Pharmacogenomic and High-Throughput Data Analysis to Overcome Triple Negative Breast Cancers Drug Resistance

Benjamin Sadacca

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Pharmacogenomic and high-throughput data analysis to overcome triple negative breast cancers drug resistance

Thèse de doctorat de l'Université Paris-Saclay
Préparée à L'Institut Curie - laboratoire RT2 & Université d'Évry - équipe Stat & Génome.

École doctorale n°582 Cancérologie : biologie - médecine - santé
Spécialité de doctorat : Sciences de la vie et de la santé

Thèse présentée et soutenue à Paris, le 15 Décembre 2017, par

Benjamin SADACCA

Composition du Jury :

Jean-Yves Pierga PUPH, Institut Curie Directeur du département d'oncologie médicale Laboratoire des Biomarqueurs Tumoraux Circulants	Président
Aurélien de Reyniès Directeur Scientifique, La ligue contre le cancer Programme Cartes d'Identité des Tumeurs	Rapporteur
Lodewyk Wessels Group leader and Head of Division, Netherlands Cancer Institute Computational Cancer Biology Division of Molecular Carcinogenesis	Rapporteur
Emmanuel Barillot Directeur d'unité, Institut Curie (U900) Département de Bioinformatique et des Systèmes Computationnels de Biologie du Cancer	Examinateur
Alexandra Leary Praticien Hospitalier, Gustave Roussy Cancer Campus (U981) Équipe Biomarqueurs et nouvelles stratégies thérapeutiques	Examinateur
Fabien Reyal Praticien Hospitalier, Institut Curie Chef de l'équipe résidus tumoral & réponse au traitement (RT2Lab) Chef de l'Unité du Sein, du Cancer gynécologique et de la chirurgie reconstructive	Directeur de thèse
Pierre Neuvial Chargée de recherche, CNRS - Institut de Mathématiques de Toulouse	Co-Directeur de thèse

Contents

Synthese

Les cancers du sein triple négatif (TNBC) se réfèrent à des cancer du sein qui n'expriment pas les gènes du récepteur des œstrogènes (ER), du récepteur de la progestérone (PR) et Her2. Ce sous-groupe présente des caractéristiques pathologiques agressives et un taux élevé d'événements métastatiques précoces (avant cinq ans depuis le diagnostic initial). Ils représentent entre 10 et 20% des carcinomes canalaires invasif et sont principalement liés à une perturbation du mécanisme de réparation de l'ADN. Ces TNBCreprésentent un défi clinique majeur, car les traitements n'ont pas été améliorés depuis des dizaines années. En effet, à ce jour, aucune thérapie ciblée n'a été approuvée, et la chimiothérapie cytotoxique reste le traitement standard. Devant le grand nombre de tumeurs du sein triple négatif résistant aux chimiothérapies, il est essentiel de comprendre les mécanismes de résistance et de trouver de nouvelles molécules efficaces.

Le taux d'attrition très élevé est un problème majeur dans le développement de nouveaux médicaments anticancéreux . Le processus de repositionnement des médicaments propose de trouver de nouvelles indications thérapeutiques à des médicaments déjà connus. Pour cela, de nouvelles méthodes analytiques sont nécessaires pour optimiser l'information présente dans les ensembles de données pharmacogénomiques à grande échelle. Nous avons analysé les données de deux ensembles de données pharmacogénomiques à grande échelle : le Genomics of Drug Sensitivity in Cancer et le Cancer Cell Line Encyclopedia. En nous centrant sur les lignées cellulaires (n = 471) et les molécules (n =15) testées en commun entre les deux ensembles, nous proposons une nouvelle classification basée sur les profils transcriptomiques des lignées cellulaires, selon un processus de sélection de gènes basé sur des réseaux biologiques. Notre classification moléculaire montre une plus grande homogénéité de la sensibilité aux médicaments que lorsque les lignées cellulaires sont groupées en fonction de leur tissu d'origine. Nous avons ensuite identifié des associations statistiquement significatives entre les groupes de lignées cellulaires et la réponse aux médicaments. Ces associations sont retrouvées de manière robuste entre les deux ensembles de données. Nous démontrons la pertinence de notre méthode en analysant deux ensembles de données supplémentaires, utilisant des mesures de sensibilité distinctes. Nous montrons que notre clustering de lignées cellulaires est capable de trouver des associations significatives avec l'efficacité des médicaments dans quatre ensembles de données différents, malgré les grandes variations entre les données pharmacologiques et les mesures de réponse aux médicaments. Cette étude définit une

classification moléculaire robuste des lignées cellulaires cancéreuses qui pourraient être utilisées pour trouver de nouvelles indications thérapeutiques à des composés connus.

