similar line of reasoning but focusing specifically on the damages induced by antineoplastic drugs, Pich *et al.* (2019) were able to characterize precisely the footprints left by six widely used anticancer therapies and presented novel signatures, most notably the signature of exposure to capecitabine/5-fluorouracil chemotherapies. This signature is dominated by T>G transversions in CpTpT contexts and was first described in a landmark study of esophageal adenocarcinoma (the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium *et al.* 2016) but without proposed origin. It is now known as SBS17b is the reference set.

From a clinical perspective, the analysis of mutational signatures has been successfully used to draw clinically-relevant information from NGS experiments as reviewed in (Koh *et al.* 2021). The accurate prediction by the `HRDetect` tool (Davies *et al.* 2017) using WGS data - 98% specificity! - of HRD caused by *BRCA1/BRCA2* deficiencies in breast cancers, which is informative of sensitivity to a specific class of treatments known as *PARP* inhibitors, was one of the first and most impactful applications. Predictions of MMRd or its closely-related condition MSI serve as other successful applications of mutational signatures in the clinic. These conditions predominantly arise from mutations in mismatch repair (MMR) genes such as *MLH1*, *MSH2*, *MSH6*, and *PMS2* or hypermethylation of *MLH1*, causing errors in DNA replication to go unchecked, buildup of mutations, and instability of microsatellites, and are of particular interest for therapeutic decisions in colorectal and endometrial cancers but also for all solid tumor types as illustrated by the 2017 FDA approval of immune checkpoint inhibitor pembrolizumab for any MMR-deficient solid tumor. The peculiarities of the footprints left on the genome by MMRd - see signatures SBS6,14,15,20,21,26 and 44, or MSI have allowed the development of different predictors using NGS data such as `MMRDetect` (Zou *et al.* 2021).

The reader is referred to the excellent review (Koh *et al.* 2021) and references therein for more examples of clinical applications of mutational signatures.

2.2.4. Extension to other types of alterations

The core idea behind the extraction of mutational signatures has been extended to larger nucleotide contexts but also to totally different classes of events, including doublet-base substitutions (DBSs), indels, CNAs, and, more generally, structural rearrangements. Already in one of the founding articles of mutational signatures (Australian Pancreatic Cancer Genome Initiative *et al.* 2013), authors used a 192-class classification of SBS identified in transcribed regions of well-annotated protein-coding genes by considering the transcription status - transcribed or untranscribed - of the strand on which the mutation occurred. This consideration highlighted a significant *transcriptional strand bias* in several signatures, such as signatures SBS4 and SBS7, for which a suggested cause for this phenomenon was transcription-coupled NER that operates predominantly on the transcribed strand of genes. This bias was also observed in signatures SBS12 and SBS16, but the potential role of the latter mechanism was unclear for these signatures. In this same work, authors reran the extraction of mutational signatures by adding two classes to the 96-class classification to incorporate two types of indels - short nucleotide repeats or with overlapping microhomology at breakpoint junctions - and found an association of these with signatures SBS3,6 and 15. In the study of the landscape of 560 breast cancer genomes by Nik-Zainal, Davies, *et al.* (2016), two indel and six rearrangement signatures were extracted *de novo*, along with twelve substitution signatures. However, the subclassification underlying indel signatures extraction is not clearly specified in the article nor in the methods, which only mention a classification "according to whether they were repeat-mediated, microhomology-mediated or neither."

In 2016, Aggarwala and Voight conducted a comprehensive analysis of substitutions by extending their investigation to encompass 7-nucleotide contexts. Their findings revealed a significant increase in the proportion of explained variation within the distributions of SBSs occurring in intergenic non-coding regions - these regions were selected to mitigate the effects of natural selection. The explained variation surged from 30% - a measure derived from considering mutations solely within trinucleotide contexts, as originally done - to an impressive 84% Aggarwala & Voight (2016). In a parallel vein, Alexandrov and his colleagues embarked on additional exploration of substitution classifications in their recent study involving an extensive dataset of 4,645 whole genomes and 19,184 exomes PCAWG Mutational Signatures Working Group *et al.* (2020). This analysis involved the consideration of two nucleotides upstream and downstream of each mutation, leading to the creation of a 1,536-class classification. The factorization of these extended profiles yielded signatures that were broadly consistent with the signatures detected from trinucleotide contexts. However, multiple signatures showed nonrandom sequence at -2 and +2 contexts. Notably, signatures SBS2 and SBS13 appeared under two different five-nucleotide contexts that could potentially reflect the differential activities of cytidine deaminases *APOBEC3A* and *APOBEC3B*.

In addition to the exploration of broader sequence contexts, the PCAWG Mutational

Signatures Working Group *et al.* (2020) considered alternative classification schemes for alternative events. More specifically, they formulated 78-class and 83-class classifications for the expansive set of more than 800,000 somatic DBSs and four million somatic indels detected across the samples they analyzed, respectively. The analysis of doublet-base signatures was motivated by the observation that the number of detected DBSs markedly exceeded what would be expected from the random adjacency of unrelated SBSs. This observation suggested the involvement of specific underlying mechanisms. The application of the signature extraction methodology revealed eleven DBS signatures. Correlative analyses of the activities of these signatures with that of SBS signatures and clinical parameters unveiled meaningful associations between certain DBS signatures and established mutagenic factors, such as UV-light exposure (DBS1), tobacco smoke exposure (DBS2), and treatment with platinum compounds (DBS5). However, despite these insightful correlations, the biological underpinnings of half of these signatures remain enigmatic at the present time[30]. Turning attention to indels, an in-depth analysis of the 83-channel profiles yielded a collection of 17 signatures. As for DBS, signatures, some of these signatures exhibited links to known mutagens, previously associated with SBS or DBS signatures. For instance, signature ID3 displayed a positive correlation with tobacco smoke exposure and was detected concomitantly with tobacco-associated signatures SBS4 and DBS2. For a comprehensive and updated compilation of signatures and their proposed etiologies, interested readers are directed to the COSMIC website[31].

More recently, structural rearrangements, particularly CNAs, have also been subjected to deconvolution techniques inspired by the framework used for deriving SBS signatures. Various methodologies have been developed, all rooted in the utilization of NMF for the extraction of signatures but using different classification schemes. In the notable study Nik-Zainal, Davies, *et al.* (2016), the authors employed a 32-class division system of rearrangements. This classification split events first by separating regionally-clustered events from unclustered ones, followed by further distinctions into four primary event types: deletions, tandem duplications, inversions, and translocations. Lastly, the size of the rearranged segment was categorized into five possible ranges, except in the case of translocations where size was not considered. The analysis of 560 breast cancer WGS led to the extraction of six rearrangement signatures, revealing varying correlations with factors such as *BRCA1*/*BRCA2* mutations or hypermethylation of their promoter regions, as well as estrogen and hormone receptor statuses.

Subsequently, Macintyre *et al.* (2018) derived a more intricate 36-class classification system for rearrangements. This method involved modeling the distribution of six distinct characteristics of CNAs as mixtures of Gaussian or Poisson distributions. The approach considered breakpoint counts per 10 Mb windows, segment copy numbers, copy number changes, breakpoint counts per chromosome arm, lengths of segments with oscillating copy numbers, and lengths of genomic segments. Each set of these distributions, derived from the shallow WGS data of 91 high-grade ovarian carcinomas, was treated as a mixture. The individual components of each mixture then served to the establishment of a profile

[30] https://cancer.sanger.ac.uk/signatures/dbs/
[31] https://cancer.sanger.ac.uk/signatures

encompassing 36 components. Applying NMF to these profiles revealed seven rearrangement signatures, with associated mechanisms proposed, including those related to *BRCA1/BRCA2*-mediated or -non-mediated HRD, WGD due to cell cycle control failure and PI3K inactivation, and tandem duplication arising from *CDK12* inactivation. Building upon this work, S. Wang *et al.* (2021) introduced a modified representation of copy numbers, resulting in an 80-class classification that encompassed eight general characteristics. This revised method, implemented through a practical tool named `sigminer`, offers biological interpretability for each component, thus departing from hard-to-interpret and tumor type-specific mixture components from the study of Macintyre and colleagues.

Last but not least, Steele *et al.* (2022) introduced a general framework for summarizing allele-specific copy-number profiles. The framework was utilized to create a catalog of 21 copy-number signatures from the nearly 10,000 SNP Affymetrix 6.0 microarrays of TCGA. This comprehensive framework first categorizes CNA events into three categories: homozygous deletion, loss-of-heterozygosity (LOH), and heterozygous segments. Further sub-classifications are established based on total copy number ranges and segment size, resulting in a set of 48-classes allele-specific copy-number profiles. The application of NMF to these profiles generated a reference catalog of 21 signatures, with only three of them remaining of unknown origin. This catalog was subsequently incorporated into the version 3.3 of the COSMIC collection of mutational signatures[32].

2.3. Are all alterations causing cancer?

2.3.1. Genomic heterogeneity

2.3.1.1. Germline variants

The great extent of genetic variation in the human population is the basis of the great phenotypic diversity of between individuals but also the basis for genetic disorders including hereditary cancer syndromes. Each human genome is made unique - except for monozygotic twins - by the combination of the germline DNA variants they inherit from their father and mother's DNAs during fertilization. Large sequencing efforts like the 1000G have allowed to quantify precisely the number and types of germline variations found in human populations from different ancestries. In their reference study of human genetic variation, The 1000 Genomes Project Consortium (2015) estimated that a typical genome differs from the reference human genome by 4.1 to 5 million sites and that about 20 million additional bases are affected by SVs of all types. These estimates are approximately in line with earlier analyses of the complete set of genetic variants detected in the genome of Craig Venter - the first published personal genome; Levy *et al.* (2007) - which estimated that this particular genome departed from the reference sequence by about 1.6% split between short indels and CNAs (1.2%), inversions (0.3%) and SNPs (0.1%) (Pang *et al.* 2010).

Although about 99.9% of germline variants are SNPs or short indels, with SNP being

[32] https://cancer.sanger.ac.uk/signatures/cn/

about ten times more frequent than indels, they affect less bases than the 2,100 to 2,500 SVs typically encountered in a human genome (The 1000 Genomes Project Consortium 2015). Sequencing efforts have revealed that SNPs are encountered at a rate above one in 1,000 bases meaning that any two random individuals differ by about 3 to 4 million SNPs. However, the total number of SNPs ever encountered in the human population exceeds by far this number and keeps increasing as more and more sequencing data becomes available. The first draft of the human genome revealed the locations of 2.1 million SNPs, 50% of which only were already catalogued in dbSNP, a database established in 1998 to host the most comprehensive catalogue of SNPs found in human individuals. This database has been continually updated as more and more data was analyzed. Most variants in dbSNP are rare and not true polymorphisms in the sense that their frequency in the general population does not exceed the 1% threshold generally regarded. The gnomAD database, which started as the ExAC database before expanding to whole genomes as technologies evolved, is another widely used resource for describing variants encountered in the general population. It now includes in its version 3.1 about 644 million short variants passing quality filters observed in the whole genomes of 76,156 human samples (S. Chen, Francioli, *et al.* 2022). Interestingly, about 96% to 99% of the millions of SNPs observed in a typical genome are common in the sense that the MAF exceeds 0.5% (The 1000 Genomes Project Consortium 2015). This observation tells us that almost all SNPs of any individual genome can now be described with the current state of the reference databases.

The analyses of large populations have also revealed that SNPs are not distributed randomly across the genome. Indeed, the number of SNPs found in our coding DNA - 20,000 to 30,000 - is lower than what would be expected assuming a uniform distribution of SNPs across the genome - 30,000 to 40,000. Additionally, the observed distribution of the consequences of these SNPs does not reflect what would be expected by chance considering the genetic code and highlights instead a bias towards under-representation of non-synonymous variants. More specifically, we should expect non-synonymous mutations to represent 80% of all single-base mutations given the genetic code but they only make up about half of the about 20,000 to 30,000 SNPs found in exonic regions, indicating a bias towards mutations that preserve the amino acid sequences. These observations put the number of SNPs affecting proteins at about 1% of all SNPs, a figure that was already reported in the first release of the human reference genome (Venter *et al.* 2001). Among these 1% protein-affecting SNPs and the protein-affecting short indels, about 100 to 200 variants cause protein truncation (The 1000 Genomes Project Consortium 2015). They consist mostly of *nonssense* mutations or *frameshift* indels which introduce a stop codon at the site of the variant or shortly after, but also of *splice* variants which usually lead to improper intron removal and protein malfunction. Although not all of these protein-truncating variants are associated to diseases, they are commonly considered as the most damaging events and actually make up the majority of the germline variants associated with predisposition to specific conditions. In their landmark study of genetic predispositions to cancer using TCGA samples, K.-l. Huang *et al.* (2018) reported pathogenic or likely pathogenic variants in about 8% of all samples with remarkable variation across tumor types. About 87% of these variants were nonsense SNVs, frameshift indels, or variants affecting splice sites across 99 different genes. A handful of infamous genes have

now been known for many years to predispose to cancer, particularly *TP53* (Malkin 1994), *BRCA1* (Easton *et al.* 1995), *BRCA2* (Hopper *et al.* 1999), and *RET* (Mulligan *et al.* 1993). Cancer-predisposing germline variants are mostly affecting genes involved in DNA repair pathways namely BER, NER, MMR, HR, non-homologous end joining (NHEJ). Germline mutations in genes involved in each of these pathways have been implicated in cancer. Notable examples include the loss-of-function in HR-implicated *BRCA1* and *BRCA2* genes, which render susceptibility to breast cancer (Easton *et al.* 1995; Hopper *et al.* 1999). Mutations in XPC, linked to NER, are responsible for syndromes such as xeroderma pigmentosum with a highly accrued risk of developing cancer (Cleaver 2005). Mutations in *MLH1*, *MSH2*, or *MSH6* genes, all involved in the MMR machinery, cause Lynch syndrome and are associated with an increased risk of ovarian or endometrial cancer. Mutations in *MUTYH* gene, which is implicated in BER, serve as another prime example of DNA repair defects associated with increased cancer risk, notably colorectal cancer (Al-Tassan *et al.* 2002).

2.3.1.2. Somatic variants

Cancer genomes are themselves extremely heterogeneous. The cancer classifications described in Chapter 1, though already complex and constantly evolving as knowledge and evidence accumulate, only provide little insight into the genomic heterogeneity of cancers. As the results of the first systematic studies of cancer organized per clinically-defined tumor type were revealed in the late 2000s, it became clear that the set of genomic alterations encountered in human cancers was highly variable across and within tumor types. The TCGA series of tumor type-specific comprehensive molecular portraits has allowed to quantify precisely this heterogeneity across the most common tumor types. This pioneering series of studies has revealed that a few genes are commonly mutated across cancer types, pointing to common roles in the initiation and progression towards a malignant state. To name a few only, *TP53* and *CDKN2A* are the most commonly altered tumor suppressor genes whereas *KRAS*, *EGFR*, *PIK3CA* and *CCND1* are among the most frequently mutated oncogenes. Secondly and more importantly, these studies have allowed to grasp the wide variety of alterations encountered and the great heterogeneity in the number of alterations detected across individual genomes. Figure 2.7, which displays the mutational burden measured as the number of somatic mutations - SNVs and indels - per megabase of sequenced genome in three different cohorts, provides an illustration of the variation in the number of detected somatic mutations across and within tumor types. While the most mutagenic tumor types, namely skin cutaneous melanomas (SKCMs) and LUSCs, have median mutational burdens of nearly 10 mut/Mb, the burden of the least mutagenic types, namely pheochromocytoma and paragangliomas (PCPGs) and thyroid carcinomas (THCAs), does not exceed 1 mut/Mb that is to say ten times less. Even more striking in this figure is the wide range of mutational burdens observed across patients in virtually all tumor types, highlighting the diversity of mutation landscapes encountered in cancer patients.

The molecular landscape of each cancer type is typically characterized by a *long-tail* distribution of alteration frequencies. This distribution comprises a small list of frequently altered and well-characterized genes and a long list of infrequently altered genes whose roles

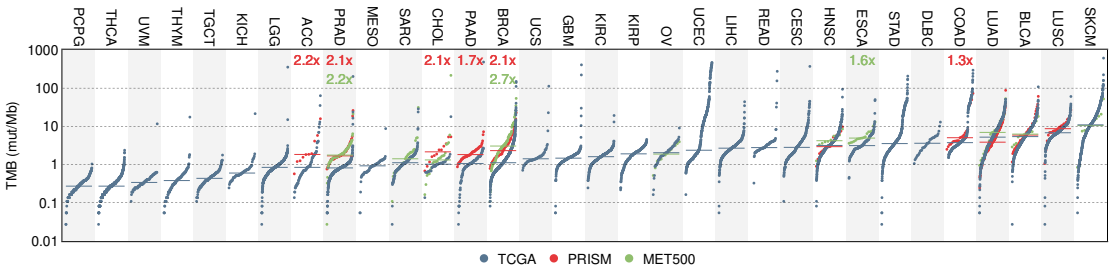


Fig. 2.7.: Mutational burden in the 32 solid tumor types represented in TCGA. Blue, red and green dots are patients from non-metastatic TCGA, metastatic META-PRISM and metastatic (MET500) pan-cancer cohorts, respectively.

in tumorigenesis are generally not clearly understood. The frequent alterations usually involve a few key genes encountered across most tumor types such as *TP53*, and a small number of genes specific to the tumor type under investigation or to a small number of tumor types. Prime examples of alterations encountered in a limited number of tumor types are *IDH1/IDH2* mutations in gliomas and AML (Dang *et al.* 2010); *RET* rearrangements in papillary thyroid carcinoma (PTC) and NSCLC (Parimi *et al.* 2023); *MYB-NFIB* gene fusion in adenoid cystic carcinomas of the breast and head and neck (Persson *et al.* 2009); *HER2* amplification in breast and gastric cancers and, to a lesser extent, gynecological, colorectal and lung cancers; *MYCN* amplification in neuroblastomas (Brodeur *et al.* 1984); deletion of chromosome 5q in MDS (Heim & Mitelman 1986).

The genetic heterogeneity of tumor cells also manifests itself at the scale of individuals and the concept of tumor *clonality* is absolutely essential to the understanding of this heterogeneity. In cellular biology, a *clone* designates a group of identical cells sharing a common ancestor. If no modification of the genome occurred during our lifetime in any of our cells, we would all be a single clone deriving from one egg cell. However, our cells are constantly exposed to endogenous or exogenous factors and natural sources of replication-associated errors that induce somatic alterations in our cells' genomes. This genomic plasticity fuels evolution but also the development of genetic disorders. While it has been commonly accepted for a long time that tumors arise from a single cell with a genetically-acquired selective advantage, this conception has been challenged a few times in the last two decades. Most notably, Pr. B. L. Parsons has discussed in details in her two reviews on the origin of tumor (B. Parsons 2008; B. L. Parsons 2018) how this conception is persisting through time although evidence has accumulated over the years in favor of the alternative hypothesis of the polyclonal origin of tumor. Whether or not a tumor initiates from a single clone or multiple clones, it is important to understand that cancer development is a gradual process where cells progressively acquire malignancy through alterations. At the cellular level, cancer development involves a multistep process of mutation and selection, leading to cells with enhanced capacity for proliferation, survival, invasion, metastasis, and resistance to drugs. This process, known as *clonal selection*, persists throughout tumor development and drives tumor evolution. The alterations acquired in the early stages of the tumor and found across all clones are said to be *clonal* in contrast

with *subclonal* alterations present only in one or a few clones. The existence of subclonal alterations is of paramount importance to cancer treatment as it is the likely reason behind most cancer treatment failures (Schmitt *et al.* 2016). In the late stages of the tumor evolution, the genomic heterogeneity is exacerbated by the independent fate of cells that detached from the primary tumor to form metastases (Yachida *et al.* 2010).

The genomic heterogeneity observed between patients from a same cancer type or between tumor cells from the same individual cancer is now substantially increased by the diversity of cancer treatments used and their effects on cells' DNA or on the clonal composition of the tumor. Treatment with antineoplastic drugs results in the destruction of tumor cells sensitive to the drug but also, all too often, to the selection and expansion of the clones with the ability to resist it. This treatment-induced selection of clones capable of evading the effects of the treatment results in a shift in the landscape of alterations in treated tumors compared to treatment-naive tumors. The amplification of androgen receptor (*AR)* gene is, for instance, almost never encountered in treatment-naive prostate cancers but is the most frequent alteration in recurrent prostate cancer (Visakorpi *et al.* 1995; Robinson *et al.* 2015). Mutations in estrogen receptor 1 (*ESR1*) gene are another example of an alteration infrequently encountered in primary breast cancers but frequently detected after treatment with aromatase inhibitors, a specific type of hormone therapy (Dustin *et al.* 2019). The study presented in Chapter 3 will provide more insights into the genomic differences between treatment-naive tumors and heavily-treated metastatic tumors.

2.3.2. Cancer drivers

2.3.2.1. Cancer hallmarks

The key capabilities that cancer cells acquire and the pathways they hijack for growing out of control are summarized in the essential concepts of *cancer hallmarks*. In their first conceptual work on the topic, Hanahan & Weinberg (2000) introduced a framework for classifying the different general principles of neoplastic processes into six cancer hallmarks thought to be shared across the majority, if not all, cancers. The six original hallmarks included the ability to *replicate indefinitely*, to *activate proliferative signaling* independently from the host, to be *insensitive to growth suppressors*, to *induce angiogenesis*, to *r*esist cell death, and, lastly, to *invade tissues and metastasize*. Tumor development is widely seen as an accelerated form of Darwinian evolution in which successive genetic events confer one or another type of growth advantage, eventually allowing tumor cells to outgrow normal cells. The malignant abilities acquired by tumor cells are the consequence of changes in their protein contents, which, for the majority, originate from alterations of the DNA molecule as subtle as point mutations and as evident as changes in the number of copies or structure of chromosome segments or entire chromosomes. Other mechanisms than mutational lesions contribute to oncogenesis, most notably epigenetic events that are heritable from one cell to another and whose role in cancer was appreciated early on (Laird 1997). However, Section 2.3.2 is focused on describing genetic variants directly implicated in tumorigenesis and will therefore not consider epigenetic events further.

In 2011, two new cancer hallmarks were added to the original six to acknowledge the role of novel biological mechanisms extensively described in the literature for their association with cancer (Hanahan & Weinberg 2011). As previously mentioned in the first version of the hallmarks, tumors are not simply a collection of tumor cells but rather complex tissues consisting of diverse cell types that form the tumor microenvironment (TME). The extensive literature on the role of the TME has led Hanahan and Weinberg to introduce two additional hallmarks: *escape from immune destruction* and *deregulation of cellular metabolism*, emphasizing the impact of the microenvironment on the tumor behavior not dictated by genetic mutations. In 2022, two more hallmarks were added, *phenotypic plasticity* and *c*ellular senescence, bringing the total number of cancer hallmarks to ten (Hanahan 2022). Phenotypic plasticity involves the dedifferentiation of cells, blocking the differentiation of progenitor cells, or transdifferentiation to alternative lineage programs. Cellular senescence, which was once viewed as a mechanism protecting against neoplasia, has been found to stimulate tumor development and malignant progression primarily through the release of chemokines and cytokines forming part of the senescence secretory phenotype.

2.3.2.2. Identification of somatic drivers

The genetic landscape of human cancers is known to be remarkably diverse, with a highly variable number of somatic mutations observed across mutated genomes. It has been established early on that only a small number of genes and mutations, referred to as cancer driver genes and mutations, are responsible for driving oncogenesis (Balmain *et al.* 1993). In a seminal study on cancer genomes, L. D. Wood *et al.* (2007) analyzed 11 breast and 11 colorectal cancer samples and concluded that no more than 15 somatic mutations out of the 80 typically observed in these individual tumors were responsible for driving initiation, progression, or maintenance of the tumor. Subsequently, Vogelstein *et al.* (2013) estimated the number of driver mutations to be between two and eight, with a time span of 20 to 30 years necessary for the occurrence of the successive genetic events required to initiate a tumor. The number of driver mutations thought to be required for driving cancer has remained relatively stable since then, as evidenced by a recent review on the topic of cancer drivers by Ostroverkhova *et al.* (2023), which puts it at between one and four depending on the tumor type. It is widely understood that tumor cells acquire malignant capabilities through different patterns of mutations in two categories of genes: oncogenes, which are activated by gain-of-function missense mutations affecting recurrent amino acid positions, and tumor suppressor genes, which are inactivated by loss-of-function mutations, mostly nonsense and frameshift events, occurring throughout the genome. Non-driver mutations, also known as passenger mutations, represent the majority of all somatic events detected in any cancer genome. They have no effect on neoplastic processes and are mostly the consequence of age-related mutagenesis or tumor-induced genomic instability. Distinguishing passenger events from driver ones has been a long-sought question that remains incompletely resolved today, although it has been addressed through many different angles as presented hereafter.

Identifying driver mutations is a daunting task due to the vast heterogeneity of genotypes

and the lack of standardized datasets for benchmarking and improvement, as noted by Ostroverkhova *et al.* (2023). The groundbreaking study by L. D. Wood *et al.* (2007) revealed that cancer genomes consist of a handful of frequently mutated genes ("mountain" genes) and a long list of rarely mutated ones ("hill" genes). As sequencing techniques improved, more comprehensive genomic coverage and increased statistical power allowed for the discovery of many more rare driver mutations. Algorithmic developments guided by refined understanding of mutagenesis have also enabled increasingly precise predictions. Computational methods for identifying driver genes fall into three broad categories: mutational burden-based, functional bias or sequenced-based, and clustering patterns-based.

Methods assessing the mutational excess over background rates, or mutational burden-based methods, have been introduced early on and improved over the years with increasingly complex models to model the background mutation rates. `Mutsig 1.0`, which computes a cancer mutation prevalence score of all observed somatic mutations using six nucleotide and dinucleotide-specific background rates, stands as a pioneering method in this category (Sjoblom *et al.* 2006). The analysis of 11 breast and 11 colorectal cancer samples through this method resulted in a list of 189 genes, a number that we now know is staggering and comprises many false positives. The method was improved years later to result in the `MutSigCV` algorithm, which attempts to better capture the variability of mutation rates across and within genomes by employing (i) patient-specific mutation frequency and spectrum and (ii) gene-specific background mutation rate incorporating regional replication time and expression level. Whereas `Mutsig 1.0` applied to 180 LUSC samples returned 450 putatively driver genes, `MutSigCV` used on the same samples lowered the list size to just eleven genes (Lawrence, Stojanov, Polak, *et al.* 2013). Other popular methods for detecting driver genes based on mutation recurrence include `Music` (Dees *et al.* 2012), which only models sequences with proper coverage in the available data and compiles p-values from three different statistical tests of mutation excess compared to nucleotide- and context-dependent background mutation rates, or nonsynonymous to synonymous ratio methods based on the assumption that synonymous mutations are less likely to be under selective pressure compared to nonsynonymous mutations. The ratio of nonsynonymous to synonymous was explored early on in the work of Greenman *et al.* (2007), who identified 119 driver genes among 578 protein kinase genes, and was expanded later on in the `dNdScv` method by Martincorena *et al.* (2017), who identified 179 genes under positive selection in 7,664 tumors sequenced by the TCGA.

The second major category of driver gene detection methods involves the analysis of sequence features, such as evolutionary conservation and the functional impact of mutations, to distinguish drivers from passengers. A majority of these methods rely on bioinformatic prediction tools that aim to quantify the functional significance of mutations through various scores such as `VEST` (H. Carter *et al.* 2013), `MutationAssessor` (Reva *et al.* 2011), `SIFT` (Ng 2003), `PolyPhen2` (Adzhubei *et al.* 2010), and `comined annotation dependent depletion (CADD)` (Kircher *et al.* 2014) scores. These scores are based on different metrics, including conservation, transcript-associated information, and protein-level scores, to predict the functional impact of missense mutations and detect genes with unusually impactful mutations. Several methods have been developed to improve the accuracy and scope of driver

gene detection. For instance, `OncodriveFM` (Gonzalez-Perez, Perez-Llamas, *et al.* 2013) uses a suite of functional impact scores, including `SIFT`, `PolyPhen2`, and `MutationAssessor`, to detect driver events in genomic regions. The method was later enhanced by (Mularoni *et al.* 2016) to identify driver events in untranslated and non-coding regions and regions with poor sample coverage. Similarly, the `RF5` method applies the random forest approach to predict three classes of genes (unknown function, oncogene, or tumor suppressor gene in the pan-cancer setting and has identified several potentially new driver genes (Kumar *et al.* 2015). More recently, the `PertInInt` method (Kobren *et al.* 2020) has emerged as a sequence-based method that combines multiple tracks of protein functionality, including interaction domain and evolutionary conservation, with a "natural variation track" to identify driverevents based on the background mutation rate of genes. These methods have advanced our understanding of driver genes and their underlying mechanisms and offer promising avenues for future research.

The last major category of tools for calling driver genes uses the clustering or spatial patterns of mutations as indicators of positive selection. The most simple of these algorithms is probably the 20/20 rule devised by Vogelstein *et al.* (2013), which builds upon their experience that driver genes are best identified through mutation patterns rather than frequency. As oncogenes are recurrently mutated at the same amino acids, known as hotspots, whereas tumor suppressor genes are affected throughout the gene length by nonsense or frameshift mutations that result in loss-of-function, driver genes of each category can, according to the authors, be easily identified by considering as oncogenes those having 20% of recurrent missense mutations for oncogenes,and tumor suppressor genes those harboring 20% of inactivating mutations. Applying this rule to the COSMIC database resulted in 71 tumor suppressor genes and 54 oncogenes. The study of the positional clustering of mutations into mutational hotspots has been modeled by `OncodriveCLUST` (Tamborero, Gonzalez-Perez, *et al.* 2013), which identifies genes with non-silent mutations that cluster together in protein sequences more than expected based on a background distribution of synonymous mutations. The methodological framework was updated with a new linear clustering algorithm to detect genomic regions with significant clustering signal, resulting in the `OncodriveCLUSTL` tool, which was shown to outperform `OncodriveCLUST` (Arnedo-Pac *et al.* 2019). Other notable methods that have harnessed the distribution of mutations inside proteins to detect signals of selection include `ActiveDriver` (Reimand & Bader 2013), which tries to identify recurrently mutated phosphorylation sites, or `e-Driver`, which identifies protein functional regions with biased mutation rates using functional networks of proteins (Porta-Pardo & Godzik 2014).

Hybrid methods have also been developed to capture signals of cancer selection derived from all three sources, namely the excess of mutation recurrence over background rates, the functional bias of mutations, or their spatial distributions along proteins. `MutSig2CV`, which combines all three signals with recurrence excess predictions from `MutSigCV`, analysis of patterns of clustering by `MutSigCL`, and assessment for enrichment in evolutionarily conserved sites by `MutSigFN`, has been applied on 4,742 samples from 21 tumor types to build two catalogs of 254 and 219 cancer genes, respectively, one being slightly more stringent than the other (Lawrence, Stojanov, Mermel, *et al.* 2014). The catalogs thus

generated are available on an interactive portal available at http://www.tumorportal.org. Likewise, `mutpanning` (Dietlein *et al.* 2020) identifies genes that are likely to be functionally relevant based on their abundance of nonsynonymous mutations and their increased number of mutations in unusual nucleotide contexts that deviate from the background mutational process. Other hybrid methods have more data-driven approaches that employ machine learning models to combine the outputs from different driver prediction methods, pathogenicity or functional impact scores, and different annotations of the sequences. A notable example is the `IntOGen` integrative tool (Gonzalez-Perez & Lopez-Bigas 2012; Martínez-Jiménez *et al.* 2020) which combines the scores of the functional impact of `SIFT`, `Polyphen2`, and `MutationAssessor` with the signals of selection predicted by `OncodriveFM` and `OncodriveCLUST` to identify drivers in a cohort of tumor samples. The application of `IntOGen` pipeline on 4,623 samples from TCGA originally uncovered a compendium of 568 cancer genes. The compendium has been expanded upon over time using the more than 28,000 cancer exomes publicly available (Martínez-Jiménez *et al.* 2020). The latest update of `IntOGen` portal, dating back to May 2023, lists 619 driver genes based on the cancer exomes and genomes of more than 33,000 samples representing 73 cancer types[33].

Recent research has revealed that the analysis of driver events in cancer may be better performed through gene networks or pathways. While a diverse range of genotypes are observed within histologically identical tumors, there are a small number of critical biological functions that are implicated in cancer. As discussed by Vogelstein *et al.* (2013), driver events function through a dozen signaling pathways that regulate three core biological processes: cell fate determination, cell survival, and genome maintenance. The success of Singh and colleagues's work (Hristov & M. Singh 2017; Hristov, Chazelle, *et al.* 2020) in employing protein-protein networks with and without prior knowledge of disease-associated genes has allowed for the implication of many lowly-mutated genes in large sets of tumors profiled by TCGA. Additionally, (Sanchez-Vega *et al.* 2018) extensive analysis of ten canonical pathways in one of the 23 papers part of the *PanCancer Atlas* has confirmed that a few critical genes accumulate the majority of driver events but also revealed a complex interplay of co-occurring and mutually exclusive alterations within and across these pathways.

Although less extensively explored, computation methods have also been developed to infer genomic segments under positive signals of selection for changes in the number of their copies, either through loss or amplification. `STAC` tool (Diskin *et al.* 2006)) is a method developed on data generated by CGH arrays which utilizes two complementary statistics to identify non-random gains and losses. The `GISTIC` method released one year later by Beroukhim, Mermel, *et al.* (2010), and its updated version `GISTIC2.0` (Mermel *et al.* 2011) which more finely models the background rates of CNAs according to different genomic characteristics, stands as the method of choice for detecting chromosome arm-level or focal CNA events under positive selection in any set of tumor samples.

[33] https://www.intogen.org/search

2.3.2.3. Databases of somatic drivers

The predictions made by the different algorithms calling somatic drivers on cancer samples have served to build reference databases of cancer driver genes and, increasingly more, databases of cancer driver mutations. These databases often incorporate different protocols and grading scales to quantify scientific confidence in the gene's driving status. The first such effort was the manually curated cancer gene census (CGC), initiated by (Futreal *et al.* 2004) and totaling 291 genesin its first version. The CGC aimed to include only genes with independent and concordant reports of causal links with cancer through mutations, excluding genes implicated in cancer only through altered expression levels or modified methylation patterns. By 2007, the CGC had grown to 350 genes, as reported in Greenman et al.'s study (Greenman *et al.* 2007). Sondka *et al.* (2018)'s re-evaluation of the CGC significantly expanded the list of cancer driver genes to 719 genes, introducing the concept of tiers to distinguish genes with documented evidence of oncogenicity (tier 1, 574 genes) from those with strong indications of playing a role in cancer but without well-established mechanistic evidence (tier 2, 145 genes). In addition to the classical dichotomy between oncogenes and tumor suppressor genes, the CGC also lists genes implicated as a partner in oncogenic fusions but not functioning as an oncogene or tumor suppressor gene alone.

Over the years, researchers have developed increasingly elaborate and accurate cancer driver gene lists using computational tools applied to large cohorts. These in-silico lists contribute to the pieces of evidences considered in the assessment of cancer-driving statuses of genes or specific mutations of the curated databases presented in the next paragraph. The Cancer5000 project identified 254 genes using `MutSig2CV` on 4,742 cancer exomes across 21 tumor types, as previously mentioned (Lawrence, Stojanov, Mermel, *et al.* 2014). The *PanCancer Atlas* released a compendium of 299 cancer genes and 3,400 driver mutations by combining the predictions of 26 computational tools on over 9,000 pan-cancer exomes (Bailey *et al.* 2018). The PCAWG project's flagship paper presented a hybrid approach to build a compendium of 744 cancer genes and 3,719 driver SNVs out of the 22,854 SNVs affecting these genes observed in the more than 2,500 whole genomes analyzed. This approach combined a rank-based approach for discovering new drivers and a rule-based approach for discovering drivers in genomic elements already implicated in cancer.

Similar to the CGC, recent efforts have focused on building high-confidence and, often, manually-reviewed databases of cancer drivers. Unlike the CGC, such efforts have sought to establish reference databases that specifically implicate only particular positions and amino acid changes in known cancer genes, answering the need to lower the resolution of the driver definition from genes to mutations as passenger events are known to occur also in cancer genes. They have additionally aimed at providing the clinical implications that these variations have in tumor type-specific contexts to help clinicians make the most of the current knowledge accumulated in the literature. The most prominent of these efforts has certainly been achieved by the MSK's precision Oncology Knowledge Base (OncoKB) database (Chakravarty *et al.* 2017; Suehnholz *et al.* 2023), which aimed at providing a unique resource for distilling curated knowledge about the oncogenicity as well as the clinical implications (diagnostic, prognostic, and therapeutic) of specific mutations in specific cancer types. As knowledge is constantly

evolving and as strong evidence sometimes only appears through the accumulation of cases or complementary studies, OncoKB employs a grading scale to quantify the current confidence level of the scientific community in each genotype-phenotype relationship. Likewise, Clinical Interpretation of Variants in Cancer (CIViC) database aims to reach similar goals but relies on the community to curate specific mutations' functional and clinical implications (Griffith *et al.* 2017). Although less rich in the total numbers of oncogenic events, CIViC contains more detailed annotations of the clinical implications of known driver mutation and can, therefore, be used in complement to OncoKB as we did for the study presented in Chapter 3. More recently, efforts from the Institute for Research in Biomedicine (Spain) have resulted in a tool called `Cancer Genome Interpreter`[34], which can be used by anyone to automate the interpretation of variants (Tamborero, Rubio-Perez, *et al.* 2018). The platform also serves as a knowledge database of known cancer driver genes and mutations and their therapeutic implications. It additionally provides in-silico predictions from a method named `OncodriveMUT` for classifying the 88% protein-affecting mutations of unknown significance falling into cancer genes. The in-silico predictions have recently been enriched with a machine learning approach named `boostDM` that exploits the patterns of mutations in 568 cancers across 28,000 cancer samples to classify all possible nucleotide changes along cancer genes as either driver or passenger, an analysis known as saturation mutagenesis (Muiños *et al.* 2021).

[34] https://www.cancergenomeinterpreter.org/home

Bibliography

1. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. en. *Nature Methods* **7,** 248–249. doi:10.1038/nmeth0410-248 (Apr. 2010).
2. Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. en. *Nature Genetics* **48,** 349–355. doi:10.1038/ng.3511 (Apr. 2016).
3. Ahn, S.-M. *et al.* The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. en. *Genome Research* **19,** 1622–1629. doi:10.1101/gr.092197.109 (Sept. 2009).
4. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. en. *Cell Reports* **3,** 246–259. doi:10.1016/j.celrep.2012.12.008 (Jan. 2013).
5. Alexandrov, L. B., Jones, P. H., *et al.* Clock-like mutational processes in human somatic cells. en. *Nature Genetics* **47,** 1402–1407. doi:10.1038/ng.3441 (Dec. 2015).
6. Alzheimers Disease Neuroimaging Initiative *et al.* Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. en. *BMC Bioinformatics* **17,** 239. doi:10.1186/s12859-016-1097-3 (July 2016).
7. Anders, S., Pyl, P. T. & Huber, W. HTSeqa Python framework to work with high-throughput sequencing data. en. *Bioinformatics* **31,** 166–169. doi:10.1093/bioinformatics/btu638 (Jan. 2015).
8. Andrews, S. *FastQC: a quality control tool for high throughput sequence data.* http://www.bioinformatics.babraham.ac.uk/projects/fastqc. 2010.
9. Angus, L. *et al.* The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. en. *Nature Genetics* **51,** 1450–1458. doi:10.1038/s41588-019-0507-7 (Oct. 2019).
10. Anzar, I., Sverchkova, A., Stratford, R. & Clancy, T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. en. *BMC Medical Genomics* **12,** 63. doi:10.1186/s12920-019-0508-5 (Dec. 2019).
11. Arbeithuber, B., Makova, K. D. & Tiemann-Boege, I. Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. en. *DNA Research* **23,** 547–559. doi:10.1093/dnares/dsw038 (Dec. 2016).
12. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. en. *Bioinformatics* **35** (ed Schwartz, R.) 4788–4790. doi:10.1093/bioinformatics/btz501 (Nov. 2019).
13. Arora, S., Pattwell, S. S., Holland, E. C. & Bolouri, H. Variability in estimated gene expression among commonly used RNA-seq pipelines. en. *Scientific Reports* **10,** 2734. doi:10.1038/s41598-020-59516-z (Feb. 2020).
14. Australian Pancreatic Cancer Genome Initiative *et al.* Signatures of mutational processes in human cancer. en. *Nature* **500,** 415–421. doi:10.1038/nature12477 (Aug. 2013).
15. Babiceanu, M. *et al.* Recurrent chimeric fusion RNAs in non-cancer tissues and cells. en. *Nucleic Acids Research* **44,** 2859–2872. doi:10.1093/nar/gkw032 (Apr. 2016).

16. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. en. *Cell* **173,** 371–385.e18. doi:10.1016/j.cell.2018.02.060 (Apr. 2018).
17. Balamurali, D. *et al.* ChiTaRS 5.0: the comprehensive database of chimeric transcripts matched with druggable fusions and 3D chromatin maps. en. *Nucleic Acids Research,* gkz1025. doi:10.1093/nar/gkz1025 (Nov. 2019).
18. Balmain, A., Barrett, J. C., Moses, H. & Renan, M. J. How many mutations are required for tumorigenesis? implications from human cancer data. en. *Molecular Carcinogenesis* **7,** 139–146. doi:10.1002/mc.2940070303 (Jan. 1993).
19. Beroukhim, R., Getz, G., *et al.* Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. en. *Proceedings of the National Academy of Sciences* **104,** 20007–20012. doi:10.1073/pnas.0710052104 (Dec. 2007).
20. Beroukhim, R., Mermel, C. H., *et al.* The landscape of somatic copy-number alteration across human cancers. en. *Nature* **463,** 899–905. doi:10.1038/nature08822 (Feb. 2010).
21. Bertucci, F. *et al.* Genomic characterization of metastatic breast cancers. en. *Nature* **569,** 560–564. doi:10.1038/s41586-019-1056-z (May 2019).
22. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. en. *Genome Medicine* **10,** 33. doi:10.1186/s13073-018-0539-0 (Dec. 2018).
23. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. en. *Bioinformatics* **30,** 2114–2120. doi:10.1093/bioinformatics/btu170 (Aug. 2014).
24. Borozan, I., Watt, S. N. & Ferretti, V. Evaluation of Alignment Algorithms for Discovery and Identification of Pathogens Using RNA-Seq. en. *PLoS ONE* **8** (ed Jordan, I. K.) e76935. doi:10.1371/journal.pone.0076935 (Oct. 2013).
25. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. en. *Nature Biotechnology* **34,** 525–527. doi:10.1038/nbt.3519 (May 2016).
26. Brodeur, G. M., Seeger, R. C., Schwab, M., Varmus, H. E. & Bishop, J. M. Amplification of N- *myc* in Untreated Human Neuroblastomas Correlates with Advanced Disease Stage. en. *Science* **224,** 1121–1124. doi:10.1126/science.6719137 (June 1984).
27. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. en. *Proceedings of the National Academy of Sciences* **101,** 4164–4169. doi:10.1073/pnas.0308531101 (Mar. 2004).
28. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. en. *BMC Bioinformatics* **11,** 94. doi:10.1186/1471-2105-11-94 (Dec. 2010).
29. Calabrese, C. *et al.* Genomic basis for RNA alterations in cancer. en. *Nature* **578,** 129–136. doi:10.1038/s41586-020-1970-0 (Feb. 2020).
30. Carrara, M. *et al.* State-of-the-Art Fusion-Finder Algorithms Sensitivity and Specificity. en. *BioMed Research International* **2013,** 1–6. doi:10.1155/2013/340620 (2013).

31. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. en. *BMC Genomics* **14,** S3. doi:10.1186/1471-2164-14-S3-S3 (May 2013).
32. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. en. *Nature Biotechnology* **30,** 413–421. doi:10.1038/nbt.2203 (May 2012).
33. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. en. *JCO Precision Oncology,* 1–16. doi:10.1200/PO.17.00011 (Nov. 2017).
34. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. en. *Bioinformatics* **34,** i884–i890. doi:10.1093/bioinformatics/bty560 (Sept. 2018).
35. Chen, S., Francioli, L. C., *et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes* en. preprint (Genetics, Mar. 2022). doi:10.1101/2022.03.20.485034.
36. Chen, Y.-C. *et al. Comprehensive Assessment of Somatic Copy Number Variation Calling Using Next-Generation Sequencing Data* en. preprint (Bioinformatics, Feb. 2021). doi:10.1101/2021.02.18.431906.
37. Chih-Jen Lin. On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization. *IEEE Transactions on Neural Networks* **18,** 1589–1596. doi:10.1109/TNN.2007.895831 (Nov. 2007).
38. Ciani, Y. *et al.* Allele-specific genomic data elucidate the role of somatic gain and copy-number neutral loss of heterozygosity in cancer. en. *Cell Systems* **13,** 183–193.e7. doi:10.1016/j.cels.2021.10.001 (Feb. 2022).
39. Cleaver, J. E. Cancer in xeroderma pigmentosum and related disorders of DNA repair. en. *Nature Reviews Cancer* **5,** 564–573. doi:10.1038/nrc1652 (July 2005).
40. Collado-Torres, L. *et al.* Reproducible RNA-seq analysis using recount2. en. *Nature Biotechnology* **35,** 319–321. doi:10.1038/nbt.3838 (Apr. 2017).
41. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. en. *Nature* **581,** 444–451. doi:10.1038/s41586-020-2287-8 (May 2020).
42. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. en. *Genome Biology* **17,** 13. doi:10.1186/s13059-016-0881-8 (Dec. 2016).
43. Corchete, L. A. *et al.* Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. en. *Scientific Reports* **10,** 19737. doi:10.1038/s41598-020-76881-x (Nov. 2020).
44. Cortés-Ciriano, I., Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M. & Park, P. J. Computational analysis of cancer genome sequencing data. en. *Nature Reviews Genetics* **23,** 298–314. doi:10.1038/s41576-021-00431-y (May 2022).
45. Cosenza, M. R., Rodriguez-Martin, B. & Korbel, J. O. Structural Variation in Cancer: Role, Prevalence, and Mechanisms. en. *Annual Review of Genomics and Human Genetics* **23,** 123–152. doi:10.1146/annurev-genom-120121-101149 (Aug. 2022).
46. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. en. *Nucleic Acids Research* **41,** e67–e67. doi:10.1093/nar/gks1443 (Apr. 2013).

47. Dang, L., Jin, S. & Su, S. M. IDH mutations in glioma and acute myeloid leukemia. en. *Trends in Molecular Medicine* **16,** 387–397. doi:10.1016/j.molmed.2010.07.002 (Sept. 2010).
48. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. en. *Nature Medicine* **23,** 517–525. doi:10.1038/nm.4292 (Apr. 2017).
49. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. en. *Genome Research* **22,** 1589–1598. doi:10.1101/gr.134635.111 (Aug. 2012).
50. Devarajan, K. Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. en. *PLoS Computational Biology* **4** (ed Bryant, B.) e1000029. doi:10.1371/journal.pcbi.1000029 (July 2008).
51. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. en. *Nature Genetics* **52,** 208–218. doi:10.1038/s41588-019-0572-y (Feb. 2020).
52. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. en. *Briefings in Bioinformatics* **14,** 671–683. doi:10.1093/bib/bbs046 (Nov. 2013).
53. Diskin, S. J. *et al.* STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. en. *Genome Research* **16,** 1149–1158. doi:10.1101/gr.5076506 (Sept. 2006).
54. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. en. *Bioinformatics* **29,** 15–21. doi:10.1093/bioinformatics/bts635 (Jan. 2013).
55. Drost, J. *et al.* Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. en. *Science* **358,** 234–238. doi:10.1126/science.aao3130 (Oct. 2017).
56. Dustin, D., Gu, G. & Fuqua, S. A. W. *ESR1* mutations in breast cancer. en. *Cancer* **125,** 3714–3728. doi:10.1002/cncr.32345 (Nov. 2019).
57. Easton, D. F., Ford, D. & Bishop, D. T. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. eng. *American Journal of Human Genetics* **56,** 265–271 (Jan. 1995).
58. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. en. *Cell Systems* **6,** 271–281.e7. doi:10.1016/j.cels.2018.03.002 (Mar. 2018).
59. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. en. *Nature Methods* **10,** 1185–1191. doi:10.1038/nmeth.2722 (Dec. 2013).
60. Ewing, B. & Green, P. Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities. en. *Genome Research* **8,** 186–194. doi:10.1101/gr.8.3.186 (Mar. 1998).
61. Fang, L. T. *et al.* An ensemble approach to accurately detect somatic mutations using SomaticSeq. en, 13 (2015).
62. Fu, J. M. *et al. RNA-seq transcript quantification from reduced-representation data in recount2* en. preprint (Genomics, Jan. 2018). doi:10.1101/247346.
63. Futreal, P. A. *et al.* A census of human cancer genes. en. *Nature Reviews Cancer* **4,** 177–183. doi:10.1038/nrc1299 (Mar. 2004).

64. Gao, Y. & Church, G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. en. *Bioinformatics* **21,** 3970–3975. doi:10.1093/bioinformatics/bti653 (Nov. 2005).
65. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. en. *Nucleic Acids Research* **40,** e169–e169. doi:10.1093/nar/gks743 (Nov. 2012).
66. Gonzalez-Perez, A., Perez-Llamas, C., *et al.* IntOGen-mutations identifies cancer drivers across tumor types. en. *Nature Methods* **10,** 1081–1082. doi:10.1038/nmeth.2642 (Nov. 2013).
67. Grant, G. R. *et al.* Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). en. *Bioinformatics* **27,** 2518–2528. doi:10.1093/bioinformatics/btr427 (Sept. 2011).
68. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. en. *Nature* **446,** 153–158. doi:10.1038/nature05610 (Mar. 2007).
69. Gregory, S. G. *et al.* The DNA sequence and biological annotation of human chromosome 1. en. *Nature* **441,** 315–321. doi:10.1038/nature04727 (May 2006).
70. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. en. *Nature Genetics* **49,** 170–174. doi:10.1038/ng.3774 (Feb. 2017).
71. Guo, M. *et al.* The landscape of long noncoding RNA-involved and tumor-specific fusions across various cancers. en. *Nucleic Acids Research* **48,** 12618–12631. doi:10.1093/nar/gkaa1119 (Dec. 2020).
72. Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. en. *Genome Biology* **20,** 213. doi:10.1186/s13059-019-1842-9 (Dec. 2019).
73. Hadi, K. *et al.* Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs. en. *Cell* **183,** 197–210.e32. doi:10.1016/j.cell.2020.08.006 (Oct. 2020).
74. Haile, S. *et al.* Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. en. *Nucleic Acids Research* **47,** e12–e12. doi:10.1093/nar/gky1142 (Jan. 2019).
75. Hanahan, D. Hallmarks of Cancer: New Dimensions. en. *Cancer Discovery* **12,** 31–46. doi:10.1158/2159-8290.CD-21-1059 (Jan. 2022).
76. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. en. *Cell* **100,** 57–70. doi:10.1016/S0092-8674(00)81683-9 (Jan. 2000).
77. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. en. *Cell* **144,** 646–674. doi:10.1016/j.cell.2011.02.013 (Mar. 2011).
78. Heim, S. & Mitelman, F. Chromosome abnormalities in the myelodysplastic syndromes. eng. *Clinics in Haematology* **15,** 1003–1021 (Nov. 1986).
79. Heim, S. & Mitelman, F. Molecular screening for new fusion genes in cancer. en. *Nature Genetics* **40,** 685–686. doi:10.1038/ng0608-685 (June 2008).
80. Hopper, J. L. *et al.* Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2. Australian Breast Cancer Family Study. eng. *Cancer Epidemiology, Biomark-*

ers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology **8,** 741–747 (Sept. 1999).

81. Hristov, B. H., Chazelle, B. & Singh, M. uKIN Combines New and Prior Information with Guided Network Propagation to Accurately Identify Disease Genes. en. *Cell Systems* **10,** 470–479.e3. doi:10.1016/j.cels.2020.05.008 (June 2020).
82. Hristov, B. H. & Singh, M. Network-Based Coverage of Mutational Profiles Reveals Cancer Genes. en. *Cell Systems* **5,** 221–229.e4. doi:10.1016/j.cels.2017.09.003 (Sept. 2017).
83. Hu, X. *et al.* TumorFusions: an integrative resource for cancer-associated transcript fusions. en. *Nucleic Acids Research* **46,** D1144–D1149. doi:10.1093/nar/gkx1018 (Jan. 2018).
84. Huang, K.-l. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. en. *Cell* **173,** 355–370.e14. doi:10.1016/j.cell.2018.03.039 (Apr. 2018).
85. Huang, W. *et al.* SMuRF: portable and accurate ensemble prediction of somatic mutations. en. *Bioinformatics* **35** (ed Birol, I.) 3157–3159. doi:10.1093/bioinformatics/btz018 (Sept. 2019).
86. India Project Team of the International Cancer Genome Consortium *et al.* Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. en. *Nature Communications* **4,** 2873. doi:10.1038/ncomms3873 (Dec. 2013).
87. Islam, S. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. en. *Cell Genomics* **2,** 100179. doi:10.1016/j.xgen.2022.100179 (Nov. 2022).
88. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. en. *Nature Communications* **6,** 8866. doi:10.1038/ncomms9866 (Dec. 2015).
89. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. en. *Nature Methods* **12,** 357–360. doi:10.1038/nmeth.3317 (Apr. 2015).
90. Kim, D., Pertea, G., *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. en. *Genome Biology* **14,** R36. doi:10.1186/gb-2013-14-4-r36 (2013).
91. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. en. *Nature Genetics* **48,** 600–606. doi:10.1038/ng.3557 (June 2016).
92. Kim, S. Y., Jacob, L. & Speed, T. P. Combining calls from multiple somatic mutation-callers. en. *BMC Bioinformatics* **15,** 154. doi:10.1186/1471-2105-15-154 (Dec. 2014).
93. Kim, Y.-A. *et al.* Mutational Signatures: From Methods to Mechanisms. en. *Annual Review of Biomedical Data Science* **4,** 189–206. doi:10.1146/annurev-biodatasci-122320-120920 (July 2021).

94. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. en. *Nature Genetics* **46,** 310–315. doi:10.1038/ng.2892 (Mar. 2014).
95. Kobren, S. N., Chazelle, B. & Singh, M. PertInInt: An Integrative, Analytical Approach to Rapidly Uncover Cancer Driver Genes with Perturbed Interactions and Functionalities. en. *Cell Systems* **11,** 63–74.e7. doi:10.1016/j.cels.2020.06.005 (July 2020).
96. Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. en. *Nature Reviews Cancer* **21,** 619–637. doi:10.1038/s41568-021-00377-7 (Oct. 2021).
97. Kou, F., Wu, L., Ren, X. & Yang, L. Chromosome Abnormalities: New Insights into Their Clinical Significance in Cancer. en. *Molecular Therapy - Oncolytics* **17,** 562–570. doi:10.1016/j.omto.2020.05.010 (June 2020).
98. Krøigård, A. B., Thomassen, M., Lænkholm, A.-V., Kruse, T. A. & Larsen, M. J. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. en. *PLOS ONE* **11** (ed Jordan, I. K.) e0151664. doi:10.1371/journal.pone.0151664 (Mar. 2016).
99. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. en. *Cell* **177,** 821–836.e16. doi:10.1016/j.cell.2019.03.001 (May 2019).
100. Kumar, R. D., Searleman, A. C., Swamidass, S. J., Griffith, O. L. & Bose, R. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. en. *Bioinformatics* **31,** 3561–3568. doi:10.1093/bioinformatics/btv430 (Nov. 2015).
101. Laird, P. W. Oncogenic mechanisms mediated by DNA methylation. en. *Molecular Medicine Today* **3,** 223–229. doi:10.1016/S1357-4310(97)01019-8 (May 1997).
102. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. en. *Nature* **409,** 860–921. doi:10.1038/35057062 (Feb. 2001).
103. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. en. *Nature Methods* **9,** 357–359. doi:10.1038/nmeth.1923 (Apr. 2012).
104. Laver, T. *et al.* Assessing the performance of the Oxford Nanopore Technologies MinION. en. *Biomolecular Detection and Quantification* **3,** 1–8. doi:10.1016/j.bdq.2015.02.001 (Mar. 2015).
105. Lawrence, M. S., Stojanov, P., Mermel, C. H., *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. en. *Nature* **505,** 495–501. doi:10.1038/nature12912 (Jan. 2014).
106. Lawrence, M. S., Stojanov, P., Polak, P., *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. en. *Nature* **499,** 214–218. doi:10.1038/nature12213 (July 2013).
107. Lee, D. D. & Seung, H. S. Algorithms for Non-negative Matrix Factorization. en, 7 (2001).
108. Lee, D. *et al.* SUITOR: Selecting the number of mutational signatures through cross-validation. en. *PLOS Computational Biology* **18** (ed Panchenko, A. R.) e1009309. doi:10.1371/journal.pcbi.1009309 (Apr. 2022).

109. Levy, S. *et al.* The Diploid Genome Sequence of an Individual Human. en. *PLoS Biology* **5** (ed Rubin, E. M.) e254. doi:10.1371/journal.pbio.0050254 (Sept. 2007).
110. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. en. *BMC Bioinformatics* **12,** 323. doi:10.1186/1471-2105-12-323 (Dec. 2011).
111. Li, H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* arXiv:1303.3997 [q-bio]. May 2013.
112. Li, H. & Durbin, R. Fast and accurate short read alignment with BurrowsWheeler transform. en. *Bioinformatics* **25,** 1754–1760. doi:10.1093/bioinformatics/btp324 (July 2009).
113. Li, P., Piao, Y., Shon, H. S. & Ryu, K. H. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. en. *BMC Bioinformatics* **16,** 347. doi:10.1186/s12859-015-0778-7 (Dec. 2015).
114. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. en. *Nature* **578,** 112–121. doi:10.1038/s41586-019-1913-9 (Feb. 2020).
115. Lin, Y. *et al.* Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. en. *BMC Genomics* **17,** 28. doi:10.1186/s12864-015-2353-z (Dec. 2016).
116. Liu, S. *et al.* Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. en. *Nucleic Acids Research* **44,** e47–e47. doi:10.1093/nar/gkv1234 (Mar. 2016).
117. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. en. *Genome Biology* **15,** 550. doi:10.1186/s13059-014-0550-8 (Dec. 2014).
118. Macintyre, G. *et al.* Copy number signatures and mutational processes in ovarian carcinoma. en. *Nature Genetics* **50,** 1262–1270. doi:10.1038/s41588-018-0179-8 (Sept. 2018).
119. Malkin, D. Germline p53 mutations and heritable cancer. en. *Annual Review of Genetics* **28,** 443–465. doi:10.1146/annurev.ge.28.120194.002303 (1994).
120. Manders, F. *et al.* MutationalPatterns: the one stop shop for the analysis of mutational processes. en. *BMC Genomics* **23,** 134. doi:10.1186/s12864-022-08357-3 (Dec. 2022).
121. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10. doi:10.14806/ej.17.1.200 (May 2011).
122. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. en. *Cell* **171,** 1029–1041.e21. doi:10.1016/j.cell.2017.09.042 (Nov. 2017).
123. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. en. *Nature Reviews Cancer* **20,** 555–572. doi:10.1038/s41568-020-0290-x (Oct. 2020).
124. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. en. *Proceedings of the National Academy of Sciences* **74,** 560–564. doi:10.1073/pnas.74.2.560 (Feb. 1977).
125. Maza, E., Frasse, P., Senin, P., Bouzayen, M. & Zouine, M. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of

relative size of studied transcriptomes. en. *Communicative & Integrative Biology* **6,** e25849. doi:10.4161/cib.25849 (Nov. 2013).

126. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. en. *Genome Biology* **12,** R41. doi:10.1186/gb-2011-12-4-r41 (Apr. 2011).
127. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. en. *Nature Methods* **5,** 621–628. doi:10.1038/nmeth.1226 (July 2008).
128. Muiños, F., Martínez-Jiménez, F., Pich, O., Gonzalez-Perez, A. & Lopez-Bigas, N. In silico saturation mutagenesis of cancer genes. en. *Nature* **596,** 428–432. doi:10.1038/s41586-021-03771-1 (Aug. 2021).
129. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. en. *Genome Biology* **17,** 128. doi:10.1186/s13059-016-0994-0 (Dec. 2016).
130. Mulligan, L. M. *et al.* Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A. en. *Nature* **363,** 458–460. doi:10.1038/363458a0 (June 1993).
131. Nam, J.-Y. *et al.* Evaluation of somatic copy number estimation tools for whole-exome sequencing data. en. *Briefings in Bioinformatics* **17,** 185–192. doi:10.1093/bib/bbv055 (Mar. 2016).
132. Ng, P. C. SIFT: predicting amino acid changes that affect protein function. en. *Nucleic Acids Research* **31,** 3812–3814. doi:10.1093/nar/gkg509 (July 2003).
133. Nik-Zainal, S., Alexandrov, L. B., *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. en. *Cell* **149,** 979–993. doi:10.1016/j.cell.2012.04.024 (May 2012).
134. Nik-Zainal, S., Davies, H., *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. en. *Nature* **534,** 47–54. doi:10.1038/nature17676 (June 2016).
135. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. en. *Nature* **511,** 428–434. doi:10.1038/nature13379 (July 2014).
136. Novo, F. J., de Mendíbil, I. O. & Vizmanos, J. L. TICdb: a collection of gene-mapped translocation breakpoints in cancer. en. *BMC Genomics* **8,** 33. doi:10.1186/1471-2164-8-33 (Dec. 2007).
137. Nurk, S. *et al.* The complete sequence of a human genome. en. *Science* **376,** 44–53. doi:10.1126/science.abj6987 (Apr. 2022).
138. O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. en. *Genome Medicine* **5,** 28. doi:10.1186/gm432 (2013).
139. Oh, E. *et al.* Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples. en. *PLOS ONE* **10** (ed Castresana, J. S.) e0144162. doi:10.1371/journal.pone.0144162 (Dec. 2015).

140. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. en. *Biostatistics* **5,** 557–572. doi:10.1093/biostatistics/kxh008 (Oct. 2004).
141. Omichessan, H., Severi, G. & Perduca, V. Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. en. *PLOS ONE* **14** (ed Galli, A.) e0221235. doi:10.1371/journal.pone.0221235 (Sept. 2019).
142. Oshlack, A. & Wakefield, M. J. Transcript length bias in RNA-seq data confounds systems biology. en. *Biology Direct* **4,** 14. doi:10.1186/1745-6150-4-14 (2009).
143. Ostroverkhova, D., Przytycka, T. M. & Panchenko, A. R. Cancer driver mutations: predictions and reality. en. *Trends in Molecular Medicine* **29,** 554–566. doi:10.1016/j.molmed.2023.03.007 (July 2023).
144. Panagopoulos, I. & Heim, S. Interstitial Deletions Generating Fusion Genes. en. *Cancer Genomics - Proteomics* **18,** 167–196. doi:10.21873/cgp.20251 (2021).
145. Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. en. *Genome Biology* **11,** R52. doi:10.1186/gb-2010-11-5-r52 (2010).
146. Parimi, V. *et al.* Genomic landscape of 891 RET fusions detected across diverse solid tumor types. en. *npj Precision Oncology* **7,** 10. doi:10.1038/s41698-023-00347-2 (Jan. 2023).
147. Parsons, B. Many different tumor types have polyclonal tumor origin: Evidence and implications. en. *Mutation Research/Reviews in Mutation Research* **659,** 232–247. doi:10.1016/j.mrrev.2008.05.004 (Sept. 2008).
148. Parsons, B. L. Multiclonal tumor origin: Evidence and implications. en. *Mutation Research/Reviews in Mutation Research* **777,** 1–18. doi:10.1016/j.mrrev.2018.05.001 (July 2018).
149. Patel, R. K. & Jain, M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. en. *PLoS ONE* **7** (ed Liu, Z.) e30619. doi:10.1371/journal.pone.0030619 (Feb. 2012).
150. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. en. *Nature Methods* **14,** 417–419. doi:10.1038/nmeth.4197 (Apr. 2017).
151. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. en. *Nature Biotechnology* **32,** 462–464. doi:10.1038/nbt.2862 (May 2014).
152. PCAWG Mutational Signatures Working Group *et al.* The repertoire of mutational signatures in human cancer. en. *Nature* **578,** 94–101. doi:10.1038/s41586-020-1943-3 (Feb. 2020).
153. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. en. *Bioinformatics* **34** (ed Hancock, J.) 867–868. doi:10.1093/bioinformatics/btx699 (Mar. 2018).

154. Persson, M. *et al.* Recurrent fusion of *MYB* and *NFIB* transcription factor genes in carcinomas of the breast and head and neck. en. *Proceedings of the National Academy of Sciences* **106,** 18740–18744. doi:10.1073/pnas.0909114106 (Nov. 2009).
155. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. en. *Nature Biotechnology* **33,** 290–295. doi:10.1038/nbt.3122 (Mar. 2015).
156. Pich, O. *et al.* The mutational footprints of cancer therapies. en. *Nature Genetics* **51,** 1732–1740. doi:10.1038/s41588-019-0525-5 (Dec. 2019).
157. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. en. *Bioinformatics* **30,** 3109–3114. doi:10.1093/bioinformatics/btu499 (Nov. 2014).
158. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. en. *Nature* **575,** 210–216. doi:10.1038/s41586-019-1689-y (Nov. 2019).
159. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. en. *Cell* **174,** 758–769.e9. doi:10.1016/j.cell.2018.06.039 (July 2018).
160. Quinn, T. P., Crowley, T. M. & Richardson, M. F. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. en. *BMC Bioinformatics* **19,** 274. doi:10.1186/s12859-018-2261-8 (Dec. 2018).
161. Rabbitts, T. H. Chromosomal translocations in human cancer. en. *Nature* **372,** 143–149. doi:10.1038/372143a0 (Nov. 1994).
162. Rahman, M. *et al.* Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. en. *Bioinformatics* **31,** 3666–3672. doi:10.1093/bioinformatics/btv377 (Nov. 2015).
163. Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. en. *Molecular Systems Biology* **9,** 637. doi:10.1038/msb.2012.68 (Jan. 2013).
164. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. en. *Nucleic Acids Research* **39,** e118–e118. doi:10.1093/nar/gkr407 (Sept. 2011).
165. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. en. *Nature* **578,** 102–111. doi:10.1038/s41586-020-1965-x (Feb. 2020).
166. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. en. *Nature Methods* **10,** 71–73. doi:10.1038/nmeth.2251 (Jan. 2013).
167. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. en. *Cell* **161,** 1215–1228. doi:10.1016/j.cell.2015.05.001 (May 2015).
168. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. en. *Nature Genetics* **52,** 306–319. doi:10.1038/s41588-019-0562-0 (Mar. 2020).
169. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA re-

pair deficiencies and patterns of carcinoma evolution. en. *Genome Biology* **17,** 31. doi:10.1186/s13059-016-0893-4 (Dec. 2016).

170. Sahraeian, S. M. E. *et al.* Deep convolutional neural networks for accurate somatic mutation detection. en. *Nature Communications* **10,** 1041. doi:10.1038/s41467-019-09027-x (Mar. 2019).
171. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. en. *Cell* **173,** 321–337.e10. doi:10.1016/j.cell.2018.03.035 (Apr. 2018).
172. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. en. *Proceedings of the National Academy of Sciences* **74,** 5463–5467. doi:10.1073/pnas.74.12.5463 (Dec. 1977).
173. Schmitt, M. W., Loeb, L. A. & Salk, J. J. The influence of subclonal resistance mutations on targeted cancer therapy. en. *Nature Reviews Clinical Oncology* **13,** 335–347. doi:10.1038/nrclinonc.2015.175 (June 2016).
174. Schweiger, M. R. *et al.* Genome-Wide Massively Parallel Sequencing of Formaldehyde Fixed-Paraffin Embedded (FFPE) Tumor Tissues for Copy-Number- and Mutation-Analysis. en. *PLoS ONE* **4** (ed Jordan, I. K.) e5548. doi:10.1371/journal.pone.0005548 (May 2009).
175. Seaby, E. G., Pengelly, R. J. & Ennis, S. Exome sequencing explained: a practical guide to its clinical application. en. *Briefings in Functional Genomics* **15,** 374–384. doi:10.1093/bfgp/elv054 (Sept. 2016).
176. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. en. *Nucleic Acids Research* **44,** e131–e131. doi:10.1093/nar/gkw520 (Sept. 2016).
177. Singh, S. *et al.* The landscape of chimeric RNAs in non-diseased tissues and cells. en. *Nucleic Acids Research* **48,** 1764–1778. doi:10.1093/nar/gkz1223 (Feb. 2020).
178. Sjoblom, T. *et al.* The Consensus Coding Sequences of Human Breast and Colorectal Cancers. en. *Science* **314,** 268–274. doi:10.1126/science.1133427 (Oct. 2006).
179. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. en. *Nature Reviews Cancer* **18,** 696–705. doi:10.1038/s41568-018-0060-1 (Nov. 2018).
180. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. en. *F1000Research* **4,** 1521. doi:10.12688/f1000research.7563.2 (Feb. 2016).
181. Srivastava, A. *et al.* Alignment and mapping methodology influence transcript abundance estimation. en. *Genome Biology* **21,** 239. doi:10.1186/s13059-020-02151-8 (Dec. 2020).
182. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. en. *Nature Reviews Genetics* **20,** 631–656. doi:10.1038/s41576-019-0150-2 (Nov. 2019).
183. Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. en. *Nature.* doi:10.1101/2021.04.30.441940 (June 2022).
184. Suehnholz, S. P. *et al.* Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. en. *Cancer Discovery.* doi:10.1158/2159-8290.CD-23-0467 (Oct. 2023).

185. Tamayo, P. *et al.* Metagene projection for cross-platform, cross-species characterization of global transcriptional states. en. *Proceedings of the National Academy of Sciences* **104,** 5959–5964. doi:10.1073/pnas.0701068104 (Apr. 2007).
186. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. en. *Bioinformatics* **29,** 2238–2244. doi:10.1093/bioinformatics/btt395 (Sept. 2013).
187. Tamborero, D., Rubio-Perez, C., *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. en. *Genome Medicine* **10,** 25. doi:10.1186/s13073-018-0531-8 (Dec. 2018).
188. Tan, R. *et al.* An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. en. *Human Mutation* **35,** 899–907. doi:10.1002/humu.22537 (July 2014).
189. Tan, V. Y. F. & Févotte, C. Automatic Relevance Determination in Nonnegative Matrix Factorization. en, 6 (2009).
190. Al-Tassan, N. *et al.* Inherited variants of MYH associated with somatic G:CT:A mutations in colorectal tumors. en. *Nature Genetics* **30,** 227–232. doi:10.1038/ng828 (Feb. 2002).
191. Teng, M. *et al.* A benchmark for RNA-seq quantification pipelines. en. *Genome Biology* **17,** 74. doi:10.1186/s13059-016-0940-1 (Dec. 2016).
192. The 1000 Genomes Project Consortium. A global reference for human genetic variation. en. *Nature* **526,** 68–74. doi:10.1038/nature15393 (Oct. 2015).
193. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. en. *Nature* **578,** 82–93. doi:10.1038/s41586-020-1969-6 (Feb. 2020).
194. the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium *et al.* Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. en. *Nature Genetics* **48,** 1131–1141. doi:10.1038/ng.3659 (Oct. 2016).
195. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. en. *Bioinformatics* **25,** 1105–1111. doi:10.1093/bioinformatics/btp120 (May 2009).
196. Trapnell, C., Williams, B. A., *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. en. *Nature Biotechnology* **28,** 511–515. doi:10.1038/nbt.1621 (May 2010).
197. Uhrig, S. *et al.* Accurate and efficient detection of gene fusions from RNA sequencing data. en. *Genome Research* **31,** 448–460. doi:10.1101/gr.257246.119 (Mar. 2021).
198. Venter, J. C. *et al.* The Sequence of the Human Genome. en. *Science* **291,** 1304–1351. doi:10.1126/science.1058040 (Feb. 2001).
199. Visakorpi, T. *et al.* In vivo amplification of the androgen receptor gene and progression of human prostate cancer. en. *Nature Genetics* **9,** 401–406. doi:10.1038/ng0495-401 (Apr. 1995).
200. Vogelstein, B. *et al.* Cancer Genome Landscapes. en. *Science* **339,** 1546–1558. doi:10.1126/science.1235122 (Mar. 2013).

201. Wang, J. *et al.* The diploid genome sequence of an Asian individual. en. *Nature* **456,** 60–65. doi:10.1038/nature07484 (Nov. 2008).
202. Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. en. *Nucleic Acids Research* **38,** e178–e178. doi:10.1093/nar/gkq622 (Oct. 2010).
203. Wang, M. *et al.* SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. en. *Scientific Reports* **10,** 12898. doi:10.1038/s41598-020-69772-8 (July 2020).
204. Wang, S. *et al.* Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. en. *PLOS Genetics* **17** (ed Gordenin, D. A.) e1009557. doi:10.1371/journal.pgen.1009557 (May 2021).
205. Weirather, J. L. *et al.* Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. en. *Nucleic Acids Research* **43,** e116–e116. doi:10.1093/nar/gkv562 (Oct. 2015).
206. Weischenfeldt, J. *et al.* Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. en. *Nature Genetics* **49,** 65–74. doi:10.1038/ng.3722 (Jan. 2017).
207. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. en. *Nature* **452,** 872–876. doi:10.1038/nature06884 (Apr. 2008).
208. Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing. en. *Genome Biology* **22,** 323. doi:10.1186/s13059-021-02533-6 (Dec. 2021).
209. Wood, D. E. *et al.* A machine learning approach for somatic mutation discovery. en. *Science Translational Medicine* **10,** eaar7939. doi:10.1126/scitranslmed.aar7939 (Sept. 2018).
210. Wood, L. D. *et al.* The Genomic Landscapes of Human Breast and Colorectal Cancers. en. *Science* **318,** 1108–1113. doi:10.1126/science.1145720 (Nov. 2007).
211. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. en. *Nature* **467,** 1114–1117. doi:10.1038/nature09515 (Oct. 2010).
212. Yuan, C. *et al.* It Is Imperative to Establish a Pellucid Definition of Chimeric RNA and to Clear Up a Lot of Confusion in the Relevant Research. en. *International Journal of Molecular Sciences* **18,** 714. doi:10.3390/ijms18040714 (Mar. 2017).
213. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. en. *Nature Genetics* **45,** 1134–1140. doi:10.1038/ng.2760 (Oct. 2013).
214. Zare, F., Dow, M., Monteleone, N., Hosny, A. & Nabavi, S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. en. *BMC Bioinformatics* **18,** 286. doi:10.1186/s12859-017-1705-x (Dec. 2017).
215. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & Von Schack, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. en. *Scientific Reports* **8,** 4781. doi:10.1038/s41598-018-23226-4 (Mar. 2018).

216. Zheng, H., Brennan, K., Hernaez, M. & Gevaert, O. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. en. *GigaScience* **8,** giz145. doi:10.1093/gigascience/giz145 (Dec. 2019).
217. Zou, X. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. en. *Nature Cancer* **2,** 643–657. doi:10.1038/s43018-021-00200-0 (Apr. 2021).
218. Zverinova, S. & Guryev, V. Variant calling: Considerations, practices, and developments. en. *Human Mutation* **43,** 976–985. doi:10.1002/humu.24311 (Aug. 2022).

3. The landscape of refractory metastatic tumors

Contents

Abstract Chapter 3

In this chapter, we will describe the integrative analyses of the genomic and transcriptomic profiles of refractory metastatic patients that I have led throughout my PhD. The work presented in this chapter relies on the concepts and tools presented in Chapters 1 and 2. Although metastatic relapse after treatment is the leading cause of cancer mortality, known resistance mechanisms are missing for most treatments administered to patients. To bridge this gap, we analyzed a pan-cancer cohort - named META-PRISM - of 1,031 refractory metastatic tumors profiled via whole-exome and transcriptome sequencing and included in precision medicine trials led at Gustave Roussy. META-PRISM tumors, particularly prostate, bladder, and pancreatic types, displayed the most transformed genomes compared with primary untreated tumors. Standard-of-care resistance biomarkers were identified only in lung and colon cancers (i.e 9.6% of META-PRISM tumors), indicating that too few resistance mechanisms have received clinical validation. In contrast, we verified the enrichment of multiple investigational and hypothetical resistance mechanisms in treated compared with untreated patients, thereby confirming their putative role in treatment resistance. The analysis of treatment resistances will be presented in depth in Chapter 4. Additionally, we demonstrated that molecular markers improve 6-month survival prediction, particularly in patients with advanced breast cancer. Our analysis establishes the utility of the META-PRISM cohort for investigating resistance mechanisms and performing predictive analyses in cancer.

PRIMARY untreated tumors have been extensively studied through genomic and transcriptomic profiling, yielding valuable insights into their heterogeneity (Hoadley *et al.* 2018; Sanchez-Vega *et al.* 2018; Ding *et al.* 2018) and demonstrating the utility of molecular profiling for precision oncology (Dietel *et al.* 2015; Yates *et al.* 2018; Malone *et al.* 2020). Recently, several studies have further investigated the utility of molecular profiling in advanced cancers (Karapetis *et al.* 2008; Le *et al.* 2015; Drilon *et al.* 2018; Le Tourneau *et al.* 2019; Sicklick *et al.* 2019; Rothwell *et al.* 2019; Rodon *et al.* 2019) or the genomic landscapes of pan-cancer (Zehir *et al.* 2017; Robinson *et al.* 2017; Priestley *et al.* 2019; Nguyen *et al.* 2022) or tumor type-specific (Gundem *et al.* 2015; Naxerova *et al.* 2017; Bertucci *et al.* 2019) metastatic cohorts and shown that genomic differences between metastatic tumors and primary non-metastatic ones. Indeed, mutagenesis is somewhat different between advanced tumors and primary tumors partly due to the impact of some widely used antineoplastic treatments, such as mutagenic platinum-based drugs (Szikriszt, Póti, Pipek, *et al.* 2016; Pich *et al.* 2019). Moreover, the therapeutic pressure imposes additional constraints that contribute to tumor evolution and acquisition of resistance (Poon *et al.* 2014). Consequently, patients who developed resistance to multiple lines of therapies have limited treatment options and dismal prognosis. However, these patients might benefit from phase I clinical trials aiming to find new therapeutic strategies that do not cause severe side effects. An accurate estimation of the expected survival time is vital to determine patients eligibility for these

clinical trials.

Over the course of the last ten years, a series of precision medicine trials have been led at Gustave Roussy that aimed at investigating the clinical benefits of selecting treatments according to the genomic alterations detected from sequencing experiments in patients with advanced tumors (Massard *et al.* 2017; Recondo *et al.* 2020; Berger *et al.* 2022; Bayle *et al.* 2022). Gustave Roussy, as a leading cancer center in France and Europe, has conducted large studies wherein systematic the exome and transcriptome were profiled either at entry into the study or, for the most recent studies, both at entry and at resistance to one or multiple innovative drugs. The retrospective study presented in this chapter is based on the data derived from two large pan-cancer studies led at the institute (Massard *et al.* 2017; Recondo *et al.* 2020) that have allowed us to establish a uniquely large and detailed database containing the clinical and molecular profiles for more than a thousand metastatic patients. Although a few other pan-cancer studies of comprehensive genomic profiles have been released recently, treatment history information is often lacking, which limits the extent of the analysis of the interplay between cancer treatments and the genomic alterations of tumors (Szikriszt, Póti, Pipek, *et al.* 2016; Pich *et al.* 2019; Pleasance, Bohm, *et al.* 2022).

This work introduces META-PRISM, a pan-cancer cohort of 1,031 tumors that progressed under at least one line of treatment or have no approved therapy options. The cohort was analyzed to answer the following three primary objectives:

1. Delineate the genetic alterations landscape in advanced metastatic cancers that have demonstrated resistance to treatment and compare it with treatment-naive early-stage tumors.
2. Ascertain the utility of conducting molecular profiling to risk stratify patients and enhance the precision of eligibility criteria for phase I clinical trials.
3. Explore established or emerging mechanisms of resistance in relation to patients' treatment histories and potentially uncover novel ones.

Somatic variations, germline mutations, and tumor microenvironments were analyzed via WES and RNA-seq. Identified genetic markers were compared with tumor-type matched untreated primary tumors from TCGA (The Cancer Genome Atlas Research Network *et al.* 2013), focusing on functional pathogenic variants associated with resistance, and validated using metastatic tumors from the external MET500 cohort (Robinson *et al.* 2017). We further investigated the utility of genomic and transcriptomic markers to improve the prediction of survival time from objective clinical variables in the META-PRISM cohort such as lactate dehydrogenase (LDH) levels, serum albumin, neutrophil-to-lymphocyte ratio, or the number of metastatic sites (Arkenau *et al.* 2009; Bigot *et al.* 2017).

3.1. The META-PRISM database

The META-PRISM project started in 2020 with the aim to retrospectively analyze the molecular profiles of more than a thousand metastatic patients that have been enrolled in

MOSCATO 1/2 (Massard *et al.* 2017) and MATCH-R (Recondo *et al.* 2020) precision medicine trials at Gustave Roussy. Although the patients analyzed were included in previous studies that have led to publications, only minimal patient and tumor information was readily available for analysis. Therefore, the project's first step consisted of querying databases from multiple hospital services to aggregate the complete clinical history of the patients and exhaustive details about the pathology and the biopsies to correlate the tumors' genotypes with the clinical observations as precisely as possible. The second step of the project involved a complete reanalysis of the archived sequencing files to extract exhaustive lists of genomic alterations and gene expression modifications and ensure the uniform processing of sequencing data. This bioinformatic processing was not limited to the META-PRISM cohort but also included the data from TCGA and MET500 cohorts, which were used for comparison and validation purposes, respectively. The processing of these external cohorts is presented in Section 3.2. The current section details how the database of META-PRISM patients underlying this large study has been assembled, curated, processed, and structured to prepare for the correlative and predictive analyses.

3.1.1. Data retrieval and curation

3.1.1.1. Biopsy and sequencing

The META-PRISM initiative was designed to leverage the wealth of sequencing and clinical data accumulated over the past decade through several large precision medicine trials conducted by Gustave Roussy. Specifically, the cohort encompassed all adult patients with solid tumors who had been offered WES or RNA-seq as part of the concluded MOSCATO 1 trial, as well as the then ongoing MOSCATO 2 and MATCH-R trials, and who had provided informed written consent. One of the participating medical oncologists compiled an initial list of patients, outlining the primary characteristics of their cancer and providing minimal details about their received treatments. This preliminary dataset comprised 1,044 patients, roughly categorized into groups mirroring the 33-type classification employed by TCGA to facilitate tumor-type specific comparisons. While the data included standard clinical information such as gender and date of birth, as well as somewhat standardized descriptions of the primary biopsy site and the site subjected to sequencing, further information was necessary to establish links with the sequencing files stored in archival repositories, as well as to facilitate biopsy selection for patients who had undergone multiple biopsies. To address this, I requested supplementary data from various hospital services and received data tables in diverse formats. As I began assimilating and aggregating these tables, I uncovered discrepancies, prompting a more exhaustive inquiry through the implementation of automated scripts.

The systematic cross-comparison of tables prior to their integration unveiled numerous data discrepancies and, of greater concern, recurrent inconsistencies within the chain of identifiers linking sequencing files to clinical data. These initial findings marked the beginning of a lengthy and *labor-intensive cycle* involving the solicitation of data, comparison with existing data, documentation of discrepancies, and their resolution through manual review with colleagues to establish a meticulously curated clinical, biopsy, and sequencing database.

As we delved into the various hospital databases to refine our own repository, we gathered more comprehensive information regarding the biopsies dates and sites. Additionally, we collected information on the performance status and blood test results obtained through assessments conducted within one month of the biopsies whenever such records were accessible. Concurrently, we undertook the task of compiling the complete history of antineoplastic treatments administered to each patient. This last task proved to be highly demanding and is expounded upon in Section 3.1.1.3.

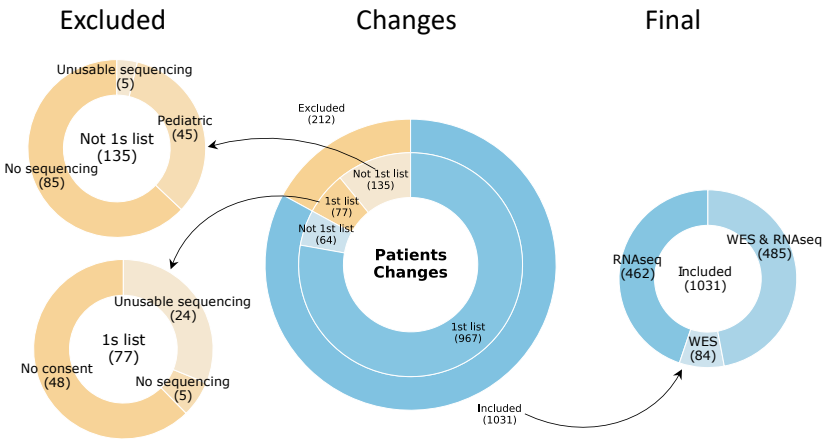


Fig. 3.1.: Changes in patient selection for the META-PRISM cohort. The final list of 1,031 patients is compared to the first list of 1,044 patients compiled at the start of the project.

The examination of the library preparation databases preceding the sequencing phase revealed a notable portion of sequenced samples that had not been initially included in the study. The inclusion of new sequencing data and of more than 50 patients with pancreatic and prostate cancers midway through the project caused shifts in the patient cohort, as delineated in Figure 3.1. Compared to the first list, 77 patients were excluded due to the absence of informed consent upon verification by ethics services 18 months into the project or the absence of sequencing data meeting the required quality control. Indeed, as explained in Chapter 2, the handling of raw sequencing data involves a succession of quality assessments. These assessments serve to flag tumor or blood samples exhibiting poor sequencing quality, characterized by shallow or heterogeneous coverage, an elevated count of duplicates, or sample contamination. In infrequent instances, some samples fail processing by one or multiple algorithms and are consequently disregarded. Subsequent to the processing of the sequencing data, additional expert quality checks are required to discard tumor samples characterized by excessively low tumor purity or samples from instances of patient mismatch between the tumor and blood samples. After all the sequencing data was retrieved, processed, and quality-controlled, a total of 1,031 patients having at least one good-quality WES (569 patients) or RNA-seq sample (937 patients) were retained and considered further (Figure 3.1).

For the study's stated purposes, we selected and analyzed exactly one biopsy per patient. We, therefore, established a set of rules to select one biopsy for patients who had undergone

several. These rules depended on the type of sequencing performed, the biopsy date, and, in case of ties, the sample tumor purity as estimated by a trained histopathologist. Given the substantial number of discrepancies identified in the initial data, we undertook a comprehensive review of the dates and locations of all the biopsies subjected to sequencing within the selected trials.

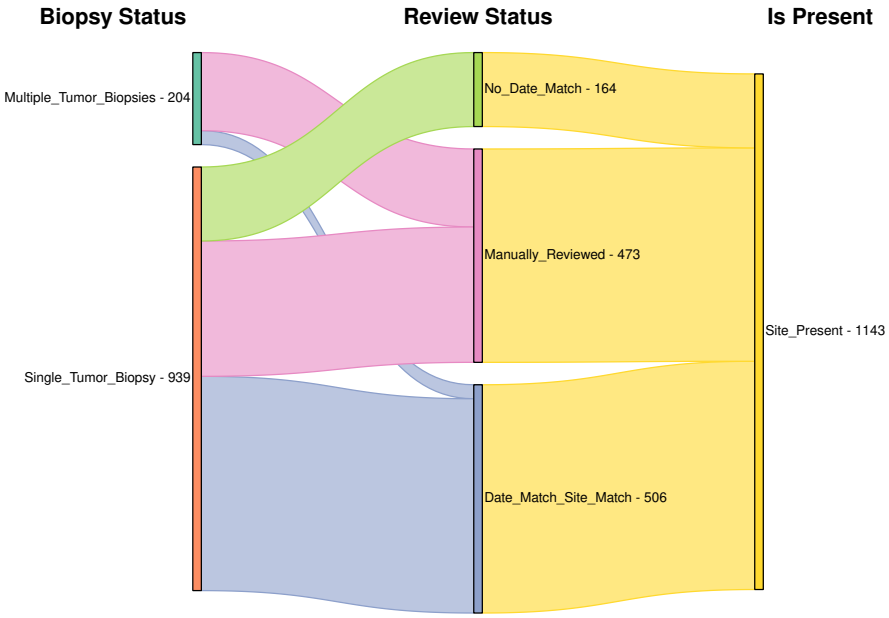


Fig. 3.2.: Review of the tumor biopsies dates and sites to allow for the selection of one biopsy per patient and enrich analyses.

As with all other review tasks, our initial approach sought to extract the necessary data from electronic databases in order to minimize the need for manual review. Regarding biopsy dates and locations, the data could be partially obtained by executing an automated query within the anatomical pathology department databases via Dr Warehouse, a digital database and query tool recently deployed at Gustave Roussy[1]. The resulting table featured the patient identifier, the date of the document extraction, the biopsy site - presented in a non-standardized French format - and a code derived from the French ADICAP nomenclature denoting the nature of the patient's pathology or sample. I processed this table, translating and standardizing the biopsy site designations into the ICD-O-3 nomenclature before cross-referencing against the existing biopsy records also standardized into the ICD-O-3 nomenclature. A match in patient identifier, biopsy date, and site was deemed sufficient to consider that the data could be trusted. Nevertheless, the patient's information was absent from the automatically retrieved table in numerous instances, or no corresponding biopsy date or location could be identified. All such instances necessitated manual review and represented 473 out of the 1,143 biopsies under consideration for the study, as illustrated in Figure 3.2.

[1] https://www.gustaveroussy.fr/fr/recueil-et-utilisation-des-donnees-des-patients-au-sein-de-dr-warehouse

Once we had collected complete information about all the biopsies that were molecularly profiled for the 1,031 patients of the study, we proceeded to apply rules for the 92 patients who had multiple biopsies to choose from, with a median of 2 biopsies (range 2-5) per patient. Given the richness of the genetic information that can be extracted from the availability of WES and RNA-seq, preference was accorded to biopsies that underwent both experiments. In cases where this scenario did not materialize, priority was given to biopsies subjected to WES over RNA-seq (Rule 1). In the event of ties, the subsequent rules were applied: selection of biopsies obtained during the selected trials or shortly before (Rule 2), preference for the most recent biopsies in chronological order (Rule 3), prioritization of biopsies with the highest estimated purity (Rule 4), and ultimately, if any ties persisted, random selection (Rule 5). The latter occurred in instances of multi-site biopsies or single-site fractionated biopsies collected on the same date, undergoing identical sequencing, and having identical estimated tumor purities. The number of biopsies selected according to each of these rules is depicted in Figure 3.3.

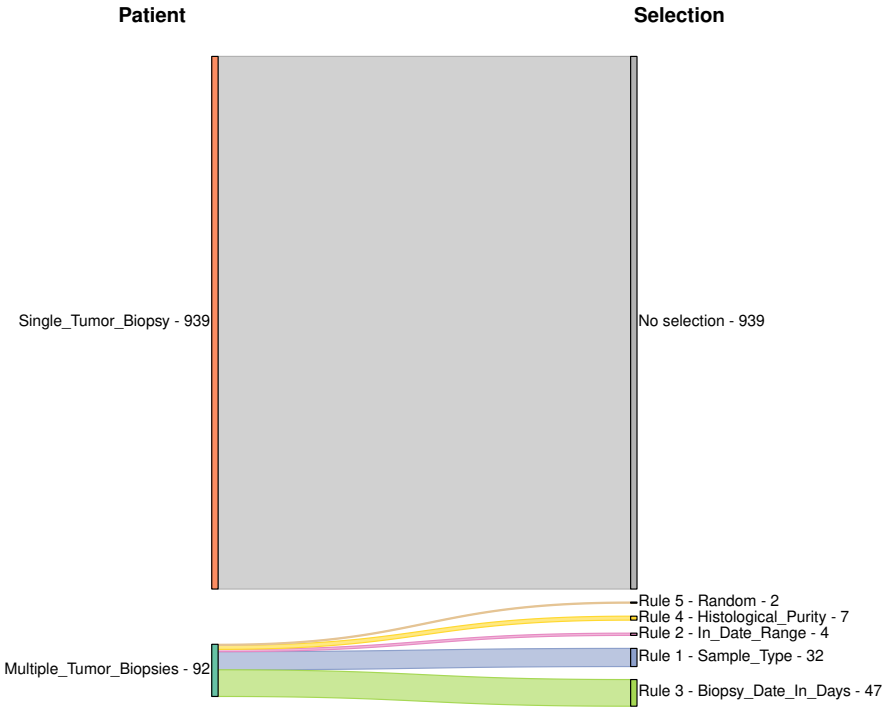


Fig. 3.3.: Application of rules to select biopsies in patients who have multiple biopsies available.

3.1.1.2. Cancer characteristics

The two defining characteristics, namely the primary site and histology, of the cancers from the 1,031 META-PRISM patients were also subjected to meticulous reviews throughout the project. As highlighted in Chapter 1 of this thesis, precise cancer classification holds immense significance for both clinicians making treatment decisions and researchers utilizing

it to categorize patients in their analyses.

Within the context of the META-PRISM study, the imperative for an accurate determination of primary tumor sites and histologies was exacerbated due to the planned comparison with TCGA, where patients are categorized into 33 specific tumor types. The initial data tables already provided a preliminary classification of each META-PRISM patient into TCGA-like classes, which was derived from the automated processing of French descriptions of the tumor. However, this automated process exhibited imperfections, and in rare instances, conflicting French descriptions of the patient's pathology coexisted. Regarding the biopsies, we embarked on a manual review of the primary site and histology of the tumor whenever we deemed it necessary to do so. This manual review took place over no less than 21 in-person meetings with participating oncologists throughout the study, spanning a period of more than two years (Figure 3.4).

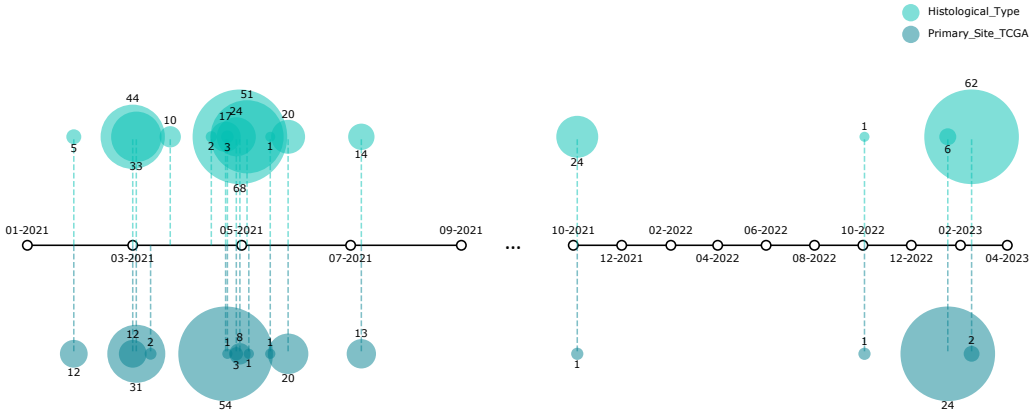


Fig. 3.4.: Timeline of all the manual reviews of patients and biopsies data performed during the course of the project.

In the study, a total of 548 patients out of the 1,031 participants underwent a thorough manual examination of at least one of the two class-defining characteristics of their tumor. During this manual review, all the preexisting and updated records of the primary site and histological subtypes were translated into the ICD-O-3 topographical and morphological tables, aligning with the reporting guidelines established by the NCI[2] and mirroring the reporting standards used by TCGA. All manual reviews were done in accordance with these guidelines.

The manual review resulted in updates to the primary site and histological information for 180 and 315 patients, respectively (Figure 3.5). A significant number of these reviews involved refining the site location, such as distinguishing between the bladder and upper urothelial tract in cases where the initial designation was bladder, or designating specifically the primary site in all patients with cancers of the oral cavity or upper digestive areas, as these types of cancer often have loosely described sites. Similar attention was given to histologies, particularly for the 122 patients initially labeled as having "carcinoma, not otherwise specified" who underwent manual review to provide precise descriptions. This effort was extended to

[2] https://seer.cancer.gov/icd-o-3/

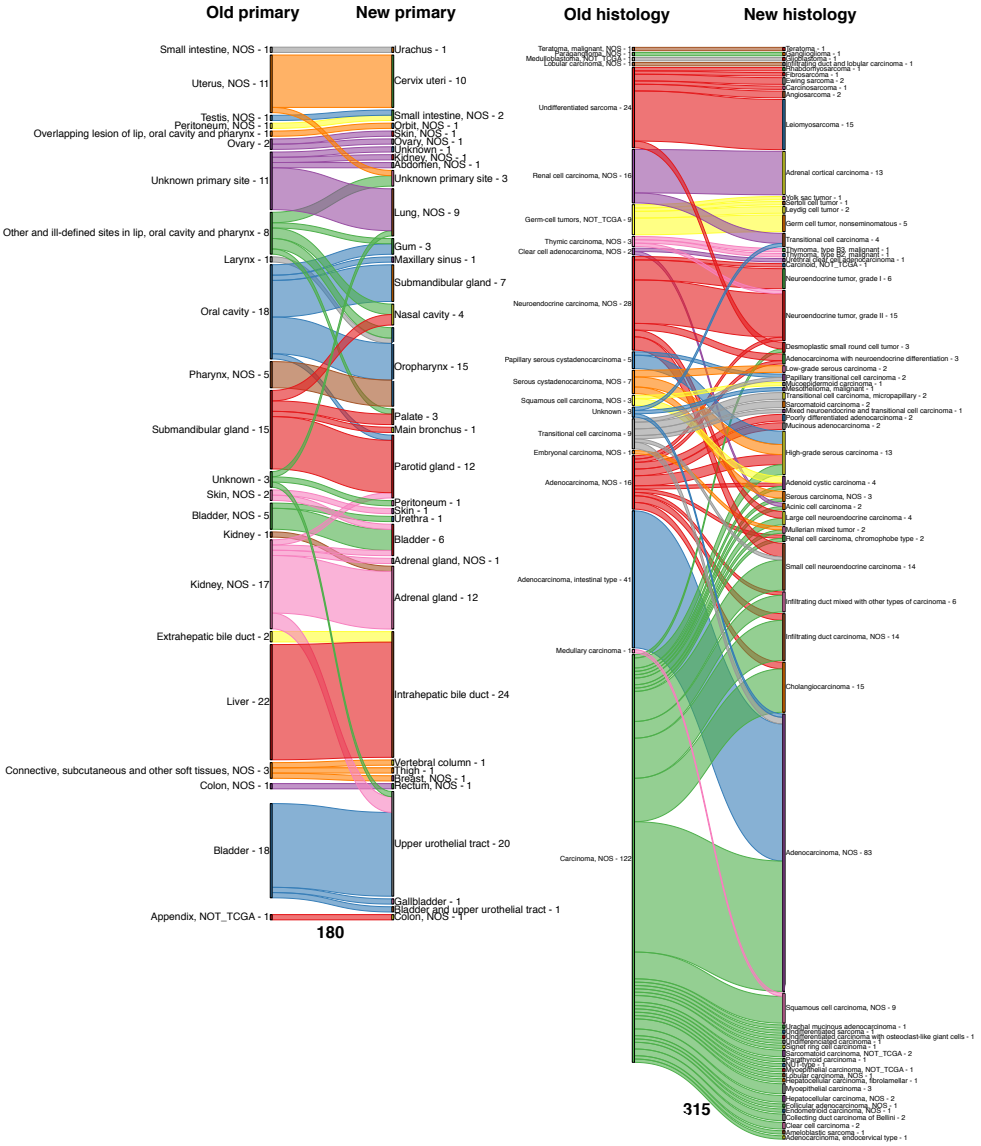


Fig. 3.5.: Changes in the tumors histologies and primary sites of META-PRISM patients upon manual review. The histologies and primary sites were updated in 180 patients and 315 patients, respectively.

select tumor types, particularly sarcomas, a notably diverse subset of cancers encompassing nearly 200 recognized histologies (Section 1.3.1.4), as well as NETs and NECs, where tumor grade information was included whenever available in the records.

In total, the cancer type assignments, which form the basis of many of the study analyses, were adjusted for 241 patients. Among these, 88 were determined to be unclassifiable within any of the 33 TCGA subtypes and were consequently excluded from all tumor-type-dependent analyses (Figure 3.6). The exclusion of these 88 patients, in addition to the previously

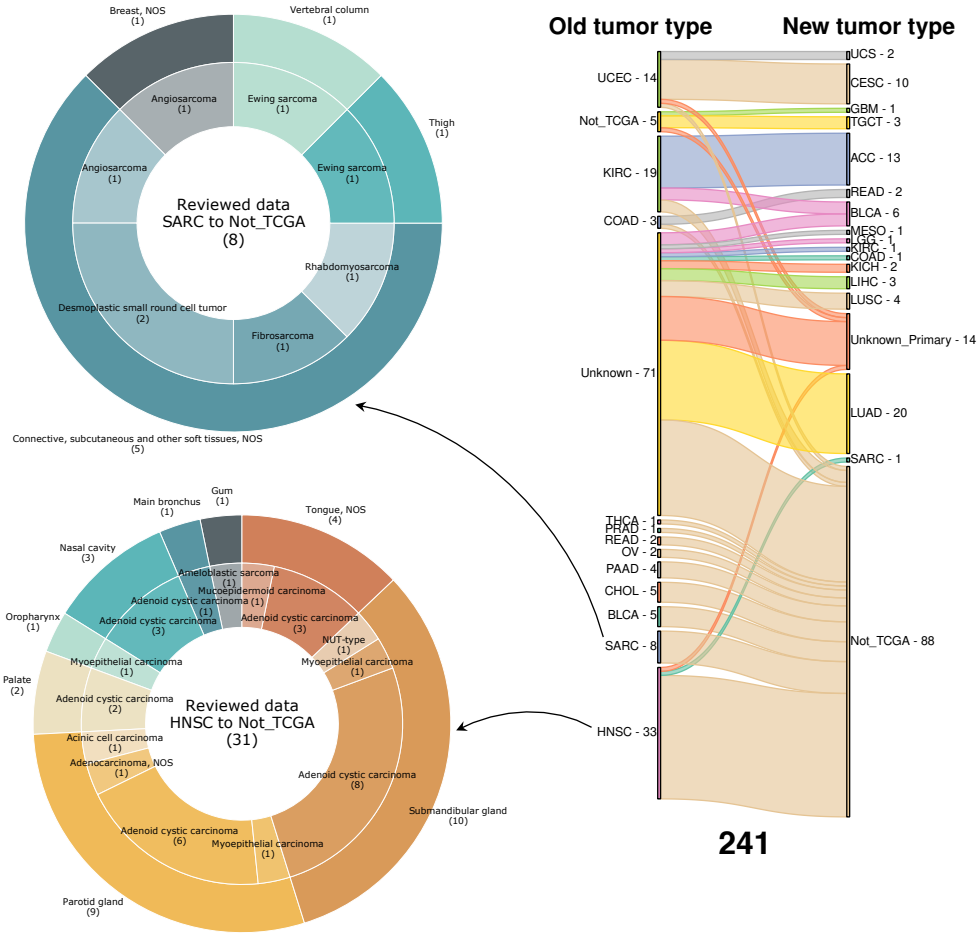


Fig. 3.6.: Changes in TCGA-like tumor type assignment in 241 META-PRISM patients upon manual review. The two left donut plots provide details about the changes in the tumor histology and primary site that motivated the changes in group assignment for mislabelled TCGA-like sarcomas - sarcoma (SARC) - and head and neck squamous cell carcinomas - HNSC.

identified 50 patients not falling within these classes, highlights the imperative nature of this review process and the importance of considering precise histologies to perform clinically meaningful group comparisons. Notable examples of cancer type assignment changes include all adenoid cystic carcinomas of the head and neck, initially classified into the HNSC TCGA study (Figure 3.6, bottom-left), which we considered as a separate entity, and all sarcoma subtypes not part of the six subtypes included in the TCGA SARC study[3] (Figure 3.6, top-left).

[3] https://www.cancer.gov/ccg/research/genome-sequencing/tcga/studied-cancers/sarcoma-study

3.1.1.3. Treatment history

Concurrently to the manual review of patients' and biopsies' core characteristics, we set out to retrieve the complete histories of the treatments received before the biopsies so as to correlate genomic alterations with treatment exposure and, more importantly, resistance. Once again, we extracted automatically lists of treatments using the Dr Warehouse database deployed at the institute. As the task of retrieving treatment histories was done concurrently with the revision of the patient list, we performed two automatic queries in Dr Warehouse, resulting in two tables of treatments referred to as "Trt_1" and "Trt_2" in Figure 3.7. As the "Trt_1" data table is a strict subset of "Trt_2", it will not be discussed much further.

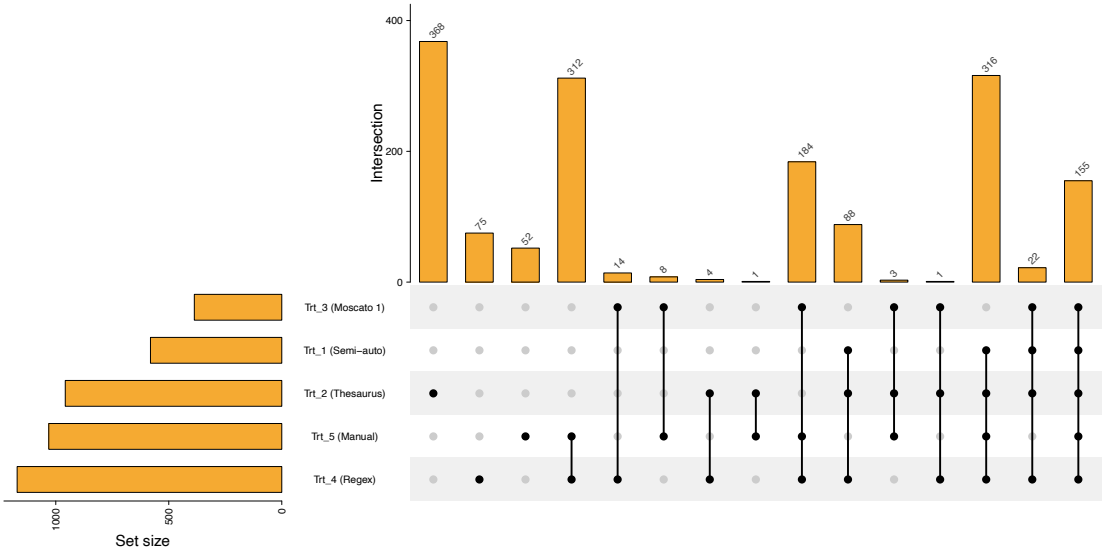


Fig. 3.7.: Patient overlap between the five different sources of treatment data.

The results of these automatic extractions included treatments administered intravenously. Depending on the extraction mode chosen, each line in the extraction tables reported either a single treatment name (Trt_2) or an excerpt of text that contained the treatment name(s), administration method, and dosage information (Trt_1). Although the "Trt_1" table was not utilized in the final analyses, I developed automated scripts to process this data, extracting detailed information, including the first and last administration dates, the number of administrations, and the total dose received for each treatment-patient combination. This information may prove valuable in subsequent analyses on the cohort.

While the resulting data was rich in information, it lacked crucial details about drug doses and treatment regimens, precluding us from distinguishing cases where a drug was used over a long period of time as a single line from cases where it was administered over short period times in different treatment lines. Moreover, the extractions made using Dr. Warehouse did not capture orally administered treatments or those administered outside Gustave Roussy. Consequently, we decided to supplement this initial treatment history data with data manually collected for the `MOSCATO` study (Trt_3). Unfortunately, no such data was available for the

MATCH-R study during the META-PRISM project. Additionally, the data collection task for the MOSCATO study aimed at listing only the last anticancer drug administered, and therefore did not align with our desire to obtain complete treatment histories.

To complement our database of treatment histories, we also attempted to apply a comprehensive and flexible regex to all electronic health records stored in Dr Warehouse (Trt_4). This regex was constructed by aggregating all encountered treatment names, including both commercial names and international common denominations (DCI). Additionally, we incorporated treatments from a table of antineoplastic drugs generously shared by a clinical collaborator. The resulting regex is shown in Annex A.3.1. As anticipated with this approach, when we compared the dates and the number of matches from the regex to the actual treatment courses for intravenous treatments from the "Trt_2" table, it became apparent that there were numerous false positives in the sense that the regex had identified the treatment on many more occasions and over much longer timeframes than the actual treatment duration and the number of administrations (as depicted in Figure 3.8).

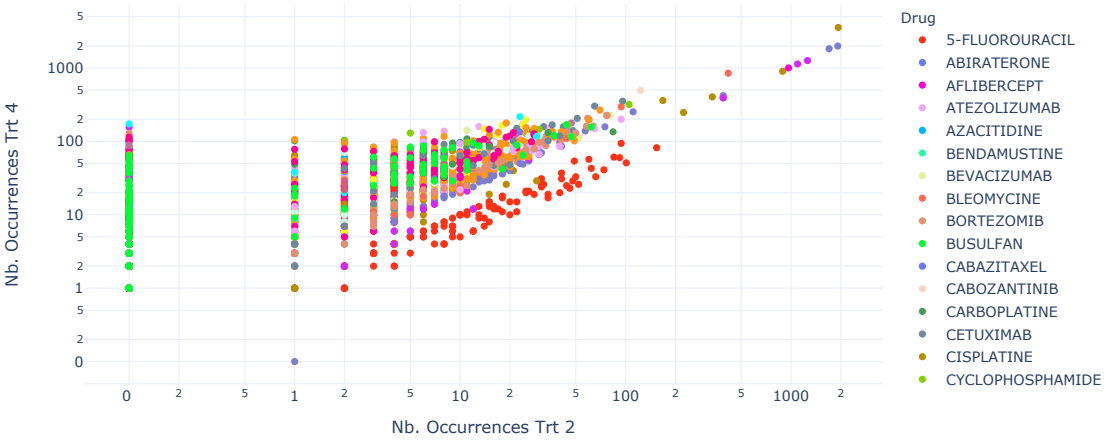


Fig. 3.8.: Comparison between the number of treatment administrations from Trt_2 table and the number of times the treatment name was matched via a regex as reported in Trt_4 table in the 586 patients and 70 drugs common between these two tables.

These false positives arose due to the numerous reasons for mentioning a treatment name in a report besides the actual treatment administration, such as summarizing the treatments received thus far or discussing potential drugs to be administered in the future. Furthermore, this approach is susceptible to false negatives because not all electronic documents can be subjected to regexes, especially in the case of scanned documents. To address this, an additional effort involving the extraction of text from images, which may require the use of one of the AI models specifically designed for this purpose, would be necessary.

Given the importance of having trustworthy treatment data for the patients, and because we wanted to focus on the treatments for which resistance was diagnosed before or shortly after the biopsy we analyzed, we requested help from three different medical oncologists to manually review the accumulated treatment data. To guide this review process, we provided the oncologists with pre-filled treatment tables, where treatment names were arranged in

columns, and patients in rows. For each cell within this pre-filled table, we populated it with the first and last dates of treatment extracted from the "Trt_2" and "Trt_3" tables. In cases where this information was not available from these tables, we included the first and last dates of regex matches as a guide for the manual review process. The specific task assigned to the oncologists was to add a binary indicator in cells where the patient had received the treatment and met the resistance criteria, and to leave the cells blank or erase pre-filled dates if the criteria weren't met (Trt_5).

The use of these pre-filled tables significantly facilitated the oncologists' work, enabling them to complete the manual review for all 1,031 patients in less than two months. It's important to note that *only the drugs listed during this manual review* were utilized in the correlative analyses against the molecular profiles of the tumors. In Table 3.1, we provide two sample treatment histories, including details of which treatments were reported from each of the five sources of treatment data aggregated throughout the project.

	Trt_2 (Thesaurus)		Trt_3 (Moscato 1)		Trt_4 (Regex)			Trt_5 (Manual)	
	Found	Dates	Found	Dates	Found	Dates	Occurrences	Found	Dates
MP-0002									
5-FLUOROURACIL	✗		✗		✗			✓	Before 25/11/2014
BEVACIZUMAB	✗		✗		✓	20/11/2013_to_02/12/2015	29	✓	Before 25/11/2014
CAPECITABINE	✗		✓	12/11/2014_to_03/12/2014	✓	03/11/2014_to_02/12/2015	23	✓	Before 25/11/2014
CYCLOPHOSPHAMIDE	✓	18/10/2011_to_24/01/2012	✗		✓	19/03/1996_to_05/11/2015	53	✓	Before 25/11/2014
DOCETAXEL	✓	18/10/2011_to_24/01/2012	✗		✓	01/09/2011_to_05/11/2015	48	✓	Before 25/11/2014
DOXORUBICINE	✗		✗		✓	19/03/1996_to_16/07/1996	5	✗	
EPIRUBICINE	✗		✗		✗			✓	Before 25/11/2014
LETROZOLE	✗		✗		✓	24/02/2012_to_05/11/2015	29	✓	Before 25/11/2014
NAB-PACLITAXEL	✗		✗		✓	26/11/2014_to_05/11/2015	6	✗	
PACLITAXEL	✗		✗		✓	20/11/2013_to_02/12/2015	50	✓	Before 25/11/2014
PEMBROLIZUMAB	✗		✓	19/12/2014_to_	✓	30/10/2014_to_02/12/2015	34	✗	
TAMOXIFENE	✗		✗		✓	13/05/1997_to_02/12/2015	39	✓	Before 25/11/2014
TRASTUZUMAB	✗		✗		✓	01/09/2011_to_23/09/2011	3	✗	
MP-0176									
AFATINIB	✗		✗		✓	23/01/2015_to_26/12/2018	27	✓	Before 07/12/2017
BEVACIZUMAB	✗		✗		✓	24/07/2013_to_26/12/2018	29	✓	Before 07/12/2017
BRIGATINIB	✗		✗		✓	14/02/2018_to_14/02/2018	1	✗	
CARBOPLATINE	✗		✗		✓	24/07/2013_to_26/12/2018	29	✓	Before 07/12/2017
CISPLATINE	✗		✗		✓	26/09/2012_to_26/12/2018	31	✓	Before 07/12/2017
DOCETAXEL	✗		✗		✓	24/07/2013_to_26/12/2018	29	✓	Before 07/12/2017
ERLOTINIB	✗		✗		✓	26/09/2012_to_26/12/2018	31	✓	Before 07/12/2017
EXEMESTANE	✗		✗		✓	24/07/2013_to_26/12/2018	30	✗	
GEMCITABINE	✗		✗		✓	26/09/2012_to_26/12/2018	31	✓	Before 07/12/2017
NAB-PACLITAXEL	✗		✗		✓	24/10/2012_to_30/05/2017	4	✗	
OSIMERTINIB	✗		✗		✓	23/04/2015_to_26/12/2018	35	✓	Before 07/12/2017
PACLITAXEL	✗		✗		✓	24/07/2013_to_26/12/2018	29	✓	Before 07/12/2017
PEMETREXED	✗		✗		✓	24/07/2013_to_26/12/2018	29	✓	Before 07/12/2017
TUSAMITAMAB RAVTANSINE	✗		✗		✓	13/06/2018_to_13/06/2018	1	✗	

Table 3.1.: Examples of treatment histories of two META-PRISM patients resulting from the parsing of four data sources.

3.1.1.4. Summary figure

Figure 3.9, which was used as the first figure of the published paper, summarizes the META-PRISM cohort and the main characteristics of the patients and biopsies.

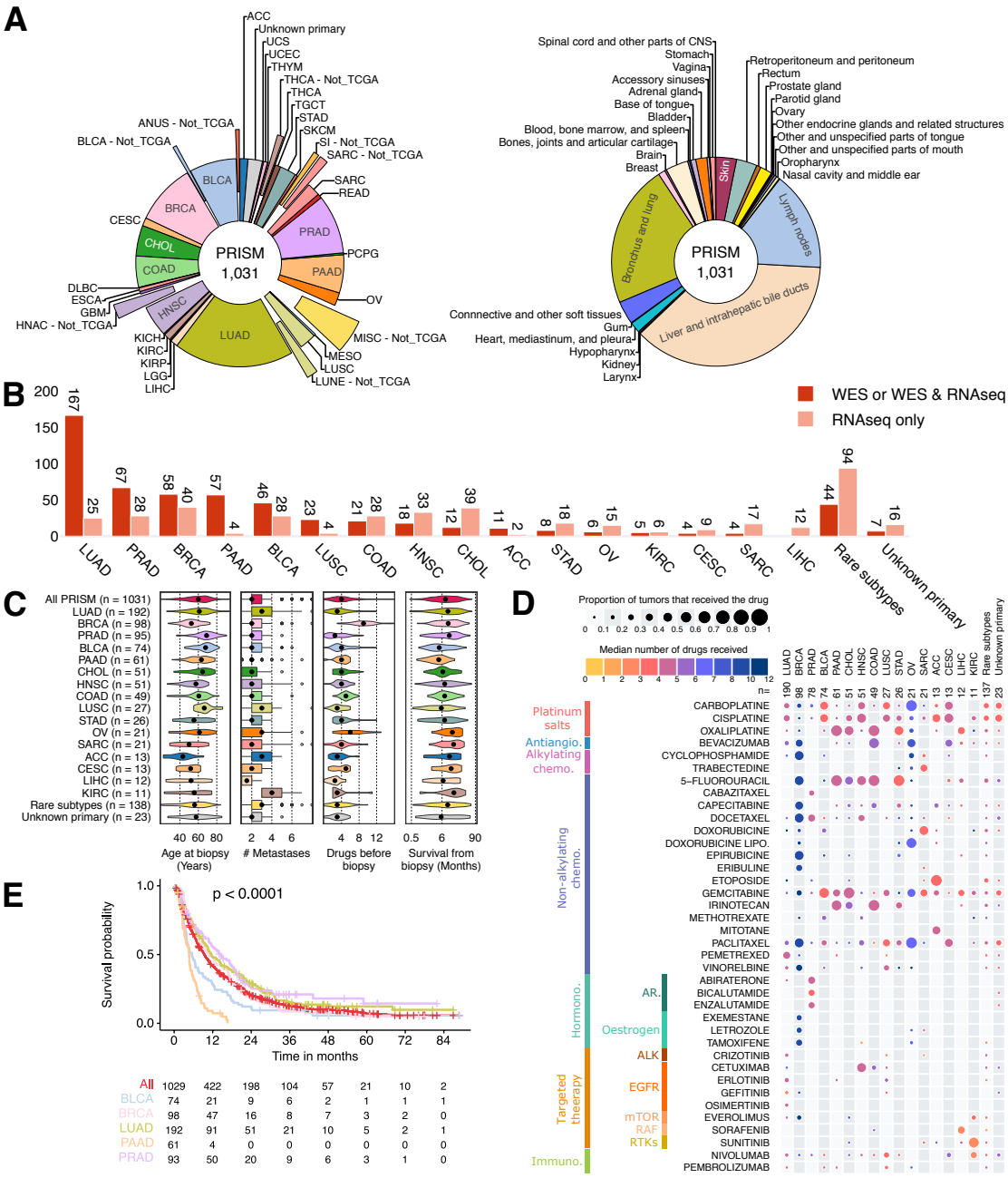


Fig. 3.9.: Clinical characteristics, sequencing data, treatment history, and outcomes of the META-PRISM cohort.

Figure 3.9.**A** presents two pie charts, one for the distribution of tumor types following TCGA classification (left), and one for the distribution of the biopsy sites according to ICD-O-3 nomenclature (right). Tumors that could not fit into TCGA classification were classified into tumor types suffixed by Not_TCGA and are shown as exploded slices. The cohort included 39 cancer types, with 5 of them represented by more than 60 tumors: 192 lung adenocarcinomas (LUADs), 98 BRCAs, 95 prostate adenocarcinomas (PRADs), 74 bladder urothelial carcinomas (BLCAs), and 61 pancreatic adenocarcinomas (PAADs). Moreover, 138 patients harbored tumors of rare subtypes and for 23 additional patients the primary site of the tumor was unknown. The delineation of the tumor type acronyms and the number of corresponding META-PRISM patients are provided in Table A.5. Figure 3.9.**B** shows the number tumors per tumor type with either RNA-seq only, WES only, or WES and RNA-seq. Tumors with WES only or WES and RNA-seq were grouped together. Only tumor types represented by at least 10 patients in META-PRISM are shown. The violin plots in Figure 3.9.**C** describe the age distribution, the number of detected metastatic sites at the time of the biopsy, the number of treatments received before the biopsy, and the survival time from the biopsy date for each tumor type. Only treatments for which resistance was diagnosed before or shortly after the biopsy are listed. Only the survival times of deceased patients are used in the violins (909/1,031). The heat map in Figure 3.9.**D** shows the types of treatment (rows) administered per tumor type (columns). The circle size encodes the percentage of patients who received treatment. The circle color encodes the median number of treatments received by these patients. Treatments are grouped by families as indicated on the left bars. Only the treatments received by at least 25 patients or by at least half of the patients of any of the displayed tumor type are shown. Lastly, Figure 3.9.**E** shows Kaplan-Meier survival curves of the whole META-PRISM cohort and the five most represented tumor types. The p-value was calculated using a log-rank test. As can be seen from this figure, the median survival time for the META-PRISM cohort stands at 7.8 months with PAADs and BLCAs having the most dismal prognosis.

3.1.1.5. Data organization

Throughout the project, all the data files and programming scripts were hosted on two remote servers, with one dedicated to data storage and the other to code hosting. This setup ensured easy access for the numerous project participants and enabled the automatic synchronization of code, data tables, and results in real time. In a study of this size, a multitude of data files were either generated internally or obtained from external sources to support our various analyses. To maintain order, all the raw and processed files were meticulously organized within a structured database hosted on a remote server, which was part of the cloud solution implemented at Gustave Roussy, known as *Nextcloud*. The use of a cloud-based solution proved to be highly convenient, as it allowed for the automatic synchronization among all users connected to the server. Furthermore, all the scripts created for tasks like data downloading, curation, processing, and analysis were synchronized with a dedicated remote server for code versioning. This server was a GitHub repository hosted

within the institute's space[4].

Raw files received from collaborators were consistently stored, and a record was appended to a summary table that tracked the date of data receipt and its source. A similar tracking system was applied to all external files acquired from sources such as published papers or data portals like the GDC data portal. All updates and changes made to the data were stored in separate files, which were then used by various curation scripts to generate the final tables of curated data that supported our analyses. This systematic organization played a crucial role in tracking the origin of each file and addressing any data errors encountered. It also allowed us to provide feedback to the institute's data management services when errors were identified, thereby contributing to the correction of data sources as needed. This setup was instrumental to the organization of the data, the analyses, and ensured the *reproducibility* and *integrity* of our research, a point that has gained a lot of attention from the scientific community and regulatory bodies in the recent years.

We utilized the cloud server not only for hosting various data files but also for storing the data generated by the bioinformatic pipelines applied to the raw sequencing data. The raw sequencing data itself is housed in sequencing archives physically located on the institute's premises and can be accessed through the high-performance computing cluster deployed at the institute. All bioinformatic pipelines were executed on this computing cluster, and the results of manageable file sizes, such as tables of gene expression, gene fusions, somatic mutations, copy-number segments, and more, were subsequently uploaded to the *Nextcloud* server. The detailed bioinformatic processing of the sequencing files for META-PRISM patients is further elaborated upon in the following section.

3.1.2. Bioinformatic analyses

The bioinformatic processing of the files originating from the sequencing of tumor and blood samples was split between WES and RNA-seq. Among the samples selected for the study, 569 blood samples were subjected to WES, 485 tumor samples to WES and RNA-seq, 84 tumor samples to WES only, and 462 tumor samples to RNA-seq only. Consequently, the WES pipeline was applied to process 569 pairs of tumor and matched blood samples, while the RNA-seq pipeline was used for 947 tumor samples.

3.1.2.1. WES pipeline

The WES pipeline was initially constructed by building upon preexisting code developed by a former member of Dr. Nikolaev's team. This code was then expanded to meet the specific requirements of the study and adhere to the best practices of the GATK[5]. The WES workflow can be divided into two main parts. The first part begins with raw, unaligned reads produced by the sequencing machines in the form of FASTQ files and concludes with filtered and aligned reads in the form of BAM files. The second part is a collection of independent

[4] https://github.com/gustaveroussy

[5] https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows

sub-workflows, each of which utilizes the BAM files to identify, filter, and, if applicable, annotate, either short germline variants (SNPs and indels), short somatic variants (SNVs and indels), somatic CNAs, or MSI.