Dans un second travail, nous étudions une cohorte de 16 patients atteints d'un cancer du sein triple négatif ayant résisté à la chimiothérapie néoadjuvante (anthracyclines-taxanes). Cette chimiothérapie néoadjuvante (c'est-à-dire avant la chirurgie) représente une formidable opportunité pour étudier et surveiller in vivo la sensibilité au traitement des tumeurs. L'observation d'une réponse pathologique complète (pCR) après la chimiothérapie néoadjuvante (NAC) est un indicateur fort du bénéfice de la chimiothérapie sur le pronostic des patients. Malgré leur chimiosensibilité relative, on observe une pCR dans seulement 30% des patients TNBC traités par NAC en routine. L'observation d'une maladie résiduelle suivant une chimiothérapie néoadjuvante est au contraire un indicateur de mauvais pronostic. L'objectif de cette étude est d'identifier des biomarqueurs associées à la résistance aux traitements néoadjuvant à partir de données RNAseq et Whole Exome Sequencing obtenus sur des échantillons avant et apres traitement. Une classification non supervisée de nos échantillons révèle quatre groupes de patients enrichis pour les gènes exprimés dans les modules suivants: "matrice extracellulaire", "lymphocytes T", "métabolisme et régulation des niveaux hormonaux" et "processus du système nerveux". Cette classification ne permet pas de séparer les différents échantillons avant après traitement de chaque patient. Nous constatons une forte variabilité inter-tumorale, comparé à la variabilité intra-tumorale, indiquant que le traitement néoadjuvant a un effet relativement faible par rapport à la variabilité entre les patients. Ces résultats sont corroborés par le petit nombre de gènes différentiellement exprimés entre tumeurs primaires et résidu tumoral. Ces gènes sont enrichis dans des processus induisant une réponse aux médicaments et aux composés organophosphorés ou contenant de la purine, en accord avec l'exposition des échantillons à l'épirubicine et au cyclophosphamide. Bien que nous observons une évolution clonale sous traitement, aucun mécanisme récurrent de résistance n'a pu être identifié. Nos résultats suggèrent que chaque tumeur possède un profil moléculaire unique et qu'il est important d'étudier de grandes séries de tumeurs afin de mettre en évidence des profils moléculaire de résistance.

Enfin, nous proposons d'améliorer un outil appelé rMAPS qui est conçu pour identifier les sites de liaison des protéines de liaison à l'ARN autour des exons. Nous proposons d'utiliser une approche innovante pour contrôler la proportion de faux positifs qui n'est pas réalisé par l'algorithme existant. De plus ,nous montrons l'efficacité de notre approche en utilisant deux séries de données différentes.

General overview

We detail below the structure of the thesis.

- **Chapter 1** is an introduction to the relevant concepts that will be used and developed in the following work.
- **Chapter 2** presents a novel analysis of two large scale pharmacogenomic datasets, the Cancer Cell Line Encyclopedia (CCLE) and the Genomics of Drug Sensitivity in Cancer (GDSC). This work provides a new analytical method to study cell lines drug sensitivity. This study was accepted for publication in the journal Scientific Report.
- **Chapter 3** presents our work on the molecular mechanisms underlying treatment-resistant triple-negative breast cancer. We performed complete DNA and RNA analyzes on the core biopsy and residual disease of 16 clinically-defined, triple-negative breast cancers resistant to neoadjuvants, including 6 matched lymph nodes. Even in a series of homogeneous clinical samples, we observed a large molecular heterogeneity where few recurrent alterations or biological pathways can be identified. Further analysis remains to be done before publishing this work.
- **Chapter 4** constitutes a methodological contribution to motif enrichment analysis. In this project, we propose to use an innovative technique to control the proportion of false discoveries. We demonstrate the relevance of our approach through two different datasets. One of them correspond to the data introduced in chapter 3, the other is an external collaboration on Ewing's sarcoma. This contribution will be released and made available online as an R package.
- **Chapter 5** describes our collaborations within RT2 lab.
- **Chapter 6** describes our collaborations outside of RT2 lab.
- **Chapter 7** discusses our results and gives perspectives to continue and expand our work.
- For ease of reading, we decided to put important supplementary information at the end of each chapter. The information we consider less important for the understanding has been placed in the **appendix** at the end of the manuscript. The page number of each additional piece of information has been referenced each time.