While the study's first bioinformatician had already implemented the steps up to the point just before annotation with knowledge databases, the code had become overly complex and crucial filtering and annotation components were still in the process of being developed when he left. As a result, I took the initiative to assume responsibility for the WES pipeline. I comprehensively reorganized the code to align with best practices and incorporated the additional steps required for filtering and annotating somatic alterations (short indels and CNAs). Aside from the technical knowledge gained through this undertaking, it provided valuable insights into the various bioinformatic processing steps. In my view, a solid understanding of these steps is crucial for running meaningful downstream analyses and analyze them critically. The final structure of the code and the underlying rules are the result of extensive work and a series of rule updates, conducted in parallel to the harmonization of pipelines (Section 3.2.2) and somatic variant annotation (Section 3.1.2.3) work.

As explained in Chapter 2, the processing of raw sequencing data invariably begins with a series of quality control and read filtering procedures. These steps are essential to eliminate low-quality reads and reduce sources of artifacts, such as the presence of PCR-induced duplicate reads, which can skew estimations of VAFs and CNAs. Quality control of paired-end reads was conducted using `FastQC` v0.11.8, followed by the use of `Fastp` v0.20 (Chen *et al.* 2018) to trim adapters and polynucleotide tracts from reads exceeding 25 nucleotides in length. The resulting cleaned FASTQ files were subsequently aligned to the reference human genome GRCh37 using `BWA-MEM` v0.7.17 (Li & Durbin 2009). Intermediate BAM files underwent further processing, including read deduplication `MarkDuplicates` from `Picard` v2.20.3, coordinate sorting using `SAMtools` v1.9 (Li, Handsaker, *et al.* 2009), and base quality recalibration using `BaseRecalibrator` and `ApplyBQSR`. All these tools are included in the GATK bundle v4.1.8.1 and are considered part of the best practices (DePristo *et al.* 2011). Alignment quality was assessed using three different algorithms: `mosdepth` v0.2.5 (Pedersen & Quinlan 2018), `flagstat` from `SAMtools` v1.9 (Li, Handsaker, *et al.* 2009), and `CollectHsMetrics` from GATK v4.1.8.1.

Germline SNVs and indels were identified using `HaplotypeCaller` (Poplin *et al.* 2017). After the initial call, putative germline variants underwent a rigorous filtering process, including hard thresholds for various parameters: `QualByDepth` (QD $>$ 2), genotype quality (`QUAL` $>$ 30), `FisherStrand` (`FS` $<$ 60 for SNPs, $<$ 200 for indels), `ReadPosRankSumTest` (`ReadPosRankSum` $>$ -8 for SNPs, $>$ -20 for indels), `RMSMappingQuality` (`MQ` $>$ 40 for SNPs only), and `MappingQualityRankSumTest`($>$ -12.5 for SNPs only), all in accordance with the GATK best practices[6]. Variants that successfully passed all the filters were then annotated using `ANNOVAR` (K. Wang *et al.* 2010).

Germline variants associated with cancer predisposition were identified within a list of 130

[6] https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants

genes curated from the original list of 152 genes published by Huang and colleagues (Huang *et al.* 2018). Only variants with a maximum general population allele frequency of 5% in the gnomAD v2.1.1 exome database (Q. Wang *et al.* 2020), and annotated as "Pathogenic," "Likely_pathogenic," or "Pathogenic/Likely_pathogenic" in the ClinVar database (Landrum, Lee, *et al.* 2014; Landrum, Chitipiralla, *et al.* 2020), were retained. In total, 93 cancer-predisposing variants were identified in 73 out of 569 META-PRISM samples with available germline data, and these were included in the analyses.

For somatic joint mutations and small indels, `Mutect2` (Cibulskis *et al.* 2013) was employed. To mitigate artifacts and false positives, a *panel of normal* samples was created from the blood samples and integrated into the Mutect2 calling process, following the GATK best practices guidelines[7]. Putative variants then underwent an analysis for read orientation artifact and sample contamination, which was conducted by running GATK `LearnReadOrientationModel` and `CalculateContamination`, aligning once again with best practices.

A set of filters was applied to the somatic variants, including:

- Not being filtered by `Mutect2` (MUTECT_FILTERS).
- A minimum VAF of 5% (LOW_VAF).
- A minimum sequencing coverage of 20X in the tumor sample (LOW_COVERAGE_TUMOR).
- A minimum sequencing coverage of 10X in the normal sample (LOW_COVERAGE_NORMAL).
- Being located inside exonic regions, as defined by the canonical transcripts used by variant effect predictor (VEP) v104 on the GRCh37 assembly (NOT_EXONIC).
- An allele frequency of less than 0.04% across all gnomAD v2.1 exome subpopulations (COMMON_VARIANT). This rule was not applied for driver mutations (Section 3.1.2.3).
- Being within the META-PRISM target region, which is defined by the intersection of the capture regions of all four different kits used (OFF_TARGETS_INTERSECTION). This region spans approximately 36.6 Mb.

The impact of each filter, individually and in combination with others, is summarized in Figure 3.10. In total, 117,747 somatic substitutions and small indels were utilized for analysis. All mutations underwent annotation using VEP release 104 (McLaren *et al.* 2016) on canonical transcripts, with additional metrics and pathogenicity scores obtained from the dbNSFP v4.2 database integrated into the annotations through the plugin feature of VEP.

CNAs, tumor purity, and average tumor ploidy were assessed using the `FACETS` R package v0.5.14 (Shen & Seshan 2016) with parameters cval_pre=25 and cval_pro=500. To minimize the impact of segmentation errors, only gene CNAs stemming from segments spanning less than 10 megabases were taken into account for subsequent analyses. Automated scripts were executed to detect issues including low tumor purity, hypersegmentation, exceedingly large

[7] https://gatk.broadinstitute.org/hc/en-us/articles/360035890631-Panel-of-Normals-PON-

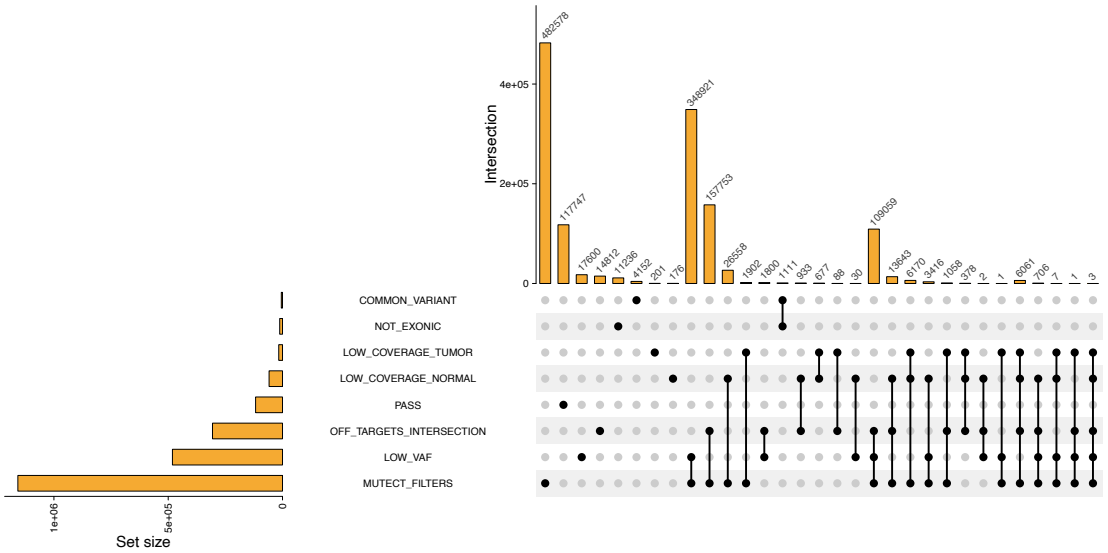


Fig. 3.10.: Upset plot showing the number of mutations filtered out individually by each filtering criteria and in combination with other criteria in META-PRISM WES samples. Mutations that passed all filters are described in the PASS set.

deletions, or incorrect positioning of the tumor diploidy level in the FACETS-generated profiles. Flagged profiles were manually reviewed and were either excluded or, if feasible, corrected.

Following this, each CNA was categorized into one of six categories (Table A.6). Only high-level focal amplifications or homozygous focal deletions were retained in the list of alterations used in the various analyses of the study. The presence and number of WGDs were determined based on the lowest positive integer 'k' that satisfied the condition that at least 11 autosomes had undergone 'k' duplications. This occurred when the major allele ploidy was strictly greater than $1.5 \times 2^{k-1}$ on at least half of the chromosome length. If this condition could not be met with $k = 1$, it was assumed that no WGD had occurred.

MSI analysis was conducted using MANTIS v1.0.3, following the same procedure as described in the original study (Kautto *et al.* 2017). To run MANTIS, a list of microsatellite loci needs to be compiled in a 6-column BED file, including genomic coordinates, the motif, and its count on the reference genome. In this analysis, the same BED file as used in the TCGA study (Cortes-Ciriano *et al.* 2017) was employed, comprising 2,530 loci. The method examines repetitive regions in aligned reads from both tumor and normal BAM files, one locus at a time. Per-locus read counts for each repeat motif are calculated in both the tumor and normal samples and are used to compute an instability score. Subsequently, the average of all locus instability scores is calculated to generate a final score for the tumor/normal pair. The reported scores range from 0.0, indicating complete stability, to 2.0, indicating complete instability. Samples were classified as MSI-high if they had a final score exceeding the default threshold of 0.4.

A summary graph of the WES pipeline workflow is provided in Figure 3.11.

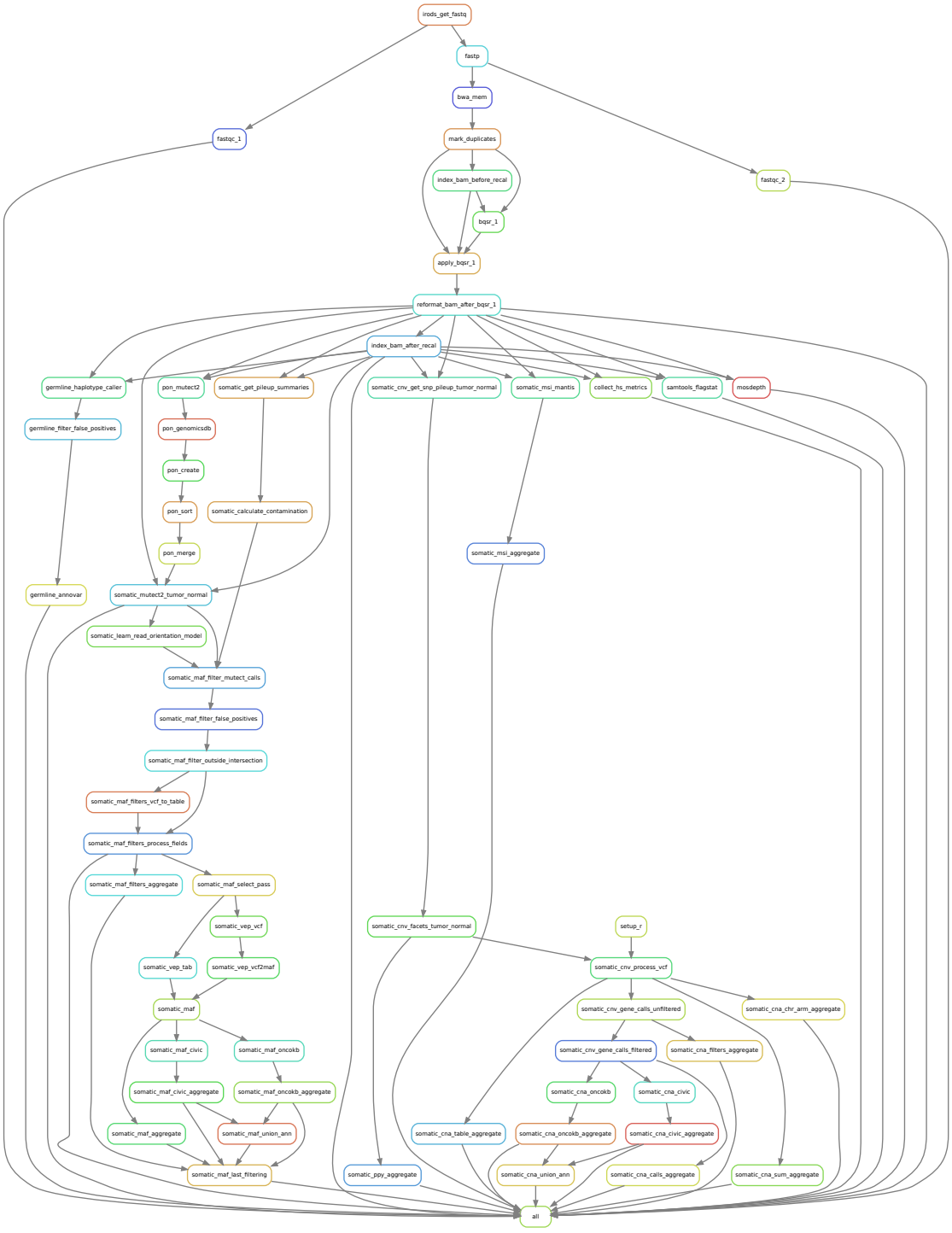


Fig. 3.11.: Rule graph of the WES pipeline implemented using Snakemake.

3.1.2.2. RNAseq pipelines

The processing of RNA-seq files largely relied on pre-existing pipelines developed by third parties. Two pipelines were utilized: one for quantifying gene expression and another for identifying, filtering, and annotating gene fusions.

In a manner similar to the WES pipeline, the processing of raw sequencing reads from RNA-seq experiments also begins with quality control and read filtering procedures. In contrast to WES experiments, PCR duplicates are typically retained in RNA-seq data. However, similar to WES, control of PCR duplication is assessed through metrics and reports generated by tools like `FastQC` or similar tools. In this study, quality control for paired-end reads was systematically conducted using `FastQC`, and adapter sequences were removed using `Trim Galore` v0.4.4.

The gene expression quantification in our study employed a pseudoalignment method, which offers significantly faster performance compared to alignment-based quantification methods, as detailed in Section 2.1.4.3. We chose to quantify gene expression in our META-PRISM samples by adapting the Nextflow pipeline available at https://github.com/gevaertlab/RNASeq_pipeline. The main reason for this choice was the public availability of the quantification tables produced by this pipeline for TCGA samples, which were released as supplementary data to the work by Zheng *et al.* (2019) presenting extensive comparative analyses of various quantification methods. Specifically, the sequencing reads were pseudoaligned to the human transcriptome from GENCODE version 27 (58,288 genes) using `Kallisto` v0.44.0 (Bray *et al.* 2016). Subsequently, transcript-level estimates were aggregated to the gene level using `TxImport` v1.16.0 (Soneson *et al.* 2016), and the resulting gene quantifications, in raw counts and TPM formats, were recorded.

Putative gene fusions were called by six different calling algorithms using the Nextflow nf-core/rnafusion pipeline v1.2.0[8]. This pipeline was developed by nf-core community, which is dedicated to creating a "curated set of analysis pipelines built using Nextflow" (Ewels *et al.* 2020). Calls generated by two of these algorithms were excluded from consideration. One of these exclusions was due to the tool's high failure rate on our RNA-seq samples (`FusionCatcher`, Nicorici *et al.* (2014)), while the other was due to the observation that none of the putative fusions reported by this algorithm were corroborated by the other tools (`SQUID`, Ma *et al.* (2018)).

For the analyses presented in the paper, gene fusions were identified on the GRCh38 reference genome by applying a filtering process and consolidating the fusions predicted by four callers, namely `Arriba` v1.2.0 (Uhrig *et al.* 2021), `EricScript` v0.5.5 (Benelli *et al.* 2012), `Pizzly` v0.37.3 (Melsted *et al.* 2017) , and `STAR-Fusion` v1.8.1 (Haas *et al.* 2019). In total, 707,027 fusions were detected by at least one of these four fusion-calling algorithms (Figure 3.12).

[8] https://github.com/nf-core/rnafusion/tree/1.2.0

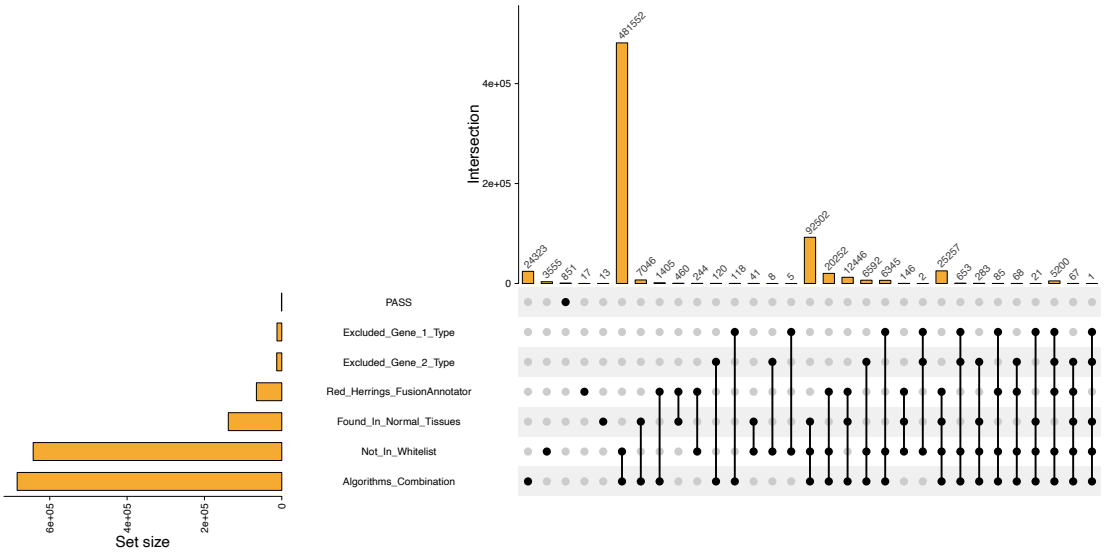


Fig. 3.12.: Upset plot showing the number of gene fusions filtered out individually by each filtering criteria and in combination with other criteria in META-PRISM RNA-seq samples. Gene fusions that passed all filters are described in the PASS set.

Each individual caller's set of putative gene fusions was refined by excluding fusions previously reported in studies of normal tissues (blacklists) and retaining only those fusions reported in studies of cancer tissues or those involving cancer driver genes (whitelists). Additionally, only gene fusions identified by both `Arriba` and `EricScript` or both `Pizzly` and `STAR-Fusion`, irrespective of the predicted breakpoints, were retained (a detailed rationale for this specific combination is provided in Section 3.2.2). Following these filtering criteria, a total of 851 gene fusions (without considering breakpoints) were identified in 445 out of 944 samples with RNA-seq data. It should be noted that three out of 947 RNA-seq samples repeatedly failed in one or more of the fusion-calling algorithms and were consequently excluded from analyses utilizing gene fusions.

3.1.2.3. Catalog of oncogenic events

We compiled a list of 360 cancer driver genes by intersecting the list of driver genes (Tiers 1 and 2) from COSMIC census v92 with the list of genes annotated in the OncoKB database as of July 2021 (Chakravarty *et al.* 2017). Oncogenic events were identified by intersecting somatic substitutions, indels, focal high-level gains and homozygous deletions from segments spanning less than 10 Mb, and gene fusions with the OncoKB and CIViC (Griffith *et al.* 2017) databases. OncoKB constitutes a literature- and knowledge-based database that is accessible through an application programming interface (API) after having registered an account and requested a token. In accordance with the type of event to be annotated, different scripts from `oncokb-annotator`[9] were used:

[9] https://github.com/oncokb/oncokb-annotator

1. `MafAnnotator.py` for substitutions and indels
2. `CnaAnnotator.py` for CNAs
3. `FusionAnnotator.py` for gene fusions

CIViC is also a literature and knowledge-based database, but no annotation script was available at the time of our analyses. As a consequence, I developed in-house scripts[10](Section 4.2.1.2) to annotate substitutions, indels, gene fusions, and CNAs using the table of clinical evidence summaries 01-Jan-2022-ClinicalEvidenceSummaries.xlsx from the January 2022 release. A minor number of errors were manually curated from the table. Additionally, missing genomic coordinates for mutations were manually filled where possible.

Importantly, the majority of OncoKB and CIViC annotations are tumor-type specific. Consequently, annotation with these databases require a description of the tumor type alongside each variant to allow for precise *on-label* annotations. OncoKB uses the MSK's oncotree nomenclature, whereas CIViC uses designations that do not follow specific rules and have a varying degree of specificity. As a consequence, we performed a *thorough work of tumor-type matching* with the help of oncologists to navigate between TCGA types, MSK's oncotree, and CIViC designations.

Depending on the type event, all annotated variants were retained or additional filtering was applied. CNA and gene fusions annotated in OncoKB *or* CIViC databases were all retained. However, for mutations we applied additional filtering. Firstly, only mutations annotated by OncoKB were retained due to the fact that the manual inspection of CIViC-only events revealed many false matches caused by unspecific variant descriptions in CIViC. Secondly, among the mutations that were annotated by `oncokb-annotator`, events with `MUTATION_EFFECT` as `likely neutral`, `neutral`, or `unknown` were discarded unless the `ONCOGENIC` field reported `likely oncogenic`, `predicted oncogenic`. Lastly, only mutations that were classified as either `Missense_Mutation`, `Frame_Shift_Del`, `Frame_Shift_Ins`, `In_Frame_Del`, `In_Frame_Ins`, `Nonsense_Mutation`, `Splice_Site`, or `Translation_Start_Site` were retained.

3.2. Comparison and validation cohorts

3.2.1. TCGA and MET500 cohorts

The molecular profiles of the refractory advanced tumors sampled in META-PRISM patients were compared against the profiles of treatment-naive non-metastatic tumors from TCGA (The Cancer Genome Atlas Research Network *et al.* 2013). The tumor type-specific and -agnostic comparisons were systematically repeated on the MET500 cohort (Robinson *et al.* 2017) whenever sufficient numbers of patients were available so as to validate candidate differences.

[10] https://github.com/ypradat/CivicAnnotator

3.2.1.1. TCGA

The TCGA consortium and its associated data have already been described in Section 1.1.2. It is important to note that for the purpose of this comparative study, we refined the patient selection to exclude metastatic patients identified by a stage IV classification in the TNM classification system, as well as patients documented with an "unacceptable history of prior treatment" as stipulated in the accompanying notes retrieved from the GDC data portal. This information was obtained using the R package `GenomicDataCommons` alongside other clinical and biological metadata. Additional data tables were acquired from supplementary tables included in publications released by the consortium or from the publicly accessible PanCanAtlas page[11].

In brief, we retrieved metadata from 11,315 patient samples and 34,815 aliquots of interest for our study, encompassing 23,346 WES and 11,469 RNA-seq data. Considering the available data and the most recent patient annotations, a total of 875 patients were excluded. The reasons and numbers of patients excluded for each reason are detailed in Figure 3.14. A total of 10,440 patients for whom we had one or multiple types of molecular data available and had no reason for exclusion were considered for analysis (Figure 3.13). Though counterintuitive, we reclassified some TCGA patients in "Not_TCGA" types based on the review notes of the AWG pathology review file merged_quality_annotations.tsv[12]. As detailed hereafter, data from each modality (somatic mutations, germline mutations, MSI, somatic CNAs, gene expression, and gene fusions) were limited to these 10,440 patients and further reduced in each modality to exclude samples that did not pass exclusion criteria or quality controls.

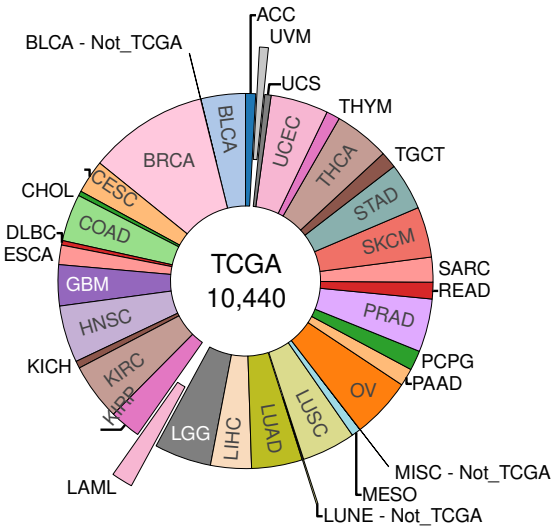


Fig. 3.13.: Cancer types in TCGA.

Somatic mutations The TCGA somatic mutation catalog was taken from the controlled-access MAF file mc3.v0.2.8.CONTROLLED.maf.gz, downloaded with permission. All filters described in the original "FILTER" column were applied as well as the following additional filters:

- Minimum VAF of 5% (LOW_VAF).
- Minimum sequencing coverage of 20X in the tumor sample (LOW_COVERAGE_TUMOR).
- Minimum sequencing coverage of 10X in the normal sample (LOW_COVERAGE_NORMAL).
- Localized within the META-PRISM target region (OFF_TARGETS_INTERSECTION).

[11] https://gdc.cancer.gov/about-data/publications/pancanatlas
[12] https://gdc.cancer.gov/about-data/publications/pancanatlas

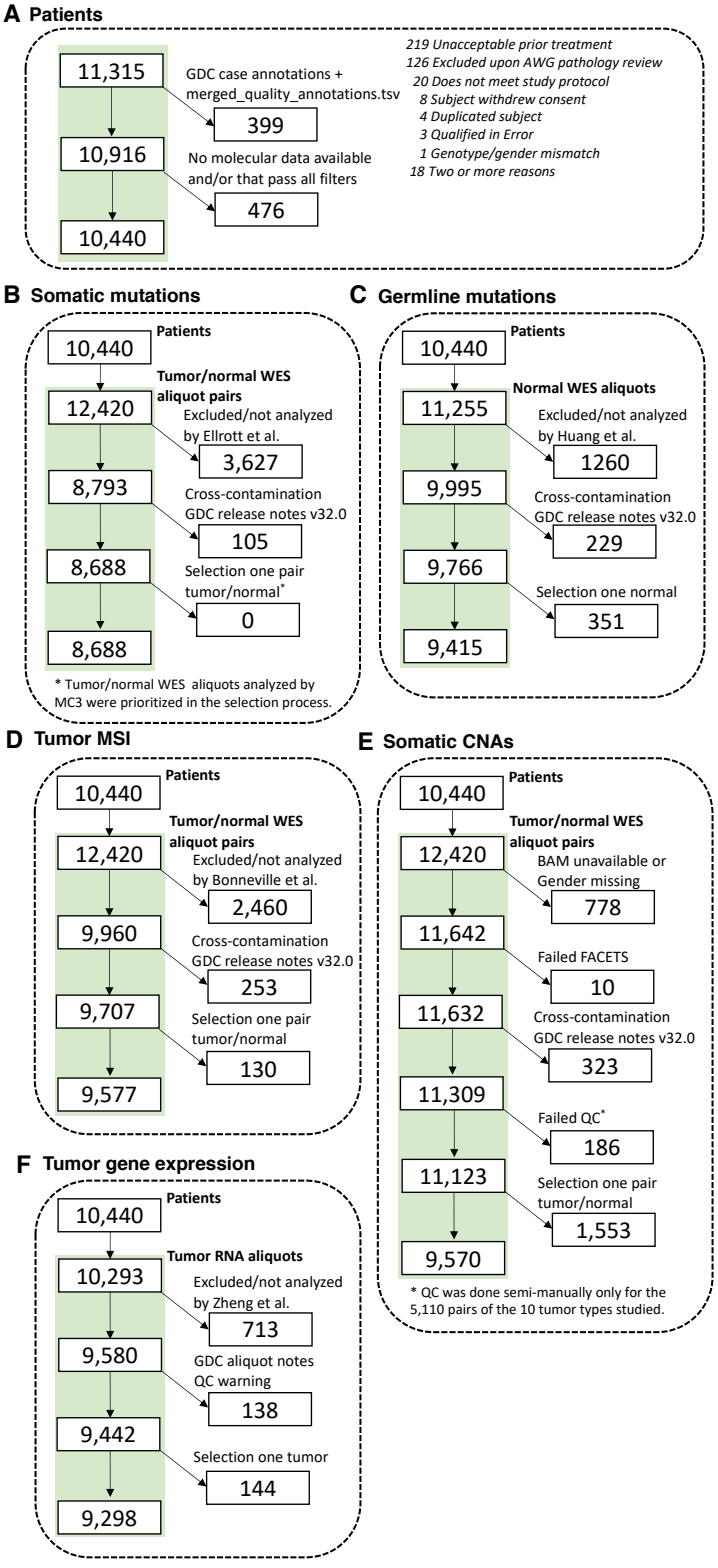


Fig. 3.14.: TCGA patients and samples selection for each type of data considered.

- Located inside exonic regions as defined by the set of canonical transcripts used by VEP v104 on GRCh37 assembly (NOT_EXONIC).
- Allele frequency across all gnomAD v2.1 exome subpopulations is <0.04% (COMMON_VARIANT). This rule is not applied for driver mutations (Section 3.1.2.3).
- Only SNVs and multi-nucleotide variants identified by at least two of the five callers were retained. Likewise, only indels identified by `Indelocator` or `Varscan` (INDELOCATOR or VARSCANI tags in the "CENTERS" column from the controlled-access MAF file) were retained (INDEL/SNP_CALLING_ALIGNMENT). The rationale for these selection criteria is expanded upon in Section 3.2.2.

A total of 2,109,671 somatic mutations and small indels were used for analysis in the 8,688 patients for whom somatic mutation data were available and no reason for exclusion existed. The filtering that resulted in this list of somatic calls is summarized in Figure A.1. All PASS mutations from the refiltered mutations file were split into individual VCF files for each tumor/normal pair and were subsequently annotated using the same annotation pipeline as used on META-PRISM data.

Germline mutations The list of cancer-risk germline variants detected in TCGA samples was taken from the file PCA_pathVar_ integrated_filtered_adjusted.tsv[13] (Huang *et al.* 2018). Reducing the original list of 1,393 mutations to only mutations in the 10,440 patients considered for analysis and excluding samples flagged as contaminated in GDC release notes v32.0 (Figure 3.14) results in 1,342 pathogenic germline variants detected in 1,253 patients. After running the germline mutation filtering procedure described in Section 3.1.2.1, 1,082/1,253 variants were retained and used for analysis.

Microsatellite instability MSI scores were retrieved from supplementary data of Bonneville *et al.* (2017). Samples flagged as contaminated in GDC release notes v32.0 were excluded, resulting in an available MSI score for 9,298 patients among the 10,440 TCGA patients we considered (Figure 3.14).

Somatic CNAs Somatic CNAs, WGD status, chromosome arm somatic CNAs, sample purity, and ploidy, were derived by applying our somatic CNA pipeline to TCGA sequencing files. The pipeline was executed on all accessible WES files from TCGA patients with known gender (gender needs to pre-specified to the pipeline), which comprised 21,987 BAM files originating from 10,332 patients having at least one pair of tumor/normal WES files (totaling 12,129 pairs). To ensure the reliability of the results, we excluded samples that were flagged as contaminated in GDC release notes version 32.0, as well as those failing the `FACETS` analysis or not meeting the quality control criteria for `FACETS` profiles. As a result of this rigorous filtering process, somaticCNAs were available for analysis in a cohort of 9,570 patients (Figure 3.14). This reanalysis was deemed essential to avoid substantial batch effects when

[13] https://gdc.cancer.gov/about-data/publications/PanCanAtlas-Germline-AWG

comparing data derived from distinct methodologies for assessing copy number variations, in particular WES-based versus microarray-based CNA-calling as done initially in the project. The execution of the pipeline was carried out via the GCE with technical support from the Institute for Systems Biology Cancer Genomics Cloud and financial support from the NCI. Panel 3.1 presents in more details how I achieved this computationally intensive task in a short time and cost-effectively.

Gene expression Gene-level and transcript-level expression tables for all TCGA RNA-seq samples were retrieved from the supplementary data of Zheng *et al.* (2019). This pipeline is identical to the pipeline that we have used on META-PRISM samples, thereby minimizing technical differences between the cohorts. Samples flagged with quality control warnings in GDC aliquot-level notes were excluded, resulting in available RNA-seq profiles for 9,298 patients among the 10,440 TCGA patients we considered (Figure 3.14).

Gene Fusions Three independent and publicly available lists of TCGA gene fusions were retrieved from the following sources:

`PRADA` X. Hu *et al.* (2018). Supplementary Table nar-02671-data-e-2017-File007.xlsx
`StarFusion` Gao *et al.* (2018). Supplementary Table S1
`DEEPEST` Dehghannasiri *et al.* (2019) Supplementary Table pnas.1900391115.sd01.xlsx

Different combination of these lists were assessed against different combinations of calls from the four callers used on META-PRISM samples after applying the different filtering steps detailed in Section 3.1.2.2. The best agreement was obtained when selecting gene fusions reported by `StarFusion` or by both `DEEPEST` and `PRADA` on the side of TCGA. The overlap between the three external lists of gene fusions detected in TCGA samples, prior to any filtering, is shown in Figure A.2. Once again, the rationale for this combination is elucidated in Section 3.2.2.

Panel 3.1: Efficient processing of TCGA data on cloud services

To analyze somatic CNAs within the TCGA samples, I leveraged the availability of TCGA BAM files hosted on GDC-controlled Google buckets. This enabled me to process the data rapidly and cost-effectively by utilizing the Google Cloud Engine (GCE). While the use of the GCE involves expenses, these costs can be minimized by making the most of GCE options and optimizing the efficiency of the pipeline.

Besides cost-efficiency, employing the GCE is also highly efficient in terms of data handling and execution time in comparison to the conventional approach of downloading and processing TCGA BAM files locally on a high-performance cluster. The conventional method is notably slow due to the limited transfer bandwidth relative to the extensive size of the more than 20,000 WES BAM files produced by TCGA, amounting to over 200 terabytes. Furthermore, local storage and computational resources are often inadequate for such an undertaking.

The GCE, in contrast, allows for quick data transfers between buckets and between buckets and virtual machines (VMs). The resources available on the GCE are substantial, enabling a single user to execute a multitude of tasks in parallel. To enhance the cost-efficiency of the pipeline on the GCE, I activated the preemption option, which offered discounts of up to 91% on running costs[a]. However, it's important to note that activating the preemption option means that VMs may be terminated by the GCE at any time. In order to handle this, I developed a set of Bash scripts to automatize the deployment of VMs, execute the pipeline in batches across thousands of VMs, and continuously monitor their status. This allowed for prompt reactivation or recreation of VMs in response to preemptions or failures. To maintain precise control over the system's behavior, I elaborated a sophisticated logging system to address the various failure scenarios encountered in some batches. When necessary, these batches were restarted with increased per-job resources to ensure error-free processing. I additionally optimized the CNA-calling sub-workflow from the WES pipeline presented in Section 3.1.2.1 and Figure 3.11. This optimization involved designing code that could withstand unforeseen disruptions, something I achieved by using Snakemake and saving intermediate results and log files on Google buckets to avoid the unnecessary execution of rules already ran before the disruption.

The entirety of the code necessary for coordinating the creation of multiple VMs, overseeing their parallel execution, and storing results and logs on Google buckets is accessible at the GitHub repository https://github.com/ypradat/TCGA_Facets. This code can be executed from a local machine, provided a stable internet connection is maintained. Remarkably, the processing of over 20,000 WES BAM files generated by TCGA was successfully completed in *less than three days*, incurring a *cost of approximately $700*. This cost represents roughly half of the $1,500 credits allocated to me by the NCI for this project.

[a] https://cloud.google.com/compute/docs/instances/preemptible

3.2.1.2. MET500

The cohort presented by Robinson *et al.* (2017) served as a validation cohort for our project, involving 500 metastatic patients and their associated samples.

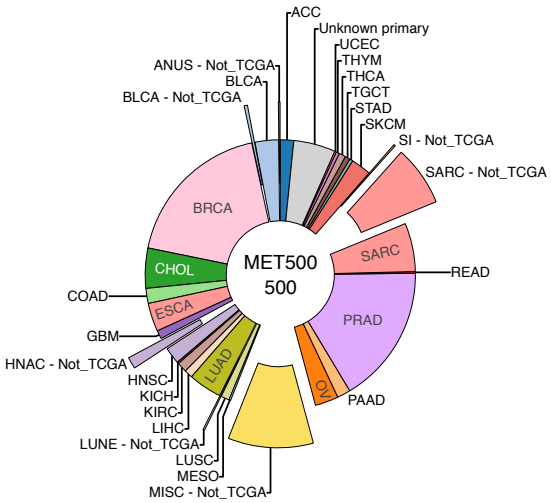


Fig. 3.15.: Cancer types in MET500.

Metadata were sourced with permission from dbGaP under the study identifier phs000673.v4.p1. Supplementary Tables S1 and S2 from the publication were also utilized. All relevant variables were subjected to processing and merged into curated tables, serving as the data source for the analyses. Particular attention was paid to the variables defining the classification of patients into distinct cancer types. For consistency with META-PRISM and TCGA data, descriptors such as primary site, biopsy site, and tumor histology were harmonized and standardized according to the ICD-O-3 nomenclature. These standardized values were then employed to categorize patients into classes mirroring the methodology applied in the META-PRISM cohort. The resulting distribution of cancer types in the MET500 cohort is depicted in Figure 3.15.

Somatic mutations Raw sequencing files were downloaded with permission and processed with our internal pipeline, as described in Section 3.1.2.1. Only samples from patients included in the publication by Robinson *et al.* (2017) were considered. A total of 106,341 somatic mutations and small indels were used for analysis after applying the filtering procedure as done for META-PRISM samples. The filtering that resulted in this list of somatic calls is summarized in Figure A.3. Of note, the OFF_TARGETS_INTERSECTION filter does not appear in the figure as, for these samples, the BED file containing the positions in the intersection of all capture kits used on META-PRISM samples was provided as input to the intervals parameter of `Mutect2`.

Germline mutations Raw sequencing files were processed with our internal pipeline, and germline mutations were called using `HaplotypeCaller` as done for META-PRISM samples. A total of 71 germline cancer-predisposing variants were detected in the 500 MET500 samples after applying the filtering procedures.

Microsatellite instability Raw sequencing files were processed with our internal pipeline and MSI was called using `MANTIS` as described previously.

Somatic CNAs Raw sequencing files were processed with our internal pipeline, and somatic CNAs, WGD status, chromosome arm somatic CNAs, purity, and ploidy were identified using

our internal CNA pipeline based on `FACETS` as done for META-PRISM and TCGA samples.

Gene expression RNA-seq data files were downloaded for 497 samples. The majority of samples had multiple files available produced from either polyA+ selection, hybridization capture, or both. As all RNA-seq files were produced from polyA+ in META-PRISM, we prioritized polyA+ whenever possible, resulting in 386 polyA+ and 111 hybridization capture RNA-seq libraries. Gene expression was then quantified using the `Kallisto/TxImport` pipeline as used on META-PRISM samples.

Gene fusions Raw sequencing files were processed with our internal pipeline, and gene fusions were identified using our internal gene fusion-calling pipeline as done for META-PRISM samples. A total of 731 gene fusions (disregarding breakpoints) were retained in 308 out of 497 samples with RNA-seq data as depicted in Figure A.4.

3.2.2. Pipelines harmonization

In any comparative study involving data originating from diverse institutions and processed using different pipelines, it is crucial to *carefully consider the potential impact of technical disparities* on the comparisons. In the META-PRISM study, we dedicated substantial efforts to mitigating the sources of these technical differences. Given that we lack control over technical variations that may arise prior to sequencing, our primary focus was on standardizing the bioinformatic pipelines. However, the systematic reprocessing of raw sequencing files is a laborious and costly solution (although cost-effective solutions exist, see Panel 3.1). Additionally, such reprocessing may be unnecessary in some instances, especially when teams from other institutes have already conducted high-quality bioinformatic analyses.

As described in the previous section, we opted for the utilization of identical pipelines for analyzing somatic CNAs and quantifying gene expression across all three cohorts: META-PRISM, MET500, and TCGA. This uniform bioinformatic processing approach extends to other data types in the META-PRISM and MET500 cohorts but not in the case of TCGA. The somatic mutation catalog for TCGA was derived from the data released by the TCGA-led MC3 project (Ellrott *et al.* 2018), while germline mutations were sourced from the Pan-Cancer Atlas paper by Huang *et al.* (2018), and gene fusions were obtained from three distinct sources as described in Section 3.2.1.1. For somatic mutations and gene fusions, we acquired raw sequencing data from a small subset of TCGA samples to fine-tune our filtering criteria and achieve optimal alignment between the data processed by our pipelines and the data released by third-party entities.

The alignment of somatic mutation calling between META-PRISM and TCGA was performed by running our internal pipeline on a test set of 58 TCGA raw WES files downloaded with permission from the GDC data portal. Various levels of stringency were explored, considering common metrics such as sequencing coverage depth at mutation sites and VAF. Moreover, given that the MC3 project incorporated five different callers for single-

or multi-base substitutions and five other callers for indels, we treated substitutions and indels separately. For each mutation type, we explored numerous combinations of callers and filters in comparison to the calls generated by `Mutect2` from our own pipeline. We employed the Dice-Sorensen coefficient (DSC) to quantify the concordance between the filtered calls reported by our internal pipeline and the filtered mutations reported by MC3.

The highest score (DSC=0.917) for substitutions was achieved by considering the variants detected by two callers or more from the five used by the MC3, at positions covered by at least 20 reads in both tumor and blood samples, and with a VAF no less than 10% on both sides. However, we determined that this VAF threshold was overly stringent for our study objectives. Consequently, we chose the best combination of filtering criteria under the constraint of a minimal VAF of 5% in both lists. Under these conditions, the most favorable agreement was achieved by variants detected by at least two callers on the MC3 side, occurring at positions covered by at least 20 readsin tumor samples and 10 reads in blood samples on both sides. This configuration yielded a DSC of 0.896, as documented in Table 3.2. The table additionally provides details on the true positive and false positive rates (TPR and FPR, respectively), as well as the numbers of events retained from each source. True mutations are defined as the mutations reported in the filtered MC3 mutations file.

Mutations	MC3 Callers	DSC	TPR	FPR	Internal	MC3
SNVs/MNVs	2 or more	0.896	0.91	0.121	5,686	5,860
Indels	`INDELOCATOR` or `VARSCANI`	0.840	0.864	0.193	228	241

Table 3.2.: Quantitative metrics for assessing the overlap between the mutation lists from the filtered MC3 table and from the reprocessing of 58 TCGA WES files using our internal pipeline.

In a similar manner, various criteria for aligning indel calls were explored. Maintaining a consistent VAF threshold of at least 5%, we identified the optimal agreement by applying the same minimum coverage criteria as for substitutions and considering indels reported by either of two specific callers on the MC3 side, namely `INDELOCATOR` or `VARSCANI`. This configuration yielded a DSC of 0.840, as documented in Table 3.2. It is important to note that this DSC is lower compared to substitutions, but this is expected due to the inherent challenges associated with the accurate calling of indels. Figures A.5 and A.6 show the alignment between the number of substitutions and indels called using our internal pipeline and using the list of variants reported by MC3.

In order to adjust the filtering criteria and align the nf-core fusion pipeline used for META-PRISM fusions to the three different pipelines used for the three published lists of TCGA gene fusions (see Section 3.2.1.1), we downloaded with permission RNA-seq FASTQ files for 69 TCGA samples from the GDC data portal and analyzed them using the nf-core fusion pipeline. All 69 samples were successfully processed by the six different callers available in the pipeline, namely, `Arriba` v1.2.0 ("AR", Uhrig *et al.* (2021)), `EricScript` v0.5.5 ("ES", Benelli *et al.* (2012)), `FusionCatcher` v1.20 ("FC", Nicorici *et al.* (2014)), `Pizzly` v0.37.3 ("PZ", Melsted *et al.* (2017)), `SQUID` v1.5 ("SQ", Ma *et al.* (2018)), and `STAR-Fusion` v1.8.1 ("SF", Haas

et al. (2019)). All detected fusions were then annotated with `FusionAnnotator` which connects with databases of gene fusions detected in cancer or normal tissues[14].

Figure 3.16 summarizes the overlap between the fusions detected by these six algorithms. The fusions predicted by `SQUID` were not corroborated by other callers and were therefore discarded. Additionally, even though `FusionCatcher` was successful on all 69 TCGA samples, it repeatedly failed on some META-PRISM samples and was therefore not considered further. Consequently, only fusion calls from `Arriba`, `EricScript`, `Pizzly`, and `STAR-Fusion` were used.

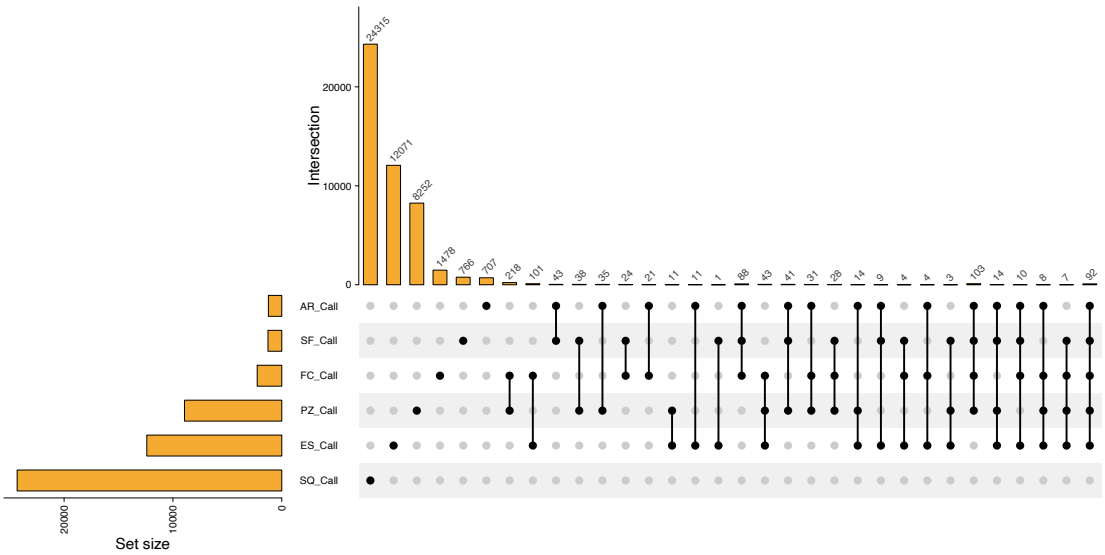


Fig. 3.16.: Upset plot showing the number of fusions identified individually by each algorithm and in combination with other algorithms. AR, ES, FC, PZ, SF, SQ stand for Arriba, EricScript, FusionCatcher, Pizzly, STAR-Fusion, and SQUID, respectively.

As our study only aimed at describing the variations relevant to cancer, only fusions known in cancer or involving at least one oncogenic partner (Section 3.1.2.3) were analyzed. We therefore limited the lists of gene fusions as detailed hereafter.

Firstly, we removed all fusions that have been previously reported in studies of normal tissues. More specifically, fusions were removed if they met any of the following criteria:

- Are in Babiceanu_Normal list (Babiceanu *et al.* 2016).
- Are in ChimerSeq_Normal_v4.0 list (available upon request to the authors (Jang *et al.* 2019)), which was established from the analysis of 1,144 TCGA normal samples and curated in order to remove well-known fusions (e.g., *TMPRSS2-ERG*) sometimes seen in normal samples.
- Are in GTEX_V6 Supplementary Table S3 (Singh *et al.* 2020).
- Have at least one Red Herring flag (`FusionAnnotator` annotations) among the following: GTEx_recurrent_StarF2019, BodyMap, DGD_PARALOGS, HGNC_GENEFAM,

[14] https://github.com/FusionAnnotator

Greger_ Normal, ConjoinG.
- One of the partners is not protein coding (Section 3.1.2.3).

Secondly, only gene fusions that met one of the following criteria were retained:

- Are in COSMIC v95 list of fusions[15].
- Are in ChimerKB v4.0 list (Jang *et al.* 2019).
- Are in Chitars Cancer v5.0 list (Singh *et al.* 2020).
- Are in TIC v3.3 list[16] (Novo *et al.* 2007).
- One of the partners is a cancer driver.

Lastly, after filtering the gene fusion calls for the 69 samples in each of the three TCGA published lists and in each of the lists of fusions predicted by the four callers considered, we looked for the combinations of calls that showed the best agreement. The best agreement was obtained between the following combinations:

- For the three published TCGA lists: fusions seen by `StarFusion` or by both `DEEPEST` and `PRADA`.
- For the four callers in our pipeline: fusions seen by both `Arriba` and `EricScript` or by both `Pizzly` and `STAR-Fusion`.

Agreement results are summarized in Table 3.3. True fusions designate fusions from the combination of the three published TCGA fusion lists. The DSC was 0.90 if fusion calls were assessed to be concordant regardless of the breakpoint prediction and was 0.77 if the predicted breakpoint was required to be identical. Applying the above rules, we retained tcga fusions seen by `StarFusion` alone (8,194 fusions), by `DEEPEST` and `PRADA` exactly (2,604 fusions), or by all three methods (9,989 fusions) as depicted in Figure A.2.

DSC	TPR	FPR	Internal	3 lists	Use breakpoints
0.77	0.76	0.22	166	169	Yes
0.90	0.88	0.08	140	145	No

Table 3.3.: Quantitative metrics for assessing the overlap between the fusion lists from the combination and filtering of three external fusion tables and from the reprocessing of 69 TCGA RNA-seq files using our internal pipeline.

3.3. Genomic profiles

We focused our genetic analysis on the tumor types (excluding tumors of unknown origin) that were represented by at least 10 sequenced tumors, resulting in 10 types for WES (META-PRISM WES, 83.8% of all DNA samples).

[15] https://cancer.sanger.ac.uk/cosmic/fusion
[16] https://genetica.unav.edu/TICdb

3.3.1. Mutational burden and signatures

WES analysis revealed a significant increase of somatic mutations in META-PRISM versus TCGA for 6 of 10 studied tumor types. Tumors from the MET500 cohort demonstrated a tumor mutational burden (TMB) similar to that of META-PRISM tumors. We observed the most significant increase of TMB in low-burden tumor types BRCA (2.1X fold in META-PRISM, 2.7X fold in MET500), PRAD (2.1X fold in META-PRISM, 2.2X in MET500), and PAAD (1.7X fold in META-PRISM), whereas no such increase was observed in high-burden tumor types, namely, BLCA, LUAD, and LUSC (Figure 3.17.A).

The delineation of the signatures activities in the genomes harboring 50 somatic substitutions or more was achieved by projecting onto the COSMIC v3.2 catalog of reference SBS signatures using `MutationalPatterns` R package (Blokzijl *et al.* 2018). Signature activities were further refined by using part of the Julia reimplementation of `SigProfiler`[17] from Pich *et al.* (2019) to run the sparsity-inducing step included in the original method.

The deconvolution of mutational signatures revealed similar signature compositions between metastatic META-PRISM, MET500, and primary TCGA cancers in all studied tumor types. However, a notable and consistent difference in META-PRISM versus TCGA across tumor types was the presence of signatures associated with platinum treatments (SBS31 and SBS35), reflecting that the majority of META-PRISM tumors (691 of 1,011 with known drug history) were pretreated with platinum compound therapies. Signatures SBS31 and SBS35 were detected in more than 50% of tumors in six META-PRISM tumor types [LUAD, BLCA, LUSC, HNSC, cholangiocarcinoma (CHOL), and adrenocortical carcinoma (ACC)] and contributed significantly more mutations compared with TCGA in four of them (Figure 3.17.B). Tumor types varied in frequency and types of received platinum drugs. For example, BRCA and PRAD rarely received platinum treatments and concordantly demonstrated a very low platinum-associated mutational signature.

We then investigated the association of SBS31 and SBS35 with three main platinum drugs in our cohort: cisplatin, carboplatin, and oxaliplatin. Among these drugs, cisplatin had the strongest association with SBS31 and had an association with SBS35 that was comparable with carboplatin, as revealed by logistic regression (Figure 3.17.B). More specifically, among tumors harboring at least 50 somatic substitutions and treated with cisplatin and no other platinum compound ($n = 95$), 40% harbored SBS31 and 17% SBS35, whereas in patients treated only with carboplatin ($n = 80$), 17% had a detectable activity of SBS35. In contrast, these two signatures were rarely detectable in tumor types predominantly treated with oxaliplatin (colon adenocarcinoma (COAD) and PAAD, Figure 3.17.B).

[17] https://bitbucket.org/bbglab/sigprofilerjulia

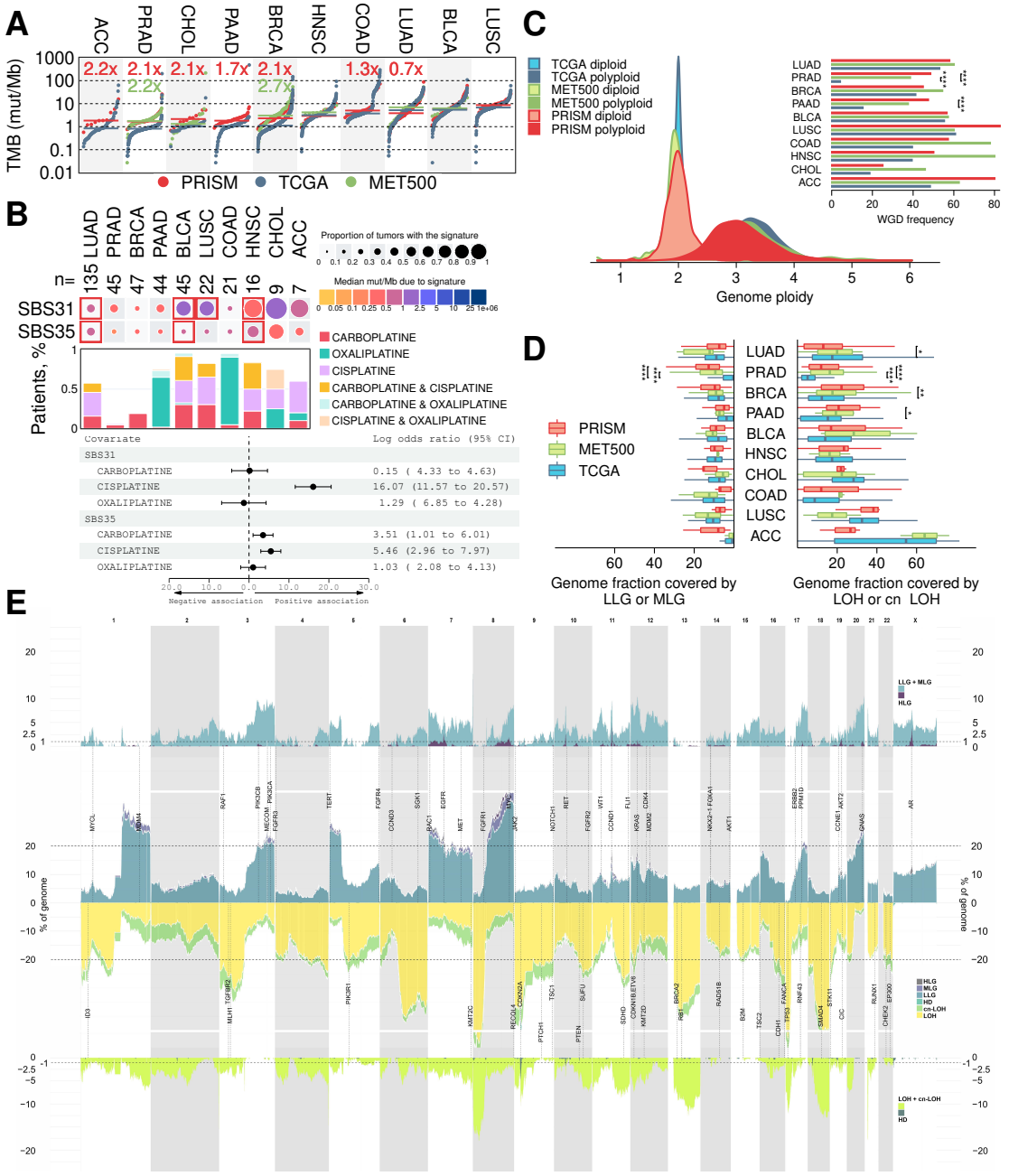


Fig. 3.17.: The genomic landscape of META-PRISM tumors.[18]

[18] **A**, Distribution of the mutational burden for each cohort and each tumor type of the META-PRISM WES subcohort (red), the MET500 cohort (green), and the TCGA cohort (blue). The fold changes and the p-values from Mann-Whitney U tests represent comparisons between mutational burdens of META-PRISM and TCGA (red), and MET500 and TCGA (green). mut/Mb, mutations per megabase. **B**, Top, detection level of mutational signatures (rows) in tumor types (columns). Colors indicate the median number of mutations contributed by the signature in samples harboring the signature. Red frames indicate significant Mann-Whitney U tests comparing META-PRISM with TCGA. Middle, information about the platinum drugs used in each tumor type. Bottom, log odds ratio from

3.3.2. Somatic copy-number alterations

WGD is a frequent event in cancer involving doubling the chromosome complement. We detected WGDs in 53.7% of META-PRISM WES tumors, which was comparable with MET500 (50.2%) but significantly higher than in TCGA (39.9%; $P < 0.001$ for META-PRISM vs TCGA and MET500 vs TCGA, Fisher-Boschloo tests). The fraction of WGDs varied between tumor types, ranging from 25.0% in CHOL to 82.6% in LUSC in META-PRISM. The most striking increase of WGD events in metastatic cancers compared with TCGA was observed in PRAD (META-PRISM 48.5%, 12X fold increase, $P < 0.001$; MET500 38.6%, 9.4X fold increase, $P < 0.001$) and PAAD (META-PRISM 47.4%, 3.1X fold increase, $P < 0.001$; MET500 37.5%, 2.5X fold increase, not significant due to small size; Fisher-Boschloo tests, Benjamini-Hochberg correction; Figure 3.17.C).

We next investigated the landscape of somatic CNAs in META-PRISM WES tumors without WGD. Using `FACETS` (Shen & Seshan 2016), we categorized CNAs in META-PRISM, MET500, and TCGA as copy gains and losses. Copy gains were subdivided into either low-level (LLG), middle-level (MLG), or high-level (HLG) gains. Copy losses included LOH, copy-neutral LOH (cn-LOH), and focal homozygous deletions (HD). LLGs, MLGs, LOH, and cn-LOH often spanned large regions or covered full chromosome arms, whereas HDs and HLGs were almost always focal. The fraction of the genome covered by low- to medium-level gains or LOH was most drastically increased in metastatic PRAD tumors. Additionally, LOH events were significantly more frequent in META-PRISM BRCA and PAAD tumors (Figure 3.17.D). The CNA profiles of META-PRISM tumors were overall similar to that of TCGA tumors for both the tumor type-adjusted dataset (Figure 3.17.E) and specific tumor types. The frequency of the majority of large CNAs did not differ significantly.

In META-PRISM WES tumors without WGD, an average of 10.7% and 19.7% of the genome were covered by copy gains and copy losses, respectively. No significant increase in this type of instability was observed in most studied tumor types except for PRAD, which demonstrated a dramatic increase in metastatic tumors compared with primary tumors, and for PAAD, in which the increase was limited to copy losses (Figure 3.17.D). Three chromosome arm gains (5p, 7p, and 8q) and seven losses (6q, 8p, 9p, 13q, 17p, 18p, and 18q) were observed in more than 20% of the non-WGD META-PRISM cohort. However, their frequency was not significantly different from TCGA. The majority of these chromosome regions enriched with

two logistic regression models predicting the presence or absence of SBS31 and SBS35 in all WES samples from META-PRISM with at least 50 somatic substitutions. CI, confidence interval. **C**, Density plot depicting the distribution of the estimated average ploidy in META-PRISM, MET500, and TCGA tumors for samples with and without WGD. The bar plot shows the proportion of polyploid tumors in the full cohorts and per tumor type. **D**, Double box plot describing the genome fraction covered by gains (low and middle level; left) and losses (LOH or cn-LOH; right) per tumor type, considering only tumors without WGD. Comparisons of META-PRISM vs. TCGA and MET500 vs. TCGA were performed using MannWhitney U tests (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$). **E**, Middle, fraction of tumors harboring different types of copy gains and losses across the genome. Top and bottom, excess of copy gains and losses in META-PRISM compared with TCGA. Somatic CNAs were classified into three types of copy gains, low, middle, and high level, and three types of copy losses, LOH, cn-LOH, and HD. The vertical dotted lines align to the loci where selected oncogenes (above) and tumor suppressor genes (below) are located. Only tumor types from the META-PRISM WES subcohort (10 tumor types) are represented in this figure. All p-values were adjusted for multiple testing using the BenjaminiHochberg procedure. P-values in B were adjusted by considering all the tests on the complete list of deconvoluted signatures.

gains and losses events were shared between several tumor types, whereas some chromosome regions were tumor-type specific: for instance, +16p, -16q, -22q in BRCA, +20p, +20q, -9q, and -11p in BLCA (Figure A.7). META-PRISM tumors with WGD did not demonstrate significant differences in arm-level copy-number losses, focal amplifications, focal deletions, or the number of driver mutations as compared with tumors without WGD in the five main tumor types of the WES subcohort.

High-level amplifications and homozygous deletions were rare in the META-PRISM WES tumors, spanning an average of 0.15% and 0.05% of the genome, respectively. However, highly amplified and homozygously deleted genes were detected on average 63 and 13 times per tumor. The most frequent of these events included amplification of *CCND1* (8.2%), *AR* (4.4%), 19q13 genes (2.7%), *EGFR* (2.7%), *MYC* (2.7%), and *KRAS* (2.5%; Figure A.8) and losses of *CDKN2A* (13%), *FAM106A/LGALS9C* (5.0%), Killer Ig-like receptors genes (3.4%), *RHD/RSRP1* (2.9%), and *PTEN* (2.7%; Figure A.8).

3.3.3. Incidence of cancer driver mutations

The discovery of significantly mutated genes with `Mutpanning` (Dietlein *et al.* 2020) confirmed previously reported cancer drivers (Figure A.9). Interestingly, this analysis highlighted genes that were reported as drivers in advanced tumors but not in primary tumors, namely, *EP300* in META-PRISM LUAD, *HERC2* in META-PRISM BRCA, *ESR1* in META-PRISM and MET500 BRCA, and *AR* in META-PRISM and MET500 PRAD. Mutations in these drivers may have been selected by the therapies or have arisen in the late stage of tumor evolution. We next selected a list of 360 cancer genes by intersecting COSMIC census Tier 1 and Tier 2 (v92) and OncoKB-annotated genes (Chakravarty *et al.* 2017) and created a catalog of driver somatic alterations (substitutions, small indels, amplifications, and deletions) using OncoKB annotations on these genes only. These types of driver events were observed in 96% of the META-PRISM WES tumors. The most frequently altered driver genes in META-PRISM were *TP53* (55% of samples), *KRAS* (25%), *CDKN2A* (18%), and *EGFR* (14%; Figure 3.18.A). On the whole cohort level, ten oncogenes and three tumor suppressor genes were significantly enriched compared with TCGA; 40% and 100% of those genes, respectively, were also enriched in MET500. Some driver genes were significantly enriched in specific tumor types: for example, *EGFR* and *CTNNB1* in LUAD; *TP53*, *AR*, *PTEN*, and *RB1* in PRAD; *ESR1* and *CCND1* in BRCA; *TP53* and *KRAS* in PAAD; and *FGFR3* in BLCA (Figure 3.18.A, FisherBoschloo tests, Benjamini-Hochberg correction). The number of WES-derived driver events (mutations and somatic CNAs) was significantly higher in metastatic tumors as compared with primary tumors at the cohort level (means of 3.9 and 3.6 in META-PRISM and MET500 vs. 3.2 in TCGA, $P < 0.0001$) and for 5 of 10 tumor types included in the META-PRISM WES cohort (Figure 3.18.B).

Major tumor suppressor genes frequently underwent biallelic inactivation. Such inactivations were observed in 92% of *TP53*-hit tumors, 95% for *CDKN2A*, 81% for *PTEN*, 88% for *SMAD4*, 41% for *ARID1A*, 78% for *APC*, and 92% for *RB1*. However, the predominant mechanisms of biallelic inactivation differed from one gene to another: in *TP53*, it was muta-

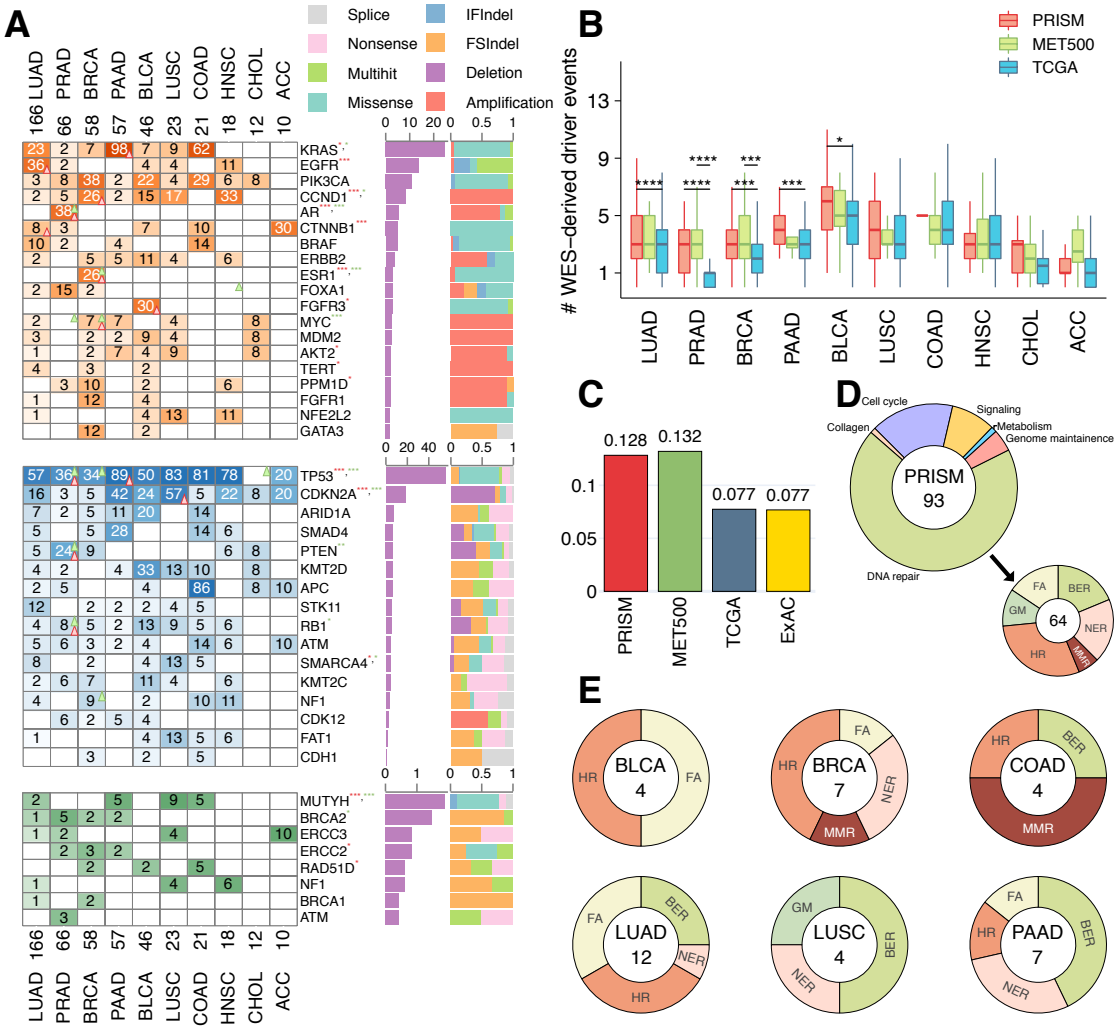


Fig. 3.18.: The landscape of cancer-associated somatic mutations, CNAs, and germline mutations in META-PRISM tumors. **A**. Heat maps depicting the percentage of tumors harboring driver events (substitutions, small indels, CNAs) in top oncogenes (top), top tumor suppressor genes (middle), and top cancer-predisposing genes (bottom). Triangle orientations (increase - triangle points up, decrease - points down) and colors (red for META-PRISM vs. TCGA, green for MET500 vs. TCGA) highlight significant changes in frequency. Similarly, stars next to the gene names represent significant changes at the cohort level using the same color code as for triangles (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$). The absolute bar plots show the percentage of tumors in META-PRISM harboring the alteration, whereas the relative bar plots show the breakdown of these alterations into different categories. Adjusted p-values per tumor type are from Fisher-Boschloo tests, whereas p-values across the cohort are from Cochran-Mantel-Haenszel tests. FS indel, frameshift insertion or deletion; IF indel, inframe insertion or deletion. **B**. Box plot of the number of driver events in META-PRISM, MET500, and TCGA. Adjusted p-values shown in the box plot are derived from Mann-Whitney U tests. **C**. Bar plot representing the incidence of cancer-risk germline variants in META-PRISM, MET500, and TCGA cancer patients and in ExAC European (non-Finnish) population. Stratified Cochran-Mantel-Haenszel test is used to account for tumor-type compositions in the cohorts except for comparison with ExAC in which the standard Fisher test is used. **D**. Pie charts representing the distribution of cancer-risk variants by pathways in META-PRISM tumors. **E**. Pie charts representing the distribution of cancer-risk variants in DNA repair pathways for each of the six tumor types harboring the most germline events. BER, base excision repair; FA, Fanconi anemia; GM, genome maintenance; HR, homologous repair; MMR, mismatch repair; NER, nucleotide excision repair. All p-values are adjusted for multiple testing using the Benjamini-Hochberg procedure. Only tumor types represented in the META-PRISM WES subcohort are shown in this figure.

tion followed by LOH; in *CDKN2A* and *PTEN*, it was homozygous deletion; and other genes demonstrated a combination of mechanisms (Figure A.10). Few oncogenes also underwent multihit events, most notably *EGFR* (57%), *AR* (12%), *PIK3CA* (7.7%), and *KRAS* (4.2%). Multihit events in *EGFR* in META-PRISM were significantly more frequent than in TCGA, likely reflecting the effect of *EGFR* inhibitors in LUAD (Figure A.11).

The WES of germline DNA from 569 META-PRISM patients was used to identify pathogenic cancer-predisposing variants (substitutions and indels). We focused on the germline variants annotated as pathogenic or likely pathogenic in the ClinVar database (Landrum, Lee, *et al.* 2014; Landrum, Chitipiralla, *et al.* 2020) or protein-disrupting and residing in genes strongly associated with cancer predisposition as described by Huang *et al.* (2018). We identified 75 patients in META-PRISM (13.1%) harboring at least one such variant. The fraction of patients with cancer-predisposing variants was similar to that in MET500 patients and was 1.75 times higher than in TCGA (8 cancer types, $P = 0.0012$, Fisher exact test) or in ExAC non-Finnish Europeans ($OR = 1.75$, $P = 0.0002$; Figure 3.18.C). Seventy-three percent of variants were in DNA repair pathway genes (Figure 3.18.D). The most frequent genes with cancer-predisposing variants in META-PRISM patients were *MUTYH* (1.9%), *BRCA2* (1.8%), *NF1* (0.9%), *ERCC2* (0.9%), *ERCC3* (0.9%), *RAD51D* (0.7%), and *FANCG* (0.7%; Figure 3.18.A). Variants in *MUTYH*, *NF1*, and *RAD51D* were significantly more frequent than in TCGA. We detected an increase of germline cancer-risk variants in most cancer types in META-PRISM. However, it reached significance only for PRAD ($P = 0.03$) and LUSC ($P = 0.004$) cancer types. Mutations in the HR pathway were the most frequent in BRCA, PRAD, LUAD, and BLCA; BER pathway in PAAD and LUSC; and MMR pathway in COAD (Figure 3.18.E). Thirty-seven percent of genes with germline cancer-risk variants harbored a somatic second-hit event, including somatic mutations in 9% and LOH resulting in the retention of the pathogenic variant in 27%.

3.4. Transcriptomic profiles

Similarly to the analyses of data from WES experiments, the analyzes of RNA-seq samples were led on all tumor types represented by at least 10 samples in META-PRISM, which corresponds to 20 tumor types (META-PRISM RNA-seq, 84.2% of all RNA samples).

Although we made a thorough effort of pipeline harmonization for the quantification of gene expression across all three cohorts (Section 3.2.2), we remained concerned about the potential impact of technical disparities on the expression profiles. We tried to measure this technical impact by using popular dimension-reduction techniques on the profiles of all three cohorts, restricting ourselves to the six most frequent tumor types in META-PRISM to limit the tumor type-related heterogeneity. As depicted in Figure A.12, some technical effects have a strong influence such as the difference between RNA samples prepared via polyA-enrichment and those prepared using hybridization capture techniques, which were exclusively utilized in some of the MET500 samples.

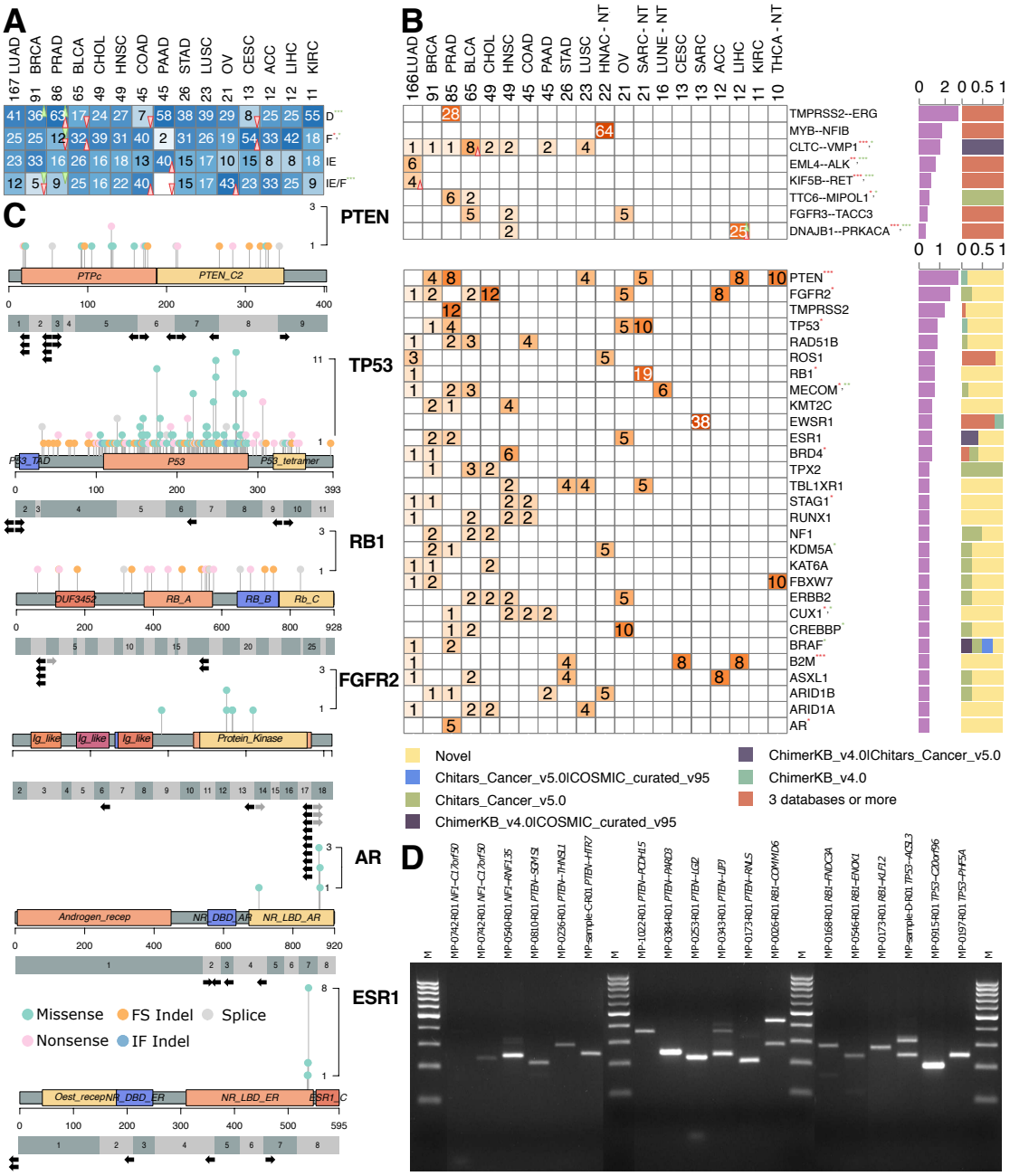


Fig. 3.19.: landscape of tumor microenvironments and cancer-associated gene fusions. [19]

[19] **A**. Heat map depicting the percentage of tumors classified in each TME subtype for each tumor type of the META-PRISM RNA-seq subcohort (non-TCGA tumor types and SARC were excluded). **B**, Heat maps depicting the percentage of tumors harboring gene fusions for known cancer gene fusions (top) or known cancer drivers (bottom). Only known oncogenic gene fusions seen in at least four samples among META-PRISM RNA-seq tumors are shown in the top. The bottom plot considers all known oncogenic gene fusions implicating drivers except for the fusions shown in the top plot that are excluded. Likewise, only drivers involved in fusions of at least four samples among META-PRISM RNA-seq tumors are shown. NT, Not_TCGA, tumor types that could not fit into TCGA classification.

We initially attempted differential gene expression analyses comparing META-PRISM samples to TCGA samples. However, these attempts were left aside due to our inability to adequately control for technical and biological disparities (library preparation protocols, sequencing devices, biopsy sites) unrelated to the differences we would have liked to investigate. Additionally, the absence in TCGA of patients with profiles comparable to META-PRISM patients, i.e., with advanced cancer and refractory to conventional treatments, or conversely, the absence of treatment-naive early cancers in MET500 and META-PRISM cohorts, impeded the use of batch-effect correction methods. Despite these challenges, we hypothesized that the impact of batch effects on the classification of TMEs could be mitigated by the use of gene expression signatures, which aggregate signals from predefined lists of genes. However, it is important to note that this hypothesis has not been empirically tested.

3.4.1. Immune characteristics

The TME is known to play a significant role in clinical outcomes and response to therapy (Galon *et al.* 2012; Goossens *et al.* 2015; Sparano *et al.* 2018; Thorsson *et al.* 2018). TMEs in META-PRISM, MET500, and TCGA patients were analyzed through the prism of the four-class classification described by Bagaev *et al.* (2021). Under this classification scheme, tumors from the tumor types considered by the authors can be categorized into four subtypes: immune-enriched and fibrotic (IE/F), immune-enriched and nonfibrotic (IE), fibrotic (F), and immune-depleted (D). This classification is based on the signature scores of 29 functional gene sets "representing the major functional components and immune, stromal, and other cellular populations of the tumor" (Bagaev *et al.* 2021).

As this categorization of TMEs is the result of a clustering analysis on TCGA samples, we had to build a classifier to predict classes for new samples. Briefly, we computed single-sample gene set enrichment scores for gene expression profiles of all three cohorts starting from the TPM tables. We then developed a classifier using the labels provided by the authors in their GitHub repository[20]. To achieve a good performance, we trained various multiclass classification models using the normalized TCGA scores, which we recalculated, as well as the scores supplied by the authors, to ensure the robustness and consistency of our approach. We compared the accuracy of different machine learning models trained using a 5-fold cross-

C. Fusions involving tumor suppressor genes, *PTEN*, *TP53*, and *RB1*, and oncogenes, *FGFR2*, *AR*, and *ESR1*. The op part of each panel indicates the protein domains (amino acid numbering) and locations and recurrence of driver somatic mutations categorized into splice site, frameshift, inframe, nonsense, and missense mutations. The bottom part shows fusion event breakpoints on the exonic structure. Only coding exons are shown. Black arrows indicate fusion breakpoints with the driver gene transcript located 5' (left arrow) or 3' (right arrow) of the breakpoint. Gray arrows indicate secondary fusions for which a principal fusion (with higher coverage) is found in the same patient and involves the same gene. FS indel, frameshift insertion or deletion; IF indel, inframe insertion or deletion. **D**. RT-PCR validation of 18 fusions in *TP53, RB1, PTEN, NF1*, and *AR*. **A** and **B**, Triangle orientations (increase - triangle points up, decrease - points down) and colors (red for META-PRISM vs. TCGA, green for MET500 vs. TCGA) highlight significant changes in subtype frequency. Similarly, stars next to the gene names represent significant changes at the cohort level using the same color code as for triangles (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$). P-values per tumor type are from Fisher-Boschloo tests, whereas p-values across the cohort are from Cochran-Mantel-Haenszel tests. **A**. P-values were not corrected for multiple testing due to the lack of independence between the tests performed within each tumor type.

[20] https://github.com/BostonGene/MFP/tree/master/Cohorts/Pan_TCGA

validation procedure. The scores reported in Table 3.4 present the average of the five test scores from the five internal cross-validation splits.

Model	Scores	Accuracy	F1 weight	Precision	Recall
AdaBoostClassifier	Authors	0.865	0.867	0.853	0.869
AdaBoostClassifier	Recomputed	0.845	0.847	0.834	0.844
KNeighborsClassifier	Authors	0.858	0.857	0.860	0.837
KNeighborsClassifier	Recomputed	0.847	0.845	0.847	0.824
LogisticRegression	Authors	0.917	0.916	0.911	0.907
LogisticRegression	Recomputed	0.897	0.897	0.888	0.884
RandomForestClassifier	Authors	0.867	0.866	0.866	0.845
RandomForestClassifier	Recomputed	0.859	0.857	0.857	0.836
SVC	Authors	0.938	0.938	0.936	0.931
SVC	Recomputed	0.911	0.911	0.906	0.900

Table 3.4.: Cross-validation performances of models trained to reproduce the four-subtype classification of tumor microenvironments from TCGA RNA-seq data.

Overall, we observed a marginal decrease in classification performance when substituting the scores computed by the authors with those we recalculated. The authors, as indicated in their GitHub repository, recommended the utilization of `KNeighbors` classifiers with a parameter value of $k = 35$. However, it is noteworthy that, among the five distinct models we evaluated, these `KNeighbors` classifiers [with best hyperparameters (tested values ranged from 2 to 80) at $k = 37$ and $k = 36$ for scores provided by the authors and recomputed by us, respectively], consistently yielded the poorest performance. For our analyses, we selected predictions generated by the logistic regression model trained on the scores that we had recalculated. It is important to note that tumor type-specific normalization parameters learned from the full TCGA dataset are required for performing predictions. As a consequence, only tumor types analyzed by Bagaev *et al.* (2021) could receive a TME subtype assignment. Notably, the authors excluded sarcomas from their analysis, which consequently prevented our examination of the TME within these tumors, as well as within rare tumor types.

No consistent difference in the distribution of TMEs across tumor types was observed when comparing META-PRISM to TCGA. However, the TME subtypes in some individual tumor types showed striking variation between the cohorts. For instance, immunosuppressive depleted (D) and fibrotic (F) subtypes were significantly increased in PRAD and BLCA, respectively, in META-PRISM compared with TCGA (63% vs. 41%, $P < 0.001$; 32% vs. 17%, $P = 0.008$; p-values are not adjusted due to the lack of independence; Figure 3.19.A). Enrichment of the depleted (D) subtype in PRAD was also significant in MET500 versus TCGA (78% vs. 41%, $P < 0.001$).

3.4.2. Known and novel driver gene fusions

We next investigated gene fusions in the 795 META-PRISM RNA-seq tumors successfully processed by the gene fusion-calling pipeline and belonging to the 20 selected tumor types. It is important to note that the identification of gene fusions was confined to two specific

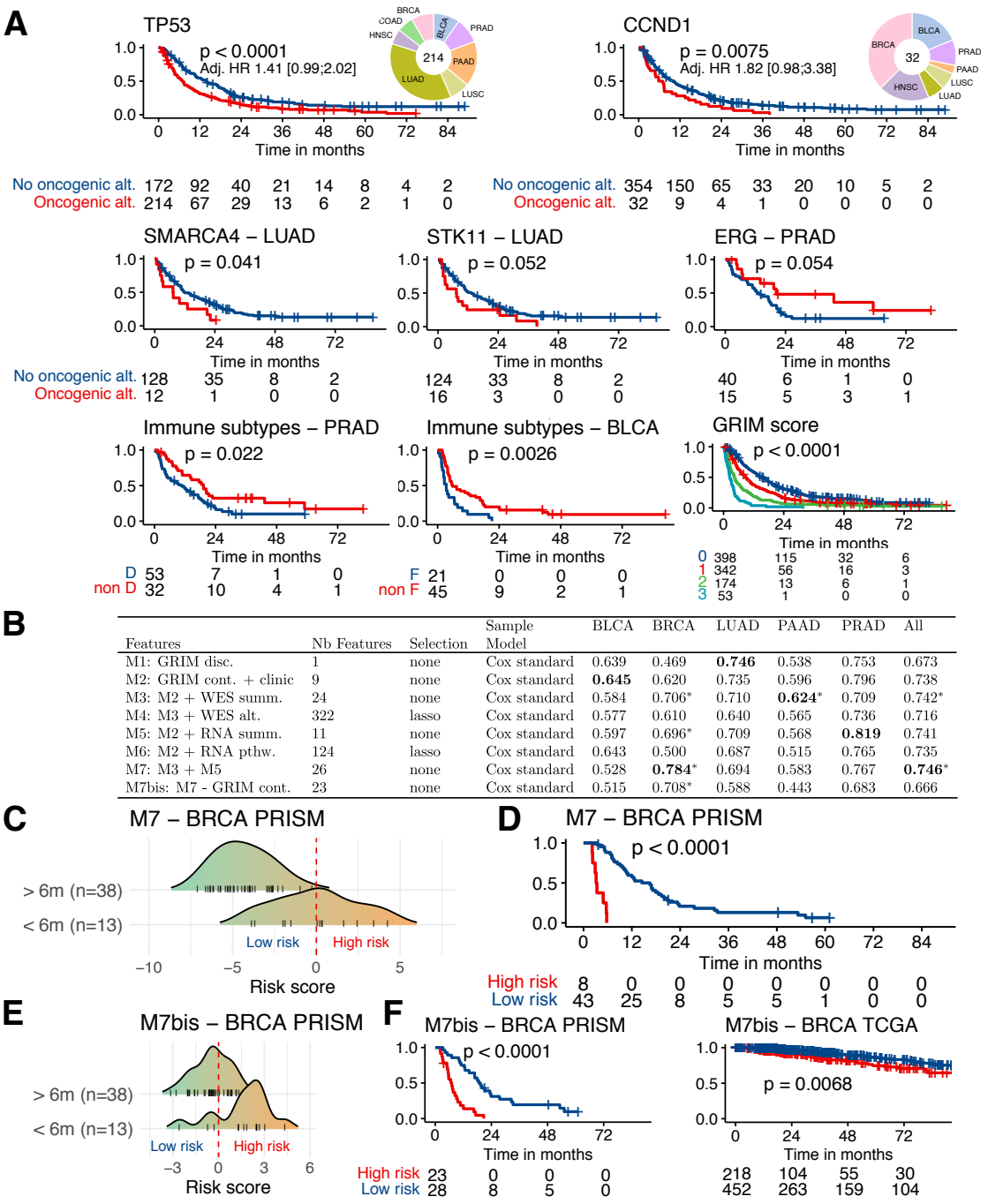
categories: known oncogenic fusions and fusions featuring a cancer driver. These categories were prioritized as they were deemed the most pertinent in the context of cancer and could be detected with a high degree of reliability across all three cohorts, as described in our methodology (Section 3.1.2.2).

A total of 432 known oncogenic gene fusions [Chimer KB v4.0, Jang *et al.* (2019); Chitars v5.0, Balamurali *et al.* (2019); COSMIC v95[21]; TIC v3.3, Novo *et al.* (2007)) were identified in META-PRISM RNA-seq tumors (34%). As previously described, well-known oncogenic gene fusions were often tumor-type specific - *TMPRSS2-ERG* 28% in PRADs, *EML4-ALK* 6% in LUADs, *DNAJB1-PRKACA* 25% in liver hepatocellular carcinoma (LIHCs), *MYB-NFIB* 64% in head and neck adenoid cystic carcinomas (HNACs) - and some were significantly enriched in META-PRISM versus TCGA (Figure 3.19.B). In addition to known oncogenic fusions, we identified 329 fusions involving cancer driver genes and promiscuous partners (29% of META-PRISM RNA-seq tumors). Among the most recurrently fused driver genes, we identified tumor suppressor genes *PTEN* (1.9%), *TP53* (0.9%), and *RB1* (0.8%). In these genes, 72% of fusion breakpoints were recurrent (Figure 3.19.C). *PTEN* fusions were observed in several tumor types in META-PRISM, but they were most prevalent in PRAD (Figure 3.19.B). The oncogene most frequently involved in fusions across different tumor types was *FGFR2* (1.5%). Interestingly, *ESR1* and *AR*, both known to be critical players in hormone therapy resistance in BRCA and PRAD, were involved in fusions in 2.2% and 4.7% of the respective tumor types in META-PRISM (Figure 3.19.B and C). Sixteen of 18 tested fusions (89%) were experimentally validated through RT-PCR and Sanger sequencing (Figure 3.19.D).

3.5. Improved survival predictions

The association of molecular markers with overall survival measured from the day of diagnosis has been the subject of extensive work (Van 'T Veer *et al.* 2002; Paik *et al.* 2004; Olivier *et al.* 2006; Yoshimoto *et al.* 2007; Goldstein *et al.* 2008; Cardoso *et al.* 2008; Riley *et al.* 2009; Ihle *et al.* 2012), some of which has become standard of care. However, the association of these markers with survival at the very late stage has been scarcely explored, partly because it is not current clinical practice to systematically profile advanced tumors and partly because the clinical value of such sequencing is still an open question. In the context of the META-PRISM cohort, where patients were often enrolled in early-phase clinical trials at the metastatic stage, tools for accurately assessing the risk profile of each individual are very much needed to assist clinical decisions. The multivariate survival analysis models built using the molecular markers described in the previous sections are an attempt to gauge the prognostic utility of molecular profiling and, most importantly, the potential additional value that such molecular data could contribute beyond the information provided by standard clinical parameters. The assessment of the models prognostic capabilities were assessed at six-month, a time horizon which used in practice to take treatment decisions.

[21] https://cancer.sanger.ac.uk/cosmic/fusion



Features	Nb Features	Selection	Sample Model	BLCA	BRCA	LUAD	PAAD	PRAD	All
M1: GRIM disc.	1	none	Cox standard	0.639	0.469	**0.746**	0.538	0.753	0.673
M2: GRIM cont. + clinic	9	none	Cox standard	**0.645**	0.620	0.735	0.596	0.796	0.738
M3: M2 + WES summ.	24	none	Cox standard	0.584	0.706*	0.710	**0.624***	0.709	0.742*
M4: M3 + WES alt.	322	lasso	Cox standard	0.577	0.610	0.640	0.565	0.736	0.716
M5: M2 + RNA summ.	11	none	Cox standard	0.597	0.696*	0.709	0.568	**0.819**	0.741
M6: M2 + RNA pthw.	124	lasso	Cox standard	0.643	0.500	0.687	0.515	0.765	0.735
M7: M3 + M5	26	none	Cox standard	0.528	**0.784***	0.694	0.583	0.767	**0.746***
M7bis: M7 - GRIM cont.	23	none	Cox standard	0.515	0.708*	0.588	0.443	0.683	0.666

Fig. 3.20.: KaplanMeier curves for significant predictors and prediction of 6-month survival using clinical characteristics and combinations of WES or RNA-seq features.[22]

[22] **A**. Kaplan-Meier survival curves of META-PRISM WES and RNA-seq tumors according to the oncogenic alterations status of *TP53* and *CCND1* (top row), of LUAD tumorsaccording to *SMARCA4* and *STK11*, of PRAD tumors - according to *ERG* (middle row), of BLCA tumors with fibrotic (F) and without fibrotic (non-F) TMEs (middle

3.5.1. Single prognostic markers

We investigated how driver genes are univariably associated with metastatic cancer patient survival in META-PRISM WES and RNA-seq tumors. Twenty-one genes altered (CNAs, mutations, gene fusions) in more than 5% of patients were considered for the analysis. Only *TP53* and *CCND1* genes showed an association with survival [adjusted hazard ratio = 1.44; confidence interval (CI) 95%, 0.992.02 and hazard ratio = 1.82; CI 95%, 0.983.38, respectively] after adjusting for the tumor-type composition (Figure 3.20.A, upper row). Analyses per tumor type revealed an association of *SMARCA4* and *STK11* events with poor prognosis in LUAD and of *ERG* events with favorable prognosis in PRAD (Figure 3.20.A, middle row).

TME classifications, which have already been shown to be prognostic (Fridman *et al.* 2012; Bagaev *et al.* 2021), were associated with survival in META-PRISM. Indeed, immune-cold subtypes (F and D) had worse survival compared with immune-enriched subtypes (IE/F and IE). Strikingly, the F subtype in BLCA and D subtype in PRAD, both enriched in META-PRISM versus TCGA, were also associated with the poorest prognosis in the corresponding tumor type (P = 0.002, F vs. non-F BLCA; P = 0.02, D vs. non-D in PRAD; Figure 3.20.A, bottom row).

3.5.2. Multivariate models

Patients with advanced metastatic cancer are characterized by severe physiologic deterioration as measured by LDH levels, serum albumin, or neutrophil-to-lymphocyte ratio. The Gustave Roussy Immune score (GRIM score) (Bigot *et al.* 2017), which combines these physiologic markers, could predict six-month survival in the META-PRISM patients with an average concordance index of 0.67 (Figure 3.20.B). We then investigated if the prediction of survival at this stage of the disease could be *improved by considering genetic markers* engineered from the previous analyses. The list of genetic markers retrieved from WES was composed of

right), of PRAD tumors with depleted (D) and without depleted (non-D) TMEs, and all META-PRISM tumors classified into four categories as determined by the Gustave Roussy Immune Score (GRIM score; bottom row). All p-values are computed from log-rank tests. Hazard ratios are computed by fitting univariate Cox models except for *TP53* and *CCND1*, in which the cancer type was included in the model. **B**. Averaged C-indexes on the 1,000 cross-validation subsamples for each combination of features analyzed. Cox models were fit on META-PRISM patients with both WES and RNA-seq data and from the five most frequent tumor types or a pan-cancer cohort composed of patients belonging to the 10 most prevalent tumor types. Asterisks indicate adjusted p-values below 0.1 (Benjamini-Hochberg correction) from one-sided Mann-Whitney U tests comparing C-indices from model M2 to C-indices from models M3 to M7bis. Bold values indicate the highest mean C-index across models for each tumor type or combination of tumor types. M1: discrete GRIM score. M2: continuous components of GRIM score, age, tumor type, and gender are considered. M3: variables from M2 and WES summary statistics are considered. M4: variables from M2 and WES-derived alterations in driver genes (lasso reduction applied). M5: variables from M2 and RNA-seq-derived fusion burden and TMEs. M6: variables from M2 and RNA-seq-derived gene expression signatures (lasso reduction applied). M7: variables included in M3 and M5 are considered. M7bis: variables M7 without the components of the GRIM score. Nb, number. **C**. Distribution of the predicted risk scores from the Cox model M7 in META-PRISM BRCA patients who had a survival time greater or lower than 6 months from the biopsy date. One patient who was censored before 6 months was included in the "> 6m" group. **D**. KaplanMeier curves for META-PRISM BRCA patients predicted to be high-risk (risk score > 0) or low-risk (risk score < 0) using the M7 model. **E**. Identical to C using the M7bis model. **F**. KaplanMeier curves for META-PRISM BRCA (left) and TCGA BRCA (right) patients predicted to be high-risk (risk score > 0) or low-risk (risk score < 0) using the M7bis model.

summary statistics, including somatic CNAs, WGD status, MSI score, TMB, and the presence of oncogenic alterations aggregated into genes or pathways discriminating between ESMO Scale for Clinical Actionability of molecular Targets (ESCAT) levels (see Chapter 4). The list of markers retrieved from RNA-seq included TME subtypes and gene expression signatures measuring the activity of immune pathways, activation of general pathways (Subramanian *et al.* 2005), or main transcription factors (Garcia-Alonso *et al.* 2019).

In order to compare the added prognostic value from different categories of markers to the objective clinical markers, we ran multiple Cox proportional hazard regressions on the META-PRISM WES and RNA-seq subcohort and each of the five main tumor types considering only the samples with both WES and RNA-seq. To provide a comparative assessment of the added and independent value of biomarkers derived from WES or RNA-seq as compared with clinical biomarkers (age, gender, global tumor characteristics, blood test results, treatment history), we experimented with multiple models that incorporated different combinations of predictors. The baseline model (M2) to be improved upon was composed of all standard clinical variables plus the continuous components of the GRIM score (LDH, albumin, neutrophil, and lymphocyte levels). Separate models were run for each of the five most represented tumor types in our cohort (BLCA, BRCA, LUAD, PAAD, and PRAD) and the META-PRISM WES and RNA-seq cohort (10 tumor types). In order to ensure a fair comparison with more complex models, only samples with complete molecular profiles (WES and RNA-seq) were used in the different Cox models.

For each combination of predictors and samples, we repeated the feature selection and coefficient estimation steps *1,000 times* to assess the selection procedure's stability, if any, and to provide robust estimates and CIs for the effect of each variable on survival. Each of the 1,000 repeats consisted of running the selection procedure, if any, and training the model on a random 80% subsample. The remaining 20% was used to assess the model quality (C-index). The C-index values reported in Figure 3.20.B are averages of the 1,000 estimates of the C-index on the 20% test subsamples. Estimates of the covariates' coefficients were computed by averaging the coefficient fitted on each 80% subsample across the 1,000 repeats. In case a selection procedure is active at each repeat, covariates were assigned an estimated coefficient of zero every time the selection procedure did not retain them. CIs at the 95% level were estimated by computing the empirical 2.5% and 97.5% quantiles from the 1,000 estimates. A selection procedure was applied in case the number of covariates was unreasonably high given the number of available observations. We experimented with univariate - fitting of a univariate Cox model for each candidate predictor and selecting only predictors for which adjusted p-values fell below a certain threshold - and multivariate selection procedures - fitting of a Cox model with lasso penalization and selecting only predictors with nonzero coefficients to identify a small set of predictors that, hopefully, is a superset of all important predictors. The Cox regressions on predictors selected by the lasso were always more prognostic (higher value of C-index) than when predictors were selected through the univariate procedure.

Each set of 1,000 repeats of Cox regressions was preceded by preprocessing steps (run once for each combination of samples and features) in which covariates were formatted, imputed, min-max transformed, and analyzed for redundancy. The imputation relied on the

`MICE` R package (Zhang 2016), which performs multiple imputations of the same dataset by iteratively learning a predictive model of each covariate with missing data from all other covariates and randomly sampling from observed values through a predictive mean matching procedure. Data tables with at least one missing value were imputed five times, and coefficient estimates across each of the 1,000 repeats were calculated by averaging estimates across the five imputed tables.

The C-index was estimated in two different ways, either using the C-statistic proposed by Harrell *et al.* (1996) or using the *inverse probability weighting estimate* presented by Uno *et al.* (2011) to account for the fact that the C-statistic proposed by Harrell and colleagues depends on the censoring distribution which in practice is rarely independent of the covariates used in the Cox regression. As survival times are subject to censoring, and because we were mainly interested in discriminating between patients having a survival longer or shorter than six months as it is the time horizon clinicians are interested in for making important decisions, we used an estimate of the C-index truncated at six months by considering only pairs of patients for which one of the two patients had a survival shorter than six months. The estimator proposed by Uno and colleagues (see formula 2.3 from their work), as well as the C-statistic proposed by Harrell and colleagues, were implemented in C using a truncation time of six months. A comparison of the estimated values from both estimators showed very little difference due to the fact that only a minority of patients were right-censored in our cohort ($<$10%). Values reported in the main text use the C-statistic from Harrell and colleagues truncated at six months.

3.5.3. Results

As expected, the models' performances based on discrete GRIM score (M1 model) were significantly improved by using the continuous metrics underlying it and by adding baseline clinical variables (age, gender, tumor type, clinical subtypes; M2 model). Nevertheless, the incorporation of WES-derived markers (M3 and M4 models), RNA-seq-derived markers (M5 and M6 models), or both (M7 model) resulted in a further increase of the C-index over the continuous GRIM score model (M2) in all analyzed tumor types and for the whole cohort except for LUAD, in which the discrete GRIM score classification was the most prognostic (Figure 3.20.B).

Strikingly in META-PRISM BRCAs, for which the GRIM score alone (M1, C-index = 0.469) had no prognostic value and the GRIM score components along with clinical variables including immunohistochemistry (IHC) subtypes was moderately prognostic (M2, C-index = 0.620), the inclusion of genetic markers engineered from WES or RNA-seq considerably improved the six-month survival prediction (C-index = 0.784). Analysis of the model's coefficients shows that multiple markers have significant prognostic value independently of the baseline clinical variables, which include the IHC subtypes, namely: (i) the total number of ESCAT Tier 1 alterations [log hazard ratio (LHR) 3.75; CI 95%, 1.076.53], the fraction of the genome covered by (ii) focal deletions (LHR -4.05; CI 95%, -8.73, -0.34), by (iii) focal amplifications (LHR 7.28; CI 95%, 2.63 - 13.55), and by (iv) middle- to low-level gains (LHR

-3.65, CI 95%, -8.18, -1.22), and (v) the TME subtypes (LHR -3.18 for F vs. D, -2.72 for IE vs. D, -4.80 IE/F vs. D; Figure A.13). The predicted risk scores from this model were significantly higher in patients with a survival time greater than six months than in patients with a survival time lower than six months from the date of the biopsy (Figure 3.20.C). By discriminating between patients having a positive risk score (poor prognosis, 8/51) and a negative risk score (good prognosis, 43/51), the model was able to split BRCA patients into two groups with very different clinical outcomes ($P < 0.0001$ log-rank test; Figure 3.20.D).

In order to validate the survival model on an external cohort, we took advantage of the large TCGA BRCA cohort and considered all patients with survival, mutation, gene expression, and gene fusion data available ($n = 670$). Due to the unavailability in TCGA of clinical variables underlying the GRIM score, we considered one extra model (M7bis) that was identical to M7, except that the GRIM score components were excluded. This model was trained on META-PRISM BRCA and validated on TCGA BRCA. It was found to be predictive of survival in META-PRISM (C-index = 0.71, cross-validation; $P < 0.0001$ log-rank test) and in TCGA (C-index = 0.65 at one year and C-index = 0.63 at six years; $P = 0.0068$; Figure 3.20.E and F).

3.6. Conclusions

The META-PRISM project provides genetic and transcriptomic variation for a large cohort of refractory tumors from 39 tumor types, 20 of which were analyzed in depth. This cohort is characterized by a short survival time after the biopsy date and by a high proportion of multiresistant tumors or rare tumors with no approved therapy options. Consequently, the genomes of these tumors represent a much-advanced evolutionary stage containing the footprints of mutagenic treatments and therapeutic pressure. To characterize the genetic traits specific to this cohort and assess how these may inform the tumor's aggressiveness and resistance to therapies, we compared META-PRISM to >10,000 primary untreated tumors from TCGA (The Cancer Genome Atlas Research Network *et al.* 2013) and validated all results using an external cohort of 500 metastatic tumors (Robinson *et al.* 2017).

The compilation, curation, and structuring of data for the META-PRISM cohort represented a critical and labor-intensive undertaking to establish a robust database with detailed and high-quality data. To ensure the seamless integration of data tables across the three cohorts in our comparative analyses, we meticulously standardized the descriptions of the tumor types, primary sites, and biopsy sites according to internationally recognized nomenclatures. We additionally performed a thorough effort of pipeline harmonization to minimize the sources of technical disparities that could introduce confounding elements into our analyses. This work entailed the reprocessing of all sequencing files from MET500 and more than 200 terabytes of TCGA sequencing of data for the description of CNAs. The high-rigor and precision with which the data has been collected and harmoniously processed have been instrumental to the overall success of this project. The clinico-biological tables and the somatic alterations of META-PRISM patients have been deposited on a public por-

tal at https://cbioportal.gustaveroussy.fr/study/summary?id=metaprism_2023 allowing any interested researcher to explore the data interactively and test new hypotheses.

Our analysis reveals that several types of genomic instability were strongly enriched in refractory cancers, particularly the mutation rate, the frequency of WGD, and the fraction of the genome covered by focal CNAs. These results are consistent with previous studies that reported increased genomic instability and mutational burden in metastases of different cancer types (Bakhoum *et al.* 2018; Priestley *et al.* 2019; Shukla *et al.* 2020; Z. Hu *et al.* 2020; Watkins *et al.* 2020; Nguyen *et al.* 2022). Correlative analyses between the mutational profiles of tumors and the history of treatments received have shown increased activity of signatures SBS31 and SBS35 in tumors treated with cisplatin and to a lesser extent with carboplatin. The high mutation rate of these two signatures supports previous observations of strong mutational footprints caused by platinum treatment in metastatic cancers (Pich *et al.* 2019) and relatively low mutagenic effects of oxaliplatin (Szikriszt, Póti, Németh, *et al.* 2021). The driver gene fusions and somatic CNAs represented 9.4% and 18.8% of all detected variation in driver genes, respectively, and were strongly enriched in META-PRISM tumors compared with TCGA tumors.

Markers of physiologic deterioration, such as levels of albumin, neutrophils, lymphocytes, and LDH, are used in objective risk scoring systems to predict patient survival, which is essential for assessing the eligibility of these patients for phase I clinical trials (Garrido-Laguna *et al.* 2012; Feng *et al.* 2020). This study shows that tumor genomic and transcriptomic features can be used to improve the accuracy of predictions based on objective risk factors at this stage of the disease. The added value of genetic markers to current prognostic scores is currently limited at the pan-cancer level, likely due to the high heterogeneity of mechanisms driving tumorigenesis in each tumor type. However, models incorporating genetic markers are significantly more accurate in predicting six-month survival in refractory BRCAs, showing that WES and RNA-seq are important for accurately establishing the individual risk profile of patients with late-stage cancer.

This study validates the previous combined genomic and transcriptomic descriptions from pretreated pan-cancer metastatic diseases (Pleasance, Titmuss, *et al.* 2020; Pleasance, Bohm, *et al.* 2022). It also highlights the feasibility of precision medicine in clinical routine in patients without standard treatments available. Benefits of molecular profiling-guided access to new drug protocols in such patients has now been demonstrated in multiple clinical studies (Massard *et al.* 2017; Rodon *et al.* 2019; Recondo *et al.* 2020; Pleasance, Bohm, *et al.* 2022; Andre *et al.* 2022). The present cohort advances translational cancer genomics by providing a unique resource combining detailed clinical data with exome and transcriptome profiling.

Bibliography

1. Andre, F. *et al.* Genomics to select treatment for patients with metastatic breast cancer. en. *Nature* **610,** 343–348. doi:10.1038/s41586-022-05068-3 (Oct. 2022).
2. Arkenau, H.-T. *et al.* Prospective Validation of a Prognostic Score to Improve Patient Selection for Oncology Phase I Trials. en. *Journal of Clinical Oncology* **27,** 2692–2696. doi:10.1200/JCO.2008.19.5081 (June 2009).
3. Babiceanu, M. *et al.* Recurrent chimeric fusion RNAs in non-cancer tissues and cells. en. *Nucleic Acids Research* **44,** 2859–2872. doi:10.1093/nar/gkw032 (Apr. 2016).
4. Bagaev, A. *et al.* Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. en. *Cancer Cell,* S1535610821002221. doi:10.1016/j.ccell.2021.04.014 (May 2021).
5. Bakhoum, S. F. *et al.* Chromosomal instability drives metastasis through a cytosolic DNA response. en. *Nature* **553,** 467–472. doi:10.1038/nature25432 (Jan. 2018).
6. Balamurali, D. *et al.* ChiTaRS 5.0: the comprehensive database of chimeric transcripts matched with druggable fusions and 3D chromatin maps. en. *Nucleic Acids Research,* gkz1025. doi:10.1093/nar/gkz1025 (Nov. 2019).
7. Bayle, A. *et al.* Liquid versus tissue biopsy for detecting actionable alterations according to the ESMO Scale for Clinical Actionability of molecular Targets in patients with advanced cancer: a study from the French National Center for Precision Medicine (PRISM). en. *Annals of Oncology* **33,** 1328–1331. doi:10.1016/j.annonc.2022.08.089 (Dec. 2022).
8. Benelli, M. *et al.* Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. en. *Bioinformatics* **28,** 3232–3239. doi:10.1093/bioinformatics/bts617 (Dec. 2012).
9. Berger, F. *et al.* Randomised, open-label, multicentric phase III trial to evaluate the safety and efficacy of palbociclib in combination with endocrine therapy, guided by ESR1 mutation monitoring in oestrogen receptor-positive, HER2-negative metastatic breast cancer patients: study design of PADA-1. en. *BMJ Open* **12,** e055821. doi:10.1136/bmjopen-2021-055821 (Mar. 2022).
10. Bertucci, F. *et al.* Genomic characterization of metastatic breast cancers. en. *Nature* **569,** 560–564. doi:10.1038/s41586-019-1056-z (May 2019).
11. Bigot, F. *et al.* Prospective validation of a prognostic score for patients in immunotherapy phase I trials: The Gustave Roussy Immune Score (GRIm-Score). en. *European Journal of Cancer* **84,** 212–218. doi:10.1016/j.ejca.2017.07.027 (Oct. 2017).
12. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. en. *Genome Medicine* **10,** 33. doi:10.1186/s13073-018-0539-0 (Dec. 2018).
13. Bonneville, R. *et al.* Landscape of Microsatellite Instability Across 39 Cancer Types. en. *JCO Precision Oncology,* 1–15. doi:10.1200/PO.17.00073 (Nov. 2017).
14. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. en. *Nature Biotechnology* **34,** 525–527. doi:10.1038/nbt.3519 (May 2016).

15. Cardoso, F. *et al.* Clinical Application of the 70-Gene Profile: The MINDACT Trial. en. *Journal of Clinical Oncology* **26,** 729–735. doi:10.1200/JCO.2007.14.3222 (Feb. 2008).
16. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. en. *JCO Precision Oncology,* 1–16. doi:10.1200/PO.17.00011 (Nov. 2017).
17. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. en. *Bioinformatics* **34,** i884–i890. doi:10.1093/bioinformatics/bty560 (Sept. 2018).
18. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. en. *Nature Biotechnology* **31,** 213–219. doi:10.1038/nbt.2514 (Mar. 2013).
19. Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. en. *Nature Communications* **8,** 15180. doi:10.1038/ncomms15180 (June 2017).
20. Dehghannasiri, R. *et al.* Improved detection of gene fusions by applying statistical methods reveals oncogenic RNA cancer drivers. en. *Proceedings of the National Academy of Sciences* **116,** 15524–15533. doi:10.1073/pnas.1900391116 (July 2019).
21. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. en. *Nature Genetics* **43,** 491–498. doi:10.1038/ng.806 (May 2011).
22. Dietel, M. *et al.* A 2015 update on predictive molecular pathology and its role in targeted cancer therapy: a review focussing on clinical relevance. en. *Cancer Gene Therapy* **22,** 417–430. doi:10.1038/cgt.2015.39 (Sept. 2015).
23. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. en. *Nature Genetics* **52,** 208–218. doi:10.1038/s41588-019-0572-y (Feb. 2020).
24. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. en. *Cell* **173,** 305–320.e10. doi:10.1016/j.cell.2018.03.033 (Apr. 2018).
25. Drilon, A. *et al.* Efficacy of Larotrectinib in *TRK* FusionPositive Cancers in Adults and Children. en. *New England Journal of Medicine* **378,** 731–739. doi:10.1056/NEJMoa1714448 (Feb. 2018).
26. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. en. *Cell Systems* **6,** 271–281.e7. doi:10.1016/j.cels.2018.03.002 (Mar. 2018).
27. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. en. *Nature Biotechnology* **38,** 276–278. doi:10.1038/s41587-020-0439-x (Mar. 2020).
28. Feng, J.-F., Wang, L., Yang, X. & Chen, S. Gustave Roussy Immune Score (GRIm-Score) is a prognostic marker in patients with resectable esophageal squamous cell carcinoma. en. *Journal of Cancer* **11,** 1334–1340. doi:10.7150/jca.37898 (2020).

29. Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. en. *Nature Reviews Cancer* **12,** 298–306. doi:10.1038/nrc3245 (Apr. 2012).
30. Galon, J. *et al.* Cancer classification using the Immunoscore: a worldwide task force. en. *Journal of Translational Medicine* **10,** 205. doi:10.1186/1479-5876-10-205 (Dec. 2012).
31. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. en. *Cell Reports* **23,** 227–238.e3. doi:10.1016/j.celrep.2018.03.050 (Apr. 2018).
32. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. en. *Genome Research* **29,** 1363–1375. doi:10.1101/gr.240663.118 (Aug. 2019).
33. Garrido-Laguna, I. *et al.* Validation of the royal marsden hospital prognostic score in patients treated in the phase I clinical trials program at the MD Anderson Cancer Center. en. *Cancer* **118,** 1422–1428. doi:10.1002/cncr.26413 (Mar. 2012).
34. Goldstein, L. J. *et al.* Prognostic Utility of the 21-Gene Assay in Hormone ReceptorPositive Operable Breast Cancer Compared With Classical Clinicopathologic Features. en. *Journal of Clinical Oncology* **26,** 4063–4071. doi:10.1200/JCO.2007.14.4501 (Sept. 2008).
35. Goossens, N., Nakagawa, S., Sun, X. & Hoshida, Y. Cancer biomarker discovery and validation. eng. *Translational Cancer Research* **4,** 256–269. doi:10.3978/j.issn.2218-676X.2015.06.04 (June 2015).
36. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. en. *Nature Genetics* **49,** 170–174. doi:10.1038/ng.3774 (Feb. 2017).
37. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. en. *Nature* **520,** 353–357. doi:10.1038/nature14347 (Apr. 2015).
38. Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. en. *Genome Biology* **20,** 213. doi:10.1186/s13059-019-1842-9 (Dec. 2019).
39. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15,** 361–387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4 (Feb. 1996).
40. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. en. *Cell* **173,** 291–304.e6. doi:10.1016/j.cell.2018.03.022 (Apr. 2018).
41. Hu, X. *et al.* TumorFusions: an integrative resource for cancer-associated transcript fusions. en. *Nucleic Acids Research* **46,** D1144–D1149. doi:10.1093/nar/gkx1018 (Jan. 2018).

42. Hu, Z., Li, Z., Ma, Z. & Curtis, C. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. en. *Nature Genetics* **52,** 701–708. doi:10.1038/s41588-020-0628-z (July 2020).
43. Huang, K.-l. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. en. *Cell* **173,** 355–370.e14. doi:10.1016/j.cell.2018.03.039 (Apr. 2018).
44. Ihle, N. T. *et al.* Effect of KRAS Oncogene Substitutions on Protein Behavior: Implications for Signaling and Clinical Outcome. en. *JNCI: Journal of the National Cancer Institute* **104,** 228–239. doi:10.1093/jnci/djr523 (Feb. 2012).
45. Jang, Y. E. *et al.* ChimerDB 4.0: an updated and expanded database of fusion genes. en. *Nucleic Acids Research,* gkz1013. doi:10.1093/nar/gkz1013 (Nov. 2019).
46. Karapetis, C. S. *et al. K-ras* Mutations and Benefit from Cetuximab in Advanced Colorectal Cancer. en. *New England Journal of Medicine* **359,** 1757–1765. doi:10.1056/NEJMoa0804385 (Oct. 2008).
47. Kautto, E. A. *et al.* Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. en. *Oncotarget* **8,** 7452–7463. doi:10.18632/oncotarget.13918 (Jan. 2017).
48. Landrum, M. J., Lee, J. M., *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. en. *Nucleic Acids Research* **42,** D980–D985. doi:10.1093/nar/gkt1113 (Jan. 2014).
49. Landrum, M. J., Chitipiralla, S., *et al.* ClinVar: improvements to accessing data. en. *Nucleic Acids Research* **48,** D835–D844. doi:10.1093/nar/gkz972 (Jan. 2020).
50. Le Tourneau, C., Borcoman, E. & Kamal, M. Molecular profiling in precision medicine oncology. en. *Nature Medicine* **25,** 711–712. doi:10.1038/s41591-019-0442-2 (May 2019).
51. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. en. *New England Journal of Medicine* **372,** 2509–2520. doi:10.1056/NEJMoa1500596 (June 2015).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with BurrowsWheeler transform. en. *Bioinformatics* **25,** 1754–1760. doi:10.1093/bioinformatics/btp324 (July 2009).
53. Li, H., Handsaker, B., *et al.* The Sequence Alignment/Map format and SAMtools. en. *Bioinformatics* **25,** 2078–2079. doi:10.1093/bioinformatics/btp352 (Aug. 2009).
54. Ma, C., Shao, M. & Kingsford, C. SQUID: transcriptomic structural variation detection from RNA-seq. en. *Genome Biology* **19,** 52. doi:10.1186/s13059-018-1421-5 (Dec. 2018).
55. Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L. & Siu, L. L. Molecular profiling for precision cancer therapies. en. *Genome Medicine* **12,** 8. doi:10.1186/s13073-019-0703-1 (Dec. 2020).
56. Massard, C. *et al.* High-Throughput Genomics and Clinical Outcome in Hard-to-Treat Advanced Cancers: Results of the MOSCATO 01 Trial. en. *Cancer Discovery* **7,** 586–595. doi:10.1158/2159-8290.CD-16-1396 (June 2017).
57. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. en. *Genome Biology* **17,** 122. doi:10.1186/s13059-016-0974-4 (Dec. 2016).

58. Melsted, P. *et al.* Fusion detection and quantification by pseudoalignment. en. doi:10.1101/166322 (July 2017).
59. Naxerova, K. *et al.* Origins of lymphatic and distant metastases in human colorectal cancer. en. *Science* **357,** 55–60. doi:10.1126/science.aai8515 (July 2017).
60. Nguyen, B. *et al.* Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. en. *Cell* **185,** 563–575.e11. doi:10.1016/j.cell.2022.01.003 (Feb. 2022).
61. Nicorici, D. *et al.* FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. en. doi:10.1101/011650 (Nov. 2014).
62. Novo, F. J., de Mendíbil, I. O. & Vizmanos, J. L. TICdb: a collection of gene-mapped translocation breakpoints in cancer. en. *BMC Genomics* **8,** 33. doi:10.1186/1471-2164-8-33 (Dec. 2007).
63. Olivier, M. *et al.* The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. en. *Clinical Cancer Research* **12,** 1157–1167. doi:10.1158/1078-0432.CCR-05-1029 (Feb. 2006).
64. Paik, S. *et al.* A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. en. *New England Journal of Medicine* **351,** 2817–2826. doi:10.1056/NEJMoa041588 (Dec. 2004).
65. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. en. *Bioinformatics* **34** (ed Hancock, J.) 867–868. doi:10.1093/bioinformatics/btx699 (Mar. 2018).
66. Pich, O. *et al.* The mutational footprints of cancer therapies. en. *Nature Genetics* **51,** 1732–1740. doi:10.1038/s41588-019-0525-5 (Dec. 2019).
67. Pleasance, E., Bohm, A., *et al.* Whole-genome and transcriptome analysis enhances precision cancer treatment options. en. *Annals of Oncology* **33,** 939–949. doi:10.1016/j.annonc.2022.05.522 (Sept. 2022).
68. Pleasance, E., Titmuss, E., *et al.* Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. en. *Nature Cancer* **1,** 452–468. doi:10.1038/s43018-020-0050-6 (Apr. 2020).
69. Poon, S., McPherson, J. R., Tan, P., Teh, B. & Rozen, S. G. Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. en. *Genome Medicine* **6,** 24. doi:10.1186/gm541 (2014).
70. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. en. *bioRxiv.* doi:10.1101/201178 (Nov. 2017).
71. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. en. *Nature* **575,** 210–216. doi:10.1038/s41586-019-1689-y (Nov. 2019).
72. Recondo, G. *et al.* Feasibility and first reports of the MATCH-R repeated biopsy trial at Gustave Roussy. en. *npj Precision Oncology* **4,** 27. doi:10.1038/s41698-020-00130-7 (Dec. 2020).
73. Riley, R. D., Sauerbrei, W. & Altman, D. G. Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. en. *British Journal of Cancer* **100,** 1219–1229. doi:10.1038/sj.bjc.6604999 (Apr. 2009).

74. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. en. *Nature* **548,** 297–303. doi:10.1038/nature23306 (Aug. 2017).
75. Rodon, J. *et al.* Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. en. *Nature Medicine* **25,** 751–758. doi:10.1038/s41591-019-0424-4 (May 2019).
76. Rothwell, D. G. *et al.* Utility of ctDNA to support patient selection for early phase clinical trials: the TARGET study. en. *Nature Medicine* **25,** 738–743. doi:10.1038/s41591-019-0380-z (May 2019).
77. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. en. *Cell* **173,** 321–337.e10. doi:10.1016/j.cell.2018.03.035 (Apr. 2018).
78. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. en. *Nucleic Acids Research* **44,** e131–e131. doi:10.1093/nar/gkw520 (Sept. 2016).
79. Shukla, A. *et al.* Chromosome arm aneuploidies shape tumour evolution and drug response. en. *Nature Communications* **11,** 449. doi:10.1038/s41467-020-14286-0 (Jan. 2020).
80. Sicklick, J. K. *et al.* Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study. en. *Nature Medicine* **25,** 744–750. doi:10.1038/s41591-019-0407-5 (May 2019).
81. Singh, S. *et al.* The landscape of chimeric RNAs in non-diseased tissues and cells. en. *Nucleic Acids Research* **48,** 1764–1778. doi:10.1093/nar/gkz1223 (Feb. 2020).
82. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. en. *F1000Research* **4,** 1521. doi:10.12688/f1000research.7563.2 (Feb. 2016).
83. Sparano, J. A. *et al.* Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. en. *New England Journal of Medicine* **379,** 111–121. doi:10.1056/NEJMoa1804710 (July 2018).
84. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. en. *Proceedings of the National Academy of Sciences* **102,** 15545–15550. doi:10.1073/pnas.0506580102 (Oct. 2005).
85. Szikriszt, B., Póti, Á., Németh, E., *et al.* A comparative analysis of the mutagenicity of platinum-containing chemotherapeutic agents reveals direct and indirect mutagenic mechanisms. en. *Mutagenesis* **36,** 75–86. doi:10.1093/mutage/geab005 (Apr. 2021).
86. Szikriszt, B., Póti, Á., Pipek, O., *et al.* A comprehensive survey of the mutagenic impact of common cancer cytotoxics. en. *Genome Biology* **17,** 99. doi:10.1186/s13059-016-0963-7 (Dec. 2016).
87. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. en. *Nature Genetics* **45,** 1113–1120. doi:10.1038/ng.2764 (Oct. 2013).
88. Thorsson, V. *et al.* The Immune Landscape of Cancer. en. *Immunity* **48,** 812–830.e14. doi:10.1016/j.immuni.2018.03.023 (Apr. 2018).
89. Uhrig, S. *et al.* Accurate and efficient detection of gene fusions from RNA sequencing data. en. *Genome Research* **31,** 448–460. doi:10.1101/gr.257246.119 (Mar. 2021).

90. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. On the Cstatistics for evaluating overall adequacy of risk prediction procedures with censored survival data. en. *Statistics in Medicine* **30,** 1105–1117. doi:10.1002/sim.4154 (May 2011).
91. Van 'T Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. en. *Nature* **415,** 530–536. doi:10.1038/415530a (Jan. 2002).
92. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. en. *Nucleic Acids Research* **38,** e164–e164. doi:10.1093/nar/gkq603 (Sept. 2010).
93. Wang, Q. *et al.* Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. en. *Nature Communications* **11,** 2539. doi:10.1038/s41467-019-12438-5 (May 2020).
94. Watkins, T. B. K. *et al.* Pervasive chromosomal instability and karyotype order in tumour evolution. en. *Nature* **587,** 126–132. doi:10.1038/s41586-020-2698-6 (Nov. 2020).
95. Yates, L. *et al.* The European Society for Medical Oncology (ESMO) Precision Medicine Glossary. en. *Annals of Oncology* **29,** 30–35. doi:10.1093/annonc/mdx707 (Jan. 2018).
96. Yoshimoto, M. *et al.* FISH analysis of 107 prostate cancers shows that PTEN genomic deletion is associated with poor clinical outcome. en. *British Journal of Cancer* **97,** 678–685. doi:10.1038/sj.bjc.6603924 (Aug. 2007).
97. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. en. *Nature Medicine* **23,** 703–713. doi:10.1038/nm.4333 (June 2017).
98. Zhang, Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. eng. *Annals of Translational Medicine* **4,** 30. doi:10.3978/j.issn.2305-5839.2015.12.63 (Jan. 2016).
99. Zheng, H., Brennan, K., Hernaez, M. & Gevaert, O. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. en. *GigaScience* **8,** giz145. doi:10.1093/gigascience/giz145 (Dec. 2019).

4. Genetic mechanisms of treatment resistance

Contents

Abstract Chapter 4

In this chapter, we will delve into the currently described drug resistance mechanisms and explore how molecular profiling of tumors in patients contributes to our understanding of the genetic resistance mechanisms against both conventional and innovative drugs. The first section offers a brief overview of various cancer treatments and the recognized mechanisms of drug resistance, setting the stage for the subsequent analyses of patients data. The second section provides a detailed account of how we integrated data from the OncoKB and CIViC knowledge databases with somatic alterations identified through WES and RNA-seq to elucidate treatment resistances observed in META-PRISM patients. The insights gained from this comprehensive project were subsequently applied to uncover resistance mechanisms associated with two innovative drugs in distinct clinical settings. The first setting involves an ongoing phase 2 clinical trial investigating the effectiveness of the antibody-drug conjugate (ADC) trastuzumab deruxtecan in breast cancers. The second setting leverages molecular profiles from urothelial cancer patients who participated to three large trials conducted at Gustave Roussy so as to decipher the mechanisms of resistance to fibroblast growth factor receptor (*FGFR*) inhibitors.

How tumor cells evade the effects of antineoplastic drugs is a vast question that remains mostly elusive to our understanding although general principles and specific biomarkers of resistance are gradually being described. Chapter 3 introduced the META-PRISM study, providing an overview from database assembly to the identification of molecular alterations in WES and RNA-seq data and their applicability in predicting survival outcomes. As detailed in Section 3.1.1.3, extensive efforts were made to collect information on past treatments for the majority of META-PRISM patients. With the exception of 20 patients, all others have available comprehensive records of the antineoplastic drugs they received and progressed upon prior to the biopsy we analyzed. This valuable dataset offers numerous analytical opportunities, the first of which is the delineation of the associations between tumor genotypes and exposure to treatments and, subsequently, resistance to these treatments.

While WES and RNA-seq provide rich information, they also have inherent limitations that constrain the depth and breadth of our explorations into resistance mechanisms. Notably, the limited depth of WES, with a median of 140x across the cohort, hampers the detection of variants present in minor proportions in the tumor. This technical constraint, coupled with the spatial heterogeneity of tumors that a single biopsy cannot fully capture (Swanton 2012; Jamal-Hanjani *et al.* 2017), restricts the exhaustive description of molecular alterations explaining the observed resistances. Nevertheless, leveraging two high-quality knowledge databases, which reports extensive drug response-molecular alterations relationships with varying degrees of confidence, allowed us to explore the potential therapeutic implications of WES and RNA-seq data. The expertise and technical insights gained through the META-PRISM project has been put into practice for analyzing genotyping data from two other translational projects, both focused on elucidating resistance mechanisms to innovative drugs and presented in the last section of this chapter.

4.1. Drugs and mechanisms of resistance

4.1.1. Classification of drugs

Nowadays there is a diverse array of antineoplastic drugs at our disposal for treating cancer patients but this hasn't always been the case. The availability of therapeutic options is not uniform across different tumor types; it logically tends to be higher for frequently encountered cancers like lung, breast, or prostate cancer and lower for rare tumor types that have been less studied and are not well understood. Figure 4.1 provides a timeline depicting all FDA-approved drugs for the treatment of lung cancer, starting from the introduction of mechlorethamine hydrochloride, commonly known as nitrogen mustard, in the 1940s.

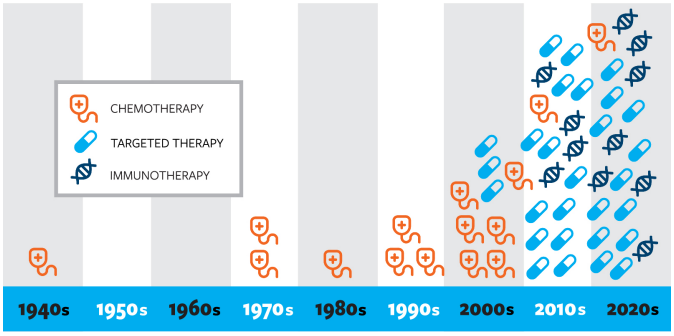


Fig. 4.1.: Timeline of FDA drug approvals for the treatment of lung cancer as of October 2023.
Source: https://www.lungcancerresearchfoundation.org

The wide diversity of cancer drugs is better understood through classification systems, most of which group together drugs sharing similar mechanisms of action for eliminating tumor cells or disrupting the signals they rely on. While numerous drug classification systems exist, describing them is beyond the scope of this introduction. Instead, we will use a simple classification that categorizes them into four overarching classes: chemotherapies, targeted therapies, hormone therapies, and immunotherapies. It is important to note that this classification is not rigid, as, for instance, some view hormone therapies and immunotherapies as types of targeted therapy.

Chemotherapies have been widely used for many decades and across many cancer type (Galmarini *et al.* 2012). As the earliest drugs developed for cancer treatment, they initially stood as the sole therapeutic option. Nowadays, they remain widely in use either in the pre-operative setting as *neoadjuvant* therapy to decrease tumor burden or after surgery as adjuvant treatment to eliminate residual tumor cells post-treatment with other drugs, or as non-curative or palliative care in patients with rare or aggressive tumors or those that have exhausted all treatment options. Additionally, they are frequently used in combination with other drugs, such as targeted inhibitors, as reviewed by (Bashraheel *et al.* 2020). The mechanisms of action of chemotherapies involve targeting rapidly dividing cells through diverse approaches, categorizing them into five main families:

1. *Alkylating agents* such as busulfan, platinum-based agents, cyclophosphamide, and temozolomide. They induce DNA damage, leading to cell cycle arrest.
2. *Antimetabolites* such as 5-fluorouracil, capecitabine, and pemetrexed. They prevent DNA replication during the S-phase.
3. *Topoisomerase I and II inhibitors* such as irinotecan and topotecan (topoisomerase I inhibitors) and doxorubicin and etoposide (topoisomerase II inhibitors) prevent DNA

resealing during replication and cause DNA damage.
4. *Antimitotic agents* such as include docetaxel, paclitaxel, and vincristine. They disrupt the normal dynamics of microtubules, leading to mitotic arrest.
5. *Tumor antibiotics* such as bleomycin, mitomycin C, and dactinomycin. They interfere with RNA synthesis or induce DNA damage through specific binding.

Targeted therapies have a relatively recent emergence compared to chemotherapies and constitute a diverse array of drugs comprising three main families: monoclonal antibodies, tyrosine kinase inhibitors (TKIs), and antibody-drug conjugates (ADCs). However, targeted therapies are more expensive than chemotherapies, particularly monoclonal antibodies, limiting their availability (Smith & Prasad 2021). Another significant major issue of targeted agents lies in their inability to comprehensively inhibit cancer-relevant signaling pathways, a limitation underscored by the high relapse rates observed in the long run (Groenendijk & Bernards 2014).

1. *Monoclonal antibodies* are designed to specifically target and bind selected proteins on the surface of cancer cells or other cells involved in cancer development. They work by triggering an immune response towards cells or by blocking the connection between cancer cells and the growth factors that these cells depend on for survival and proliferation. The classification of monoclonal antibodies naturally aligns with the specific targets they bind to. Notable early examples, approved by the FDA in the late 1990s or early 2000s, include rituximab, which targets the *CD20* antigen on B cells in non-Hodgkin lymphomas and CLLs; trastuzumab, which targets *HER2* in *HER2*-positive breast cancer; bevacizumab, which targets vascular endothelial growth factors (*VEGF*s) in various solid cancers; and cetuximab, which binds to the *EGFR* and is indicated in colorectal and head and neck cancers.
2. *Tyrosine kinase inhibitors* are a second main class of targeted therapies. These drugs act by inhibiting specific enzymes called tyrosine kinases, which relay signals from and into the cell by phosphorylating tyrosine residues on target proteins. Tyrosine kinases, a subset of protein kinases, play critical roles in various cellular functions, including cell growth, proliferation, and differentiation. Given the causal link between kinase alterations and the aberrant characteristics of tumor cells (Blume-Jensen & Hunter 2001), a multitude of drugs has been developed to inhibit these enzymes. TKIs are designed to be selective for particular tyrosine kinases or kinase families, minimizing off-target effects, but some TKIs have broad specificities. Examples of TKIs include imatinib, which targets *BCR-ABL1* in CML; erlotinib, designed to inhibit *EGFR* in NSCLC; and sunitinib, targeting *VEGFRs*, *PDGFR*, and *CSFR* in renal cell carcinoma. It's noteworthy that, for historical reasons, drugs inhibiting non-tyrosine kinases are also commonly referred to as TKIs, although they do not specifically target tyrosine kinases. Examples include *BRAF* inhibitors vemurafenib and dabrafenib in melanomas, PI3K inhibitor alpelisib with indications across various cancer types, and *PARP* inhibitors olaparib, rucaparib, and niraparib employed in breast, ovarian, pancreatic and prostate cancers.
3. *Antibody-drug conjugates* represent the third major category of targeted therapies,

presenting as an emerging class of drugs that combine the cytotoxic effectiveness of chemotherapy with the precision inherent to standard targeted therapies. They consist of engineered molecules composed of a monoclonal antibody, a cytotoxic payload, and a linker. The antibody serves to selectively deliver the drug to cells harboring the targeted antigen, such as *HER2*, *HER3*, and *EGFR* for trastuzumab, panitumumab, and cetuximab antibodies, respectively. The payload,is the "warhead" that will induce cytotoxicity when released inside recognized cells or their proximity. Currently approved ADCs use a chemotherapeutic agent for the payload but other types of payload are under development. Lastly, the linker plays a critical role in connecting the antibody and payload, ensuring the stability and tolerability of the entire molecule. Gemtuzumab ozogamicin, targeting *CD33*, marked the first approved ADC in 2000, while brentuximab vedotin, targeting *CD30*, became the second approved ADC in 2011. As of December 2021, 14 ADCs had received EMA or FDA approval including six in solid tumors, and more than 100 are currently under investigation in clinical trials (Z. Fu *et al.* 2022).

The third main class of cancer treatments encompasses *hormone therapies*, also known as endocrine therapies. The history of hormone suppression in cancer management can be traced back to the late 19th century with the seminal work of Thomas Beatson, who demonstrated the control of breast cancer through the removal of ovaries. In 1977, tamoxifen emerged as the first FDA-approved estrogen-inhibiting drug for breast cancer treatment, marking a groundbreaking development that continues to have widespread applications in managing hormone-positive breast cancers (Osborne 1998). Endocrine therapies function by suppressing hormones essential for the growth of cancer cells, achieved either by impeding the body's hormone production or inhibiting hormone receptors. They are indicated for the treatment of hormone-sensitive cancers, primarily comprising breast and prostate cancers, and to a lesser extent, other genitourinary cancers. The three major families of approved hormone therapies are antiestrogens, employed in breast cancer and further subdivided into aromatase inhibitors (anastrozole, letrozole, and exemestane), selective estrogen receptor degraders (fulvestrant, elacestrant), and selective estrogen receptor modulators (tamoxifen, toremifene) ; antiandrogens, such as bicalutamide or enzalutamide, used in prostate cancer; and GnRH agonists and antagonists, such as goserelin and degarelix, respectively, which act by desensitizing or blocking hormone receptors on the pituitary gland.

Immunotherapies have been gaining a lot of attention lately due to the success stories of long-term or complete responses achieved in clinical trials and now in clinical care. Interestingly, immune checkpoint inhibitors were described over two decades ago in seminal papers that laid the foundation for modern cancer immunotherapy (Kroemer & Zitvogel 2021). These papers detailed the targeting of three critical molecules - *CTLA-4* receptor (Krummel & Allison 1995), *PD-1* receptor, and *PD-L1* ligand (Freeman *et al.* 2000), which are the only three targets of the six currently approved immune checkpoint inhibitors (Bagchi *et al.* 2021) apart from LAG3-blocking immunotherapy relatlimab which is approved in combination with nivolumab in advanced melanomas[1]. The efficacy of immune checkpoint inhibitors in cancer treatment

[1] https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-opdualag-unresectable-or-metastatic-melanoma

was recognized after the spectacular results of the first phase III clinical trial demonstrating the anti-melanoma effects of anti-*CTLA-4* ipilimumab, reported in 2010 (Hodi *et al.* 2010). Apart from immune checkpoint inhibitors, immunotherapies include cytokine-based therapies, adoptive T-cell therapies, oncolytic viruses, and cancer vaccines, all of which are actively being investigated or already incorporated in the standard of care (Waldman *et al.* 2020; Boardman & Salles 2023). The monoclonal antibodies we described earlier as part of targeted therapies may also be considered a passive immunotherapy type. New cancer immunotherapies are continuously being developed such as bispecific antibodies with combine two binding sites targeted at two different antigens or epitopes of the same cell, or bi-specific T-cell engagers which aim at activating the immune system against the tumor by binding T-cells and cancer cells.

4.1.2. Mechanisms of resistance

Before cancer treatment begins, cancer cells may already have a way of resisting or escaping the effects of therapy, in which cases the resistance is said to be *intrinsic*. Alternatively, they might develop a new capability that allows them to *adapt* to the treatment and resist it. However, distinguishing between *acquired* and *intrinsic* resistances is challenging due to our inability to exhaustively genotype all cancer cells, particularly in advanced cases where there are multiple heterogeneous clones. *Intratumor heterogeneity*, whether genetic, epigenetic, or microenvironment-related, increases the likelihood of cancer cells adapting to the therapy-induced selective pressures, surviving therapy, and ultimately driving cancer progression (McGranahan & Swanton 2017). In heterogeneous tumors, some clones may be sensitive to the drug, while others may already have the ability to resist it. Depending on which clones are most prevalent when the drug is introduced, no clinical benefit may be observed or, on the contrary, a good response may be observed initially before the therapy-induced clonal expansion of resistant subclones drives cancer progression. The success of profiling sequencing experiments in distinguishing acquired resistances from intrinsic ones depends on their ability and sensitivity to determine the genotype and phenotype of minor clones pre- and post-treatment.

The mechanisms by which cancer cells develop resistance to drugs can be roughly categorized into six classes: changes in the drug pharmacokinetics, modifications in the target expression or molecular structure, repair of the DNA damage induced by some therapies, deactivation of cell death pathways or activation of pro-survival pathways, reactivation of the targeted pathway or activation of an alternative and functionally redundant pathway, and changes in cells phenotypes or in the TME (Gottesman 2002; Holohan *et al.* 2013; Housman *et al.* 2014; Mansoori *et al.* 2017; Nussinov *et al.* 2021). However, it is important to note that there are probably still unknown mechanisms of drug resistance, and therefore current classifications remain incomplete.

The effectiveness of drugs on cells is firstly dependent on *pharmacokinetic factors* which include drug efflux, influx, distribution, or metabolism. For many types of anticancer medications, such as chemotherapies and ADCs, their ability to enter cells and deliver their

cytotoxic agents is critical for inducing cell death. However, the presence of *ATP*-binding cassette transporter proteins, particularly those labeled as multidrug resistance proteins (MRPs) of the C family, can impact drug efflux and influx. This can lead to decreased drug effectiveness and chemotherapy resistance in cultured cells (Sodani *et al.* 2012). MRPs are known to remove various hydrophobic compounds and can bind to cytotoxic agents like taxanes or topoisomerase inhibitors. Preclinical studies have shown that elevated levels of MRPs, including *ABCB1*, *ABCC1*, and *ABCG2*, were associated with a poor response to chemotherapies (Doyle *et al.* 1998; Gottesman *et al.* 2002; Szakács *et al.* 2004). However, attempts to inhibit these proteins in combination with chemotherapy have yielded unsatisfactory results in clinical trials (Thomas & Coley 2003; Pusztai *et al.* 2005). In addition, the expression of certain genes like *NEK2* has been linked to drug resistance through increased drug efflux (Zhou *et al.* 2013). The presence or absence of specific molecules, such as GSH antioxidant or enzymes required for converting drugs like 5-fluorouracil or irinotecan into their active forms, are yet another mechanism employed by cancer cells to impact the drug effectiveness (T. Wilson *et al.* 2006).

A secondary and primary mechanism of drug resistance involves *alterations to the target*. These alterations can manifest as a significant increase in its expression levels, thereby reducing the treatment efficacy for which doses cannot be increased because of its toxicity. Alternatively, structural changes to the target may render the drug ineffective against cells harboring such modifications. In prostate cancer, one recognized mechanism of resistance to androgen receptor (*AR*) antagonists, like the first-generation drug bicalutamide (Visakorpi *et al.* 1995), is *AR* amplification, leading to heightened androgen expression. The overexpression of the *AR-v7* splice variant, lacking the binding domain targeted by second- and third-generation inhibitors enzulatamide and abiraterone, is an alternative resistance mechanisms to more recent AR inhibitors (Antonarakis *et al.* 2014). Commonly encountered mechanisms of resistance also include mutations in gatekeeper residues of oncogenic kinases. For instance, the T315I mutation in the *BCR-ABL1* fusion gene was the first mutation causally linked to imatinib resistance in CMLs (O'Hare *et al.* 2005). The third-generation *BCR-ABL1* inhibitor ponatinib, FDA-approved in 2012, stands as the sole clinically available inhibitor capable of overcoming T315I-mediated resistance. Another illustrative example is the *EGFR* T790M mutation frequently observed after initial responses to first- and second-generation *EGFR* inhibitors (gefitinib, erlotinib, afatinib, dacomitinib) in NSCLCs (Kobayashi *et al.* 2005; Pao, Miller, *et al.* 2005). The second-generation inhibitor afatinib demonstrates enhanced efficacy against various mutant forms of *EGFR* but is ineffective against those harboring T790M or exon 20 insertions (Sun *et al.* 2013). In contrast, the third-generation inhibitor osimertinib, which is active against T790M mutants and received FDA accelerated approval in 2015, is compromised by mutations affecting the C797 site (notably C797S) utilized by the drug, among other mechanisms (Papadimitrakopoulou *et al.* 2018).

DNA damage repair is a third mechanism employed by cancer cells to evade the effects of certain drugs. Numerous chemotherapeutic agents operate by inducing DNA damage in cells (Section 4.1.1), causing cell cycle arrest and ultimately resulting in either cell death through apoptosis or cell cycle resumption after damage repair. Interestingly, tumor cells frequently

exhibit modifications in critical repair pathways, particularly those involved in responding to double-strand breaks, such as HR and NHEJ. The presence of these alterations renders cancer cells heavily reliant on unaffected pathways that can, therefore, be targeted. For instance, inhibiting *PARP*-dependent single-strand break repair has proven to be a highly effective strategy in breast and ovarian cancers with HRD *BRCA1/BRCA2* mutant cells, as demonstrated by Farmer *et al.* (2005). However, it is noteworthy that inframe deletions in *BRCA2* mutant cells, which restore its DNA repair function, have been identified as a mechanism of resistance to *PARP* inhibitors (Sakai *et al.* 2008). Tumor cells can also enhance the activity of pathways repairing damages induced by alkylating chemotherapeutic agents, the so-called bulky lesions, by for instance overexpressing *ERCC1*, a key gene of NER, as evidenced by poor outcomes under cisplatin in lung cancer patients with higher expression of this gene (Ceppi *et al.* 2006).

A fourth mechanism contributing to drug resistance involves the inhibition of *apoptotic* pathways or activation of *prosurvival* pathways. Alterations in the equilibrium between pro- and anti-apoptotic factors enable tumor cells to evade apoptosis, a common objective of both conventional and innovative anticancer drugs. Overexpression of anti-apoptotic proteins within the *BCL2* family or their associated transcription factors, notably *NF-kb* and *STAT3*, represents prevalent mechanisms associated with the deregulation of apoptosis and the development of drug resistance, as discussed by Koren & Fuchs (2021). As a consequence, inhibitors targeting anti-apoptotic proteins of the *BCL2* family have been investigated, as exemplified by navitoclax, for which approval was, however, withheld due to severe side effects, or venetoclax, approved in 2019, for the treatment of adult patients with CLLs and older adult patients with AML (Juárez-Salcedo *et al.* 2019). The inactivation of *TP53*, observed in half of all cancers, or other related checkpoints, serves as a facilitator of genomic instability and resistance to DNA-damaging agents such as cisplatin, temozolomide, or doxorubicin. This resistance extends to targeted treatments due to the existing interactions between p53 and specific targets, such as *ER* for hormone therapy (e.g., tamoxifen) and *EGFR* for the monoclonal antibody cetuximab, as reviewed by Hientz *et al.* (2017). Interestingly, although small molecules capable of restoring the wild-type conformation of p53 have been identified, only few have demonstrated favorable pharmacologic profiles, and none have yet reached regulatory approval. Furthermore, the activation of *autophagy*, a process inherent in cells under metabolic stress, represents another prosurvival pathway exploited by tumor cells to resist drug-induced cell death. Most notably, the inhibition of autophagy with chloroquine or hydroxychloroquine in combination with cytotoxic drugs has shown promising results in preclinical models and early-phase clinical trials, as discussed in a review by Low *et al.* (2023).

A fifth and pivotal mechanism of resistance involves alterations that *reactivate the targeted pathway* or *upregulate a parallel pathway*, enabling cells to bypass the consequences of inhibited signaling. The MAPK-RAS-ERK and PI3K-AKT-mTOR signaling pathways play central roles in the response and resistance to targeted therapies. Reactivation of the targeted pathway can occur through various mechanisms, such as inactivating downstream phosphatases or activating upstream or downstream kinases. In diverse cancer types, reactivation of MAPK pathway signaling following inhibition has been documented through distinct mechanisms, including

KRAS mutations in *EGFR* inhibitor-exposed lung cancers (Pao, T. Y. Wang, *et al.* 2005), *RAS/MEK1* mutations or amplifications in *BRAF* inhibitor-resistant melanomas (Nazarian *et al.* 2010), squamous cell carcinomas (Su *et al.* 2012), and colorectal cancers (Ahronian *et al.* 2015), as well as RAS mutations or *BRAF* amplification in *MEK* inhibitor resistance (Caunt *et al.* 2015). Supporting RAS-mediated reactivation of MAPK signaling, the synergistic effect of the *EGFR* inhibitor cetuximab with irinotecan chemotherapy observed in *KRAS*-wild-type colorectal cancers is absent in *KRAS*-mutant cases. Regarding the PI3K-AKT-mTOR pathway, reactivation-mediated mechanisms of resistance remain less elucidated. Although loss of *PTEN* has been linked to resistance to PI3K inhibitors (Juric *et al.* 2015), *PTEN* loss alone does not induce resistance to class I PI3K inhibitors (Yang *et al.* 2019). Tumor cells frequently exhibit overexpression of redundant oncogenic signaling through multiple kinases. Cross-talk and compensatory feedback between the MAPK-RAS-ERK and PI3K-AKT-mTOR signaling pathways counteract the therapeutic effects of inhibiting tyrosine kinases in either pathway. Indeed, the upregulation of MAPK-RAS-ERK signaling leads to adaptive resistance to PI3K, *AKT*, and *mTOR* inhibitors, and conversely, the upregulation of the PI3K-AKT-mTOR pathway results in adaptive resistance to *EGFR* (Sergina *et al.* 2007), *BRAF* (Coffee *et al.* 2013), and *MEK* (Wee *et al.* 2009) inhibitors from the MAPK-ERK pathway. Inhibition of *EGFR* (MAPK-RAS-ERK pathway) may for instance be circumvented by *MET* receptor tyrosine kinase amplification or overexpression of its ligand *HGF* growth factor (Engelman *et al.* 2007; Ko *et al.* 2017), activation of PI3K-AKT signaling through *ERBB2* amplification or activation of *ERBB3* via upregulation of its ligand heregulin (Sergina *et al.* 2007; Yonesaka *et al.* 2011), or overexpression of *AXL* receptor tyrosine kinase (Zhang *et al.* 2012), among other bypass mechanisms (L. Huang & L. Fu 2015). These findings have prompted the evaluation of combined inhibition of multiple kinases in both preclinical and clinical models. While promising results have been observed (Stommel *et al.* 2007; Flaherty *et al.* 2012; Corcoran *et al.* 2018), the concern over increased toxicity is substantial, and synergies between drugs are infrequent (Jaaks *et al.* 2022).

Lastly, changes in cell *phenotype* or the composition of the *TME* have also been associated with treatment resistance. Epithelial-mesenchymal transition (EMT) is a process involving the dedifferentiation of epithelial cells toward motile states with invasive and metastatic capabilities. This process is induced by cytokines *IL8* and *TGF-*β, which activate downstream signaling pathways and transcription factors from three main families: *SNAIL*, *TWIST*, and *ZEB*. Associations between EMT and chemotherapy resistance have long been described across multiple cancer types, but mechanistic explanations of this link remain elusive as reviewed by Du & Shim (2016) and Ramesh *et al.* (2020). Given the association of EMT with drug resistance, inhibition of *TGF-*β signaling has been described as a promising avenue in preclinical models (S. Huang *et al.* 2012). Still, clinical development is progressing slowly, and several programs have been discontinued (C.-Y. Huang *et al.* 2021). Likewise, inhibition of *AXL* alone (Asiedu *et al.* 2014) or concomitantly with *MEK* (Konen *et al.* 2021) has demonstrated restored sensitivity to chemotherapy or *MEK* inhibition in preclinical models, respectively. More recently, EMT phenotype has also been associated with resistance to immunotherapies (Horn *et al.* 2020). Transitioned tumor cells have the ability to impact the TME by modulating the extracellular matrix (ECM), promoting T cell exclusion, and more

generally, promoting an immune suppressive environment through the upregulation of immune checkpoints *PD1* and *PD-L1* (Hsu *et al.* 2018). The TME, which in solid tumors consists of the ECM, cancer-associated fibroblasts (CAFs), and immune and inflammatory cells, also plays a pivotal role in drug resistance. CAFs can, for instance, promote drug resistance by secreting EMT-inducing cytokines, particularly *TGF-*β (Xu *et al.* 2009), releasing exosomes or modulating tumor immunity (Kalluri 2016). The targeting of CAF-mediated resistance through depletion, phenotype reversal toward a quiescent state, inhibition of CAF-secreted cytokines, or reduction of CAF-derived ECM proteins forming a protective barrier for the tumor have recently been reviewed by Mhaidly & Mechta-Grigoriou (2021). The immune cells of the TME have also been associated with drug resistance, such as cytotoxic CD8+ T-cells and tumor-associated macrophages in chemotherapy resistance (Wu *et al.* 2021), or immune suppressive cells in immunotherapy resistance (Sharma *et al.* 2017; Jenkins *et al.* 2018). The expression of growth factors acting as ligands to receptor tyrosine kinases targeted by drugs is another way through which cells can overcome the effects of drug-mediated inhibition. *HGF, FGF*, and *NGR1* were all shown to cause drug resistance in cell lines (T. R. Wilson *et al.* 2012). Once again, the description of all these resistance mechanisms suggests combination strategies using inhibitors of the adaptive resistance mechanisms alongside the original drug, but the efficacy of combination therapies needs to be balanced against their deleterious effects on the host (Wargo *et al.* 2015; Labrie *et al.* 2022). The interactions between the TME and therapeutic response are being continually reviewed as biological knowledge increases, experimental evidence accumulates, and new therapeutic strategies emerge (Tredan *et al.* 2007; McMillin *et al.* 2013; Q. Wang *et al.* 2023).

4.2. Annotations of genetic resistances

4.2.1. Methodology of annotation

The genetic variants identified through the analyses of WES and, to a lesser extent, RNA-seq, were compared against the contents of two widely recognized databases to describe events associated with cancer and treatment resistance. All somatic changes in the DNA molecule, encompassing point mutations (SNVs, MNVs, indels), focal CNAs (homozygous deletions and high-level gains, see Section 3.1.2.1 and 3.1.2.3), and gene fusions (Section 3.1.2.2), were included in the event list for characterizing cancer-driving and resistance-related biomarkers. Three additional specific biomarkers with established therapeutic implications were also considered: MSI and TMB quantified from WES, and the relative levels of the *AR*-v7 isoform among all *AR* isoforms quantified from RNA-seq. Notably, expression levels of cancer genes were excluded from consideration due to the complexity of determining over- or under-expression in samples derived from diverse cancer types and originating from various tissues. The possibility of identifying gene expression outliers is explored in the concluding section of this chapter.

4.2.1.1. OncoKB annotations

We first relied on the OncoKB knowledge database (Chakravarty *et al.* 2017; Suehnholz *et al.* 2023), briefly introduced in Sections 2.3.2.3 and 3.1.2.3 where it was used to determine the catalog of oncogenic events supporting the analyses on the META-PRISM cohort. As described by Chakravarty *et al.* (2017), OncoKB aims at providing an easily accessible clinical tool for distilling the current knowledge about the diagnostic, prognostic, and therapeutic implications of specific somatic alterations. It also describes the known or likely biological consequence and role in cancer for each of the alterations it includes. The database employs grading systems to quantify confidence in the strength of each relationship based on published evidence. Diagnostic and prognostic implications are categorized into three levels, while grading for therapeutic consequences differs between treatment sensitivities and resistances. Associations with treatment sensitivity have five levels of evidence, whereas associations with resistance fall into either FDA-approved or investigational levels. In the latest data release, v4.11, OncoKB incorporates 416 biomarker-tumor type-drug associations, with only 14% dedicated to resistance. This asymmetry highlights the disparity in our understanding of treatment sensitivities versus resistances. The associations are continuously updated with emerging evidence, as exemplified by the *BRAF* V600E mutation, previously a level-1 sensitivity marker only for anaplastic thyroid cancer, melanoma, and NSCLC before receiving FDA approval for all solid tumors except colorectal cancer in June 2022 and being incorporated as such in OncoKB release of September 2022.

The development and maintenance teams of OncoKB database have conveniently developed an API and Python scripts to allow anyone to add OncoKB annotations to their data programmatically, provided it adheres to the specified format constraints. In order to alleviate the computational burden and avoid unnecessary calculations, the list of all detected somatic mutations and focal CNAs (high-level gains and homozygous deletions, see Section 3.3.2) were intersected with the publicly available list of OncoKB-annotated genes[2]. As the annotations of somatic events in META-PRISM samples was conducted in early 2022, the data release from January 2022 was utilized, featuring 685 OncoKB-annotated genes. This number has subsequently been expanded and now reaches 820 genes in the latest release from November 2023.

Annotations for the three categories of somatic events (mutations, CNAs, gene fusions) were executed using dedicated Python scripts available on the GitHub repository[3]: `MafAnnotator.py`, `CnaAnnotator.py`, and `FusionAnnotator.py`. Notably, as annotations are dependent on tumor type, we supplied tumor type information for all cases in the oncotree nomenclature, which OncoKB relies on. As mentioned in Section 3.1.2.3, we painstakingly established tables for navigating between TCGA types, MSK's oncotree, and CIViC designations. These tables proved to be instrumental for running external tools like `oncokb-annotator` and for sharing our data in standardized formats. The scripts from `oncokb-annotator` conveniently generate data tables with constant number of rows and additional columns for describing the biological and oncogenic functions of submitted

[2] https://www.oncokb.org/cancer-genes
[3] https://github.com/oncokb/oncokb-annotator

alterations, as well as every level of the prognostic, diagnostic, and therapeutic grading scales. Oncogenic roles reported by OncoKB are one of `oncogenic`, `predicted oncogenic`, `likely oncogenic`, `neutral`, `likely neutral`, `inconclusive`, or `unknown`, while biological effects are categorized as `gain-of-function`, `likely gain-of-function`, `loss-of-function`, `likely loss-of-function`, `switch-of-function`, `likely switch-of-function`, `neutral`, `inconclusive`, or `unknown`. Alterations with both biological and oncogenic roles reported as `unknown` or `(likely) neutral` were excluded. In the META-PRISM (resp. MET500 and TCGA) cohort, OncoKB-annotated events represented 1.3% (resp. 1.2% and 1.2%) of all somatic mutations, 0.10% of all focal gene CNAs (resp. 0.09% and 0.16%), and 44% of all somatic cancer gene fusions (resp. 38% and 15%). Additionally, 33% (resp. 24%, 28%) of annotated mutations, 17% (resp. 17%, 16%) of annotated CNAs, and 33% (resp. 27%, 25%) of annotated gene fusions had a therapeutic implication in META-PRISM (resp. MET500 and TCGA). Figure 4.2 depicts the overlap between all OncoKB annotations and the ones with therapeutic implications in each of the three types of alteration in META-PRISM data.

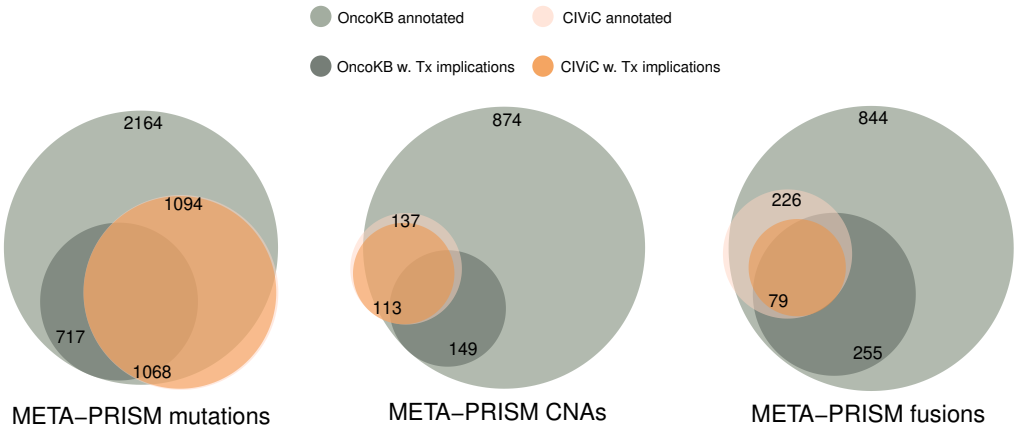


Fig. 4.2.: Overlap in META-PRISM data between OncoKB and CIViC annotations, including subsets with therapeutic implications, for each of the three category of somatic alterations.

4.2.1.2. CIViC annotations

The CIViC database, a widely recognized knowledge resource, is the second database we employed for annotating mechanisms of resistance in our study (Griffith *et al.* 2017). Similarly to OncoKB, CIViC aims to curate a comprehensive database that aggregates diagnostic, prognostic, and therapeutic implications of genomic alterations in cancer patients. CIViC also employs a grading system to convey the confidence of the expert curation team regarding the clinical significance of a variant based on evidence from published studies. While sharing similar goals with OncoKB, CIViC differs in structure and more faithfully reflects the complexity of current literature, where identical phenotype-genotype relationships may

be supported or contradicted by different studies. For instance, evidence id 977[4] suggests that *EGFR* amplification in NSCLC does not confer sensitivity to erlotinib or gefitinib, with a CIViC-assigned confidence level B. Conversely, evidence id 5924[5] supports the opposite conclusion with an identical confidence level.

As for OncoKB, CIViC exhibits a bias toward positive therapeutic implications over negative ones, i.e., resistance implications. This database bias was acknowledged in the seminal publication by (Griffith *et al.* 2017). In January 2017, CIViC contained 1,625 curated interpretations for 713 variants affecting 283 genes, with 70% relating to sensitivity and 30% to resistance. By November 2023, the curated interpretations had grown to 4,198, with a 60%-40% split between sensitivity and resistance, involving 1,683 variants in 350 genes an over two-fold increase in database size over six years.

Unlike OncoKB, CIViC lacks a tool or script for automated interaction with its API to annotate custom variant lists. Consequently, we developed an annotator akin to the `oncokb-annotator` for OncoKB. This tool, mirroring its counterpart, can annotate mutations, CNAs, and gene fusions using data formats comparable to those provided to `oncokb-annotator`. The annotator requires tumor type annotations for each submitted case to enable precise annotations based on tumor type specific relationships curated by CIViC. Python scripts were developed from scratch for this annotation, and close collaboration with Dr. Nikolaev was critical for correcting some evidences when necessary and preparing the table for automatic processing. Initial annotation rounds indeed revealed unexpected associations that were subsequently considered errors, such as evidence id 91 associating *BRAF* V600E mutation with resistance to dabrafenib, which, upon careful examination, is linked to *KRAS* G12D mutation in *BRAF*-mutant patients according to the paper supporting the evidence (Rudin *et al.* 2013). Similarly, evidence id 1775 erroneously claims that *BRCA1* mutations in breast cancer are associated with resistance to the *PARP* inhibitor olaparib, whereas the study only reported that no response to olaparib was observed in breast cancer patients regardless of *BRCA1* status[6].

The challenge with using CIViC also stems from the lack of standardized entries for variant descriptions or tumor types. We addressed the tumor-type matching issue by using our exhaustive tables to navigate between nomenclatures (Section 3.1.2.3). For variant entries, the `CivicAnnotator` tool allows flexible matching of patterns, considering genomic coordinates, variants in HGVS protein or coding formats, and accommodating partial matches when only amino acid positions are specified. Exotic descriptions were manually reformatted, and genomic coordinates of specific alterations were completed. Notably, the exon boundaries of all variants describing an exon number were completed. Additionally, each genotype-phenotype relationship was classified into either a mutation (m), a CNA (c), or a gene fusion (f). Some relationships involve multiple alterations that may combine alterations from different classes. We also classified these cases using combinations of the three letters (m), (c), and (f) and gave the annotator the capability to consider such combinations of

[4] https://civicdb.org/evidence/977/summary
[5] https://civicdb.org/evidence/5924/summary
[6] https://civicdb.org/evidence/1775/summary

events in the annotation process. However, the therapeutic implications supported by such combinations were not included in the analyses of the META-PRISM study.

The `CivicAnnotator` tool, available on my GitHub page at https://github.com/ypradat/CivicAnnotator, was utilized to annotate META-PRISM patients, and results were crosschecked against OncoKB annotations. The agreement between the annotations from the two databases was excellent, as depicted in Figure 4.2, enhancing our confidence in the reliability of our work. Figure 4.2 also illustrates that most events annotated in CIViC are also annotated in OncoKB, particularly mutations, but also that a significant number of alterations are only annotated by OncoKB. However, a non-negligible number of somatic mutations were annotated only by our annotator. Manual inspection revealed many instances of matching arising from unspecific variant descriptions in CIViC tables. As examples are *EGFR* exon 19 frameshift deletions in lung cancer patients, which are not annotated by OncoKB (only inframe deletions are) but which were matched to the CIViC molecular profile `EGFR Exon 19 Deletion` by the annotator. Due to this observation, mutations annotated only by CIViC were not considered further in our analysis of treatment resistances.

In the META-PRISM (resp. MET500 and TCGA) cohort, CIViC-annotated events represented 0.7% (resp. 0.5% and 0.5%) of all somatic mutations, 0.02% (resp. 0.01% and 0.02%) of all focal gene CNAs, and 9% (resp. 8%, 2%) of all somatic cancer gene fusions. Additionally, 98% (resp. 97%, 95%), 82% (resp. 91%, 88%), and 57% (resp. 24%, 27%) of CIViC-annotated somatic mutations, focal gene CNAs, and gene fusions had a therapeutic implication in META-PRISM (resp. MET500 and TCGA).

4.2.1.3. ESCAT tiers and emerging markers

As mentioned previously, OncoKB classifies therapeutic implications using five-level and two-level confidence scales for treatment sensitivities and resistances, respectively. CIViC employs similar grading systems but uses five-level confidence scales for both directions. To create a unified system, the grading systems of both databases were harmonized according to ESCAT guidelines (Mateo *et al.* 2018). ESCAT, developed through a collaborative effort led by ESMO, is a ranking system designed to establish a standardized vocabulary for interpreting genomic reports and describing the clinical significance of alterations. In the United States, a multidisciplinary working group from AMP/ASCO/CAP proposed a highly similar tier classification system shortly before ESCAT was introduced in Europe (M. M. Li *et al.* 2017).

The ESCAT ranking system is the result of an ESMO-led collaborative project that aimed at developing a harmonized vocabulary for interpreting genomic reports and describing the clinical value of detected alterations. Since its introduction, the ESCAT classification has gained significant attention and is now utilized to inform clinical decisions. Notably, ESCAT scores have been incorporated over the last years into the ESMO treatment recommendations for various cancers. Another exemplary demonstration of the impact of the ESCAT scale is the systematic reimbursement in Italy of targeted therapies supported by genetic alterations

with an ESCAT tier 1 level of confidence since February 2023 [7].

In the META-PRISM study, sensitivity and resistance confidence scales from OncoKB and CIViC were harmonized into the first three tiers of the ESCAT ranking.

ESCAT Tier 1 level, which includes standard-of-care biomarkers already used in clinical practice, was used for alterations falling into one of

- CIViC Level A (proven/consensus association in human medicine)
- OncoKB Level 1 (FDA-recognized biomarker predictive of response to an FDA-approved drug in this indication)
- OncoKB Level 2 (standard-of-care biomarker recommended by the National Comprehensive Cancer Center or other professional guidelines predictive of response to an FDA-approved drug in this indication)
- OncoKB Level R1 (standard-of-care biomarker predictive of resistance to an FDA-approved drug in this indication)

ESCAT Tier 2 level (investigational), which designates investigational targets that likely define a patient population that benefits from a given drug but for which additional data are needed, encompassed alterations with a therapeutic implication among

- CIViC Level B (clinical trial or other primary patient data support association)
- OncoKB Level 3A (compelling clinical evidence supports the biomarker as being predictive of response to a drug in this indication)
- OncoKB Level R2 (compelling clinical evidence supports the biomarker as being predictive of resistance to a drug)

Lastly, ESCAT Tier 3 level (hypothetical) was set to include all targets that have demonstrated a clinical impact on other tumor types or that are supported by scarce data (case reports). These were identified as:

- CIViC Levels C, D, and E (supported by case study, preclinical, and inferential data, respectively)
- OncoKB Level 3B (standard-of-care or investigative biomarker predictive of response to an FDA-approved drug in another indication)
- OncoKB Level 4 (compelling biological evidence supports the biomarker as being predictive of response to a drug)

An additional annotation tier was introduced to our analyses to encompass biomarkers that had not been incorporated into the two knowledge databases we relied on at the time of the study. The annotation of drug resistances or sensitivities associated with these *emerging* markers was conducted by a cancer genomics expert possessing extensive clinical experience, providing a nuanced understanding of the intricate relationships between genomic alterations and responses to antineoplastic drugs, particularly targeted therapies. All reported claims regarding the associations between therapy responses and *emerging* biomarkers in the META-PRISM study are supported by at least one peer-reviewed publication available in the current

[7] https://www.esmo.org/policy/esmo-scale-for-clinical-actionability-of-molecular-targets-escat

literature, as detailed in supplementary table 10 accompanying the paper. Importantly, the literature search was biased towards events observed in patients that are known to be resistant to the treatment of interest, effectively disregarding potential emerging markers not detected in our patients.

Of note, to maintain consistency with the analysis of oncogenic events detailed in Chapter 3, only therapeutic implications backed by alterations involving a gene listed in the 360-genes catalog of cancer drivers from Section 3.1.2.3 were retained for the subsequent analyses presented in the following section.

4.2.2. The current knowledge gap

We used the OncoKB and CIViC annotations harmonized against the ESCAT ranking system to describe the frequency at which therapeutic implications, particularly indications of treatment resistance, of each tier were encountered across the 10 tumor types of the META-PRISM WES subcohort and in each of three cohorts analyzed.

At the patient-level, we observed at least one tier 1 resistance and one tier 1 sensitivity biomarker in 9.6% and 47.5% of META-PRISM WES and RNA-seq tumors, respectively (Figure 4.3 and A.14), whereas biomarkers of any level were found in 74.9% and 88.4% of these tumors, respectively. Tier 1 resistance biomarkers were detected in only three tumor types: LUAD (*EGFR* 17%), LUSC (*EGFR* 5%), and COAD (*KRAS* 65%, *NRAS* 6%), whereas investigational Tier 2 and hypothetical Tier 3 biomarkers were rather frequent in the majority of tumor types (Figure 4.3). META-PRISM WES and RNA-seq tumors harbored significantly more resistance biomarkers of all three tiers compared with TCGA: Tier 1, common odds ratio (cOR) 7.5 [CI 95%, 3.715.2, CochranMantelHaenszel test stratified by tumor type]; Tier 2, cOR 1.7 (CI 95%, 1.32.2); and Tier 3, cOR 2.2 (CI 95%, 1.72.8). The increase was also replicated in MET500 for all three tiers (Tier1 cOR 4.7; CI 95%, 1.317.3; Tier2 cOR 2.2, CI 95%, 1.62.9; Tier3 cOR 3.2, CI 95%, 2.44.3). Some common sensitivity biomarkers were frequent in META-PRISM and often enriched versus TCGA, most notably *EGFR* L858R/exon 19 del (first- and second-generation *EGFR* inhibitors), *EGFR* T790M (third-generation *EGFR* inhibitors), and *ALK* oncogenic mutations or fusions (*ALK* inhibitors) in LUAD; *PTEN* fusions or loss-of-function mutations (mTOR inhibitors) in PRAD; and *FGFR3* p.R248C, p.S249C, p.G372C, and p.Y375C mutations (*FGFR* inhibitors) in BLCA. On top of that, we detected 13% of hypermutated tumors (>10 mut/Mb) and 3% of MSI tumors in META-PRISM, which are indications for treating patients with immunotherapies (Figure A.14).

At the treatment level, we described all the biomarkers that could explain the treatment resistances diagnosed in each patient. Unfortunately, we could identify resistance mechanisms only for a minority of treatments (Figure 4.4A-G). More specifically, standard-of-care resistance events were detected only for first- and second-generation *EGFR* inhibitors in LUAD for 40% of treated patients and *EGFR* antibodies in COAD for one in five treated patients (Figure 4.4A and D). In contrast, investigational and hypothetical resistance markers substantially increase the fraction of observed treatment resistances that may be explained. In LUAD, Tier 2 resistance alterations were identified in 8 of 28 patients treated with third-generation *EGFR*

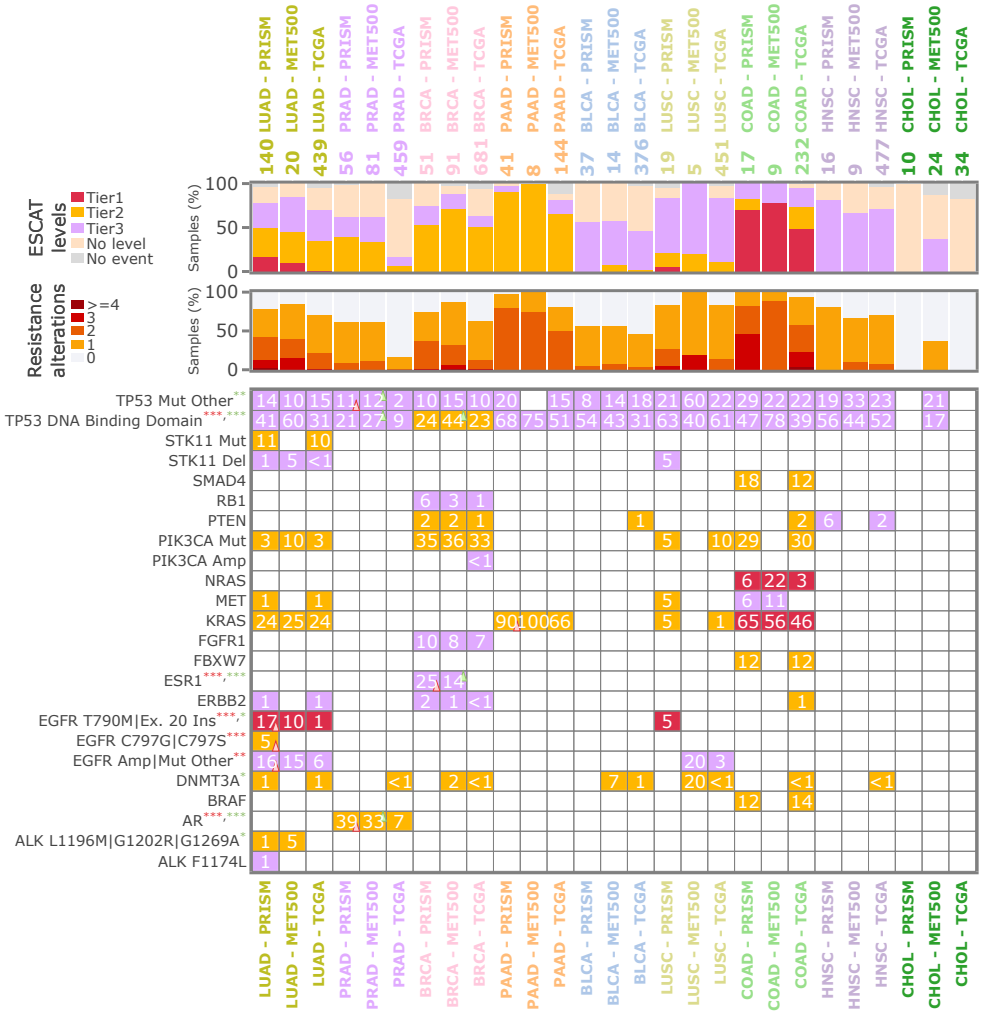


Fig. 4.3.: Known genetic markers of treatment resistance in META-PRISM, MET500, and TCGA by tumor type. **Top**, fractions of tumors harboring resistance markers split by tier (only the best tier is shown for each tumor). **Middle**, fractions of tumors with multiple resistance markers. **Bottom**, heat map showing the most frequent resistance-associated variants. Triangle orientations (increase - triangle points up, decrease - points down) and colors (red for META-PRISM vs. TCGA, green for MET500 vs. TCGA) highlight significant changes in prevalence. Similarly, stars next to the gene alterations represent significant changes at the cohort level using the same color code as for triangles (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$). P-values per tumor type are from Fisher-Boschloo tests, and p-values across the cohort are from Cochran-Mantel-Haenszel tests. All p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.

inhibitors, in 2 of 17 patients who progressed on first- or second-generation *ALK* inhibitors, and in 5 of 27 patients who progressed on immunotherapies (Figure 4.4A). BRCA patients treated with hormone therapy harbored Tier 3 resistance through ESR1 mutations in 33% of cases (Figure 4.4B). PRAD patients treated with hormone therapy harbored Tier 2 resistance through *AR* alterations in 31% of cases, 25% of which were the overexpression of *AR*-V7 splice isoform (Figure 4.4C). In COAD, Tier 2 resistance alterations could explain resistance to *EGFR* antibodies in two of five treated patients (Figure 4.4D). In PAAD, we did not identify any resistance markers for the drugs patients received. Additionally, we were able to associate *emerging* resistance mechanisms with a considerable fraction of resist- ance to targeted and hormonal therapies. For example, 13%, 32%, and 67% of LUAD patients treated with first- or second-generation *EGFR* inhibitors, third-generation *EGFR* inhibitors, and BRAF inhibitors (dabrafenib), respectively, harbored *emerging* resistance mechanisms with literature support (Figure 4.4A). Interestingly, one out of the two patients who progressed on lorlatinib (third-generation *EGFR* inhibitor) harbored an *EML4-ALK* fusion and a double *ALK* mutation (p.F1174L and p.G1202R) as previously described (40). In HNSC, we associated driver mutations in *EGFR*, *NF1*, *PIK3CA*, and *PTEN* to the *EGFR* inhibitor cetuximab (Figure 4.4E). In BLCA and CHOL, we were able to associate mutations in *FGFR1, FGFR2, PIK3CA, KRAS*, and *TSC1* with resistance to *FGFR* inhibitors (Figure 4.4F and G). The observations above show that standard-of-care *resistance biomarkers were associated with only 1.6% of received treatments*, whereas investigational, hypothetical, and *emerging* mechanisms could further explain 2.7%, 2.3%, and 7% of resistance, respectively. Strikingly, except for *BRAF* V600E mutation in COAD, no known or *emerging* mechanism of resistance was observed in our cohort for chemotherapies or antiangiogenic treatments, though these two types of treatments account for 54% of all treatments administered.

4.3. Two studies of genetic resistances to innovative drugs

4.3.1. Trastuzumab deruxtecan in breast cancers

4.3.1.1. The DAISY trial

Antibody-drug conjugates are, as mentioned in the introductive Section 4.1.1, an emerging class of antineoplastic drugs that deliver cytotoxic agents much more selectively than conventional chemotherapies. This selectivity makes ADC a promising type of targeted therapy and is made possible by the bioengineering of molecules linking an antibody that serves to select specific cells to a payload that delivers the cytotoxic effect upon release in targeted cells. Trastuzumab deruxtecan (T-DXd), also known as DS-8201a or enhertu, is an antineoplastic drug combining the long-known *HER2*-targeting trastuzumab antibody with a

[8] **A**. The left bar plot indicates the percentage of LUAD patients in the META-PRISM WES and RNA-seq subcohort that received each treatment or group of treatments. Middle error bar plot represents the OR and 95% CI estimates for the enrichment of resistance markers of Tier 1, Tier 2, and Tier 3 or extracted from the literature in the patients who received the corresponding treatment. The right group of four bar plots provides details about the identity of the markers associated with resistance in treated patients at each of the four confidence levels. Del, deletion. **BG**, Same as **A** but for BRCA, PRAD, COAD, HNSC, BLCA and CHOL patients, respectively.

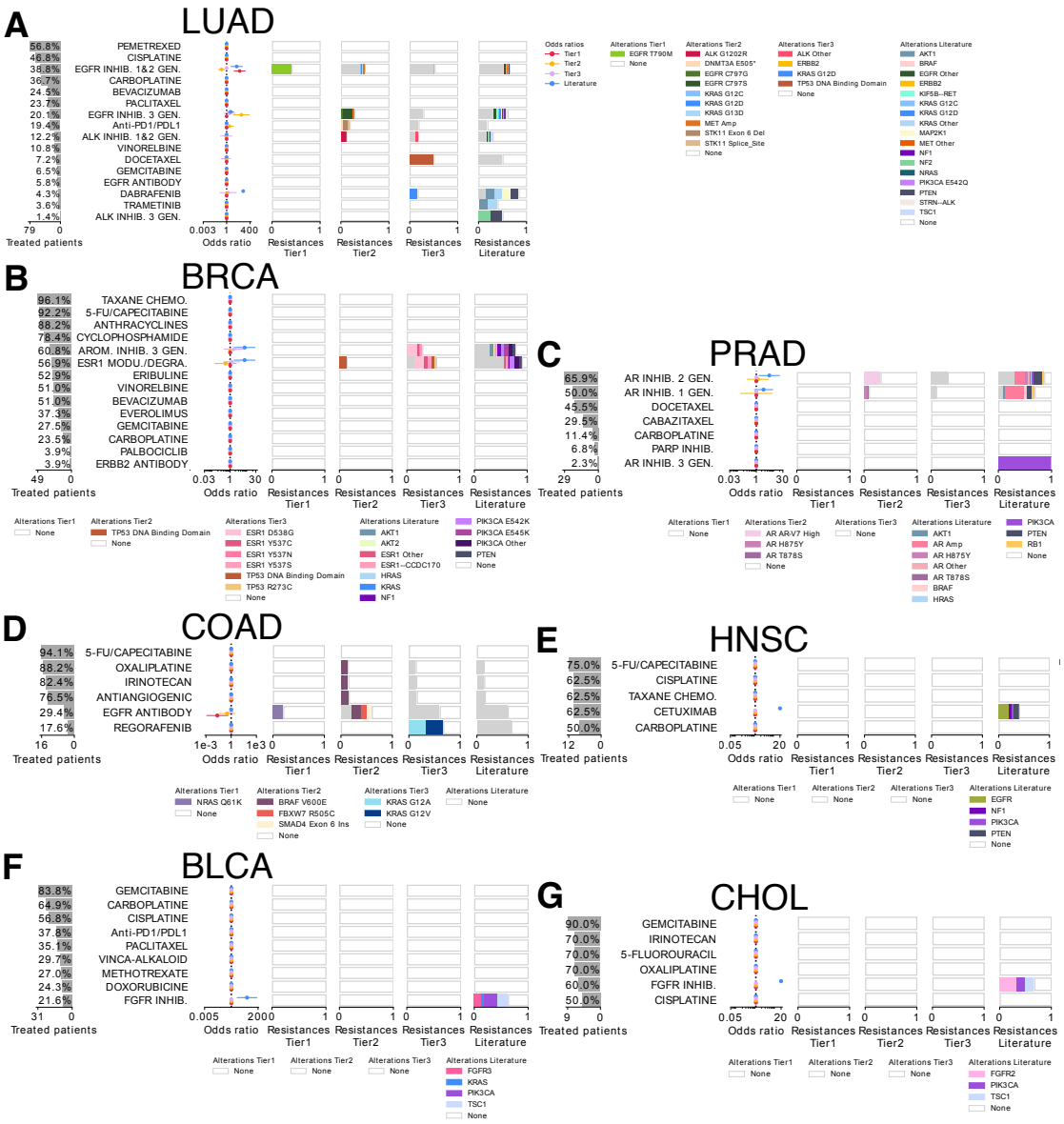


Fig. 4.4.: Associations between treatments received and molecular markers in META-PRISM patients. [8]

new topoisomerase I inhibitor deruxtecan, a new member in the topoisomerase I inhibitors family alongside irinotecan, topotecan, or exatecan. The trastuzumab antibody is attached to deruxtecan payload by a tetrapeptide-based linker which is cleaved specifically by lysosomal enzymes found in cells.

T-DXd received FDA-accelerated approval in December 2019 following the positive results of the phase 2 multicentric DESTINY-Breast01 clinical trial in which 184 *HER2*-positive, pretreated (at least two lines of anti-*HER2* therapies) and metastatic breast cancer patients demonstrated an overall response rate of 60%. The phase 3 confirmatory trial DESTINY-Breast03, which enrolled over 500 *HER2*-positive breast cancer patients with similar inclusion criteria, demonstrated the superiority of T-DXd over trastuzumab emtansine (T-DM1), the first approved ADC for breast cancers, in 2013, and the then standard-of-care for second-line therapy in metastatic HER2-positive breast cancers until that moment (Cortés *et al.* 2022). These results led to the FDA approval of T-DXd in the settings of the DESTINY-Breast03 trial in May 2022 in the second line and beyond, replacing T-DM1. The results of the complementary phase 3 study DESTINY-Breast04, comparing T-DXd to standard chemotherapy in *HER2-low* - defined as IHC1+ or IHC2+/in situ hybridization (ISH)-, unresectable or metastatic breast cancer patients, led to the FDA approval of the drug in this hard-to-treat category of breast cancer patients in August 2022 (Modi *et al.* 2022).

The DAISY trial (NCT04132960) is a phase 2 French multicentric study aiming to assess the efficacy of T-DXd across all *HER2* spectrums of advanced breast cancers. Unicancer sponsored and coordinated the study, while Gustave Roussy provided patients as one of the 21 participating centers and centralized all collected biospecimens. The study included a total of 186 patients subdivided into three cohorts based on the expression level of *HER2* on cancer cells as measured by IHC or ISH assays (Figure 4.5). Apart from the traditional *HER2*-positive and *HER2*-negative (IHC 0) groups, the third group included so-called *HER2-low* breast cancers as defined in the DESTINY-Breast04 study, i.e., patients classified IHC1+ or IHC2+/ISH-. Of note, the relevance of this new clinical entity of breast cancers has been the subject of much debate recently[9].

[9] https://www.aacr.org/blog/2023/02/01/sabcs-2022-the-uncharted-territory-of-her2-low-breast-cancer/

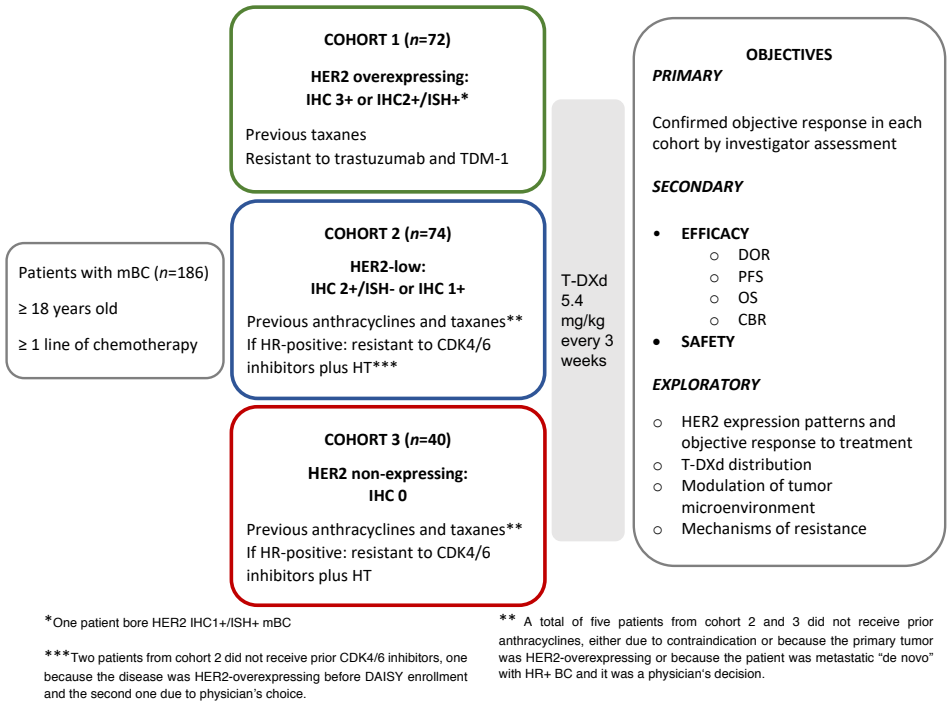


Fig. 4.5.: Overview of the patients enrolled in the DAISY-trial.

DAISY was conceptualized to explore various questions regarding the effectiveness of T-DXd in advanced breast cancers, with a primary focus on unraveling the mechanisms of drug sensitivity and resistance. Despite the enhanced survival outcomes associated with T-DXd in advanced breast cancers, most of the patients inevitably face disease progression during treatment and, ultimately, death. The search for biomarkers influencing patient responses to the drug relied on diverse data modalities extracted from tissue and blood biopsies sampled at baseline, on-treatment, or upon disease progression. The results from the clinical and the investigation biomarkers were published in 2023 in Nature Medicine (Mosele *et al.* 2023).

Firstly, *HER2* IHC assays were performed at baseline and resistance for twenty-five patients, including five that were IHC 0 at baseline. The comparison in HER2-expressing at baseline revealed that 13 patients exhibited a decrease in *HER2* IHC status at resistance. Subsequently, digitalized *HER2* slides from FFPE tumor biopsy samples were algorithmically analyzed. Only baseline tissue slides from patients in cohorts 1 and 2 (Figure 4.5) were included. In a nutshell, an AI model-assisted unsupervised clustering analysis was applied to the *HER2*-stained slides to unveil spatial patterns that could distinguish between responders and non-responders. The AI model identified eight distinct patterns, each varying in proportion across individual slides. Analyses revealed a subtle correlation between the fraction of one of these eight clusters and drug response, though this association was observed exclusively within cohort 1. Lastly, fresh frozen baseline and progression tumor biopsies, as well as frozen baseline blood samples, were subjected to WES. The results are presented in the next section.

4.3.1.2. Drug response vs genotypes

Following stringent control of sequencing and sample qualities, most notably the tumor content, 110 tumor samples and 85 blood samples originating from 99 patients were included in the analysis as depicted in Figure 4.6. Of particular interest were the 21 WES profiles from samples collected at progression, with a focus on the 11 cases where a matched WES profile from the baseline biopsy was also available.

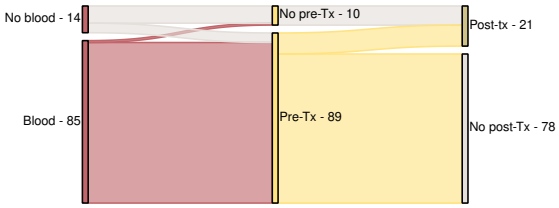


Fig. 4.6.: DAISY WES samples

Somatic substitutions, small indels, and CNAs were detected using the end-to-end WES pipeline described in Section 3.1.2.1. Changes were introduced in the pipeline to handle cases never encountered in META-PRISM, most notably the unavailability of a matched blood sample for some biopsies and the presence of longitudinal data. Most algorithms within the pipeline have the option to run in *tumor-only* mode, meaning without a paired normal sample. `Mutect2`, the mutation-calling algorithm, supports this mode with a note that additional filtering is necessary due to the highly-likely inclusion of germline events in the list of putative somatic mutations. On the other hand, `FACETS` does not directly allow the omission of a normal sample. Still, the documentation offers a workaround by utilizing a "fake" normal sample created by pooling normal samples from various unrelated patients. In this case, the user specifies the `-unmatched` option, altering the log odds-ratios calculations. However, the precise quantification of MSI is unfeasible without a matched normal sample, as it is crucial for accurately measuring variations in the length of microsatellites.

The presence of longitudinal samples called for additional pipeline modifications so as to carefully distinguish acquired or lost mutations from persisting ones. Specific mutation-calling rules were consequently applied to tumor samples from patients with WES available both at baseline and progression (11 patients). More particularly, for each baseline (and progression, respectively) sample, `SAMtools mpileup` version 1.9 (H. Li *et al.* 2009) was run on the positions where `Mutect2` identified and retained mutations in the corresponding progression (and baseline, respectively) sample to rule out incorrect claims of mutation acquisition or loss caused by conservative filtering or non-detection by `Mutect2`. If a mutation detected by `Mutect2` in a sample at a given timepoint was also seen in the sample from the other timepoint with sufficiently many reads supporting the alternative allele (at least one read if coverage $<$100, two reads if 100 $\leq$ coverage $<$ 500 and three reads if coverage $>$500), the mutation was also called in the latter sample. Additionally, in patients without a matched blood sample, any mutation identified as germline at any of the two timepoints was discarded from both samples. After all the filtering, 20,469 somatic substitutions and small indels were considered in the analysis of the 110 WES samples (89 at baseline, 21 at resistance).

In the quest for biomarkers predicting drug response, we examined the correlation between tumor genotypes at baseline and clinical responses. Patients were classified between confirmed responders if a complete or partial response was observed in one evaluation and confirmed in the subsequent evaluation, following the Response Evaluation Criteria in Solid Tumors

(RECIST) v1.1 guidelines; otherwise, they were considered non-responders. Figure 4.7 summarizes the landscape of known breast cancer-associated alterations in baseline samples, displaying only genes altered in at least three patients. The left-side bar plot facilitates a quick comparison between responders and non-responders, revealing no statistically significant markers of drug response, except for the expected *ERBB2* amplification, given the mechanism of action of T-DXd.

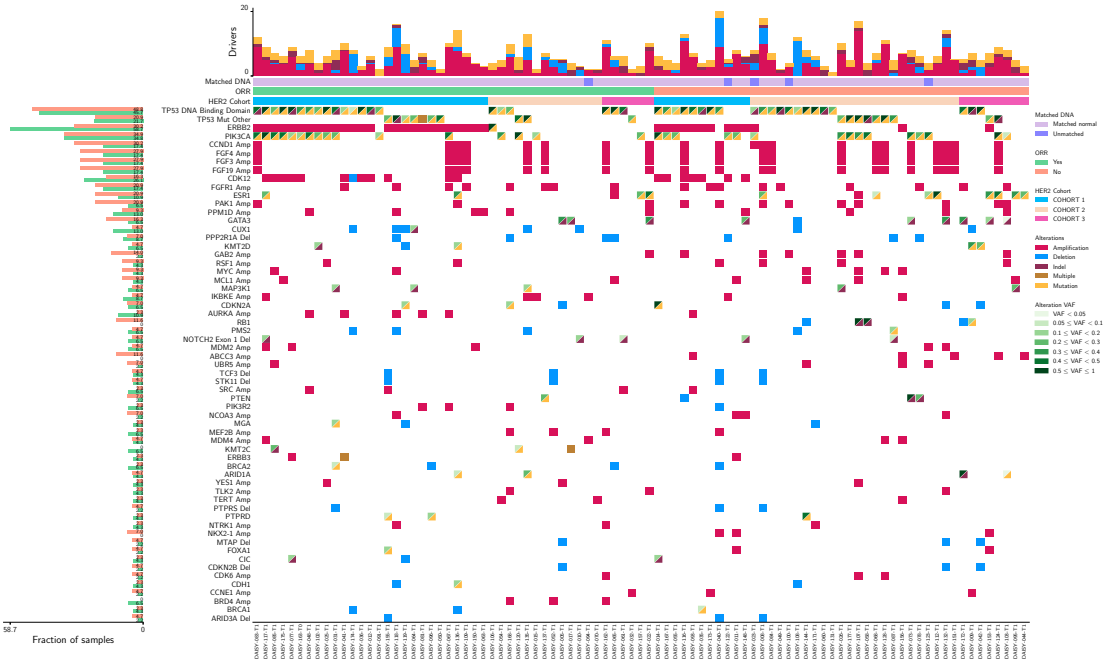


Fig. 4.7.: Oncoplot of driver mutations and CNAs identified in at least 3% of tumor biopsies at baseline (n = 89). Blood samples were available for analyses in 84 patients. If a gene has at least one driver mutation or CNA in at least 3% of pre-treatment biopsies, any other driver alteration of the same gene is shown, regardless of its frequency.

To explore the genetic mechanisms of acquired resistance, we analyzed changes in tumor genotypes upon progression under treatment. This analysis utilized paired pre- and post-treatment WES profiles from 11 patients with available data. Figure 4.8 summarizes the alterations in tumor genotypes, featuring genes that were unaltered in any of the 11 pre-treatment samples but acquired an alteration in at least two of the post-treatment samples at resistance (three samples in cases where all events were CNAs). None of the alterations displayed notable biological significance. Of note, 14% of the samples obtained at resistance to T-DXd (3 out of 21) presented an *SLX4* mutation. One of these mutations was not observed in the matched baseline sample, the second was present in the baseline biopsy, and for the third, the baseline biopsy was not available. Two of these mutations were classified as deleterious according to CADD and sorting intolerant from tolerant (SIFT); however, no evidence of loss of the second allele was found.

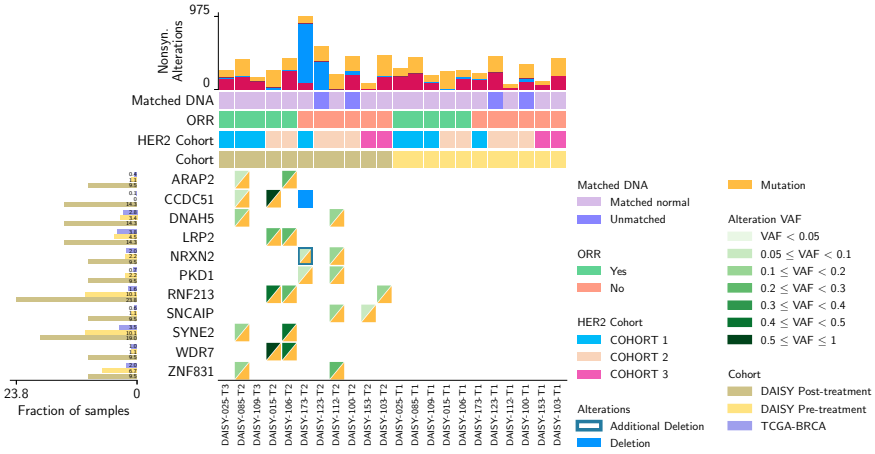


Fig. 4.8.: Oncoplot of acquired genetic alterations identified at resistance (n = 11). Eleven biopsies at resistance (on the left) were matched with pret-reatment biopsies (on the right) from the same patient. Only genes that were not altered in any of the 11 pretreatment samples and that acquired an alteration in at least two samples at resistance (three samples in case all events were CNAs) are shown.

4.3.2. FGFR inhibitors in urothelial cancers

*FGFR*s inhibitors are another important class of targeted therapies, and more particularly of TKIs, which act by blocking receptor tyrosine kinases of the four-member *FGFR* family: *FGFR1, FGFR2, FGFR3*, and *FGFR4*. They have been recognized early on as important therapeutic targets given their regulating role in fundamental cell signaling pathways and their recurrent alteration in cancer (Turner & Grose 2010). *FGFR* aberrations are indeed encountered in 7.1% of all cancers, mainly through amplifications, with strong enrichment in urothelial (32%), breast (18%), endometrial (18%), and squamous lung (13%) cancers (Helsten *et al.* 2016).

Erdafitinib was the first FDA-approved *FGFR* inhibitor following the positive results of the phase 2 trial, which reported a 40% rate of confirmed responses (Loriot *et al.* 2019). The drug was granted accelerated approval in April 2019 for treating advanced bladder cancers harboring actionable *FGFR2* or *FGFR3* alterations. One year later, pemigatinib became the second FDA-approved *FGFR* inhibitor for CHOLs with *FGFR2* rearrangement or gene fusion. Infigratinib and futibatinib were the third and fourth FDA-approved *FGFR* inhibitors, respectively, sharing analogous indications with pemigatinib. Infigratinib has, however, been discontinued by the manufacturer in March 2023 due to difficulties in enrolling patients for the confirmatory trial needed to receive full FDA clearance.

In bladder and upper tract urothelial cancers, oncogenic activation of the *FGFR* receptors occurs mainly through *FGFR3* mutations, and in 2 to 3% of cases through gene fusions, with *TACC3* being the most frequent partner. As of 2023, a very small number of studies had investigated the potential mechanisms of resistance to *FGFR* inhibitors, with the most notable being an analysis of ctDNA in 22 patients who progressed under infigratinib. This study reported acquired mutations in the tyrosine kinase domain of *FGFR3* at progression (Pal *et al.* 2018). We utilized the molecular profiles of the 47 patients successfully profiled through

WES out of the 56 with advanced or metastatic urothelial cancer included in MOSCATO or MATCH-R studies to draw the molecular landscape in non-localized disease, contrasting with other molecular portraits already reported. The global landscape is depicted in Figure 4.9 and shows no significant novelty compared with what was previously described.

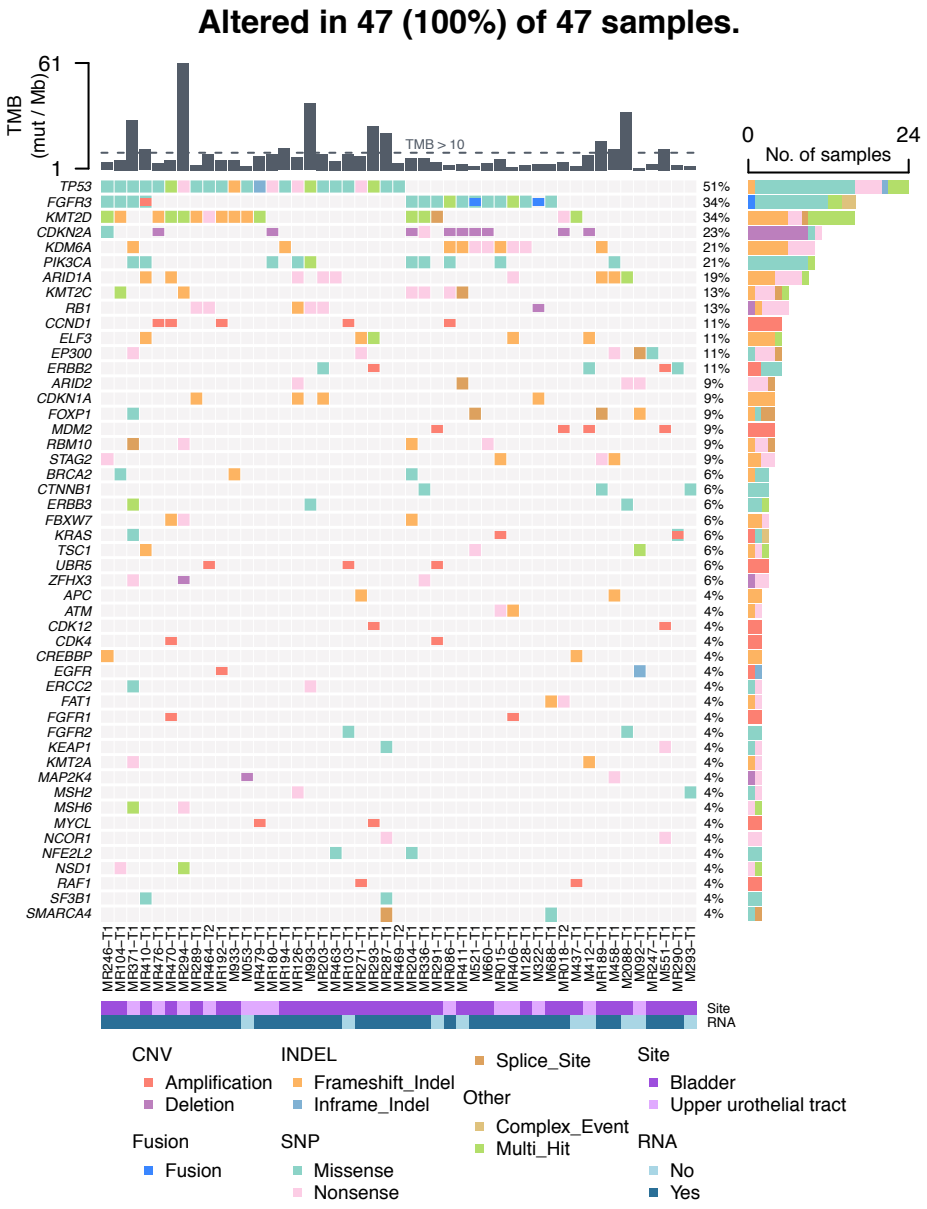


Fig. 4.9.: Mutational landscape of advanced or metastatic urothelial cancers included in MATCH-R or MOSCATO studies with WES performed.

In this context, we at Gustave Roussy undertook to investigate the genomic profiles of *FGFR*-driven bladder or upper tract urothelial cancer patients who received *FGFR* inhibitors

in one of three large precision medicine trials led at the institute (MOSCATO - NCT01566019, MATCH-R - NCT02517892, and STING - NCT04932525) to identify potential mechanisms of resistance to these novel therapeutic agents. Twenty-one patients with *FGFR*-driven urothelial cancer were included in the analyses. All patients had post-treatment samples available, analyzed either through WES (12/21), panel sequencing (3/21), or ctDNA (18/21). Five patients had longitudinal data available, including one with pre and post-treatment data from all three profiling techniques as well as RNA-seq. Nineteen of the 21 patients considered harbored alterations implicating *FGFR3*, with eleven having *FGFR3* S249C hotspot mutation, five *FGFR3::TACC3* gene fusion, and three *FGFR3* Y373C hotspot mutation. The two other patients harbored *FGFR2::FAM83H-AS1* gene fusion and *FGFR4* D276N mutation, respectively. Post-treatment samples were available upon progression to erdafitinib in fourteen cases, futibatinib in four cases, and pemigatinib for the three other cases. All patients presented with advanced disease and most received the *FGFR* inhibitor as a second line of treatment.

As discussed in Section 4.1.2, there are different categories of resistance mechanisms. For targeted therapies, the most commonly implicated events involve alterations either in the target itself or in proteins from the same pathway or a parallel pathway capable of circumventing the inhibitory effects. Firstly, among the 19 patients with tumors carrying *FGFR3* alterations prior to treatment, seven showed acquired mutations in the kinase domain of *FGFR3*, affecting amino acids in the kinase domain (N540, V553, V555, E589, or L608). Secondly, activating alterations in the PI3K-AKT-mTOR pathway were identified in 11 out of the 19 patients with *FGFR3*-altered tumors upon disease progression. Among them, three patients had pre-existing *PIK3CA* mutations before treatment, with one experiencing primary resistance. Conversely, two patients acquired *PIK3CA* mutations, resulting in the substitution of the amino acid glutamate (E) with lysine (K) at positions 545 and 726 in the protein sequence. While position 545 is a known hotspot, the 726 mutation, although less frequent, is not uncommon[10]. These mutations lead to a constitutively active form of the p110α protein, causing dysregulated signaling and heightened downstream activation of the PI3K pathway.

In addition to *PIK3CA*, the PI3K-AKT-mTOR pathway was implicated through inactivating mutations in *TSC1*, *TSC2*, or *PTEN*. Five out of the 19 patients with *FGFR3*-altered tumors demonstrated acquired nonsense or frameshift mutations in any of these three genes upon disease progression. Both the *TSC1-TSC2* protein complex and *PTEN* proteins act as negative regulators of the PI3K pathway. The *TSC1-TSC2* complex inhibits the *mTORC1* enzyme, a promoter of protein synthesis and cell growth, while *PTEN*, as a phosphatase, counteracts the activity of PI3K, primarily by dephosphorylating *PIP3* protein.

Additional evidence supporting the suggested resistance mechanism outlined in the previous paragraphs was obtained through experiments with cell lines. To validate the impact of mutations in the tyrosine kinase domain of *FGFR3*, Ba/F3 cells with *FGFR3::TACC3* mutation and the same mutations as observed in patients were exposed to various concentrations of multiple *FGFR* inhibitors. These experiments demonstrated that all tested *FGFR3* kinase

[10] https://www.oncokb.org/gene/PIK3CA

domain mutations required a substantial increase in drug dose to achieve control over cell growth, except for the V553L mutation. Furthermore, the hypothesized role of the *PIK3CA* E545K mutation in resistance was examined in a patient-derived xenograft model. This model revealed that the combined inhibition of *FGFR* and PI3K was necessary to inhibit tumor growth. Lastly, in another patient-derived xenograft model where no novel genetic aberration was detected upon progression, hyperphosphorylation of *EGFR* was identified. The driving role of *EGFR* was confirmed by the synergistic effect observed using a combination of *FGFR* and *EGFR* inhibitors.

4.4. Conclusions

Cellular plasticity is a formidable tool that our cells are equipped with to to adapt and transform into different states or phenotypes in response to environmental stresses or changes in conditions. This adaptability is crucial for the survival and normal functioning of cells in various physiological or pathological situations. It is, however, a double-edged sword, as cellular plasticity is also a critical enabler of cancer progression and treatment resistance. The first section has provided a brief overview of the variety of biological mechanisms cancer cells mobilize to escape the effects of cytotoxic therapies. Cancer cells demonstrate phenomenal plasticity, whether through depletion of drug-enabling enzymes, overactivation of pro-survival, DNA repair, or antiapoptotic pathways, reactivation or bypass of inhibited pathways, phenotypical changes, or microenvironment modifications. Intra-tumor heterogeneity is also a critical enabler of treatment resistance as it increases the chance that one of the many cancerous genotypes is equipped to genetically or phenotypically escape treatment. Although already diverse, the mechanisms we described are most probably an incomplete picture of treatment resistance. However, our understanding of resistance mechanisms will hopefully continue to improve as research progresses.

In the META-PRISM study, we took advantage of the clinical richness of the database we assembled to draw a detailed landscape of our current understanding of treatment resistances across a wide range of tumor types. Due to limited cohort sizes, only nine tumor types were investigated in depth, but many statistics encompass all the patients with complete molecular profiles available regardless of their tumor type. Whenever available, we harnessed WES and RNA-seq data to describe three types of somatic events fixed in DNA, namely mutations (substitutions and indels), focal CNA, and gene fusions. Given their known therapeutic importance, three additional specific biomarkers were also added: high TMB, MSI, and *AR*-v7 isoform levels. All these somatic events were compared against the contents of two high-quality knowledge databases, OncoKB and CIViC, in a tumor type-specific manner to comprehensively describe the currently known therapeutic implications among all detected alterations.

Annotated events implicating one of the 360 cancer genes we focused on represented 2066 events across the 485 META-PRISM patients with both WES and RNA-seq available. Mutations represented 67.7% of events but 91.8% of alterations with drug resistance implications.

The driver gene fusions and CNAs represented 9.4% and 18.8% of events, respectively, and were strongly enriched in META-PRISM tumors compared with TCGA tumors. However, 0% and 4.8% of these two types of events were associated with treatment resistance, respectively, reflecting the scarcity of current knowledge about genetic events conferring drug resistance.

Our research also reveals that current annotations of resistance mechanisms can only account for a small proportion of all observed treatment resistances, as resistance markers were found in 74.9% of patients, but for patients who received a given treatment, clinically-validated markers could explain only 1.6% of all resistance; investigational, hypothetical, and *emerging* mechanisms could further explain 2.7%, 2.3%, and 7% of resistances. These low percentages were observed even though the possibility of using innovative treatments drove patient inclusion in `MOSCATO` and `MATCH-R` trials and even though META-PRISM represents a cohort of uniquely aggressive tumors, more so than the tumors included in the MET500 cohort. These data highlight the unmet need for large-scale efforts that combine molecular profiling with exhaustive clinical annotations to fill our current lack of understanding of resistance mechanisms in cancer.

As outlined in Section 4.2.1, the exploration of known or emerging biomarkers to explain treatment resistances observed in META-PRISM patients did not involve the individual gene expression levels, despite the CIViC database providing evidences based on gene over- or under-expression. A potential approach to identify *gene expression outliers* could be the adoption of the comprehensive rules established by Pleasance and colleagues (Pleasance *et al.* 2022). Their gene expression outlier analysis involves comparing the gene expression of the investigated case with comparator datasets sourced from TCGA, TreeHouse[11], GTEx, TARGET, MET500, and HMF. For each case, they select a primary disease comparator for analysis that most closely reflects the diagnosis of the case and use any additional disease comparators that may be useful, such as a normal tissue comparator that most closely reflects the tissue type of origin and a biopsy tissue comparator that most closely reflects the biopsy site. Comparison methods include percentile rank, number of interquartile ranges, or Z score, with dataset-tuned thresholds for reporting outliers. It's noteworthy that in their assessment of the clinical relevance of WGS and RNA-seq for treatment decisions, Pleasance *et al.* (2022) found that RNA expression was the most informative data type overall, and was the sole therapy decision guide in 25% of cases. This underscores the importance of RNA-seq in treatment guidance. However, the role of RNA expression in explaining treatment resistance requires further investigation. Additionally, the response to this question may be particularly sensitive to the timing of the biopsy in relation to the treatment course, as signaling pathways activated or deactivated by tumor cells to resist drugs may return to normal levels in the absence of therapeutic stress.

In the last section, we delineated how tumor genotyping can serve as a crucial tool in deciphering resistance mechanisms to novel drugs administered within clinical trials. The `DAISY` study specifically focused on assessing the efficacy and mechanism of action of the ADC trastuzumab deruxtecan in advanced breast cancers without prior selection based on *HER2*

[11] https://treehousegenomics.soe.ucsc.edu/public-data/

expression levels. The availability of post-treatment WES profiles for 21 patients, including 11 with matched baseline WES profiles, allowed for a preliminary exploration of potential genetic mechanisms underlying acquired resistance. While no prominent genetic mechanisms emerged from these initial analyses, this trial, alongside initiatives like *ICARUS* (NCT04965766 and NCT04940325)[12], which samples tumor tissue at multiple treatment stages (pre, during, and post) with other deruxtecan-based ADCs in 200 breast and lung cancer patients, will provide invaluable data. The unraveling of the mechanisms of action for these innovative drugs and the identification of predictive or prognostic biomarkers are particularly crucial given the potential transition of cancer treatments toward this new generation of antineoplastic drugs, eventually replacing conventional chemotherapies in cases where low tolerance to chemotherapy toxicity prevents their use (Shastry *et al.* 2023).

The retrospective analysis of advanced urothelial cancers treated with *FGFR* inhibitors represents another significant study, being among the first to decipher resistance mechanisms to these newly approved drugs, with the first FDA approval occurring only in 2019. Presently, the important secondary effects observed in patients treated with *FGFR* inhibitors, coupled with nearly systematic disease progression due to drug resistance (typical for most targeted therapy), limits their clinical utility (Kommalapati *et al.* 2021). However, a more profound understanding of resistance mechanisms holds the potential to identify patients who can derive maximum benefit from targeted therapy or formulate combination treatment strategies effectively inhibiting both the target and the mechanisms by which cancer cells evade the consequences of such inhibition, as demonstrated in the cell line experiments of the study.

[12] https://www.gustaveroussy.fr/fr/programme-innocare

Bibliography

1. Ahronian, L. G. *et al.* Clinical Acquired Resistance to RAF Inhibitor Combinations in *BRAF* -Mutant Colorectal Cancer through MAPK Pathway Alterations. en. *Cancer Discovery* **5,** 358–367. doi:10.1158/2159-8290.CD-14-1518 (Apr. 2015).
2. Antonarakis, E. S. *et al.* AR-V7 and Resistance to Enzalutamide and Abiraterone in Prostate Cancer. en. *New England Journal of Medicine* **371,** 1028–1038. doi:10.1056/NEJMoa1315815 (Sept. 2014).
3. Asiedu, M. K. *et al.* AXL induces epithelial-to-mesenchymal transition and regulates the function of breast cancer stem cells. en. *Oncogene* **33,** 1316–1324. doi:10.1038/onc.2013.57 (Mar. 2014).
4. Bagchi, S., Yuan, R. & Engleman, E. G. Immune Checkpoint Inhibitors for the Treatment of Cancer: Clinical Impact and Mechanisms of Response and Resistance. en. *Annual Review of Pathology: Mechanisms of Disease* **16,** 223–249. doi:10.1146/annurev-pathol-042020-042741 (Jan. 2021).
5. Bashraheel, S. S., Domling, A. & Goda, S. K. Update on targeted cancer therapies, single or in combination, and their fine tuning for precision medicine. en. *Biomedicine & Pharmacotherapy* **125,** 110009. doi:10.1016/j.biopha.2020.110009 (May 2020).
6. Blume-Jensen, P. & Hunter, T. Oncogenic kinase signalling. en. *Nature* **411,** 355–365. doi:10.1038/35077225 (May 2001).
7. Boardman, A. P. & Salles, G. CAR Tcell therapy in large B cell lymphoma. en. *Hematological Oncology* **41,** 112–118. doi:10.1002/hon.3153 (June 2023).
8. Caunt, C. J., Sale, M. J., Smith, P. D. & Cook, S. J. MEK1 and MEK2 inhibitors and cancer therapy: the long and winding road. en. *Nature Reviews Cancer* **15,** 577–592. doi:10.1038/nrc4000 (Oct. 2015).
9. Ceppi, P. *et al.* ERCC1 and RRM1 gene expressions but not EGFR are predictive of shorter survival in advanced non-small-cell lung cancer treated with cisplatin and gemcitabine. en. *Annals of Oncology* **17,** 1818–1825. doi:10.1093/annonc/mdl300 (Dec. 2006).
10. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. en. *JCO Precision Oncology,* 1–16. doi:10.1200/PO.17.00011 (Nov. 2017).
11. Coffee, E. M. *et al.* Concomitant BRAF and PI3K/mTOR Blockade Is Required for Effective Treatment of *BRAFV600E* Colorectal Cancer. en. *Clinical Cancer Research* **19,** 2688–2698. doi:10.1158/1078-0432.CCR-12-2556 (May 2013).
12. Corcoran, R. B. *et al.* Combined BRAF, EGFR, and MEK Inhibition in Patients with *BRAF* V600E-Mutant Colorectal Cancer. en. *Cancer Discovery* **8,** 428–443. doi:10.1158/2159-8290.CD-17-1226 (Apr. 2018).
13. Cortés, J. *et al.* Trastuzumab Deruxtecan versus Trastuzumab Emtansine for Breast Cancer. en. *New England Journal of Medicine* **386,** 1143–1154. doi:10.1056/NEJMoa2115022 (Mar. 2022).
14. Doyle, L. A. *et al.* A multidrug resistance transporter from human MCF-7 breast cancer cells. en. *Proceedings of the National Academy of Sciences* **95,** 15665–15670. doi:10.1073/pnas.95.26.15665 (Dec. 1998).

15. Du, B. & Shim, J. Targeting EpithelialMesenchymal Transition (EMT) to Overcome Drug Resistance in Cancer. en. *Molecules* **21,** 965. doi:10.3390/molecules21070965 (July 2016).
16. Engelman, J. A. *et al. MET* Amplification Leads to Gefitinib Resistance in Lung Cancer by Activating ERBB3 Signaling. en. *Science* **316,** 1039–1043. doi:10.1126/science.1141478 (May 2007).
17. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. en. *Nature* **434,** 917–921. doi:10.1038/nature03445 (Apr. 2005).
18. Flaherty, K. T. *et al.* Combined BRAF and MEK Inhibition in Melanoma with BRAF V600 Mutations. en. *New England Journal of Medicine* **367,** 1694–1703. doi:10.1056/NEJMoa1210093 (Nov. 2012).
19. Freeman, G. J. *et al.* Engagement of the Pd-1 Immunoinhibitory Receptor by a Novel B7 Family Member Leads to Negative Regulation of Lymphocyte Activation. en. *The Journal of Experimental Medicine* **192,** 1027–1034. doi:10.1084/jem.192.7.1027 (Oct. 2000).
20. Fu, Z., Li, S., Han, S., Shi, C. & Zhang, Y. Antibody drug conjugate: the biological missile for targeted cancer therapy. en. *Signal Transduction and Targeted Therapy* **7,** 93. doi:10.1038/s41392-022-00947-7 (Mar. 2022).
21. Galmarini, D., Galmarini, C. M. & Galmarini, F. C. Cancer chemotherapy: A critical analysis of its 60 years of history. en. *Critical Reviews in Oncology/Hematology* **84,** 181–199. doi:10.1016/j.critrevonc.2012.03.002 (Nov. 2012).
22. Gottesman, M. M. Mechanisms of Cancer Drug Resistance. en. *Annual Review of Medicine* **53,** 615–627. doi:10.1146/annurev.med.53.082901.103929 (Feb. 2002).
23. Gottesman, M. M., Fojo, T. & Bates, S. E. Multidrug resistance in cancer: role of ATPdependent transporters. en. *Nature Reviews Cancer* **2,** 48–58. doi:10.1038/nrc706 (Jan. 2002).
24. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. en. *Nature Genetics* **49,** 170–174. doi:10.1038/ng.3774 (Feb. 2017).
25. Groenendijk, F. H. & Bernards, R. Drug resistance to targeted therapies: Déjà vu all over again. en. *Molecular Oncology* **8,** 1067–1083. doi:10.1016/j.molonc.2014.05.004 (Sept. 2014).
26. Helsten, T. *et al.* The FGFR Landscape in Cancer: Analysis of 4,853 Tumors by Next-Generation Sequencing. en. *Clinical Cancer Research* **22,** 259–267. doi:10.1158/1078-0432.CCR-14-3212 (Jan. 2016).
27. Hientz, K., Mohr, A., Bhakta-Guha, D. & Efferth, T. The role of p53 in cancer drug resistance and targeted chemotherapy. en. *Oncotarget* **8,** 8921–8946. doi:10.18632/oncotarget.13475 (Jan. 2017).
28. Hodi, F. S. *et al.* Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. en. *New England Journal of Medicine* **363,** 711–723. doi:10.1056/NEJMoa1003466 (Aug. 2010).

29. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. en. *Nature Reviews Cancer* **13,** 714–726. doi:10.1038/nrc3599 (Oct. 2013).
30. Horn, L. A., Fousek, K. & Palena, C. Tumor Plasticity and Resistance to Immunotherapy. en. *Trends in Cancer* **6,** 432–441. doi:10.1016/j.trecan.2020.02.001 (May 2020).
31. Housman, G. *et al.* Drug Resistance in Cancer: An Overview. en. *Cancers* **6,** 1769–1792. doi:10.3390/cancers6031769 (Sept. 2014).
32. Hsu, J.-M. *et al.* STT3-dependent PD-L1 accumulation on cancer stem cells promotes immune evasion. en. *Nature Communications* **9,** 1908. doi:10.1038/s41467-018-04313-6 (May 2018).
33. Huang, C.-Y. *et al.* Recent progress in TGF- inhibitors for cancer therapy. en. *Biomedicine & Pharmacotherapy* **134,** 111046. doi:10.1016/j.biopha.2020.111046 (Feb. 2021).
34. Huang, L. & Fu, L. Mechanisms of resistance to EGFR tyrosine kinase inhibitors. en. *Acta Pharmaceutica Sinica B* **5,** 390–401. doi:10.1016/j.apsb.2015.07.001 (Sept. 2015).
35. Huang, S. *et al.* MED12 Controls the Response to Multiple Cancer Drugs through Regulation of TGF- Receptor Signaling. en. *Cell* **151,** 937–950. doi:10.1016/j.cell.2012.10.035 (Nov. 2012).
36. Jaaks, P. *et al.* Effective drug combinations in breast, colon and pancreatic cancer cells. en. *Nature* **603,** 166–173. doi:10.1038/s41586-022-04437-2 (Mar. 2022).
37. Jamal-Hanjani, M. *et al.* Tracking the Evolution of NonSmall-Cell Lung Cancer. en. *New England Journal of Medicine* **376,** 2109–2121. doi:10.1056/NEJMoa1616288 (June 2017).
38. Jenkins, R. W., Barbie, D. A. & Flaherty, K. T. Mechanisms of resistance to immune checkpoint inhibitors. en. *British Journal of Cancer* **118,** 9–16. doi:10.1038/bjc.2017.434 (Jan. 2018).
39. Juárez-Salcedo, L. M., Desai, V. & Dalia, S. Venetoclax: evidence to date and clinical potential. *Drugs in Context* **8,** 1–13. doi:10.7573/dic.212574 (Oct. 2019).
40. Juric, D. *et al.* Convergent loss of PTEN leads to clinical resistance to a PI(3)K inhibitor. en. *Nature* **518,** 240–244. doi:10.1038/nature13948 (Feb. 2015).
41. Kalluri, R. The biology and function of fibroblasts in cancer. en. *Nature Reviews Cancer* **16,** 582–598. doi:10.1038/nrc.2016.73 (Sept. 2016).
42. Ko, B., He, T., Gadgeel, S. & Halmos, B. MET/HGF pathway activation as a paradigm of resistance to targeted therapies. *Annals of Translational Medicine* **5,** 4–4. doi:10.21037/atm.2016.12.09 (Jan. 2017).
43. Kobayashi, S. *et al. EGFR* Mutation and Resistance of NonSmall-Cell Lung Cancer to Gefitinib. en. *New England Journal of Medicine* **352,** 786–792. doi:10.1056/NEJMoa044238 (Feb. 2005).
44. Kommalapati, A., Tella, S. H., Borad, M., Javle, M. & Mahipal, A. FGFR Inhibitors in Oncology: Insight on the Management of Toxicities in Clinical Practice. en. *Cancers* **13,** 2968. doi:10.3390/cancers13122968 (June 2021).
45. Konen, J. M. *et al.* Dual Inhibition of MEK and AXL Targets Tumor Cell Heterogeneity and Prevents Resistant Outgrowth Mediated by the Epithelial-to-Mesenchymal Transi-

tion in NSCLC. en. *Cancer Research* **81,** 1398–1412. doi:10.1158/0008-5472.CAN-20-1895 (Mar. 2021).

46. Koren, E. & Fuchs, Y. Modes of Regulated Cell Death in Cancer. en. *Cancer Discovery* **11,** 245–265. doi:10.1158/2159-8290.CD-20-0789 (Feb. 2021).
47. Kroemer, G. & Zitvogel, L. Immune checkpoint inhibitors. en. *Journal of Experimental Medicine* **218,** e20201979. doi:10.1084/jem.20201979 (Mar. 2021).
48. Krummel, M. F. & Allison, J. P. CD28 and CTLA-4 have opposing effects on the response of T cells to stimulation. en. *The Journal of experimental medicine* **182,** 459–465. doi:10.1084/jem.182.2.459 (Aug. 1995).
49. Labrie, M., Brugge, J. S., Mills, G. B. & Zervantonakis, I. K. Therapy resistance: opportunities created by adaptive responses to targeted therapies in cancer. en. *Nature Reviews Cancer* **22,** 323–339. doi:10.1038/s41568-022-00454-5 (June 2022).
50. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. en. *Bioinformatics* **25,** 2078–2079. doi:10.1093/bioinformatics/btp352 (Aug. 2009).
51. Li, M. M. *et al.* Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. en. *The Journal of Molecular Diagnostics* **19,** 4–23. doi:10.1016/j.jmoldx.2016.10.002 (Jan. 2017).
52. Loriot, Y. *et al.* Erdafitinib in Locally Advanced or Metastatic Urothelial Carcinoma. en. *New England Journal of Medicine* **381,** 338–348. doi:10.1056/NEJMoa1817323 (July 2019).
53. Low, L. E. *et al.* Hydroxychloroquine: Key therapeutic advances and emerging nanotechnological landscape for cancer mitigation. en. *Chemico-Biological Interactions* **386,** 110750. doi:10.1016/j.cbi.2023.110750 (Dec. 2023).
54. Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S. & Baradaran, B. The Different Mechanisms of Cancer Drug Resistance: A Brief Review. en. *Advanced Pharmaceutical Bulletin* **7,** 339–348. doi:10.15171/apb.2017.041 (Sept. 2017).
55. Mateo, J. *et al.* A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). en. *Annals of Oncology* **29,** 1895–1902. doi:10.1093/annonc/mdy263 (Sept. 2018).
56. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. en. *Cell* **168,** 613–628. doi:10.1016/j.cell.2017.01.018 (Feb. 2017).
57. McMillin, D. W., Negri, J. M. & Mitsiades, C. S. The role of tumour-stromal interactions in modifying drug response: challenges and opportunities. en. *Nature Reviews Drug Discovery* **12,** 217–228. doi:10.1038/nrd3870 (Mar. 2013).
58. Mhaidly, R. & Mechta-Grigoriou, F. Role of cancerassociated fibroblast subpopulations in immune infiltration, as a new means of treatment in cancer. en. *Immunological Reviews* **302,** 259–272. doi:10.1111/imr.12978 (July 2021).
59. Modi, S. *et al.* Trastuzumab Deruxtecan in Previously Treated HER2-Low Advanced Breast Cancer. en. *New England Journal of Medicine* **387,** 9–20. doi:10.1056/NEJMoa2203690 (July 2022).

60. Mosele, F. *et al.* Trastuzumab deruxtecan in metastatic breast cancer with variable HER2 expression: the phase 2 DAISY trial. en. *Nature Medicine* **29,** 2110–2120. doi:10.1038/s41591-023-02478-2 (Aug. 2023).
61. Nazarian, R. *et al.* Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. en. *Nature* **468,** 973–977. doi:10.1038/nature09626 (Dec. 2010).
62. Nussinov, R., Tsai, C.-J. & Jang, H. Anticancer drug resistance: An update and perspective. en. *Drug Resistance Updates* **59,** 100796. doi:10.1016/j.drup.2021.100796 (Dec. 2021).
63. O'Hare, T. *et al. In vitro* Activity of Bcr-Abl Inhibitors AMN107 and BMS-354825 against Clinically Relevant Imatinib-Resistant Abl Kinase Domain Mutants. en. *Cancer Research* **65,** 4500–4505. doi:10.1158/0008-5472.CAN-05-0259 (June 2005).
64. Osborne, C. K. Tamoxifen in the Treatment of Breast Cancer. en. *New England Journal of Medicine* **339** (ed Wood, A. J.) 1609–1618. doi:10.1056/NEJM199811263392207 (Nov. 1998).
65. Pal, S. K. *et al.* Efficacy of BGJ398, a Fibroblast Growth Factor Receptor 13 Inhibitor, in Patients with Previously Treated Advanced Urothelial Carcinoma with *FGFR3* Alterations. en. *Cancer Discovery* **8,** 812–821. doi:10.1158/2159-8290.CD-18-0229 (July 2018).
66. Pao, W., Miller, V. A., *et al.* Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain. en. *PLoS Medicine* **2** (ed Liu, E. T.) e73. doi:10.1371/journal.pmed.0020073 (Feb. 2005).
67. Pao, W., Wang, T. Y., *et al.* KRAS Mutations and Primary Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib. en. *PLoS Medicine* **2** (ed Herbst, R.) e17. doi:10.1371/journal.pmed.0020017 (Jan. 2005).
68. Papadimitrakopoulou, V. *et al.* Analysis of resistance mechanisms to osimertinib in patients with EGFR T790M advanced NSCLC from the AURA3 study. en. *Annals of Oncology* **29,** viii741. doi:10.1093/annonc/mdy424.064 (Oct. 2018).
69. Pleasance, E. *et al.* Whole-genome and transcriptome analysis enhances precision cancer treatment options. en. *Annals of Oncology* **33,** 939–949. doi:10.1016/j.annonc.2022.05.522 (Sept. 2022).
70. Pusztai, L. *et al.* Phase II study of tariquidar, a selective P-glycoprotein inhibitor, in patients with chemotherapy-resistant, advanced breast carcinoma. en. *Cancer* **104,** 682–691. doi:10.1002/cncr.21227 (Aug. 2005).
71. Ramesh, V., Brabletz, T. & Ceppi, P. Targeting EMT in Cancer with Repurposed Metabolic Inhibitors. en. *Trends in Cancer* **6,** 942–950. doi:10.1016/j.trecan.2020.06.005 (Nov. 2020).
72. Rudin, C. M., Hong, K. & Streit, M. Molecular Characterization of Acquired Resistance to the BRAF Inhibitor Dabrafenib in a Patient with BRAF-Mutant NonSmall-Cell Lung Cancer. en. *Journal of Thoracic Oncology* **8,** e41–e42. doi:10.1097/JTO.0b013e31828bb1b3 (May 2013).

73. Sakai, W. *et al.* Secondary mutations as a mechanism of cisplatin resistance in BRCA2-mutated cancers. en. *Nature* **451,** 1116–1120. doi:10.1038/nature06633 (Feb. 2008).
74. Sergina, N. V. *et al.* Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. en. *Nature* **445,** 437–441. doi:10.1038/nature05474 (Jan. 2007).
75. Sharma, P., Hu-Lieskovan, S., Wargo, J. A. & Ribas, A. Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. en. *Cell* **168,** 707–723. doi:10.1016/j.cell.2017.01.017 (Feb. 2017).
76. Shastry, M. *et al.* Rise of Antibody-Drug Conjugates: The Present and Future. en. *American Society of Clinical Oncology Educational Book,* e390094. doi:10.1200/EDBK_390094 (May 2023).
77. Smith, C. E. P. & Prasad, V. Targeted Cancer Therapies. eng. *American Family Physician* **103,** 155–163 (Feb. 2021).
78. Sodani, K., Patel, A., Kathawala, R. J. & Chen, Z.-S. Multidrug resistance associated proteins in multidrug resistance. *Chinese Journal of Cancer* **31,** 58–72. doi:10.5732/cjc.011.10329 (Feb. 2012).
79. Stommel, J. M. *et al.* Coactivation of Receptor Tyrosine Kinases Affects the Response of Tumor Cells to Targeted Therapies. en. *Science* **318,** 287–290. doi:10.1126/science.1142946 (Oct. 2007).
80. Su, F. *et al. RAS* Mutations in Cutaneous Squamous-Cell Carcinomas in Patients Treated with BRAF Inhibitors. en. *New England Journal of Medicine* **366,** 207–215. doi:10.1056/NEJMoa1105358 (Jan. 2012).
81. Suehnholz, S. P. *et al.* Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. en. *Cancer Discovery.* doi:10.1158/2159-8290.CD-23-0467 (Oct. 2023).
82. Sun, J.-M., Ahn, M.-J., Choi, Y.-L., Ahn, J. S. & Park, K. Clinical implications of T790M mutation in patients with acquired resistance to EGFR tyrosine kinase inhibitors. en. *Lung Cancer* **82,** 294–298. doi:10.1016/j.lungcan.2013.08.023 (Nov. 2013).
83. Swanton, C. Intratumor Heterogeneity: Evolution through Space and Time. en. *Cancer Research* **72,** 4875–4882. doi:10.1158/0008-5472.CAN-12-2217 (Oct. 2012).
84. Szakács, G. *et al.* Predicting drug sensitivity and resistance. en. *Cancer Cell* **6,** 129–137. doi:10.1016/j.ccr.2004.06.026 (Aug. 2004).
85. Thomas, H. & Coley, H. M. Overcoming Multidrug Resistance in Cancer: An Update on the Clinical Strategy of Inhibiting P-Glycoprotein. en. *Cancer Control* **10,** 159–165. doi:10.1177/107327480301000207 (Mar. 2003).
86. Tredan, O., Galmarini, C. M., Patel, K. & Tannock, I. F. Drug Resistance and the Solid Tumor Microenvironment. en. *JNCI Journal of the National Cancer Institute* **99,** 1441–1454. doi:10.1093/jnci/djm135 (Oct. 2007).
87. Turner, N. & Grose, R. Fibroblast growth factor signalling: from development to cancer. en. *Nature Reviews Cancer* **10,** 116–129. doi:10.1038/nrc2780 (Feb. 2010).

88. Visakorpi, T. *et al.* In vivo amplification of the androgen receptor gene and progression of human prostate cancer. en. *Nature Genetics* **9,** 401–406. doi:10.1038/ng0495-401 (Apr. 1995).
89. Waldman, A. D., Fritz, J. M. & Lenardo, M. J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. eng. *Nature Reviews. Immunology* **20,** 651–668. doi:10.1038/s41577-020-0306-5 (Nov. 2020).
90. Wang, Q. *et al.* Role of tumor microenvironment in cancer progression and therapeutic strategy. en. *Cancer Medicine* **12,** 11149–11165. doi:10.1002/cam4.5698 (May 2023).
91. Wargo, J. A., Reuben, A., Cooper, Z. A., Oh, K. S. & Sullivan, R. J. Immune Effects of Chemotherapy, Radiation, and Targeted Therapy and Opportunities for Combination With Immunotherapy. en. *Seminars in Oncology* **42,** 601–616. doi:10.1053/j.seminoncol.2015.05.007 (Aug. 2015).
92. Wee, S. *et al.* PI3K Pathway Activation Mediates Resistance to MEK Inhibitors in KRAS Mutant Cancers. en. *Cancer Research* **69,** 4286–4293. doi:10.1158/0008-5472.CAN-08-4765 (May 2009).
93. Wilson, T., Longley, D. & Johnston, P. Chemoresistance in solid tumours. en. *Annals of Oncology* **17,** x315–x324. doi:10.1093/annonc/mdl280 (Sept. 2006).
94. Wilson, T. R. *et al.* Widespread potential for growth-factor-driven resistance to anti-cancer kinase inhibitors. en. *Nature* **487,** 505–509. doi:10.1038/nature11249 (July 2012).
95. Wu, P. *et al.* Adaptive Mechanisms of Tumor Therapy Resistance Driven by Tumor Microenvironment. *Frontiers in Cell and Developmental Biology* **9,** 641469. doi:10.3389/fcell.2021.641469 (Mar. 2021).
96. Xu, J., Lamouille, S. & Derynck, R. TGF--induced epithelial to mesenchymal transition. en. *Cell Research* **19,** 156–172. doi:10.1038/cr.2009.5 (Feb. 2009).
97. Yang, J. *et al.* Targeting PI3K in cancer: mechanisms and advances in clinical trials. en. *Molecular Cancer* **18,** 26. doi:10.1186/s12943-019-0954-x (Dec. 2019).
98. Yonesaka, K. *et al.* Activation of ERBB2 Signaling Causes Resistance to the EGFR-Directed Therapeutic Antibody Cetuximab. en. *Science Translational Medicine* **3.** doi:10.1126/scitranslmed.3002442 (Sept. 2011).
99. Zhang, Z. *et al.* Activation of the AXL kinase causes resistance to EGFR-targeted therapy in lung cancer. en. *Nature Genetics* **44,** 852–860. doi:10.1038/ng.2330 (Aug. 2012).
100. Zhou, W. *et al.* NEK2 Induces Drug Resistance Mainly through Activation of Efflux Drug Pumps and Is Associated with Poor Prognosis in Myeloma and Other Cancers. en. *Cancer Cell* **23,** 48–62. doi:10.1016/j.ccr.2012.12.001 (Jan. 2013).

Postface

Cancer is a devastating disease that presents a significant challenge to our societies, affecting people from all walks of life in unpredictable ways. Family histories of cancer have provided the first indication of its genetic origins. In the 1970s, pioneering studies identified significant changes in cancerous genomes (Knudson 1971; Stehelin *et al.* 1976), confirming the genetic origin of cancer. Early studies of cancer cell lines and samples revealed the existence of oncogenes (Cooper 1982; Land *et al.* 1983) and tumor suppressor genes (Lane & Crawford 1979; Lee *et al.* 1987) that code for proteins that we now know play critical roles in cellular pathways exploited by cancer cells to outcompete healthy cells and form tumors that can be lethal if left untreated. With the advent of advanced profiling technologies, the genome, epigenome, transcriptome, and proteome of cancer cells have been characterized in great detail, with data collected from tens of thousands of patients with various types of cancer. The completion of the human genome project in 2003 and the introduction of the first next-generation sequencing devices two years later have indeed ushered in a new era of cancer research that uses the vast amount of data generated to understand the biological mechanisms behind cancer onset, progression, and response to treatment. Increasing molecular profiling of the tumors of patients across many cancer types has led to the development of specialized algorithmic tools and statistical models that have given rise to a new field of research called *computational oncology*.

Realizing the promises brought by the technological improvements and the rapid decrease in sequencing costs, international consortia have been assembled to bring these techniques to thousands of patients with cancer and to establish comprehensive molecular landscapes. The TCGA consortium, initiated in 2006 and completed in 2018 with the public release of the *PanCancer Atlas* (Hoadley *et al.* 2018; Ding *et al.* 2018; Sanchez-Vega *et al.* 2018) comprising 23 papers[13], stands as a pioneering initiative. The TCGA research network has released over the years comprehensive molecular landscapes for 33 cancer types, starting with a first molecular portrait of glioblastomas (The Cancer Genome Atlas Research Network 2008). The analyses of the more than 2.5 petabytes of data generated have provided a comprehensive catalog of genomic alterations in a wide range of cancer types, identified molecular subtypes with distinct clinicopathologic characteristics, facilitated the identification of potential therapeutic targets by uncovering genes and pathways that are frequently altered in specific cancer types, and integrated data from multiple omics platform to provide more comprehensive views of the molecular landscapes of several cancer types. The legacy of TCGA continues to shape the field of cancer research and inform ongoing efforts. Other international efforts, such as the PCAWG project co-led by ICGC and TCGA (The ICGC/TCGA

[13] https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html

Pan-Cancer Analysis of Whole Genomes Consortium 2020), or national programs such as the 100,000 genomes project led by Genomics England (Torjesen 2013), have delved into the poorly characterized areas of non-coding DNA and complex structural variants using WGS technology and uncovered many cancer-associated or cancer-driving variants. Many more ongoing efforts are addressing specific questions and characterizing in depth the molecular landscape of patients with cancer, such as the extensive work of the HMF in the Netherlands (Priestley *et al.* 2019), which has established a comprehensive database featuring whole-genome sequencing data and detailed clinical information for thousands of metastatic patients with cancer. The META-PRISM cohort presented in this thesis is a complementary work that also investigated patients with cancer at the late stage of their disease but who are additionally known to be refractory to conventional treatments. By considering in detail the complete treatment history of these patients and the genetic variants detected in their biopsies, we were able to draw a detailed inventory of the current knowledge of the genetic mechanisms of resistance to standard as well as some innovative therapies.

The first chapter of this work has presented fundamental concepts about cancer, provided insights into the genetic and epigenetic mechanisms underlying the acquisition and maintenance of malignant capabilities, and depicted the current general classification system employed in cancer research and care. These concepts are essential knowledge that all cancer researchers should have at their disposal to help them navigate the complex translational and clinical research landscapes. The last section of this chapter has additionally introduced the reader to considerations about the growing place of molecular profiling. The recent FDA approvals of multiple tumor-agnostic drugs, currently six in number (four targeted therapies directed against *NTRK* rearrangements, *RET* rearrangements, or *BRAF* V600E mutation, and two immunotherapies indicated in highly mutagenic or unstable genomes), have spun a new era of drug development and clinical trial designs where the histology and tissue of origin are being superseded by molecular considerations. Consequently, new clinical trial designs, such as basket trials and platform trials, have emerged to assess multiple therapies across different tumor types according to specific molecular alterations with the primary goal of demonstrating efficacy (Yates *et al.* 2018; Ravi & Kesari 2022). The MOSCATO and MATCH-R trials conducted at Gustave Roussy are non-standard precision medicine trials with some characteristics of platform trials, although they did not investigate novel therapies. These trials ran for several years and aimed to prove the efficacy of administrating innovative drugs according to the detectable molecular alterations, among other objectives. While actionable alterations were detected in only a subset of patients in these trials, this subset demonstrated prolonged survival (Massard *et al.* 2017; Recondo *et al.* 2020). This first chapter is the result of numerous readings undertaken throughout my Ph.D., the many interactions I had with various colleagues, and my attendance at high-quality conferences that have described many new and promising techniques and clinical outcomes.

The second chapter provided a technical overview of the analysis of data generated by high-throughput sequencing technologies. It includes a description of sample preparation and sequencing experiments, as well as an overview of how bioinformatics workflows are organized and the variety of tools available for any given task. The chapter emphasized

the challenging decisions that bioinformaticians face in selecting methods to detect genetic variants or quantify gene expression, and the impact of those choices on downstream analyses. Understanding the biases and underlying hypotheses of each tool choice is essential for interpreting the data tables generated by bioinformatic workflows. This is especially important given the increasing availability of data-rich datasets for researchers to test hypotheses with greater statistical power. However, technical differences in how shared data are generated can hinder meaningful analyses, and researchers must understand and, if possible, overcome these differences by either reprocessing raw data uniformly or using batch-effect correction techniques. Batch-effect corrections are crucial for addressing the influences of experimental settings in a variety of sequencing techniques, most particularly RNA expression profiles (Tung *et al.* 2017). However, we still need to have calibrated tools that can adapt to the many sources of batch effects and correct them while preserving biological signals. Ignoring the importance of these batch effects and failing to control for them can overshadow any biological difference and confound comparative analyses. The chapter also presented some standard analyses that are commonly performed on high-throughput DNA sequencing data, most notably the analysis of mutational signatures and, since the recent extensions that have been presented, the analysis of CNA and SV signatures. The mathematical framework originally devised by Nik-Zainal *et al.* (2012) is presented in detail so as to raise awareness about methodological hypotheses that may need reassessment in light of the still significant number of reference signatures with unknown etiology or associated with putative sequencing artifacts.

The third chapter of the thesis presents the main result of this thesis work which is a comprehensive evaluation of the exome and transcriptome sequencing profiles of 1,031 metastatic patients who had become refractory to conventional treatments. This cohort, named META-PRISM, was assembled by considering all adult patients with solid tumors that benefited from whole-exome sequencing (WES) or RNA sequencing (RNA-seq) profiling as part of the MOSCATO and MATCH-R precision medicine trials conducted at Gustave Roussy. The primary objective of the analyses presented in this chapter was two-fold. Firstly, to elucidate the genomic differences between the tumors of refractory metastatic patients and tumors from treatment-naive non-metastatic patients, taken from TCGA. Secondly, to quantify the additional clinical utility of WES and RNA-seq sequencing over standard clinical variables for risk-stratifying patients and guiding therapeutic decisions. This chapter builds upon the fundamental concepts and tools presented in the first two chapters and establishes a robust database with harmonized tumor type classifications and a list of molecular alterations detected across META-PRISM, TCGA, and MET500 cohorts, with MET500 serving as the validation cohort (Robinson *et al.* 2017). Significant efforts were dedicated to reprocess raw sequencing files from all cohorts when data tables generated by the bioinformatic pipelines were not available, as in the case of CNA calls in TCGA. Comparative analyses of the tumor genotypes revealed an overall enrichment in the mutational burden, driver mutation incidence, and genomic instability in metastatic patients with specific enrichments in some tumor types, such as whole-genome duplication events in metastatic prostate cancers. The analysis of blood tissues additionally pointed out a significant enrichment of deleterious mutations affecting cancer-predisposing genes, opening the possibility of germline counseling for patients at accrued risk of developing aggressive forms of cancer. To assess the potential

prognostic power of the genetic variants identified, survival models were run on different combinations of biomarkers and clinical parameters and compared to the current baseline model utilizing the GRIM score. A significant increase in risk prediction performance was observed in patients with breast cancer for which the GRIM score was not informative at all in this cohort. Although this model could not be validated in patients with metastatic breast cancer from MET500 due to the unavailability of survival data, it was demonstrated to be predictive in treatment-naive TCGA patients with breast cancer. This finding enhances our confidence in its utility.

The fourth and last chapter provided a general introduction to the realm of cancer drugs and the different currently known mechanisms through which cancer cells evade therapies and described two example translational studies of resistance mechanisms to innovative targeted therapies. To determine the therapeutic implications of genetic variants detected from WES and RNA-seq, we used the OncoKB (Chakravarty *et al.* 2017; Suehnholz *et al.* 2023) and CIViC (Griffith *et al.* 2017) databases which are the result of extensive curation efforts and are continually expanding as more experimental evidence accumulates. The therapeutic implications drawn from the confrontation of the variants we detected and the contents of these knowledge databases were compared against the resistance histories of META-PRISM patients to determine the prevalence of confirmed, investigative, and putative biomarkers of resistance. This comparison revealed a significant gap in our current understanding of treatment resistances and the heterogeneity of this understanding across classes of therapies, with targeted therapies and particularly TKIs concentrating most of the known resistance mechanisms. Two additional studies were presented to analyze putative mechanisms of resistance to innovative drugs, the first focusing on ADC trastuzumab deruxtecan in breast cancers using an AI model and standard WES, and the second investigating on-target and bypass resistance mechanisms to *FGFR* inhibitors in urothelial cancers using various genotyping methods. The rapid accumulation of knowledge about genotype-treatment relationships through studies as the two presented, coupled with improvements in diagnostic testing, is driving the current cancer care towards the era of personalized treatments that are tailored to the specific molecular characteristics of each cancer.

The different translational projects I participated in throughout my PhD have allowed me to interrogate data derived from whole-exome and bulk-transcriptome sequencing. However, other sequencing technologies such as epigenomics, proteomics, and metabolomics are available. Technological advancements have led to improvements in WES and RNA-seq, the decrease in whole-genome sequencing costs, and the development of technologies that can assess multiple types of omics data at single-cell (Tang *et al.* 2009) or even spatially-resolved subcellular resolutions (Crosetto *et al.* 2015). These different methods allow us to better understand cancer cells by profiling them in various ways. Combining the information we gather from these different technologies may be the key to unlocking the determinants of cancer cells' behavior. Researchers have made headways in this direction, contributing to various improvements in tasks like risk classification, subtype discovery, and survival prediction. For example, I participated in a study (Benkirane *et al.* 2023) where we used variational autoencoders for learning compact representations of patients profiled through multiple omics

technologies and used this representation for predicting survival, shedding light on the promises of deep-learning approaches. Other integrative approaches relying on standard and deep networks, Bayesian models, factorisation techniques, and feature extraction or transformation methods have also been explored and reviewed (Huang *et al.* 2017; Subramanian *et al.* 2020; Nicora *et al.* 2020; Reel *et al.* 2021). The future of these approaches depends on their ability to demonstrate better performance in different clinically relevant tasks, as well as the practical feasibility of profiling tumors through various omic assays.

The data generated by TCGA has also shed light on the extensive heterogeneity of cancer cells both at the intra- and inter-tumor levels. Understanding that tumor heterogeneity plays a crucial in cancer, notably through its association with poor prognosis and poor response to many cancer treatments, pioneering initiatives such as TracerX (Jamal-Hanjani *et al.* 2014) have allowed to precisely quantify the spatial and longitudinal heterogeneity of tumors. Although we did not present a detailed analysis of intra-tumor heterogeneity in META-PRISM patients, this is part of the future analysis plans, mainly through a new project that is just starting and will involve performing comprehensive comparisons of pre- and, when available, post-treatment biopsies. This project will aim to elucidate precisely the intricate relationships between tumor genotypes and treatment response, a question for which intra-tumor heterogeneity will most likely be central to the analyses.

Finally, I cannot overstate how fortunate I have been to have joined Paul-Henry Cournède's team at a time when a remarkable collaboration between a top-tier university and a world-class cancer center was just beginning. Since the start of my PhD journey, this partnership has evolved into a robust and dynamic source of collaborations, fostering innovative research that bridges the gap between basic and experimental sciences. The successes achieved through this collaboration have been numerous, and they are poised to accelerate further, thanks to the sustained support from funding bodies for the diverse projects undertaken by hybrid research teams comprising dedicated scientists with complementary expertise. Being part of an interdisciplinary project like META-PRISM has played a pivotal role in my scientific growth. It provided a unique opportunity to establish fruitful collaborations with experts from various disciplines, allowing me to glean insights from each of them and shaping me into the cancer research scientist I am today.

Bibliography

1. Benkirane, H., Pradat, Y., Michiels, S. & Cournède, P.-H. CustOmics: A versatile deep-learning based strategy for multi-omics integration. en. *PLOS Computational Biology* **19** (ed Noble, W. S.) e1010921. doi:10.1371/journal.pcbi.1010921 (Mar. 2023).
2. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. en. *JCO Precision Oncology,* 1–16. doi:10.1200/PO.17.00011 (Nov. 2017).
3. Cooper, G. M. Cellular Transforming Genes. en. *Science* **217,** 801–806. doi:10.1126/science.6285471 (Aug. 1982).
4. Crosetto, N., Bienko, M. & Van Oudenaarden, A. Spatially resolved transcriptomics and beyond. en. *Nature Reviews Genetics* **16,** 57–66. doi:10.1038/nrg3832 (Jan. 2015).
5. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. en. *Cell* **173,** 305–320.e10. doi:10.1016/j.cell.2018.03.033 (Apr. 2018).
6. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. en. *Nature Genetics* **49,** 170–174. doi:10.1038/ng.3774 (Feb. 2017).
7. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. en. *Cell* **173,** 291–304.e6. doi:10.1016/j.cell.2018.03.022 (Apr. 2018).
8. Huang, S., Chaudhary, K. & Garmire, L. X. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics* **8,** 84. doi:10.3389/fgene.2017.00084 (June 2017).
9. Jamal-Hanjani, M. *et al.* Tracking Genomic Cancer Evolution for Precision Medicine: The Lung TRACERx Study. en. *PLoS Biology* **12,** e1001906. doi:10.1371/journal.pbio.1001906 (July 2014).
10. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. en. *Proceedings of the National Academy of Sciences* **68,** 820–823. doi:10.1073/pnas.68.4.820 (Apr. 1971).
11. Land, H., Parada, L. F. & Weinberg, R. A. Cellular Oncogenes and Multistep Carcinogenesis. en. *Science* **222,** 771–778. doi:10.1126/science.6356358 (Nov. 1983).
12. Lane, D. P. & Crawford, L. V. T antigen is bound to a host protein in SY40-transformed cells. en. *Nature* **278,** 261–263. doi:10.1038/278261a0 (Mar. 1979).
13. Lee, W.-H. *et al.* Human Retinoblastoma Susceptibility Gene: Cloning, Identification, and Sequence. en. *Science* **235,** 1394–1399. doi:10.1126/science.3823889 (Mar. 1987).
14. Massard, C. *et al.* High-Throughput Genomics and Clinical Outcome in Hard-to-Treat Advanced Cancers: Results of the MOSCATO 01 Trial. en. *Cancer Discovery* **7,** 586–595. doi:10.1158/2159-8290.CD-16-1396 (June 2017).

15. Nicora, G., Vitali, F., Dagliati, A., Geifman, N. & Bellazzi, R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Frontiers in Oncology* **10,** 1030. doi:10.3389/fonc.2020.01030 (June 2020).
16. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. en. *Cell* **149,** 979–993. doi:10.1016/j.cell.2012.04.024 (May 2012).
17. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. en. *Nature* **575,** 210–216. doi:10.1038/s41586-019-1689-y (Nov. 2019).
18. Ravi, R. & Kesari, H. V. Novel Study Designs in Precision Medicine Basket, Umbrella and PlatformTrials. en. *Current Reviews in Clinical and Experimental Pharmacology* **17,** 114–121. doi:10.2174/1574884716666210316114157 (July 2022).
19. Recondo, G. *et al.* Feasibility and first reports of the MATCH-R repeated biopsy trial at Gustave Roussy. en. *npj Precision Oncology* **4,** 27. doi:10.1038/s41698-020-00130-7 (Dec. 2020).
20. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. en. *Biotechnology Advances* **49,** 107739. doi:10.1016/j.biotechadv.2021.107739 (July 2021).
21. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. en. *Nature* **548,** 297–303. doi:10.1038/nature23306 (Aug. 2017).
22. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. en. *Cell* **173,** 321–337.e10. doi:10.1016/j.cell.2018.03.035 (Apr. 2018).
23. Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. en. *Nature* **260,** 170–173. doi:10.1038/260170a0 (Mar. 1976).
24. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. en. *Bioinformatics and Biology Insights* **14,** 117793221989905. doi:10.1177/1177932219899051 (Jan. 2020).
25. Suehnholz, S. P. *et al.* Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. en. *Cancer Discovery*. doi:10.1158/2159-8290.CD-23-0467 (Oct. 2023).
26. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. en. *Nature Methods* **6,** 377–382. doi:10.1038/nmeth.1315 (May 2009).
27. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. en. *Nature* **455,** 1061–1068. doi:10.1038/nature07385 (Oct. 2008).
28. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. en. *Nature* **578,** 82–93. doi:10.1038/s41586-020-1969-6 (Feb. 2020).
29. Torjesen, I. Genomes of 100 000 people will be sequenced to create an open access research resource. en. *BMJ* **347,** f6690–f6690. doi:10.1136/bmj.f6690 (Nov. 2013).
30. Tung, P.-Y. *et al.* Batch effects and the effective design of single-cell gene expression studies. en. *Scientific Reports* **7,** 39921. doi:10.1038/srep39921 (Jan. 2017).

31. Yates, L. *et al.* The European Society for Medical Oncology (ESMO) Precision Medicine Glossary. en. *Annals of Oncology* **29,** 30–35. doi:10.1093/annonc/mdx707 (Jan. 2018).

A. Annexes

A.1. Annexes to chapter 1

A.1.1. Molecular classifications of the 33 TCGA studies

Abbr.	Study name	Selected subtype	Subtypes	Reference
ACC	Adrenocortical carcinoma	DNAmeth	CIMP-high, CIMP-intermediate, CIMP-low	Zheng et al. Cancer Cell 2016
BLCA	Bladder Urothelial Carcinoma	mRNA	1-4	Robertson et al. Cell 2017
BRCA	Breast invasive carcinoma	PAM50	Basal, Her2, LumA, LumB, Normal	Berger et al. Cancer Cell 2018
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	mRNA		Berger et al. Cancer Cell 2018
CHOL	Cholangiocarcinoma	mRNAseq		Farshidfar et al. Cell Reports 2017
COAD	Colon adenocarcinoma	Molecular Subtype	CIN, GS, MSI, HM-SNV, EBV	Liu et al. Cancer cell 2018
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma			
ESCA	Esophageal carcinoma	Molecular Subtype	CIN, GS, MSI, HM-SNV, EBV	Liu et al. Cancer cell 2018
GBM	Glioblastoma multiforme	Supervised DNAmeth	G-CIMP-low, G-CIMP-high, Codel, Classic-like, Mesenc-like, LGm6-GBM, PA-like	Ceccarelli et al. Cell 2016
HNSC	Head and Neck squamous cell carcinoma	mRNA	1, 2, 3, 4, 5, 6	Campbell et al. Cell Reports 2018
KICH	Kidney Chromophobe	mRNA	1, 2, 3, 4, 5, 6	Campbell et al. Cell Reports 2018
KIRC	Kidney renal clear cell carcinoma	mRNA	1, 2, 3, 4, 5, 6	Campbell et al. Cell Reports 2018
KIRP	Kidney renal papillary cell carcinoma	mRNA	1, 2, 3, 4, 5, 6	Campbell et al. Cell Reports 2018
LAML	Acute Myeloid Leukemia	mRNA	1, 2, 3, 4, 5, 6, 7	TCGA Research Network. NEJM 2013
LGG	Brain Lower Grade Glioma	Supervised DNAmeth	G-CIMP-low, G-CIMP-high, Codel, Classic-like, Mesenc-like, LGm6-GBM, PA-like	Ceccarelli et al. Cell 2016

LIHC	Liver hepatocellular carcinoma	iCluster	1-3	TCGA Research Network. Cell 2017
LUAD	Lung adenocarcinoma	iCluster	1, 2, 3, 4, 5, 6	TCGA Research Network. Nature 2014
LUSC	Lung squamous cell carcinoma	mRNA	1, 2, 3, 4, 5, 6	Campbell et al. Cell Reports 2018
MESO	Mesothelioma	mRNA cluster k4	C1 S2, C2 S4, C3 S3, C4 S1	Hmeljak et al. Cancer Discovery 2018
OV	Ovarian serous cystadenocarcinoma	OV Subtype	Differentiated, Immunoreactive, Mesenchymal, Proliferative	Berger et al. Cancer Cell 2018
PAAD	Pancreatic adenocarcinoma	mRNA Bailey Clusters (All 150 Samples) 1squamous 2immunogenic 3progenitor 4ADEX	1, 2, 3, 4	TCGA Research Network. Cancer Cell 2017
PCPG	Pheochromocytoma and Paraganglioma	mRNA Subtype Clusters	Kinase signaling, Wnt-altered, Pseudohypoxia, Cortical admixture	TCGA Research Network. Cancer Cell 2017
PRAD	Prostate adenocarcinoma	mutation/fusion	1-ERG, 2-ETV1, 3-ETV4, 4-FLI1, 5-SPOP, 6-FOXA1, 7-IDH1, 8-other	TCGA Research Network. Cancer. Cell 2015
READ	Rectum adenocarcinoma	Molecular Subtype	CIN, GS, MSI, HM-SNV, EBV	Liu et al. Cancer cell 2018
SARC	Sarcoma			
SKCM	Skin Cutaneous Melanoma	mutation	BRAF, RAS, NF1, Triple-WT	TCGA Research Network. Cancer. Cell 2015
STAD	Stomach adenocarcinoma	Molecular Subtype	CIN, GS, MSI, HM-SNV, EBV	Liu et al. Cancer cell 2018
TGCT	Testicular Germ Cell Tumors	methylation k5		Shen et al. Cell Report 2018
THCA	Thyroid carcinoma	mRNA	1-5	TCGA Research Network. Cell 2014
THYM	Thymoma			

UCEC	Uterine Corpus Endometrial Carcinoma	UCEC Histology	Endometrioid, Mixed Serous and Endometrioid, Serous	Berger et al. Cancer Cell 2018
UCS	Uterine Carcinosarcoma	mRNA	1-2	Cherniak et al. Cancer Cell 2017
UVM	Uveal Melanoma			

Table A.1.: Molecular subtypes for each of the 33 tumor types analyzed by TCGA

A.2. Annexes to chapter 2

A.2.1. Variant callers

Mutation caller	Mutation type	TN req.	Reference
Germline			
BCFtools	SNVs & Indels		Release 0.1.9, 2010. Danecek et al. GigaScience. 2021
FreeBayes	SNVs & Indels		Garrison & Marth. aRxiv. 2012
Haplotypecaller	SNVs & Indels		Poplin et al. BioRxiv. 2018
DeepVariant	SNVs & Indels		Poplin et al. Nat. Biotechnol. 2018
Somatic			
Indelocator	Indels	Yes	Chapman et al. Nature. 2011
SomaticSniper	SNVs	Yes	Larson et al. Bioinformatics. 2012
deepSNV	SNVs	Yes	Gerstung et al. Nat. Commun. 2012
JointSNVMix	SNVs & Indels	Yes	Roth et al. Bioinformatics. 2012
Strelka	SNVs & Indels	Yes	Saunders et al. Bioinformatics. 2012
EBCall	SNVs & Indels	Yes	Shiraishi et al. Nucleic Acids Res. 2013
CaVEMan	SNVs	Yes	Cancer Genome Project. Sanger. 2014
Radia	SNVs	Yes	Radenbaugh et al. PLoS One. 2014
VarDict	SNVs & Indels	No	Lai et al. Nucleic Acids Res. 2016
Mutect Mutect2	SNVs & Indels	Yes No	Cibulskis et al. Nat. Biotechnol. 2013 Benjamin et al. bioRxiv. 2019
MuSE 1.0 MuSE 2.0	SNVs	Yes Yes	Fan et al. Genome Biol. 2016 v2.0 released in 2021
DRAGEN Somatic Small Variant Caller	SNVs & Indels	No	Scheffler et al. bioRxiv. 2023
Both			
VarScan VarScan2	SNVs & Indels	No No	Koboldt et al. Bioinformatics. 2009 Koboldt et al. Genome Res. 2012
LoFreq	SNVs & Indels	No	Wilm et al. Nucleic Acids Res. 2012
Platypus	SNVs & Indels	No	Rimmer et al. Nat. Gen. 2014
Pindel	Indels	No No	Ye et al. Bioinformatics. 2009 Ye et al. Nat. Med. 2015
Strelka2	SNVs & Indels	Yes	Kim et al. Nat. Meth. 2018

Table A.2.: List of the most commonly used variant callers for identifying SNVs, MNVs, or indels of germline or somatic origin. SNV, single-nucleotide variant; indel, insertion or deletion; TN req., tumor-normal required by the tool to call somatic variants.

CNA caller	Sequencing platform	Reference
Germline		
CONTRA	Targeted, WES	Li et al. Bioinformatics. 2012
CoNIFER	WES	Krumm et al. Genome Res. 2012
XHMM	WES	Fromer et al. Am J Hum Genet. 2012
GermlineCNVCaller	Any NGS	Released in 2020 with GATK v4.0.
Somatic		
ExomeCNV	Targeted, WES	Sathirapongsasuti et al. Bioinformatics. 2011
cn.MOPS	Any NGS	Klambauer et al. Nucleic Acids Res. 2012
CoNVEX	WES	Amarasinghe et al. Bioinformatics. 2013
ADTEx	WES	Amarasinghe et al. BMC Genomics. 2014
SEQUENZA	WES	Favero et al. Ann Oncol. 2015
FACETS	Targeted, WES	Shen et al. Nucleic Acids Res. 2016
ascatNgs	WGS	Raine et al. Curr Protoc Bioninformatics. 2016
ACEseq	WGS	Kleinheinz et al. bioRxiv. 2017
CODEX CODEX2	Targeted, WES	Jiang et al. Nucleic Acids Res. 2015 Jiang et al. Genome Biol. 2018
Both		
Control-FREEC	Any NGS	Boeva et al. Bioinformatics. 2012
VarScan2	WES	Koboldt et al. Genome Res. 2012
ExomeDepth	WES	Plagnol et al. Bioninformatics. 2012
PatternCNV	WES	Wang et al. Bioninformatics. 2014
CNVkit	Targeted, WES	Talevich et al. PLoS Comput Biol. 2016
EXCAVATOR2	WES	DAurizio. Nucleic Acids Res. 2016
DRAGEN	Any NGS	Released in 2023 with GATK 4.2.

Table A.3.: Examples of popular variant callers capable of identifying CNAs from DNA-sequencing experiments

Gene fusion caller	Sequencing platform	(Pseudo)-Aligner algorithm	Assembler algorithm	Reference
TopHat	SE/PE bulk RNA	Bowtie		Trapnell et al. Bioinformatics. 2009
SpliceMap	SE/PE bulk RNA	SeqMap, ELAND		Au et al. Nucleic Acids Res. 2010
MapSplice	SE/PE bulk RNA	BWA		Wang et al. Nucleic Acids Res. 2010
SnowShoes-FTD	PE bulk RNA	Bowtie		Asmann et al. Nucleic Acids Res. 2011
FusionMap	SE/PE bulk RNA/WGS	Custom		Ge et al. Bioinformatics. 2011
chimerascan	PE bulk RNA	Bowtie		Iyer et al. Bioinformatics. 2011
TopHat-Fusion	SE/PE bulk RNA	Bowtie		Kim and Salzberg. Genome Biol. 2011
Short-Fuse	PE bulk RNA	Bowtie		Kinsella et al. Bioinformatics. 2011
FusionHunter	PE bulk RNA	Bowtie		Li et al. Bioinformatics. 2011
deFuse	PE bulk RNA	Bowtie		McPherson et al. PLoS Comput. Biol. 2011
Bellerophontes	PE bulk RNA	Bowtie	cufflinks	Abate et al. Bioinformatics. 2012
BreakFusion	PE bulk RNA	User-chosen	TIGRA-SV	Chen et al. Bioinformatics. 2012
EricScript	PE bulk RNA	BWA		Benelli et al. Bioinformatics. 2012
SOAPfuse	PE bulk RNA	SOAP2		Jia et al. Genome Biol. 2013
FusionQ	PE bulk RNA	Bowtie	cufflinks	Liu et al. Bioinformatics. 2013
FusionCatcher	PE bulk RNA	Bowtie, BLAT, STAR, Bowtie2	velvet	Nicorici et al. bioRxiv. 2014
PRADA	PE bulk RNA	BWA		Torres-Garcia et al. Bioinformatics. 2014
JAFFA	SE/PE bulk RNA	BLAT	Oases	Davidson et al. Genome Med. 2015
InFusion	PE bulk RNA	Bowtie2		Okonechnikov et al. PLoS ONE. 2016
Pizzly	PE bulk RNA	Kallisto		Melsted et al. bioRxiv. 2017
ChimPipe	PE bulk RNA	GEM		Rodríguez-Martín. BMC Genomics. 2017
SQUID	PE bulk RNA	User-chosen		Ma et al. Genome Biol. 2018
TrinityFusion	PE bulk RNA	STAR	Custom	Haas et al. Genome Biol. 2019
STAR-Fusion	PE bulk RNA	STAR		Uhrig et al. Genome Res. 2021.
Arriba	PE bulk RNA	STAR		Uhrig et al. Genome Res. 2021.
novoRNAbreak	PE bulk RNA	BWA	Custom	Tan et a. J Bioinform Syst Biol. 2023

Table A.4.: List of variant callers capable of identifying RNA fusions from short-read RNA-seq experiments. SE, single-end; PE, paired-end

A.2.2. About non-negative matrix factorisation

NMF is an algorithmic procedure that aims at decomposing any matrix with non-negative coefficients $\mathbf{M} \in {\mathbb{R}_+}^{F \times N}$ into a product of matrices $\mathbf{W} \in {\mathbb{R}_+}^{F \times K}$ and $\mathbf{H} \in {\mathbb{R}_+}^{K \times N}$ with a prespecified internal dimension K. Formally, for a fixed integer K, NMF solves the following non-convex optimization problem

$$\begin{cases} \underset{\mathbf{W},\mathbf{H} \in \mathbb{R}^{F \times K} \times \mathbb{R}^{K \times N}}{\arg\min} & d(\mathbf{M}|\mathbf{WH}) \\ \text{s.t. } W_{fk} \geq 0, H_{kn} \geq 0, & \forall f, k, n \end{cases} \tag{A.1}$$

with d a divergence function applied and summed element-wise on the matrices. The next subsections provide technical details about various objective functions that may be used and the numerical strategies that have been developed to solve (A.1). Another important point that will be discussed is the selection of an optimal value for the number K of signatures to be extracted.

A.2.2.1. The cost function

The first key parameter to NMF is the definition of the cost function in (A.1). Historically, either the Kullback-Leibler divergence (Lee & Seung 2001) or the Euclidean distance (Paatero & Tapper 1994; Lee & Seung 2001) have been used to define the objective function. However, infinitely many objective functions may be considered using, for example, the set of all α and β **divergences**. Divergences are distance-type measures used to compute the distance between two n-dimensional probability distributions $\mathbf{p} = (p_1, \cdots, p_n), \mathbf{q} = (q_1, \cdots, q_n)$. We usually consider functions that are separable in the sense that $D(\mathbf{p}||\mathbf{q}) = \sum_{i=1}^{n} d(p_i|q_i)$ and $D(\mathbf{p}||\mathbf{q}) = 0 \iff \mathbf{p} = \mathbf{q}$. Except for particular cases, divergence functions are not metrics in the mathematical sense as they are generally not symmetric nor do they satisfy the triangular inequality. In its generalization of the NMF iterative updates, Kompass (2007) introduced the α-divergence

$$d_\alpha(p|q) \overset{\text{def}}{=} \begin{cases} p\frac{p^\alpha - q^\alpha}{\alpha} + q^\alpha(q-p) & \alpha \in (0,1] \\ p(\log(p) - \log(q)) + q - p & \alpha = 0 \end{cases} \tag{A.2}$$

This divergence encompasses the Euclidean distance for $\alpha = 1$ and the Kullback-Leibler divergence for $\alpha = 0$ as specific cases. Remarkably, the author derived general multiplicative updates formulas that match exactly the formulas given by Lee & Seung (2001) for the two extreme values of α but, compared to the latter work, the work of the author additionally shows that the cost function defined using the divergence (A.2) is non-increasing under the generalized updates, thereby extending the proof of monotonicity under multiplicative updates to all values of α in $[0, 1]$. The α-divergence was also considered for NMF by Cichocki, Zdunek & Amari (2006) and Cichocki, Zdunek, Phan, *et al.* (2009) using, however, a slightly

different definition of the α-divergence given by $d_{\alpha+1}^{\text{Cichocki}}(p|q) = \frac{\alpha}{q^\alpha} d_\alpha(p|q)$. Many iterative rules for solving NMF under the α-divergence with various constraints or for many other classes of functions are given in their reference book on the topic Cichocki, Zdunek, Phan, *et al.* 2009.

The β-divergence defined in the works of Cichocki, Zdunek, Phan, *et al.* (2009) or Févotte & Idier (2011) by

$$d_\beta(p|q) \overset{\text{def}}{=} \begin{cases} \frac{1}{\beta(\beta-1)} \left(p^\beta + (\beta-1)q^\beta - \beta p q^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ p \log \frac{p}{q} - p + q & \beta = 1 \\ \frac{p}{q} - \log \frac{p}{q} - 1 & \beta = 0 \end{cases} \tag{A.3}$$

is another class of functions used in the context of NMF. Similarly to the α-divergence, it encompasses as particular cases the Euclidean distance ($\beta = 2$), Kullback-Leibler divergence ($\beta = 1$), and Itakura-Saito distance ($\beta = 0$). Févotte & Idier (2011) demonstrated in their work how generalized iterative updates could be defined to solve (A.1) using the β-divergence (A.3) and extended the theoretical results of monotonicity to all values of $\beta \in (0, 1)$ for the heuristic multiplicative updates introduced in Lee & Seung (2001) - and generalized by Kompass (2007) later on. They additionally introduced the *maximization-minimisation* iterative rules, which theoretically result in non-increasing values of the cost function for all values of β, and coincide with the heuristic multiplicative update rules for $\beta \in [1, 2]$.

As evidenced above, the problem of non-negative factorization may be formulated via infinitely many objective functions whose choice directly influences the resulting factorization and ultimately the conclusions that will be drawn from the NMF-based analysis. As a consequence, the choice of a cost function should be driven by the type of data to analyze but, although many algorithm improvements have been proposed over the years, there is only a scarce number of papers dedicated to the selection of a cost function according to the application (Févotte, Bertin, *et al.* 2009). A possible solution for choosing an optimal value of β for a β-divergence-formulated NMF may rely on using a non-maximum-likelihood estimator called *score matching* as presented in Lu *et al.* (2012). This is made possible by a Bayesian formulation of NMF that allows to link the cost function in (A.1) to specific distributions of the coefficients of the matrix to be factorized $\mathbf{M}$. See Févotte, Bertin, *et al.* (2009) for proofs of how the Itakura-Saito and Kullback-Leibler divergences may be related to statistical models that assume sums of gaussians and Poisson distributions, respectively, for the coefficients of $\mathbf{M}$.

A.2.2.2. Optimization algorithms

Multiple types of algorithms have been developed over time to solve the NMF optimization problem (A.1) and may be classified into three general classes as done by Berry and colleagues in their comprehensive review of NMF algorithms and applications (Berry *et al.* 2007):

multiplicative updates, **gradient descent**, and **alternating least squares (ALS)**. Though most of these algorithms were developed to solve NMF in the context of one or several particular cost functions, many subsequent works have been published that generalize these algorithms to general classes of functions, notably α (Cichocki, Zdunek & Amari 2006; Kompass 2007), β (Cichocki, Zdunek & Amari 2006; Févotte & Idier 2011), Bregman (Sra & Dhillon 2005), or Csiszár (Cichocki, Zdunek & Amari 2006) divergences. Solving the NMF problem is made difficult by the non-negativity constraints on the factor matrices but also by the non-convexity of the formulation which means that the solution reached by any given algorithm is never guaranteed to be the global minimum. Repeating the algorithm with different initial conditions is a commonly used technique to alleviate this issue and find a good minimum. Additionally, scaling and permutation cause uniqueness issues as for any solution $\mathbf{WH}$, an infinite set of additional solutions exist considering $\mathbf{WDD^{-1}H}$ where $\mathbf{D}$ is a non-negative invertible matrix. Therefore, the uniqueness of NMF may only be considered up to scaling and permutation and the reader is referred to Laurberg *et al.* (2008) for theoretical results on the uniqueness of NMF.

The multiplicative updates algorithms introduced by Lee and Seung for the Euclidean distance and Kullback-Leibler divergences (Lee & Seung 2001) are undoubtedly the most widely used numerical schemes in the applications of NMF. The simplicity of the update rules and the concomitant enthusiasm of the applied sciences for NMF in applied fields at the time Lee and Seung introduced their algorithms, particularly signal processing, engineering and medicine, are probably key reasons explaining this popularity. However, as noted in multiple subsequent works (see Berry *et al.* (2007) and references therein), there is no theoretical guarantee of convergence to a local minimum from these update rules and it cannot even not be proven that the algorithm converges in general to a stationary point. The only theoretical guarantee of these rules is their continual descent property. Modified rules have been subsequently proposed that resolve convergence issues Chih-Jen Lin (2007) but they often involve more work and make the multiplicative update algorithms even slower than they already were.

ALS NMF algorithms are a second class of algorithms developed to solve the NMF optimization problem. They rely on the observation that if d is a convex function, then the cost function function in (A.1) is not jointly convex in $(\mathbf{W}, \mathbf{H})$ but is convex in each matrix separately when the other is held fixed (Lee & Seung 2001). This observation has allowed for the development of various mathematical strategies that alternatively minimize between $\mathbf{W}$ and $\mathbf{H}$ (Cichocki, Zdunek & Amari 2007; H. Kim & Park 2008; Gillis & Glineur 2012). The majority of ALS-based algorithms iteratively repeat a combination of two steps that start with some least squares procedure on the first factor matrix considering the other is held fixed followed by a projection of the newly optimized matrix to set all negative elements to 0, and then the exact same operations considering the other factor matrix as a second step. This updates rules have the advantage of being very fast but little can be said about global convergence in the general cases. Using directly a NNLS procedure in the alternative minimization between the matrices $\mathbf{W}$ and $\mathbf{H}$ to avoid the projection results in update rules with better theoretical properties but at the cost of much more work for every iteration and

therefore much slower algorithms. For this reason, researchers have often settled for the simple projection to ensure non-negativity in practical uses although convergence cannot be guaranteed. More recent works have tried to improve upon the slowness of the alternative NNLS, particularly the works of Gong & Zhang (2012) or of Huang *et al.* (2015) which proposed algorithm capable of achieving fast quadratic convergence under certain conditions. However, the latter algorithms only apply to the cost function built from using the Euclidean distance, thereby restricting their possible applications only to data types that would be best analyzed through this cost function.

The third and last broad class of NMF algorithms are algorithms that apply a gradient descent of the cost function to update the factor matrices. As for every gradient descent, the key parameter to the algorithm is the size of the step taken in the direction of the negative gradient. In their seminal work, Paatero & Tapper (1994) described a gradient descent algorithm that converges in "30 to 100 steps". The algorithm presented in Hoyer (2004) for Euclidean distance-based cost function initially sets the step size to 1 and then multiply it by one-half for every iteration. Though simple and fast, gradient descent algorithms cannot generally guarantee non-negativity and therefore incorporate a projection step as for the ALS algorithms to ensure that the coefficients are non-negative after every iteration. As for ALS algorithms, little can generally be said about the convergence of gradient descent algorithms, in particular when a projection step is included as it makes formal analysis of convergence much more difficult. Of note, the multiplicative update rules derived by Lee and Seung originally derive from a gradient descent algorithm in which the step sizes were carefully selected to result in the convenient multiplicative updates.

As already mentioned earlier, many algorithmic improvements have been proposed over the years to achieve higher efficiency, better theoretical guarantees or fine-tune the algorithms to specific needs or constraints on the solutions. A property commonly sought in applications is to enforce sparsity or control over the magnitude of the factor matrices coefficients using L_1- or L_2-norm constraints. Mathematically, such constraints may be enforced by adding terms to the cost function in (A.1) in the form of

$$\underset{\mathbf{W},\mathbf{H}\in\mathbb{R}^{F\times K}\times\mathbb{R}^{K\times N}}{\arg\min} \; d(\mathbf{M}|\mathbf{WH}) + \alpha_W J_W(\mathbf{W}) + \alpha_H J_H(\mathbf{H}) \tag{A.4}$$

where J_W and J_H are usually convex functions so that theoretical guarantees of the modified update rules may be preserved. In their reference book on NMF and its applications, Cichocki, Zdunek, Phan, *et al.* (2009) present many variations of already generalized algorithms for different families of divergences that incorporate constraints imposed by the additional terms in (A.4). The book gathers and extends many of the authors works presented in different articles, particularly in Cichocki, Amari, *et al.* (2006). The reader is also referred to their very comprehensive matlab toolbox Cichocki, Zdunek, Choi, *et al.* (2003) for practical implementations of many variants of NMF algorithms. Févotte and Idier also presented modified rules in their work on β divergences to incorporate any constraint formalized as in (A.4) using convex functions for J_W and J_H and showed that the theoretical property of continual descent is preserved (Févotte & Idier 2011). Among other algorithmic improvements for NMF we can mention hierarchical NMF which iteratively refactorizes the right-most

factor starting from the first factorization $\mathbf{W_1 H_1}$ so that after L iterations the final product $\mathbf{WH}$ stems from the matrices $\mathbf{W} = \prod_{l=1}^{L} \mathbf{W_l}$ and $\mathbf{H} = \mathbf{H_L}$. Cichocki and colleagues have found that this simple procedure combined with multi-start initialization can improve the performance of most NMF algorithms (Cichocki & Zdunek 2006). Various efforts have also been dedicated to improving NMF by carefully selecting initial values for the factor matrices and have shown improved efficiency in many algorithms (Wild 2003; Boutsidis & Gallopoulos 2008).

A.2.2.3. Optimal value of K

In most applications, the number K of sources contributing additively to the observed signals is not known a priori and must therefore be estimated somehow. In the absence of a global statistical framework for the analysis performed that include a model over K, there is no natural criterion to be optimized and researchers resort to heuristics instead. In one of the first applications of NMF to genomic data, Brunet *et al.* (2004) used the stochastic nature of NMF according to initial conditions to assess the robustness of the factorization for a range of possible values of K. For each candidate rank, NMF was repeated with different initializations and a consensus matrix assessing how often any pair of observation clusters together was computed. In an ideal scenario, all coefficients of the consensus matrix would be either 0 or 1. The dispersion of the values between 0 and 1, measured by the *cophenetic correlation coefficient*, served as a quantitative assessment of the stability of the factorisations of a given rank. The value of this coefficient was then plotted against the candidate values of K and the optimal rank was chosen as the value where the magnitude of the cophenetic correlation coefficients begins to fall. This last steps involves a subjective assessment of the "falling point" and may be very hard or impossible to detect in cases where the curve is only slowly changing or on the contrary displays erratic increases and decreases.

In their seminal work on mutational signatures, Alexandrov *et al.* (2013) iteratively ran an NMF-based extraction model for many candidate values of K. For every candidate rank, two quantitative metrics assessing the stability of the NMF factorization and the quality of the reconstruction were computed using the silhouette index on repeated applications of the NMF and the Euclidean distance between the original matrix and the product of the two final factor matrices, respectively. These two metrics were then drawn against the candidate factorization sizes and the best size was chosen by a subtle trade-off between these two metrics. As for the first heuristic presented, this selection rule involves a subjective step that may not always be easily applicable and which lacks theoretical guarantees. As other tools were developed to perform *de novo* extraction of mutational signatures from positive matrices of mutation counts, other criteria were used to choose the number of signatures. `SomaticSignatures` (Gehring *et al.* 2015) simply uses the goodness-of-fit assessed via the Euclidean distance-based reconstruction error, as already done years earlier by P. M. Kim & Tidor (2003) on gene array experiments, whereas `EMu` (Fischer *et al.* 2013) and `signeR` (Rosales *et al.* 2017) make us of the Bayesian information criterion.

In their Bayesian formulation of NMF, Tan and Févotte devised a probabilistic framework

which allows to estimate the optimal number of components using a technique called *automatic relevance determination* - a technique that was already employed for Bayesian PCA for instance. Briefly, an NMF probabilistic model with a large number of factors - larger than the expected number - is fitted using "precision-like parameters" on the columns of $\mathbf{W}$ and rows of $\mathbf{H}$. These parameters are estimated a posteriori alongside the components of the factor matrices starting from pre-specified prior distributions. After fitting, a certain number of these parameters are driven to a large upper bound and serve to identify irrelevant components.

A.3. Annexes to chapter 3

A.3.1. Data retrieval and curation

The treatment histories of the META-PRISM patients were retrieved using a combination of automatic and manual techniques including the application of a large regex over all electronic health records stored in Dr Warehouse, as mentioned in Section 3.1.1.3. The regex employed is the following.

```
QARZIBA|AZD\s*\-*_*,*?8186|ERLEADA|IPILIMUMAB|ORTERONEL|MPDL\s*\-*_*,*?3281|VITRAKVI|ORMANDYL|
ZYKADIA|ATRIANCE|VORINOSTAT|ENTRECTINIB|AGI\s*\-*_*,*?5198|TORISEL|GDC\s*\-*_*,*?0068|ATEZOLIZUMAB|
CONTRACNE|OMIPALISIB|MTOR\s*\-*_*,*?INHIBITOR|PERTUZUMAB|GLIADEL|ENASIDENIB|MK\s*\-*_*,*?2206|
AZ\s*\-*_*,*?909|ASP\s*\-*_*,*?9521|BINIMETINIB|IRESSA|NOVATREX|RO\s*\-*_*,*?5083945|
XRP\s*\-*_*,*?6258|FLUOROURACIL|GLUCOVANCE|
ANTI\s*\-*_*,*?CTLA\s*\-*_*,*?4\s*\-*_*,*?MONOCLONAL\s*\-*_*,*?ANTIBODY|AURICULARUM|ONARTUZUMAB|
EMTANSINE\s*\-*_*,*?TRASTUZUMAB|5\s*\-*_*,*?AZACYTIDINE|LORVIQUA|WZ\s*\-*_*,*?4002|KIDROLASE|
MOCETINOSTAT|G007\s*\-*_*,*?LK|IDELALISIB|VOTUBIA|OPDIVO|VECTIBIX|D791LC00001|CABOMETYX|MAXIDEX|
DETURGYLONE|TAXOTERE|GSK\s*\-*_*,*?321|SU\s*\-*_*,*?5402|XELODA|METFORMINE|AG\s*\-*_*,*?120|
ALISERTIB|GANETESPIB|FRAKIDEX|GEFITINIB|STATTIC|LEE\s*\-*_*,*?011|HKI\s*\-*_*,*?272|
BMS\s*\-*_*,*?936558|ALBUMINE|AMINOGLUTETHIMIDE|CAMPTO|MEKTOVI|STAUROSPORINE|ODOMZO|
PHA\s*\-*_*,*?848125AC|AZD\s*\-*_*,*?5363|PACLITAXEL|TANESPIMYCIN|STIVARGA|FOLFOX\s*\-*_*,*?4|
AG\s*\-*_*,*?1296|PILARALISIB|LYNPARZA|ONAPRISTONE|
7\s*\-*_*,*?ETHYL\s*\-*_*,*?10\s*\-*_*,*?HYDROXYCAMPTOTHECIN|MEDI\s*\-*_*,*?4736|WNT\s*\-*_*,*?974|
ACIDE\s*\-*_*,*?FOLIQUE|RUXOLITINIB|TRISENOX|MDPL\s*\-*_*,*?3280A|DEXAFREE|AFIMOXIFENE|
BGJ\s*\-*_*,*?389|TEMSIROLIMUS|TUCATINIB|ALBUMINE\s*\-*_*,*?PACLITAXEL|PEGASYS|SAVARINE|FEMARA|
TAE\s*\-*_*,*?684|CELLTOP|VELCADE|INIPARIB|ANTI\s*\-*_*,*?CD\s*\-*_*,*?33|NUTLIN\s*\-*_*,*?3A|
TOPOTECAN|XTANDI|FASUDIL|PERJETA|REVLIMID|ONC\s*\-*_*,*?201|ATINIB|GDC\s*\-*_*,*?0941|
ANTHRACYCLINE\s*\-*_*,*?ANTINEOPLASTIC\s*\-*_*,*?ANTIBIOTIC|CAPECITABINE|FOTEMUSTINE|
ANTI\s*\-*_*,*?VEGF\s*\-*_*,*?MONOCLONAL\s*\-*_*,*?ANTIBODY|LONAFARNIB|TANDUTINIB|DAUNORUBICINE|
BEVACIZUMAB|AZ\s*\-*_*,*?628|FASLODEX|GEMZAR|VINTAFOLIDE|BAFETINIB|AZD\s*\-*_*,*?2281|BRIGATINIB|
COSMEGEN|KANJINTI|PEGARGIMINASE|SNS\s*\-*_*,*?032|SUNITINIB|NOLVADEX|DAUNORUBICIN|VANDETANIB|
AZD\s*\-*_*,*?8055|DACOMITINIB|BRIVANIB|AEE\s*\-*_*,*?788|EUCREAS|THIOGUANINE|ONTRUZANT|
ANTI\s*\-*_*,*?EGFR\s*\-*_*,*?MONOCLONAL\s*\-*_*,*?ANTIBODY|BIRABRESIB|VINFLUNINE|
EOS\s*\-*_*,*?E\s*\-*_*,*?3810|SIROLIMUS|OICR\s*\-*_*,*?9429|JQ\s*\-*_*,*?1|EYLEA|DEXAMETHASONE|
TASONERMINE|VELBE|TIPIFARNIB|VELIPARIB|AGERAFENIB|I\s*\-*_*,*?BET\s*\-*_*,*?151|TK\s*\-*_*,*?216|
BIOSIMILAIRE\s*\-*_*,*?RITUXIMAB|GO\s*\-*_*,*?6983|LY\s*\-*_*,*?3009120|VENCLYXTO|U\s*\-*_*,*?0126|
ALPHARADIN|AROMATASE\s*\-*_*,*?INHIBITOR|KOMBOGLYZE|BUPARLISIB|CERTICAN|SELUMETINIB|
PICTILISIB\s*\-*_*,*?BISMESYLATE|VENETOCLAX|IMG\s*\-*_*,*?2005\s*\-*_*,*?5|CORTISAL|PANITUMUMAB|
NIACINAMIDE|ALECTINIB|CARMUSTINE|TRAZIMERA|NEXAVAR|CRIZOTINIB|VINCRISTINE|ALPELISIB|BORTEZOMIB|
DECTANCYL|FARYDAK|NSC\s*\-*_*,*?348884|VISTUSERTIB|COBIMETINIB|G\s*\-*_*,*?573|
GSK\s*\-*_*,*?2636771|LY\s*\-*_*,*?3039478|TASELISIB|JQ1\s*\-*_*,*?COMPOUND|ZYTIGA|ASPARAGINASE|
COAPROVEL|PONATINIB|GSK\s*\-*_*,*?1120212|WYE\s*\-*_*,*?354|AS\s*\-*_*,*?602868|DACARBAZINE|
TRETINOINE|SONIDEGIB|REFAMETINIB|TOMUDEX|ZANEA|RUCAPARIB|PD1\s*\-*_*,*?INHIBITOR|ETOPOPHOS|
MK\s*\-*_*,*?3475\s*\-*_*,*?028|DEXSOL|VEMURAFENIB|KU\s*\-*_*,*?0060648|COTELLIC|JAKAVI|TECENTRIQ|
ALKERAN|AV\s*\-*_*,*?203|GDC\s*\-*_*,*?0623|METOJECT|POLYDEXA|BOSUTINIB|ETOPOSIDE|APROVEL|VESANOID|
```

```
CERUBIDINE|IBRUTINIB|HERZUMA|VINFLUNINE\s*\-*_*,*?JASINT|VTX\s*\-*_*,*?11E|S\s*\-*_*,*?49076|
SAR\s*\-*_*,*?125844|IMBRUVICA|PLAQUENIL|PAZOPANIB|OXALIPLATINE|OTX\s*\-*_*,*?015|QUARFLOXIN|
EMTANSINE|CAPMATINIB|IXAZOMIB|CISPLATINE|TAXANE\s*\-*_*,*?COMPOUND|ARQ\s*\-*_*,*?197|NILOTINIB|
CO\s*\-*_*,*?1886|AMRUBICIN|PURINETHOL|BELEODAQ|XOSPATA|GDC\s*\-*_*,*?0994|RAPAMYCINE|
PCB\s*\-*_*,*?COMET\s*\-*_*,*?1\s*\-*_*,*?XL184\s*\-*_*,*?307|PANOBINOSTAT|DECITABINE|ASCIMINIB|
LOMETREXOL|NUTLIN\s*\-*_*,*?3|SAVOLITINIB|ALEMTUZUMAB|ICOTINIB|DOXORUBICIN|
ARSENIC\s*\-*_*,*?TRIOXIDE|MIDOSTAURIN|ONIVYDE|LETROZOLE|AZD\s*\-*_*,*?4547|LESTAURTINIB|POZIOTINIB|
BPTES|BAYER\s*\-*_*,*?1394|CAPIVASERTIB|SALINOMYCIN|SIROCTID|
PEGINTERFERON\s*\-*_*,*?ALFA\s*\-*_*,*?2A|NDPL\s*\-*_*,*?3280|AMG\s*\-*_*,*?172|ACTINOMYCINE|
GSK\s*\-*_*,*?2256098|RO\s*\-*_*,*?5509554|TARCEVA|ROACCUTANE|ONATASERTIB|CAMPATH|OSIMERTINIB|
BIBW\s*\-*_*,*?2992|SORAFENIB|CONATUMUMAB|PREDNISONE|MK\s*\-*_*,*?8109|MVASI|PCV\s*\-*_*,*?REGIMEN|
ZELBORAF|LUMINESPIB|TRAMETINIB\s*\-*_*,*?DIMETHYL\s*\-*_*,*?SULFOXIDE|BGJ\s*\-*_*,*?398|CEDIRANIB|
BUSILVEX|PRALSETINIB|ZYDELIG|RAMUCIRUMAB|DACTOLISIB|CX\s*\-*_*,*?5461|FRAMYXONE|IMETH|VANFLYTA|
ADO\s*\-*_*,*?TRASTUZUMAB|MEN\s*\-*_*,*?1611|PROCUTA|INLYTA|TRASTUZUMAB|APITOLISIB|
BEVACIZUMAB\s*\-*_*,*?ROSIA|PEMIGATINIB|AZD\s*\-*_*,*?3463|ENDOXAN|PEMBROLIZUMAB|
ANTI\s*\-*_*,*?TIM\s*\-*_*,*?3\s*\-*_*,*?MONOCLONAL\s*\-*_*,*?ANTIBODY|XALUPRINE|RINDOPEPIMUT|
SPLICEOSTATIN\s*\-*_*,*?A|MEDI\s*\-*_*,*?0680|SB\s*\-*_*,*?202190|CYCLOPHOSPHAMIDE|ANASTROZOLE|
MUPHORAN|LIBTAYO|NELARABINE|NERATINIB|SALIRASIB|VOTRIENT|DURVALUMAB|ENZASTAURIN|KEYTRUDA|OFEV|
FEDRATINIB|PLX\s*\-*_*,*?4720|DELIPROCT|CAPRELSA|TEPOTINIB|BRAFTOVI|GIOTRIF|BICNU|CAELYX|
DINUTUXIMAB|EFFEDERM|SGX\s*\-*_*,*?523|BEROMUN|TUKYSA|RIDAFOROLIMUS|ARS\s*\-*_*,*?853|CHLORAMBUCIL|
METHYLPREDNISOLONE|PROLASTIN|RALTITREXED|VERZENIOS|ANGIOGENESIS\s*\-*_*,*?INHIBITOR|
RO\s*\-*_*,*?5520985|GSK\s*\-*_*,*?3377794|NU\s*\-*_*,*?7441|AZD\s*\-*_*,*?9496|IFIRMASTA|
UPROSERTIB|BOSULIF|OLAPARIB|LAROTRECTINIB|RITUXIMAB|PD\s*\-*_*,*?0325901|PROLIA|TRAMETINIB|
TRICHOSTATIN\s*\-*_*,*?A|METHOTREXATE|FISOGATINIB|RO\s*\-*_*,*?4987655|ZIRABEV|GLIVEC|JAVLOR|
IMFINZI|ALUNBRIG|PD\s*\-*_*,*?180970|IRINOTECAN|ONCOVIN|SOTRASTAURIN\s*\-*_*,*?ACETATE|
VALPROIC\s*\-*_*,*?ACID|DETICENE|MM\s*\-*_*,*?141|LORLATINIB|AFLIBERCEPT|CYRAMZA|AZD\s*\-*_*,*?1480|
IDARUBICIN|IFOSFAMIDE|ARIMIDEX|KETREL|EFUDIX|GF109203X|CUSTIRSEN|IMATINIB|BICALUTAMIDE|
O6\s*\-*_*,*?BENZYLGUANINE|NAVELBINE|CETUXIMAB\s*\-*_*,*?IFCT\s*\-*_*,*?08\s*\-*_*,*?03|EXEMESTANE|
MYLERAN|SCH\s*\-*_*,*?772984|IMATINIB\s*\-*_*,*?MESYLATE|NOMEGESTROL|BENDAMUSTINE|OZURDEX|NIVAQUINE|
ERIBULINE|INFIGRATINIB|ZOLINZA|TGX\s*\-*_*,*?221|F\s*\-*_*,*?14512|OGX\s*\-*_*,*?427|BAZEDOXIFENE|
VINBLASTINE|ZOELY|AZD\s*\-*_*,*?7762|LEUCOVORIN|TEMOZOLOMIDE|LAPATINIB|AZD\s*\-*_*,*?9291|TEPADINA|
STREPTOZOCINE|APALUTAMIDE|TYVERB|TRUXIMA|RAF\s*\-*_*,*?265|H3B\s*\-*_*,*?8800|
BET\s*\-*_*,*?INHIBITOR|LEVOLEUCOVORIN|ABRAXANE|ARACYTINE|PEMETREXED|ALECENSA|MEK\s*\-*_*,*?162|
CABOZANTINIB|ERLOTINIB|NIVOLUMAB|FOLFIRI|SAPANISERTIB|ANAMORELINE|XIGDUO|FOLFOX|CYTARABINE|
ANTI\s*\-*_*,*?CD\s*\-*_*,*?123|MERCAPTOPURINE|MODOTUXIMAB|BUSULFAN|TEGAFUR|ERIVEDGE|
REGORAFENIB\s*\-*_*,*?PBT|ZAVEDOS|AVE\s*\-*_*,*?8062|DOXORUBICINE\s*\-*_*,*?LIPOSOMALE|TEMODAL|
ABEMACICLIB|A\s*\-*_*,*?66|TASQUINIMOD|FLUDARA|JNJ\s*\-*_*,*?42756493|DAUNOXOME|MELPHALAN|
HERCEPTINJANUMET|AMUVATINIB|FLUTAMIDE|LY\s*\-*_*,*?294002|BAY\s*\-*_*,*?1125976|
PF\s*\-*_*,*?06293622|AMG\s*\-*_*,*?386\s*\-*_*,*?PBT|ICLUSIG|
TRIOXYDE\s*\-*_*,*?D\s*\-*_*,*?ARSENIC|DERUXTECAN|ACLY\s*\-*_*,*?SIRNA|TAXOL|YERVOY|
AZD\s*\-*_*,*?6738|ACIDE\s*\-*_*,*?ZOLEDRONIQUE|TAGRISSO|ISTODAX|GDC\s*\-*_*,*?0575|ALVOCIDIB|
CARBOPLATIN\s*\-*_*,*?TAXOL|PCB\s*\-*_*,*?COMET\s*\-*_*,*?1|DOCETAXEL|FORETINIB|TAFINLAR|RYDAPT|
FUTUXIMAB|ROCILETINIB|GSK\s*\-*_*,*?126|ALIMTA|ELOXATINE|EVEROLIMUS\s*\-*_*,*?CA209\s*\-*_*,*?025|
QUIZARTINIB|RETACNYL|TOBRADEX|AMETYCINE|HOLOXAN|TAS\s*\-*_*,*?120|ASPIRIN|
ETOPOSIDE\s*\-*_*,*?PHOSPHATE|IFIRMACOMBI|SGK1\s*\-*_*,*?INH|LENALIDOMIDE|ZANOSAR|NILUTAMIDE|
PANRETIN|AFINITOR|METFORMIN|FULVESTRANT|HYCAMTIN|TAMOXIFEN|AZACITIDINE|TASIGNA|CEMIPLIMAB|
IRAK\s*\-*_*,*?1/4\s*\-*_*,*?INHIBITOR|CI\s*\-*_*,*?1040|RG\s*\-*_*,*?7356|HALAVEN|AROMASINE|
RETASPIMYCIN\s*\-*_*,*?HYDROCHLORIDE|KISQALI|MELOXICAM|CILOXADEX|DOVITINIB|ALFALASTIN|TOCTINO|
CHLOROQUINE|PI\s*\-*_*,*?103|KADCYLA|ZOLEDRONIC\s*\-*_*,*?ACID|ARMISARTE|CETUXIMAB|
GDC\s*\-*_*,*?0425|CHLORAMINOPHENE|AMG\s*\-*_*,*?510|VYXEOS|DASATINIB|ALIZEM|
PHA\s*\-*_*,*?848125\s*\-*_*,*?AC|PREXASERTIB|LINSITINIB|BLEOMYCINE|ENZALUTAMIDE|RIXATHON|
TRABECTEDIN|OGIVRI|AMGMDS\s*\-*_*,*?3|CURACNE|IMMUNE\s*\-*_*,*?CHECKPOINT\s*\-*_*,*?INHIBITOR|
XL\s*\-*_*,*?184|IRBESARTAN|NINLARO|ADRIBLASTINE|GEMCITABINE|PANITIMUMAB|CARBOPLATIN|
2,4\s*\-*_*,*?PYRIMIDINEDIAMINE|XALKORI|AMG\s*\-*_*,*?386|IVOSIDENIB|E\s*\-*_*,*?7438|
MM\s*\-*_*,*?121|MEGACE|BYL\s*\-*_*,*?179|OCTREOTIDE|CABAZITAXEL|AZD\s*\-*_*,*?5438|PARAPLATINE|
VISMODEGIB|GDC\s*\-*_*,*?0449|
ANTI\s*\-*_*,*?PD\s*\-*_*,*?L1\s*\-*_*,*?MONOCLONAL\s*\-*_*,*?ANTIBODY|XGEVA|CBL\s*\-*_*,*?0137|
NINTEDANIB|JW\s*\-*_*,*?55|BAVENCIO|MEDROL|CC\s*\-*_*,*?223|PLACEBO|VEPESIDE|NIMOTUZUMAB|
ALVESPIMYCIN|GSK\s*\-*_*,*?690693|ARABINOSYLGUANINE|MEKINIST|SERIBANTUMAB|DEBIO\s*\-*_*,*?1347|
SUTENT|TEPROTUMUMAB|MOBIC|CORTANCYL|ERYLIK|VINORELBINE|LUTENYL|UNC\s*\-*_*,*?1062|VIDAZA|
RG\s*\-*_*,*?7112|DENOSUMAB|HYDROCORTANCYL|PP\s*\-*_*,*?242|GILTERITINIB|SANDOSTATINE|RESPREEZA|
TALZENNA|PATRITUMAB|MAXIDROL|ERBITUX|RAPAMUNE|ACIDE\s*\-*_*,*?FOLINIQUE|EVEROLIMUS|PERIFOSINE|
```

```
AXITINIB|SAR\s*\-*_*,*?408701|MEHD7954A|MOTESANIB|OXALIPLATIN|TALAZOPARIB|PLACEBO\s*\-*_*,*?IMPRESS|
YONDELIS|MABTHERA|IPATASERTIB|RIPRETINIB|DINACICLIB|BELINOSTAT|CERITINIB|STAGID|NEOFORDEX|
GDC\s*\-*_*,*?0879|RAPALINK\s*\-*_*,*?1|S\s*\-*_*,*?78454|PD\s*\-*_*,*?173074|DABRAFENIB|IDHIFA|
CIXUTUMUMAB|JEVTANA|ABIRATERONE|CHIBRO\s*\-*_*,*?CADRON|MITOMYCIN|
MDV3100\s*\-*_*,*?AFFIRM\s*\-*_*,*?OUVERT|ZEJULA|KW\s*\-*_*,*?2449|MARQIBO|ZALTRAP|NIRAPARIB|
POMALIDOMIDE|EPIRUBICIN|WHI\s*\-*_*,*?P\s*\-*_*,*?154|NECITUMUMAB|LEVACT|DEXRAZOXANE|THIOTEPA|
AVASTIN|BIBF\s*\-*_*,*?1120|VELMETIA|GLUCOPHAGE|DERINOX|NORDIMET|GNE\s*\-*_*,*?617|PATIDEGIB|
RILOTUMUMAB|TAK\s*\-*_*,*?733|PREXATE|CARDIOXANE|GSK\s*\-*_*,*?2118436|FIRMAGON|JSI\s*\-*_*,*?124|
IMNOVID|IFCT\s*\-*_*,*?1003\s*\-*_*,*?LADIE|CRENOLANIB|CASODEX|ENCORAFENIB|RUBRACA|
ODM\s*\-*_*,*?201|REGORAFENIB|SELPERCATINIB|IODINE|IBRANCE|LANVIS|DEGARELIX|MDV\s*\-*_*,*?3100|
MC\s*\-*_*,*?1568|SAVENE|ERDAFITINIB|TAZEMETOSTAT|9F7\s*\-*_*,*?F11|RIBOCICLIB|ANANDRON|
TAK\s*\-*_*,*?700|XL184\s*\-*_*,*?307|PLX\s*\-*_*,*?8394|AKTI\s*\-*_*,*?1/2|AVAPRITINIB|CANERTINIB|
MEGESTROL|STERDEX|SPRYCEL|SU\s*\-*_*,*?5614|AVELUMAB|FARMORUBICINE|
MONOMETHYL\s*\-*_*,*?AURISTATIN\s*\-*_*,*?E|GW\s*\-*_*,*?2580|MITOMYCINE\s*\-*_*,*?C|
CYPROTERONE\s*\-*_*,*?ACETATE|PHA\s*\-*_*,*?848125AC\s*\-*_*,*?CDKO\s*\-*_*,*?006|PALBOCICLIB|
PREDNISOLONE|MYOCET|XRP\s*\-*_*,*?6976|METFORMINE\s*\-*_*,*EMBONATE|ALITRETINOINE|ISOTRETINOINE|
NAB\s*\-*_*,*PACLITAXEL|TRASTUZUMAB\s*\-*_*,*?EMTANSINE|TRASTUZUMAB\s*\-*_*,*?DERUXTECAN|
PACLITAXEL\s*\-*_*,*?ALBUMINE|IRINOTECAN\s*\-*_*,*?LIPOSOMALE|ADRIAMYCIN|LOMUSTINE|BELUSTINE|
ODM\s*\-*_*,*?203|EMACTUZUMAB
```

The following table, published as Supplementary Table 3 alongside the paper, delineates all the tumor type abbreviations for the tumors of META-PRISM patients and provides the number of patients associated to each tumor type.

Tumor Type	Description	META-PRISM
ACC	Adrenocortical Carcinoma	13
ANUS - Not_TCGA	Anal squamous cell carcinoma	6
BLCA	Bladder Urothelial Carcinoma	74
BLCA - Not_TCGA	Bladder Non-Urothelial Carcinoma	6
BRCA	Invasive Breast Carcinoma	98
CESC	Cervical Squamous Cell Carcinoma	13
CHOL	Cholangiocarcinoma	51
COAD	Colon Adenocarcinoma	49
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma	1
ESCA	Esophageal Adenocarcinoma	6
GBM	Glioblastoma Multiforme	4
HNAC - Not_TCGA	Head and Neck Adenoid Cystic Carcinoma	22
HNSC	Head and Neck Squamous Cell Carcinoma	51
KICH	Kidney Chromophobe	2
KIRC	Renal Clear Cell Carcinoma	11
KIRP	Renal Papillary Cell Carcinoma	2
LGG	Brain lower grade glioma	1
LIHC	Liver Hepatocellular Carcinoma	12

LUAD	Lung Adenocarcinoma	192
LUNE - Not_TCGA	Lung Neuroendocrine Carcinoma	18
LUSC	Lung Squamous Cell Carcinoma	27
MESO	Mesothelioma	4
MISC - Not_TCGA	Other Tumor	53
OV	Ovarian Serous Cystadenocarcinoma	21
PAAD	Pancreatic Adenocarcinoma	61
PCPG	Pheochromocytoma and Paraganglioma	3
PRAD	Prostate Adenocarcinoma	95
READ	Rectal Adenocarcinoma	8
SARC	Dedifferentiated liposarcoma, leiomyosarcoma, undifferentiated pleomorphic sarcoma, myxofibrosarcoma, malignant peripheral nerve sheath tumor, and synovial sarcoma	21
SARC - Not_TCGA	Sarcoma other than dedifferentiated liposarcoma, leiomyosarcoma, undifferentiated pleomorphic sarcoma, myxofibrosarcoma, malignant peripheral nerve sheath tumor, and synovial sarcoma	14
SI - Not_TCGA	Small Intestine Carcinoma	9
SKCM	Cutaneous Melanoma	4
STAD	Stomach Adenocarcinoma	26
TGCT	Testicular Germ Cell Tumor	9
THCA	Thyroid Papillary, Follicular, Oxyphilic, or Nonencapsulated Sclerosing Carcinoma	1
THCA - Not_TCGA	Thyroid Medullary or Undifferentiated Carcinoma	10
THYM	Thymoma	2
UCEC	Endometrial Carcinoma	6
UCS	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor	2
Unknown_Primary	Carcinoma of Unknown Primary	23

Table A.5.: List of all 39 cancer types for the 1,031 patients included in the META-PRISM cohort. Rare tumor types represented by 5 or less tumors and not represented in TCGA were grouped into the category "MISC - Not_TCGA".

A.3.2. Bioinformatic analyses

WGD	X_Male	TCN	LCN	SCNA
0	0	0		Homozygous deletion (HD)
0	0	1	0	Loss of heterozygosity (LOH)
0	0	2	0	Copy-neutral LOH (cn-LOH)
0	0	$\geq 3, < 4$		Low-level gain (LLG)
0	0	$\geq 4, < 6$		Medium-level gain (MLG)
0	0	≥ 6		High-level gain (HLG)
0	1	0		Homozygous deletion (HD)
0	1	$\geq 2, < 3$		Low-level gain (LLG)
0	1	$\geq 3, < 4$		Medium-level gain (MLG)
0	1	≥ 4		High-level gain (HLG)
k	0	0		Homozygous deletion (HD)
k	0	$> 0, < 2^{k+1} - (k-1)$	0	Loss of heterozygosity (LOH)
k	0	$\geq 2^{k+1} - (k-1), \leq 2^{k+1} + (k-1)$	0	Copy-neutral LOH (cn-LOH)
k	0	$\geq 1 + k + 2^{k+1}, < 3 + k + 2^{k+1}$		Low-level gain (LLG)
k	0	$\geq 3 + k + 2^{k+1}, < 5 + k + 2^{k+1}$		Medium-level gain (MLG)
k	0	$\geq 5 + k + 2^{k+1}$		High-level gain (HLG)
k	1	0		Homozygous deletion (HD)
k	1	$\geq k + 2^k, < 1 + k + 2^k$		Low-level gain (LLG)
k	1	$\geq 1 + k + 2^k, < 3 + k + 2^k$		Medium-level gain (MLG)
k	1	$\geq 3 + k + 2^k$		High-level gain (HLG)

Table A.6.: CNA segments identified by FACETS were classified into one of six categories according to the estimated number WGDs (0 or k >= 1), the TCN value, and the LCN value of the segment. Empty values for LCN mean that only the TCN value was used.

A.3.3. Comparison and validation cohorts

The following series of four figures provide details about the effects of different filters on the list of putative mutations and gene fusions in TCGA (Figure A.1 and A.2) and MET500 (Figure A.3 and A.4) cohorts.

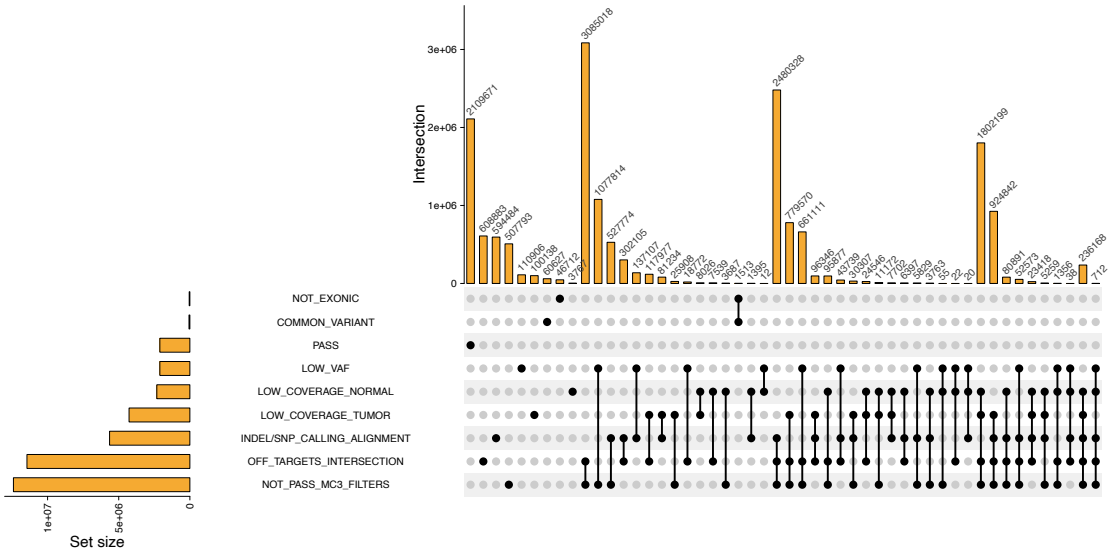


Fig. A.1.: Upset plot showing the number of mutations filtered out individually by each filtering criteria and in combination with other criteria in TCGA WES samples. Mutations that passed all filters are described in the PASS set.

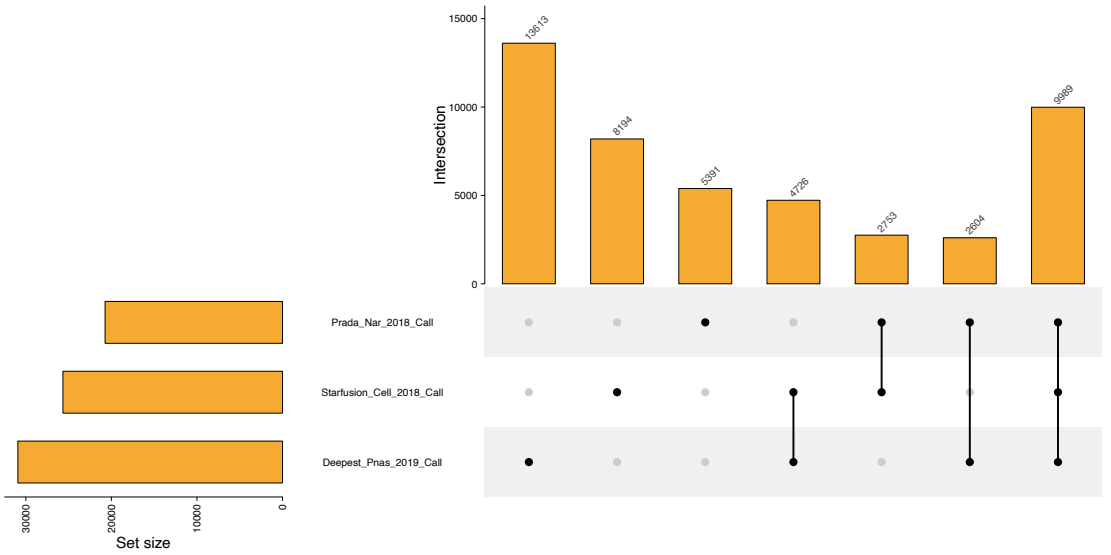


Fig. A.2.: Upset plot showing the number of gene fusions in common according to different combinations among the three external lists for TCGA RNA-seq samples.

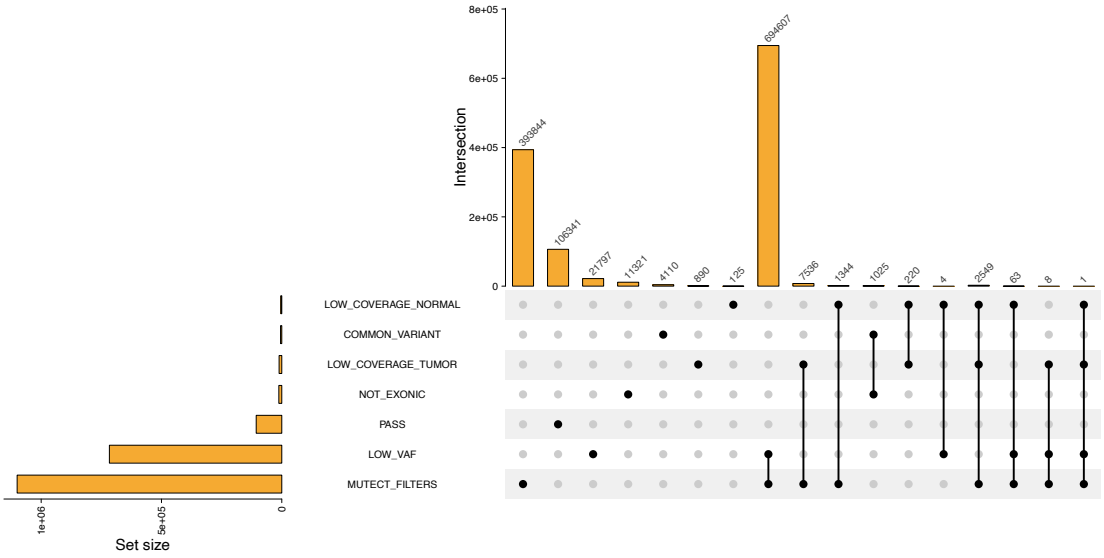


Fig. A.3.: Upset plot showing the number of mutations filtered out individually by each filtering criteria and in combination with other criteria in MET500 WES samples. Mutations that passed all filters are described in the PASS set.

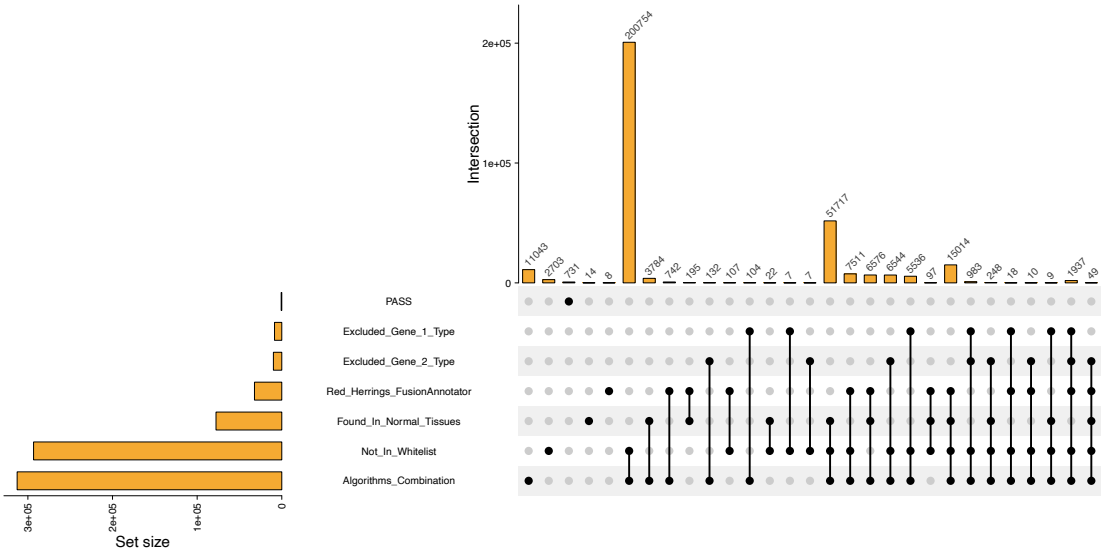


Fig. A.4.: Upset plot showing the number of gene fusions filtered out individually by each filtering criteria and in combination with other criteria in MET500 RNA-seq samples. Gene fusions that passed all filters are described in the PASS set.

The two following figures serve to show the good alignment in the calling of substitutions and indels between our internal pipeline and the application of specific filtering rules on the variants reported by MC3 project.

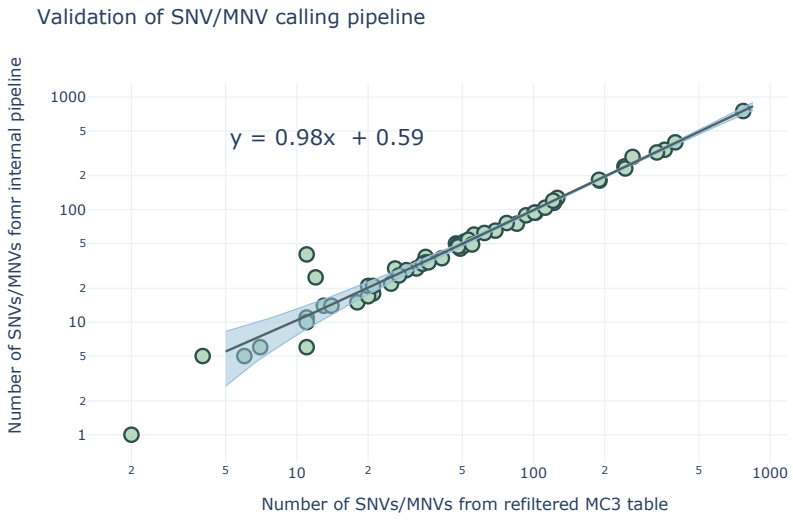


Fig. A.5.: Alignment in the calling of substitutions (SNV/MNV) between our internal pipeline and the refiltering of the MC3 table.

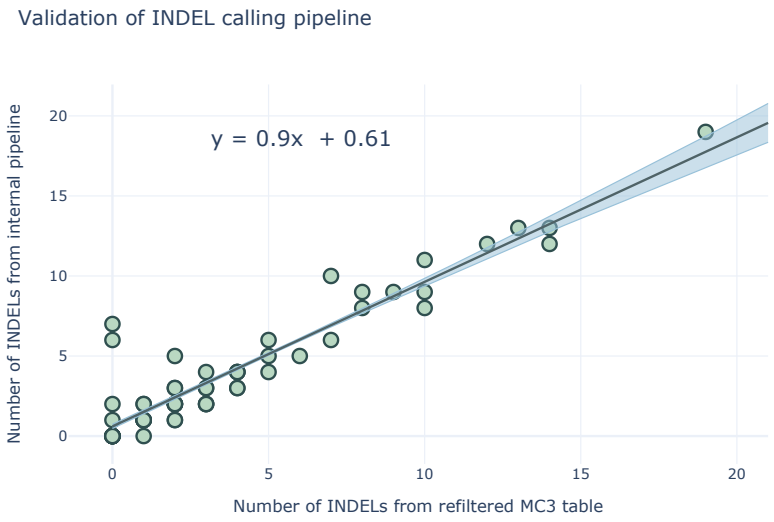


Fig. A.6.: Alignment in the calling of indels between our internal pipeline and the refiltering of the MC3 table.

A.3.4. Genomic profiles

The following list of five figures serve to illustrate some of the analyzes discussed in Section 3.3, in particular chromosome arm CNAs (Figure A.7 and A.8), discovery of cancer driver genes (Figure A.9), and the distribution of multihit events in selectedgenes (Figure A.10 and A.11).

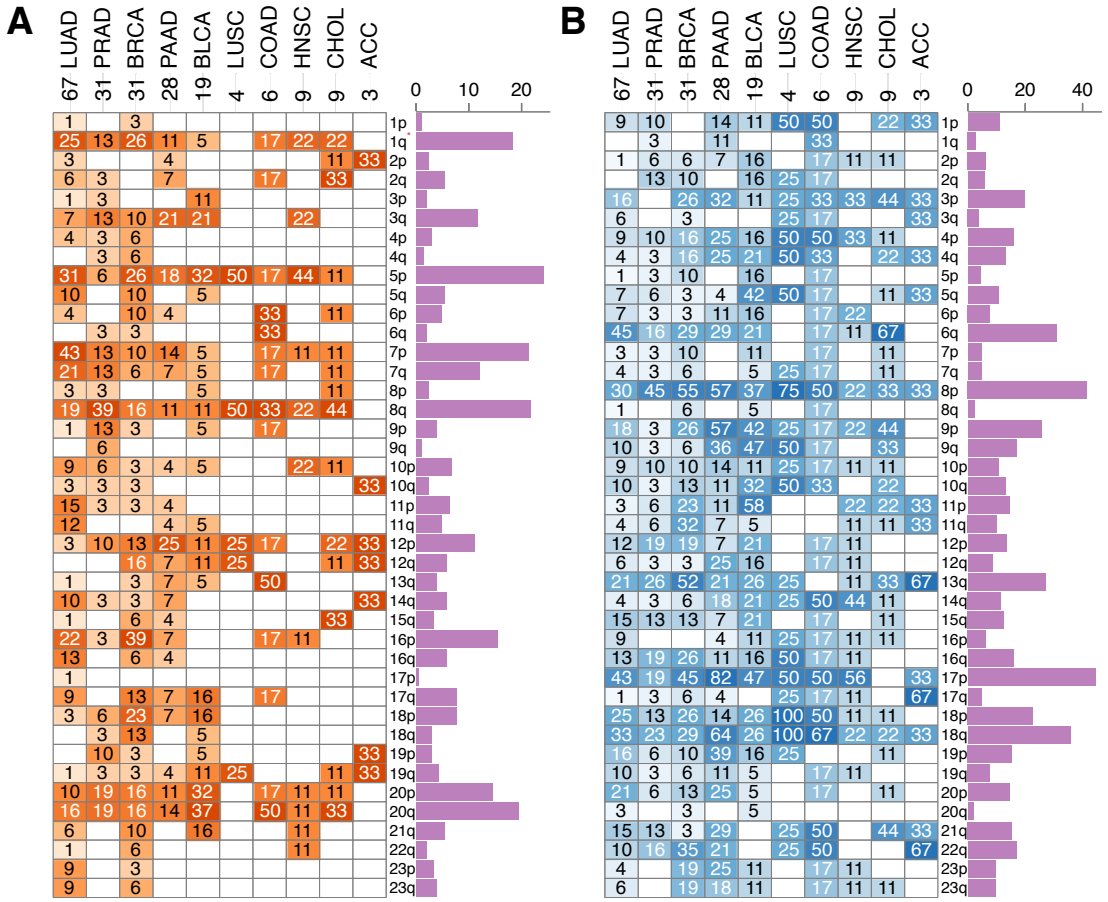


Fig. A.7.: Chromosome arm **A.** copy-gains and **B.** copy-losses in the tumor types of META-PRISM WES subcohort (10 tumor types). Heatmaps show the percentages of tumors affected by corresponding chromosome arm somatic CNAs in each tumor type. The absolute bar plots show the percentage of tumors in META- PRISM harboring the chromosome arm somatic CNA. After correction of p-values from Fisher-Boschloo tests, there was no significant change in any of the chromosome arms and tumor types.

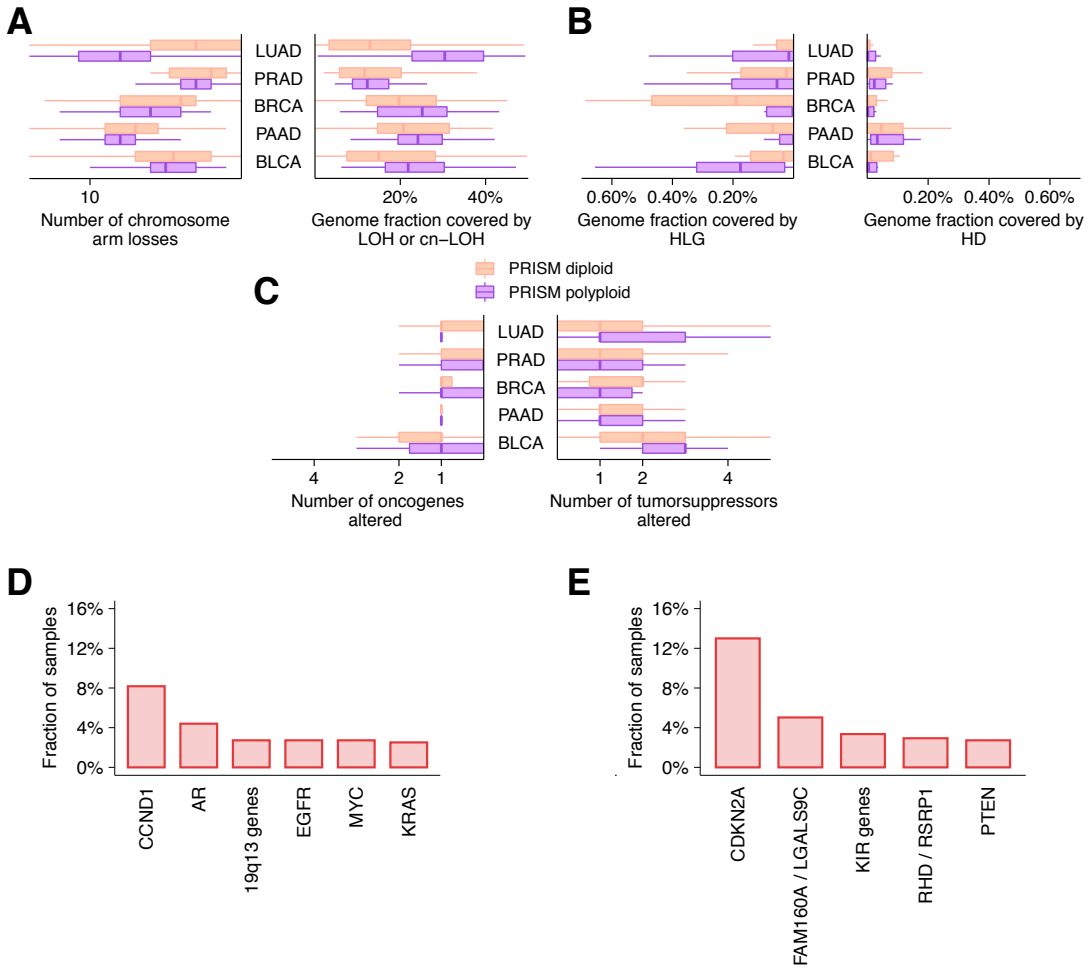


Fig. A.8.: A. Double box plot describing the number of chromosome arm losses (left) and the fraction of the genome covered by losses (loss of heterozygosity LOH and copy neutral cn-LOH; right) per tumor type. Comparisons of META-PRISM diploid vs. META-PRISM polyploid tumors were performed using Mann-Whitney U tests. No significant difference was observed. Only the 5 main tumor types are considered. **B.** Identical to **A.** but considering the fraction of the genome covered by focal high-level amplifications (HLG) and homozygous deletions (HD). **C.** Identical to **A.** but considering the number of oncogenes and tumor suppressor genes altered. D. Fraction of META-PRISM WES tumors harboring high-level gains for the genes or gene groups most frequently involved in somatic CNA. Only genes or gene groups altered in at least 2.5% of samples are shown. **E.** Identical to **D.** but considering homozygous deletions.

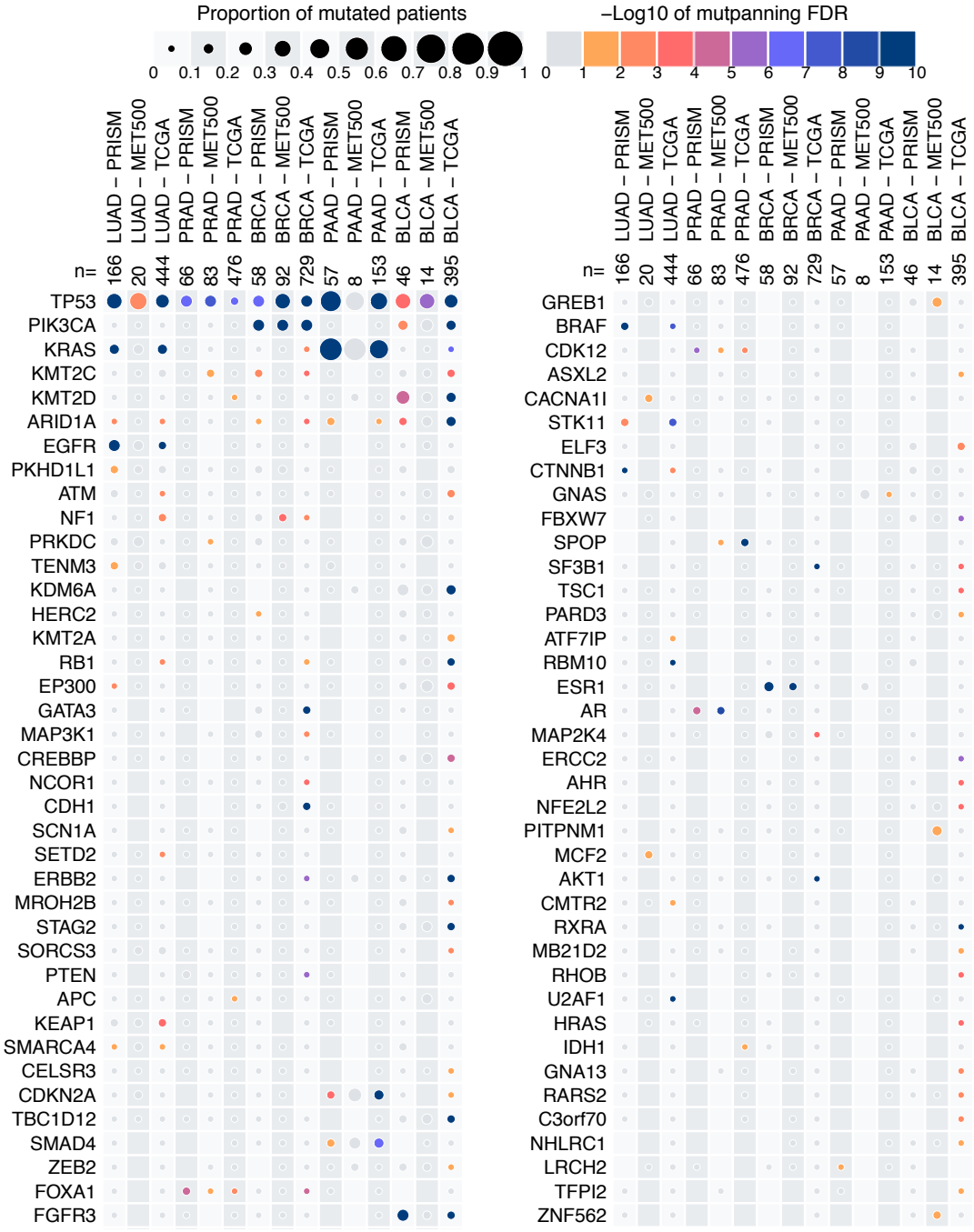


Fig. A.9.: Heatmap showing the driver genes (rows) identified by Mutpanning on the five tumor types most represented in META-PRISM WES tumors for each cohort (columns). The circle size encodes the percentage of patients harboring a somatic mutation of any type, while the circle color encodes the corrected p-value (FDR) computed by Mutpanning. Only genes reported as a driver (FDR < 0.01) in at least one subcohort are shown.

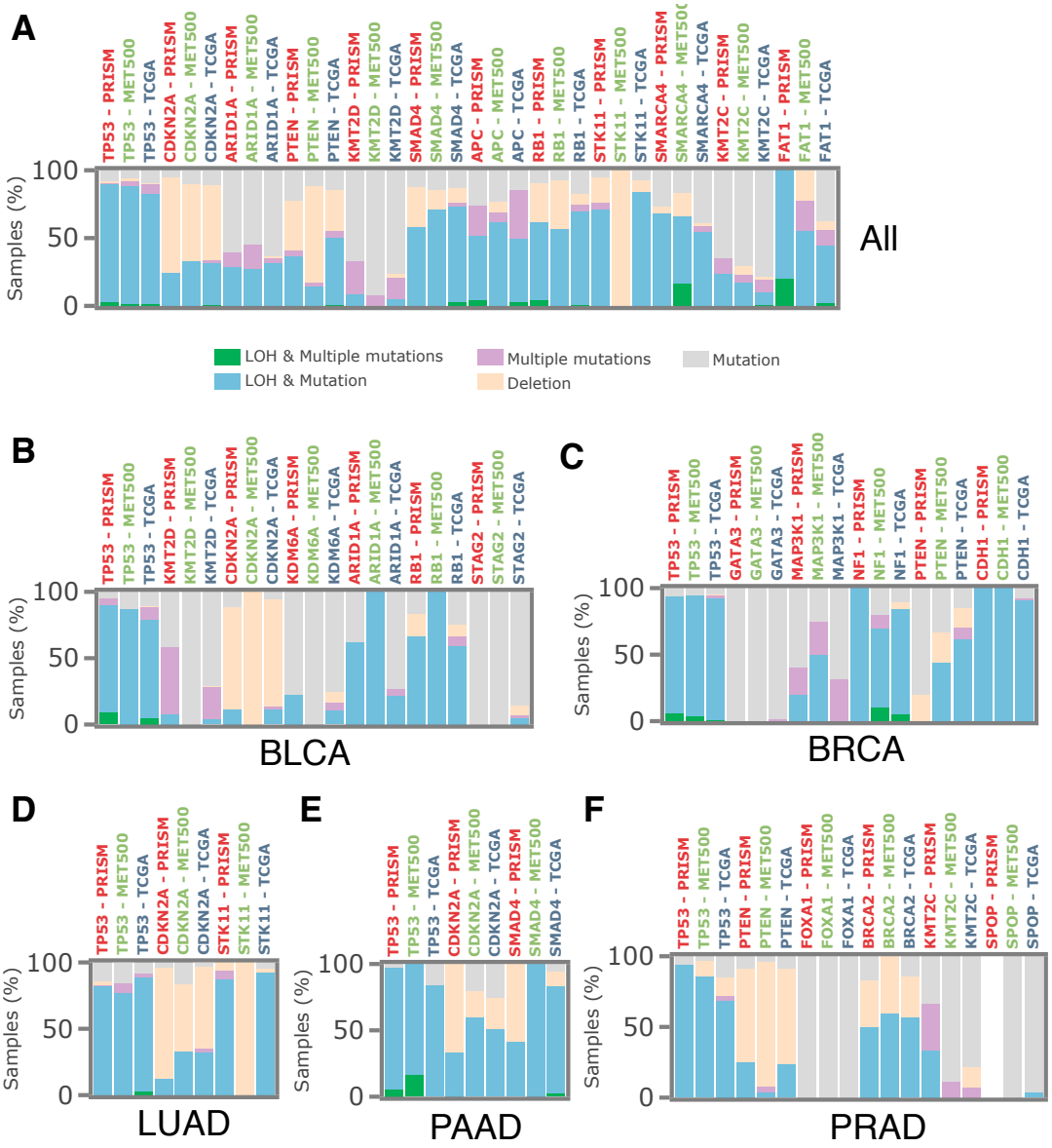


Fig. A.10.: Relative bar plots showing the relative frequency of single-hit and different types of multi-hit events in tumor suppressor genes for **A**. all 10 tumor types represented in META-PRISM WES subcohort, **B**. BLCA tumors, **C**. BRCA tumors, **D**. LUAD tumors, **E**. PAAD tumors, and **F**. PRAD tumors. For **A**., only genes altered in 5% of any cohort are included, while in single-tumor type plots the threshold was set to 10%.

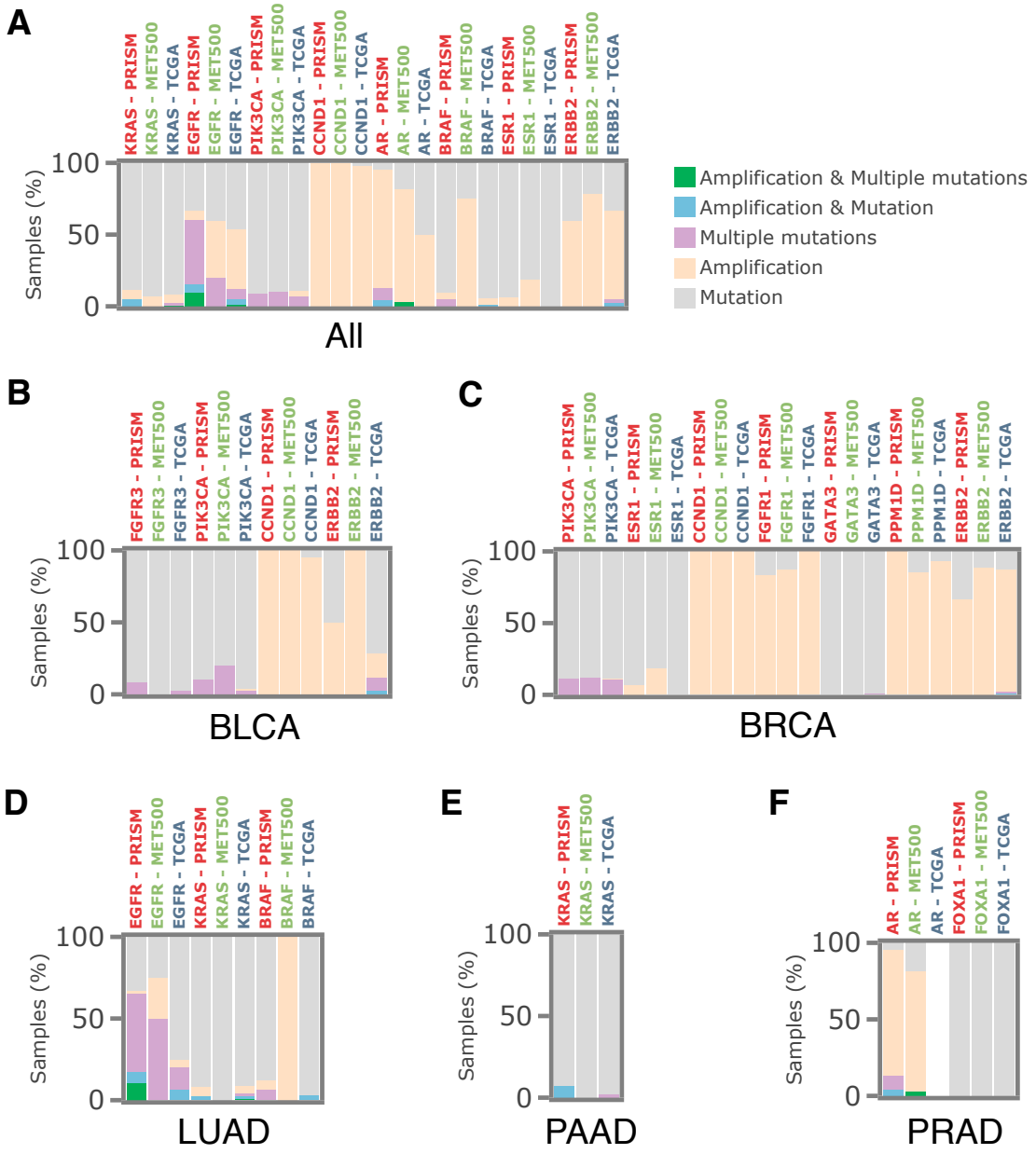


Fig. A.11.: Relative bar plots showing the relative frequency of single-hit and different types of multi-hit events in oncogenes for **A**. all 10 tumor types represented in META-PRISM WES subcohort, **B**. BLCA tumors, **C**. BRCA tumors, **D**. LUAD tumors, **E**. PAAD tumors, and **F**. PRAD tumors. For **A**., only genes altered in 5% of any cohort are included, while in single-tumor type plots the threshold was set to 10%.

A.3.5. Transcriptomic profiles

The following figure aims to assess visually the potential sources of batch effects in the RNA-seq profiles of tumors from all three cohorts of the study.

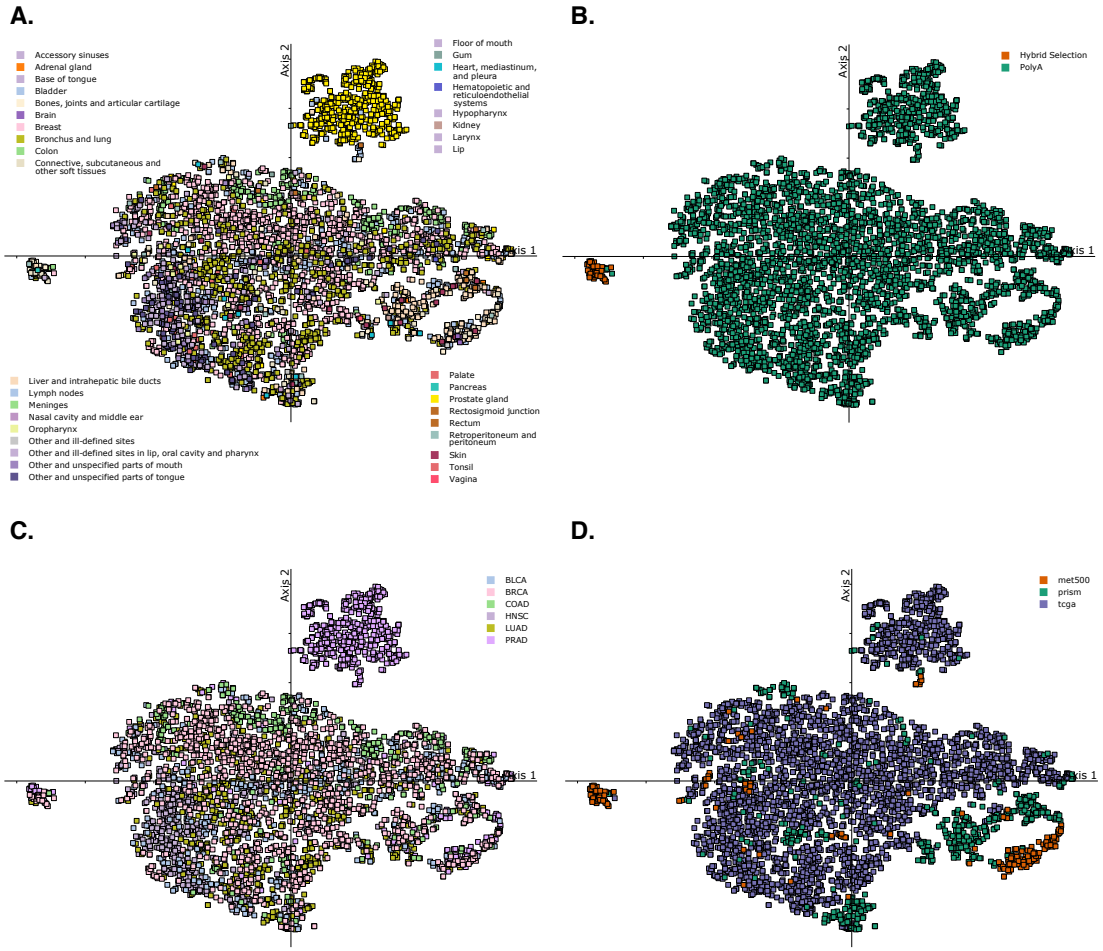


Fig. A.12.: TSNE representations of transcriptomics profiles from six tumors types and three different studies colored by **A**. biopsy site **B**. RNA selection protocol **C**. tumor type **D**. study

A.3.6. Improved survival predictions

The following figure provides details about the M7 survival model trained on BRCAs of META-PRISM cohort.

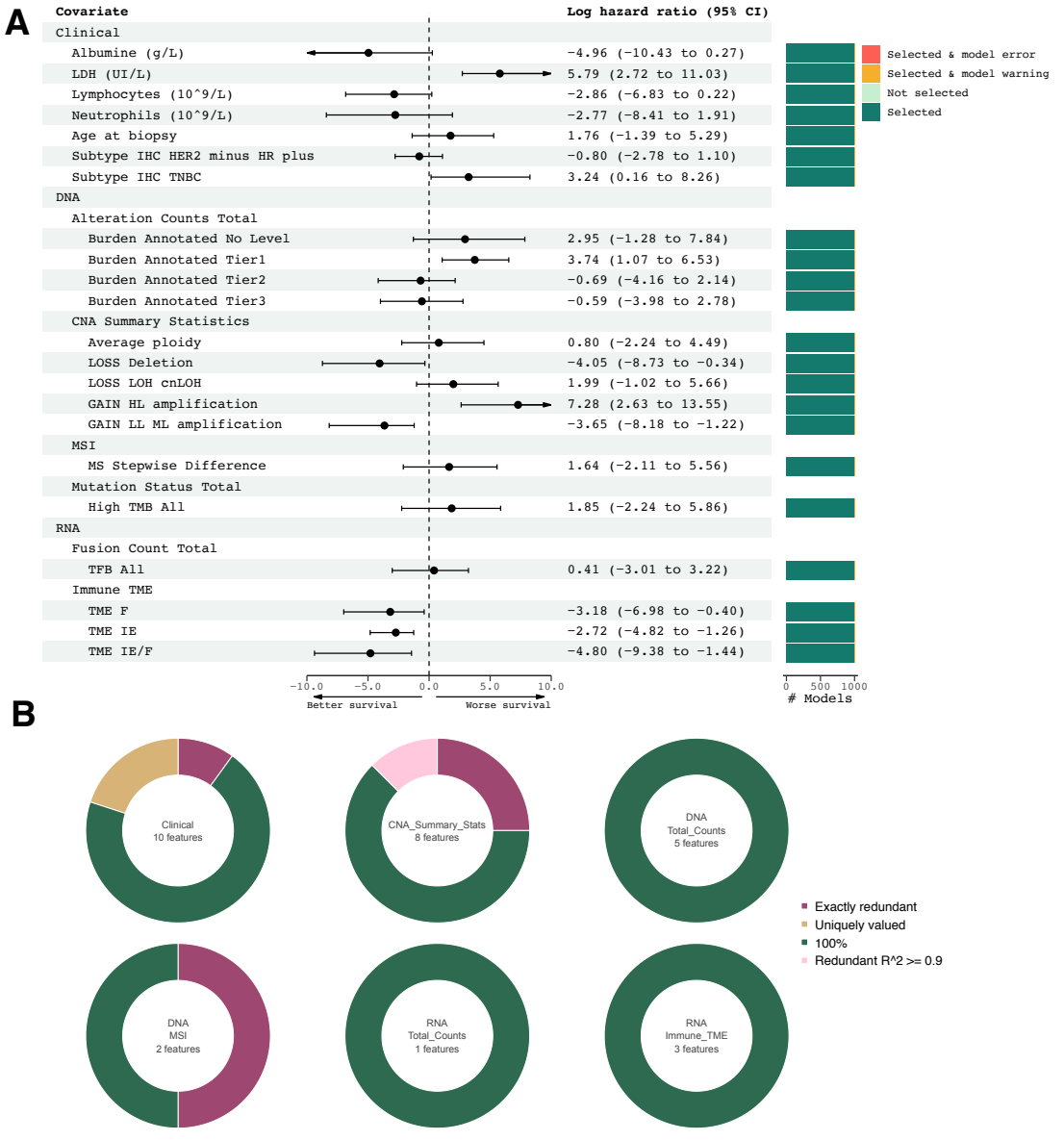


Fig. A.13.: Cox models M7 coefficients in META-PRISM BRCA tumors. **A**. Coefficients and 95% confidence intervals estimates from 1,000 Cox models using covariates from M7 model on BRCA tumors from META-PRISM WES &RNA-seq subcohort. **B**. Number of features from each category that were removed or selected during the modeling preprocessing and fitting steps. Green colors indicate the selection frequency of selected features, while non-green colors indicate reasons for removal.

A.4. Annexes to chapter 4

The following figure summarizes the genomic alterations that have positive therapeutic implications specific as annotated in OncoKB and CIViC knowledge databases.

Fig. A.14.: Known genetic markers of treatment sensitivity in META-PRISM, MET500, and TCGA by tumor type. **Top**, fractions of tumors harboring sensitivity markers split by tier (only the best tier is shown for each tumor). **Middle**, fractions of tumors with multiple sensitivity markers. **Bottom**, heat map showing the most frequent sensitivity-associated variants. Triangle orientations (increase - triangle points up, decrease - points down) and colors (red for META-PRISM vs. TCGA, green for MET500 vs. TCGA) highlight significant changes in prevalence. Similarly, stars next to the gene alterations represent significant changes at the cohort level using the same color code as for triangles (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$). P-values per tumor type are from Fisher-Boschloo tests, and p-values across the cohort are from Cochran-Mantel-Haenszel tests. All p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.

Bibliography

1. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. en. *Cell Reports* **3,** 246–259. doi:10.1016/j.celrep.2012.12.008 (Jan. 2013).
2. Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. & Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. en. *Computational Statistics & Data Analysis* **52,** 155–173. doi:10.1016/j.csda.2006.11.006 (Sept. 2007).
3. Boutsidis, C. & Gallopoulos, E. SVD based initialization: A head start for nonnegative matrix factorization. en. *Pattern Recognition* **41,** 1350–1362. doi:10.1016/j.patcog.2007.09.010 (Apr. 2008).
4. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. en. *Proceedings of the National Academy of Sciences* **101,** 4164–4169. doi:10.1073/pnas.0308531101 (Mar. 2004).
5. Chih-Jen Lin. On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization. *IEEE Transactions on Neural Networks* **18,** 1589–1596. doi:10.1109/TNN.2007.895831 (Nov. 2007).
6. Cichocki, A. & Zdunek, R. Multilayer nonnegative matrix factorisation. en. *Electronics Letters* **42,** 947. doi:10.1049/el:20060983 (2006).
7. Cichocki, A., Amari, S.-i., *et al.* en. in *Artificial Intelligence and Soft Computing ICAISC 2006* (eds Hutchison, D. *et al.*) Series Title: Lecture Notes in Computer Science, 548–562 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006). doi:10.1007/11785231_58.
8. Cichocki, A., Zdunek, R. & Amari, S.-i. in *Independent Component Analysis and Blind Signal Separation* (eds Hutchison, D. *et al.*) Series Title: Lecture Notes in Computer Science, 32–39 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006). doi:10.1007/11679363_5.
9. Cichocki, A., Zdunek, R. & Amari, S.-i. en. in *Independent Component Analysis and Signal Separation* (eds Davies, M. E., James, C. J., Abdallah, S. A. & Plumbley, M. D.) Series Title: Lecture Notes in Computer Science, 169–176 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007). doi:10.1007/978-3-540-74494-8_22.
10. Cichocki, A., Zdunek, R., Choi, S., *et al.* NMFLAB for Signal Processing (June 2003).
11. Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S.-I. *Nonnegative Matrix and Tensor Factorizations* en. doi:10.1002/9780470747278 (John Wiley & Sons, Ltd, Chichester, UK, Sept. 2009).
12. Févotte, C., Bertin, N. & Durrieu, J.-L. Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis. en. *Neural Computation* **21,** 793–830. doi:10.1162/neco.2008.04-08-771 (Mar. 2009).
13. Févotte, C. & Idier, J. Algorithms for nonnegative matrix factorization with the beta-divergence. *arXiv:1010.1763 [cs]*. arXiv: 1010.1763 (Mar. 2011).
14. Fischer, A., Illingworth, C. J., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. en. *Genome Biology* **14,** R39. doi:10.1186/gb-2013-14-4-r39 (2013).

15. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. en. *Bioinformatics* **31,** 3673–3675. doi:10.1093/bioinformatics/btv408 (Nov. 2015).
16. Gillis, N. & Glineur, F. Accelerated Multiplicative Updates and Hierarchical ALS Algorithms for Nonnegative Matrix Factorization. *Neural Computation* **24.** arXiv: 1107.5194, 1085–1105. doi:10.1162/NECO_a_00256 (Apr. 2012).
17. Gong, P. & Zhang, C. Efficient Nonnegative Matrix Factorization via projected Newton method. en. *Pattern Recognition* **45,** 3557–3565. doi:10.1016/j.patcog.2012.02.037 (Sept. 2012).
18. Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *arXiv:cs/0408058.* arXiv: cs/0408058 (Aug. 2004).
19. Huang, Y., Liu, H. & Zhou, S. Quadratic regularization projected BarzilaiBorwein method for nonnegative matrix factorization. en. *Data Mining and Knowledge Discovery* **29,** 1665–1684. doi:10.1007/s10618-014-0390-x (Nov. 2015).
20. Kim, H. & Park, H. Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method. en. *SIAM Journal on Matrix Analysis and Applications* **30,** 713–730. doi:10.1137/07069239X (Jan. 2008).
21. Kim, P. M. & Tidor, B. Subsystem Identification Through Dimensionality Reduction of Large-Scale Gene Expression Data. en. *Genome Research* **13,** 1706–1718. doi:10.1101/gr.903503 (July 2003).
22. Kompass, R. A Generalized Divergence Measure for Nonnegative Matrix Factorization. en. *Neural Computation* **19,** 780–791. doi:10.1162/neco.2007.19.3.780 (Mar. 2007).
23. Laurberg, H., Christensen, M. G., Plumbley, M. D., Hansen, L. K. & Jensen, S. H. Theorems on Positive Data: On the Uniqueness of NMF. en. *Computational Intelligence and Neuroscience* **2008,** 1–9. doi:10.1155/2008/764206 (2008).
24. Lee, D. D. & Seung, H. S. Algorithms for Non-negative Matrix Factorization. en, 7 (2001).
25. Lu, Z., Yang, Z. & Oja, E. in *Artificial Neural Networks and Machine Learning – ICANN 2012* (eds Hutchison, D. *et al.*) Series Title: Lecture Notes in Computer Science, 419–426 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012). doi:10.1007/978-3-642-33266-1_52.
26. Paatero, P. & Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. en. *Environmetrics* **5,** 111–126. doi:10.1002/env.3170050203 (June 1994).
27. Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. & Da Silva, I. T. signeR: an empirical Bayesian approach to mutational signature discovery. en. *Bioinformatics* **33** (ed Valencia, A.) 8–16. doi:10.1093/bioinformatics/btw572 (Jan. 2017).
28. Sra, S. & Dhillon, I. S. *Generalized Nonnegative Matrix Approximations with Bregman Divergences* en. in *Advances in Neural Information Processing Systems 18 - Proceedings of the 2005 Conference, NIPS 2005* (MIT Press, Vancouver British Columbia Canada, Dec. 2005), 283–290.

29. Wild, S. M. *Seeding Non-Negative Matrix Factorizations with the Spherical K-Means Clustering* MA thesis (University of Colorado, 2003).

Glossary

Symbols

1000G 1000 Genomes Project.

A

AACR American Association for Cancer Research.
ACC adrenocortical carcinoma.
ADC antibody-drug conjugate.
Adenocarcinoma a malignant tumor arising from glandular epithelial cells.
ADICAP Association pour le Développment de l'Informatique en Cytologie et Anatomie Pathologique.
ALL acute lymphocytic leukemia.
ALS alternating least squares.
Amino acid organic molecule composed of a basic amino group, an acidic carboxyl group, and a unique side chain. It is the building block of proteins.
AML acute myeloid leukemia.
API application programming interface.
ASCO American Society of Clinical Oncology.

B

BAM binary alignment map.
BCL base call format.
BER base-excision repair.
BLCA bladder urothelial carcinoma.
BRCA breast invasive carcinoma.
Breakpoint the chromosomal position at which a DNA break has occured.

C

CADD comined annotation dependent depletion.
CAF cancer-associated fibroblast.
Carcinoma a malignant tumor arising from epithelial cells.
cDNA copy DNA.
CGC cancer gene census.
CGI CpG island.

Chemotherapy a type of treatment that uses potent chemicals to destroy quickly dividing cells.
CHOL cholangiocarcinoma.
Chromosome a threadlike structure of a long polynucleotidic chain and proteins found in the nucleus of cells.
CI confidence interval.
CIN chromosomal instability.
CIViC Clinical Interpretation of Variants in Cancer.
CLL chronic lymphocytic leukemia.
CML chronic myeloid leukemia.
CNA copy-number alteration.
CNS central nervous system.
CNV copy-number variation.
COAD colon adenocarcinoma.
Codon a sequence of three nucleotides that encode for one of the twenty-two known amino acids or for the stop signal of protein translation.
COSMIC Catalogue of Somatic Mutations In Cancer.
ctDNA circulating tumor DNA.
CUP cancer of unknown primary.

D

dbGaP database of Genotypes and Phenotypes.
DBS (doublet-base substitution) mutation of two consecutive nucleotides on the DNA chain.
dbSNP database of Single Nucleotide Polymorphisms.
DNA deoxyribonucleic acid.
Driver a characteristic of a biological element or process that provides a selective growth advantage and thus promotes cancer development.
  Driver gene a gene for which one or multiple driver mutations have been described.
  Driver gene fusion a fusion gene that provides a selective growth advantage and thus promotes cancer development.
  Driver mutation a mutation that provides a selective growth advantage and thus promotes cancer development.
DSC Dice-Sorensen coefficient.

E

EACR European Association for Cancer Research.
ECM extracellular matrix.
EGA European Genome Archive.
EGFR epidermal growth factor receptor.
ELN European Leukemia Net.
EMA European Medical Agency.
EMT epithelial-mesenchymal transition.
Epigenome all the epigenetic modifications of a cell.

ESCAT ESMO Scale for Clinical Actionability of molecular Targets.
ESHG European Society of Human Genetics.
ESMO European Society of Medical Oncology.
ESP Exome Sequencing Project.
Euchromatic located in the euchromatin, the loosely packaged and genetically active region of the chromatin.
ExAC Exome Aggregation Consortium.
Exome the part of the genome consisting of exons.
Exon subsegment of a gene coding for a sequence of amino acids.

F

FASTQ text-based format for storing both a biological sequence and its corresponding quality scores using single ASCII characters.
FDA Food and Drug Administration.
FFPE formalin-fixed, paraffin-embedded.
FGFR fibroblast growth factor receptor.
FPKM (fragment per kilobase per million mapped reads) fragment counts in paired-end sequencing normalized by library size followed by length normalization in kilobase and multiplication by one million.

G

GATK Genome Analysis Tookit.
Gb gigabase.
GCE Google Cloud Engine.
GDC Genomic Data Commons.
Gene segment of DNA encoding for a specific function, majoratirily proteins, and which constitutes the unit of inheritance.
  Gene expression the set of processes through which genes are transcribed into functional gene products, either proteins or functional RNAs.
Gene fusion a hybrid gene resulting from the juxtaposition of subparts of two independent genes. It can occur as the result of a translocation, interstitial deletion, or inversion.
GENIE Genomics Evidence Neoplasia Information Exchange.
Genome all the genetic information of a cell organised in chromosomes.
Germline germline refers to sex cells, also known as germ cells, that pass on their genomes (half of it) from one generation to another in multicellular organisms with dedicated reproductive cells. Germline alterations refer to changes in cells inherited from the two parental germ cells.
GitHub a plaform and cloud-based service used to host code in private or public repositories and maintain version control using Git.
Glioma a malignant tumour of the glial tissue of the nervous system.
GLOBOCAN Global Cancer Observatory.
gnomAD Genome Aggregation Database.
GRC Genome Reference Consortium.
GRIM score Gustave Roussy Immune score.

GTEx Genotype-Tissue Expression.

H

Heterochromatic located in the heterochromatin, the densiley packaged and genetically inactive region of the chromatin.
HGP Human Genome Project.
HMF Hartwig Medical Foundation.
HNAC head and neck adenoid cystic carcinoma.
HNSC head and neck squamous cell carcinoma.
Hormone therapy a type of drug which inhibits hormones to slow or stop the growth of hormone-dependent cells.
HR homologous recombination.
HRD homologous recombination deficiency.

I

IARC International Agency for Research on Cancer.
ICD-10 International Classification of Diseases, tenth revision.
ICD-O-3 International Classification of Diseases for Oncology, third edition.
ICGC International Cancer Genome Consortium.
IHC immunohistochemistry.
Immunotherapy a type of treatment aiming at stimulating the immune response.
INCA French National Cancer Institute.
Indel insertion or deletion of nucleotides along DNA.
Intron subsegment of a gene not coding for amino acids.
ISH in situ hybridization.

K

kb kilobase.

L

LDH lactate dehydrogenase.
Leukemia type of cancer in which tumor cells circulate in the blood of the patient. It mostly arises from blood-forming cells located in the bone marrow.
LHR log hazard ratio.
LIHC liver hepatocellular carcinoma.
lncRNA (long non-coding RNA) class of RNA molecules of over 200 nucleotides that have no or limited coding capacity.
LOH (loss-of-heterozygosity) type of genomic abnormality causing a locus to lose the copy originating from one the two parents, rendering it homozygous.
LUAD lung adenocarcinoma.
LUSC lung squamous cell carcinoma.

Lymphoma type of cancer that develops in the glands or nodes of the lymphatic system.

M

MAF (minor allele frequency) the frequency at which the second most common allele occurs in a given population.
Mb megabase.
MDS myelodysplastic syndrome.
Metachronous existing or occuring at different times - in cancer with more than a six-month difference.
Metastasis the development of secondary malignant growths at a distance from a primary site of cancer.
miRNA micro-RNA.
MMR mismatch repair.
MMRd mismatch repair deficiency.
MNV (multi-nucleotide variant) mutation of a two or more consecutive nucleotides on the DNA chain rarely encountered in the general population, generally $< 1\%$. Term used in the context of somatic mutations.
Monoclonal antibody an antibody produced from a cell lineage made by cloning a unique white blood cell.
mRNA messenger RNA.
MSI microsatellite instability.
MSK Memorial Sloan Kettering.
MSK-IMPACT MSK-Integrated Mutation Profiling of Actionable Cancer Targets.
Myeloma type of cancr that arises from the plasma cells of bone marrow.

N

NCBI National Center for Biotechnology.
NCI National Cancer Institute.
NEC neuroendocrine carcinoma.
NER nucleotide-excision repair.
NET neuroendocrine tumor.
NGS next-generation sequencing.
NHEJ non-homologous end joining.
NHGRI National Human Genome Research Institute.
NHLBI National Heart, Lung, and Blood Institute.
NMF non-negative matrix factorisation.
NNLS non-negative least squares.
NSCLC non-small cell lung cancer.
Nucleic acid chemical compounds that serve as primary information-carrying molecules making up the genetic material of cells. DNA and RNA are the two primary types of nucleic acids found in living cells.

Nucleotide basic structural unit of nucleic acids composed of a nucleoside and a phosphate group.

O

Oncogene type of gene that contributes to tumor phenotype when expressed.
OncoKB MSK's precision Oncology Knowledge Base.

P

PAAD pancreatic adenocarcinoma.
Pan-cancer characterization of a study or analysis examining diverse cancer types.
PCAWG PanCancer Analysis of Whole Genomes.
PCPG pheochromocytoma and paraganglioma.
PCR polymerase chain reaction.
PI3K group of plasma membrane-associated lipid kinases existing under eight different isoforms classified in three classes. Class IA PI3K contains three isoforms 110α, 110β - 110δ encoded by *PIK3CA*, *PIK3CB*, *PIK3CD*, respectively - clearly implicated in cancer.
PRAD prostate adenocarcinoma.
Proteome all the proteins expressed in a cell.
PTC papillary thyroid carcinoma.

R

RCT randomized controlled trial.
Read a sequence of base calls produced by a machine and corresponding to all or part of a DNA fragment.
RECIST Response Evaluation Criteria in Solid Tumors.
RNA ribonucleic acid.
RNA-seq RNA sequencing.
RPKM (reads per kilobase per million mapped reads) read counts normalized by library size followed by length normalization in kilobase and multiplication by one million.
rRNA ribosomal RNA.

S

SAM sequencing alignment map.
SARC sarcoma.
Sarcoma a malignant tumor arising from supportive and connective tissue such as bone, adipose, cartilage, or muscle tissues.
SBS (single-base substitution) mutation of a single nucleotide on the DNA chain. Synonym for SNV or SNP.
SIFT sorting intolerant from tolerant.
Signature a computationally-derived molecular pattern that predicts a phenotype of interest.
  Gene expression signature a computationally-derived pattern of gene expressions predicting a particular phenotype.

Mutational signature a computationally-derived pattern of mutations characterizing the consequence of an endogenous or exogenous mutagenic agent.
SKCM skin cutaneous melanoma.
SNP (single-nucleotide polymorphism) mutation of a single nucleotide on the DNA chain commonly encountered in the general population, generally $> 1\%$. Term used in the context of germline mutations.
SNV (single-nucleotide variant) mutation of a single nucleotide on the DNA chain rarely encountered in the general population, generally $< 1\%$. Term used in the context of somatic mutations.
Somatic somatic refers to all the cells other than sperm and egg cells in multicellular organisms with dedicated reproductive cells. Somatic alterations designate alterations occurring after fertilization in somatic cells.
Squamous cell carcinoma a malignant tumor arising from squamous epithelial cells.
Substitution a type of mutation affecting one or multiple consecutive nucleotides on the DNA chain.
SV (structural variant) juxtaposition of non-contiguous chromosomal segments through a process of genomic rearrangement.
Synchronous existing or occuring at the same time - in cancer with less than a six-month difference.

T

TARGET Therapeutically Applicable Research to Generate Effective Treatments.
Targeted therapy a type of treatment targeting specific proteins which control how cells survive, grow, divide, or spread.
TCGA The Cancer Genome Atlas.
T-DM1 trastuzumab emtansine.
T-DXd trastuzumab deruxtecan.
THCA thyroid carcinoma.
TKI tyrosine kinase inhibitor.
TMB tumor mutational burden.
TME tumor microenvironment.
TMM trimmed mean of m-values.
TPM (transcripts per million maped reads) read counts normalized by gene lengths and then normalized to sum to one million.
Transcript the single-stranded RNA molecule that is produced when a gene is transcribed.
Transcription process of making an RNA copy of a segment of DNA.
Transcription factor a protein that controls the rate of transcription.
Transcriptome the set of all RNA transcripts, including coding and non-coding, in an individual or a population.
Transition single-base substitution of a purine (adenine, guanine) by a another purine (A>G of G>A) or a pyrimidine (cytosine, thymine) by another pyriminde (C>T or T>C).
Transversion single-base substitution of a purine (adenine, guanine) by a pyrimidine (cytosine, thymine) or vice-versa.

tRNA transfer RNA.
Tumor suppressor gene type of gene that negatively regulates cell proliferation and act to inhibit tumor development.
Tumorigenesis the set of all biological processes involved in the formation of a tumor.

U

UCSC University of California, Santa Cruz.

V

VAF (variant-allele frequency) number of reads supporting a mutation divided by the read depth at the genomic position.
VCF variant calling format.
VEGF vascular endothelial growth factor.
VEP variant effect predictor.
VM virtual machine.

W

WES whole-exome sequencing.
WGD whole-genome duplication.
WGS whole-genome sequencing.
WHO World Health Organisation.
WTSI Wellcome Trust Sanger Institute.

Index

E

F

G

H

I

K

L

M

N

O

P

R

S

T

U

Synthèse

À l'ère de l'acquisition et de l'analyse massives de données, notre compréhension de l'apparition et de l'évolution du cancer s'est améliorée à la lumière des résultats dérivés des analyses des portraits moléculaires de dizaines de milliers de tumeurs à travers le monde. L'avènement des technologies de séquençage de nouvelle génération dans les années 2000 a révolutionné la façon dont nous étudions les cellules tumorales des patients atteints de cancer. Ces technologies ont d'abord été utilisées pour caractériser des régions génomiques spécifiques, mais ont mûri au fil du temps pour permettre le profilage systématique de l'ensemble de l'exome, du transcriptome et même du génome entier. Étant donnée la place grandissante du séquençage dans la recherche et la pratique clinique, une compréhension complète des différents aspects de l'analyse des données de séquençage est primordiale. Bien que le séquençage à haut débit ne fasse pas encore partie du parcours clinique de tous les patients atteints de cancer, des profilages moléculaires ont été proposés à nombre de patients participant à des essais cliniques et est aujourd'hui utilisé en routine pour un certaines indications. Cette grande quantité de données est désormais disponible pour étayer de nombreuses recherches, allant de l'établissement de portraits moléculaires détaillés de groupes particuliers de patients au déchiffrement des liens entre les génotypes des tumeurs et les parcours cliniques des patients. Toutes ces recherches participent aux progrès de l'oncologie de précision.

Cette thèse couvre de nombreux aspects impliqués dans l'analyse rétrospective d'une large cohorte de patients atteints de cancer, ainsi que des revues détaillées de concepts et d'outils importants de l'oncologie moderne. Le premier chapitre introduit les concepts généraux sur la biologie et la classification du cancer, qui sont fondamentaux pour les décisions thérapeutiques mais également l'orientation et l'organisation de la recherche. La dernière section de ce premier chapitre présente en outre au lecteur des réflexions sur la place croissante du profilage moléculaire et leur impact sur les classifications et les conceptions d'essais. Le deuxième chapitre passe en revue en détail les outils informatiques et les bases de données utilisés pour analyser les données de séquençage et extraire des informations cliniquement pertinentes.

Les deux premiers chapitres servent de base à l'analyse d'une vaste cohorte de patients pan-cancer présentée dans le troisième chapitre. Cette cohorte, META-PRISM, comprend 1031 patients issus de deux de grands essais de médecine de précision menés à Gustave Roussy dans la décennie 2010-2020, dont un tiers ont bénéficié des technologies de séquençage de l'exome entier ou de l'ARN de leurs tumeurs. Comparée à d'autres analyses de grandes cohortes pan-cancer, cette étude se distingue par l'accent mis sur les patients réfractaires aux traitements, tous étant considérés comme affectés par un cancer incurable selon un comité multidisciplinaire, et par la disponibilité d'historiques cliniques détaillés des patients. Les

analyses comparatives avec une cohorte internationale de tumeurs primaires non traitées ont mis au jour des différences génétiques générales ainsi que de multiples différences spécifiques à certains types de tumeur. De plus, la modélisation prédictive de la survie des patients à l'aide de biomarqueurs moléculaires a montré que même les patients à un stade avancé peuvent bénéficier du séquençage pour des décisions thérapeutiques importantes, en particulier l'éligibilité aux essais de phases 1 ou 2.

Le quatrième et dernier chapitre se concentre sur l'analyse des marqueurs génomiques connus et émergents de résistance aux traitements dans la cohorte META-PRISM, mais également dans deux autres cohortes issues d'études cliniques récentes dirigées par Gustave Roussy, dont l'une s'intéresse à un conjugué anticorps-médicament récemment approuvé pour les cancers du sein, et l'autre à une classe particulière d'inhibiteurs pour les cancers de la vessie. Ces deux études ont démontré que les altérations de l'expression ou de la structure de la cible, ou l'activation de voies alternatives par des mutations, contribuent à la résistance aux médicaments.

université PARIS-SACLAY | ÉCOLE DOCTORALE Cancérologie : biologie - médecine - santé (CBMS)

Titre : Analyses des profils génomiques et transcriptomiques de tumeurs métastatiques issus d'essais cliniques de médecine de précision

Mots clés : Pancancer, Métastase tumorale, Sequençage de l'exome, Séquençage de l'ARN, Survie, Résistance aux traitements

Résumé : À l'ère de l'analyse des données, les connaissances sur l'apparition et la progression du cancer se sont approfondies grâce à l'analyse moléculaire de nombreuses tumeurs dans le monde. Le séquençage de nouvelle génération, apparu dans les années 2000, a transformé la recherche sur les cellules cancéreuses en permettant le profilage complet de l'exome, du transcriptome et même du génome entier. Bien que le séquençage à haut débit ne soit pas systématique dans la pratique clinique, il est couramment utilisé dans les essais thérapeutiques. Le vaste réservoir de données ainsi généré alimente de nombreuses recherches qui contribuent aux progrès de l'oncologie de précision.

Cette thèse explore l'analyse de cohortes de patients atteints de cancer et les outils modernes d'oncologie. Le premier chapitre couvre les principes essentiels de la biologie du cancer, en mettant l'accent sur le rôle évolutif du profilage moléculaire dans le traitement et la recherche. Le deuxième chapitre passe en revue les outils informatiques et les bases de données employés pour l'analyse des données de séquençage. Ces chapitres donnent les clés pour le troisième chapitre, axé sur la cohorte META-PRISM, comprenant 1 031 patients issus d'essais de médecine de précision à Gustave Roussy. Il met en évidence les spécificités génétiques des patients réfractaires et les possibilités de modélisation prédictive sur les données du séquençage haut débit. Le quatrième chapitre examine les marqueurs de résistance aux traitements connus et émergents dans la cohorte META-PRISM et dans deux études cliniques récentes, révélant des altérations de cibles et des activations de voies alternatives comme facteurs de résistance clés.

Title : Analyses of genomic and transcriptomic profiles of metastatic tumors from precision medicine clinical trials

Keywords : Pancancer, Tumor metastasis, Exome sequencing, RNA sequencing, Survival, Treatment resistance

Abstract : In the era of extensive data analysis, insights into cancer onset and progression have deepened through molecular analysis of numerous tumors globally. Next-generation sequencing, emerging in the 2000s, transformed cancer cell investigation by enabling exome, transcriptome, and now whole genome profiling. While high-throughput sequencing has not yet entered clinical pratice for all, it is commonly used in trials. The vast data pool thus generated fuels many research areas which contribute to precision oncology advancements.

This thesis explores cancer patient cohort analysis and modern oncology tools. The first chapter covers cancer biology fundamentals, emphasizing molecular profiling's evolving role in treatment and research. The second chapter reviews computing tools and databases for sequencing data analysis. These chapters set the stage for the third chapter, focusing on the META-PRISM cohort, comprising 1,031 patients from precision medicine trials at Gustave Roussy. It highlights the molecular specificities of refractory and the promises of predictive modeling based on high-throughput sequencing data. The fourth chapter delves into known and emerging treatment resistance markers in the META-PRISM cohort and two recent clinical studies, revealing target alterations and alternative pathway activations as key resistance factors.