1

Introduction

In this work, various clinical and biological aspects of cancer and computational biology are described. This introduction is not intended to be exhaustive. It aims to present the notions and concepts that we will develop in the manuscript. This thesis occur at the interface of biology, clinic and data science. We hope that it will allow each one to understand the stakes of each discipline and to exchange with the same vocabulary.
We will first introduce the topic of cancer, followed by the therapeutic strategies and pre-clinical models available today. We detail the characteristics of breast cancer and will explain why the triple-negative subtype is a major issue. Next, we will describe current high-throughput technologies that accurately characterize the molecular profiles of each tumor sample and allow to screen a large amount of drugs. We will finally discuss the statistical and bioinformatics tools that have been developed around the issues addressed.

1.1 Biology of cancer

1.1.1 Epidemiology

Worldwide in 2012, the number of new cancer cases was estimated at 14.1 million. 8.2 million of people died by cancer the same year (5). It is the second leading cause of death in developed countries after cardiovascular diseases. Prostate cancer for males and breast cancer for females are the most abundant cancers followed by lung and colorectal cancers.

Due to the growth and aging of the population and the adoption of lifestyles that are known to increase cancer risk (smoking, diet, physical inactivity), the number of new cancer cases is expected to 21.7 million and 13 million cancer deaths in 2030. It is a real public health issue and better ways of diagnosis and treatment are needed.

1.1.2 What is cancer ?

Cancer can be defined as a group of diseases in which a group of abnormal cells develop uncontrollably by ignoring the normal rules of cell division. Tumor cells differ from normal cells in many ways that have been introduced by Hanahan and Weinberg (74, 75) (Figure 1.1).

Cancer cells ignore signals that stop the cellular division. While normal cells are kept under control by growth factors, cancer cells continue to proliferate in an uncontrolled manner. During cellular division there is in average 100 errors that are introduced while copying the DNA sequence. It can be single base exchanges (single nucleotide polymorphism, SNPs), small insertion and deletion of one or more nucleotides (indels) or changes in the number of copies of DNA segment (copy number alterations, CNAs). Normal cells with damaged DNA sequence are subject to processes of DNA repair or mechanism inducing cell suicide called apoptosis. These processes prevent the propagation of errors in the genetic code that can lead to malfunction. Malignant cells are able to bypass these mechanisms, which gives them immortality associated with genome instability.

Cells need energy to sustain division and proliferation. Cancer cells are able to reprogramming energy metabolism to support their continuous cell growth and proliferation. Furthermore, they can induce the development of new blood vessels to keep their oxygen and nutrient supply. This process is called angiogenesis. The uncontrolled formation of these vessels contributes to the construction of the tumor microenvironment, which is closely related and interact constantly with the tumor.

Immune cells at various densities infiltrate nearly all tumor microenvironment This inflammation can provide molecules to the microenvironment to sustain proliferation, survival and promote angiogenesis. Immune system normally removes damaged or abnormal cells from the body. Cancer cells are able to "hide" from them and avoid immune destruction.

Finally, cancer cells may spread through the blood or lymph system. They can form new tumors (called metastases) into other parts of the body where nutrient and space are not limited. Ultimately, most of tumors will spread and 90% of deaths by cancer are due to metastases (170).

1.1.3 Cancer is a gene disease

Cancer is a genetic disease. Genetic changes that cause cancer can be inherited from our parents arise after certain environmental exposure or being the results of error during the cell division. An erroneous gene sequence will have an impact only if the protein coded by the gene is produced. Genes are then transcribed in transcripts called pre-messenger RNAs (pre-mRNAs) in a process called transcription. Before being translated into proteins, several steps are required.

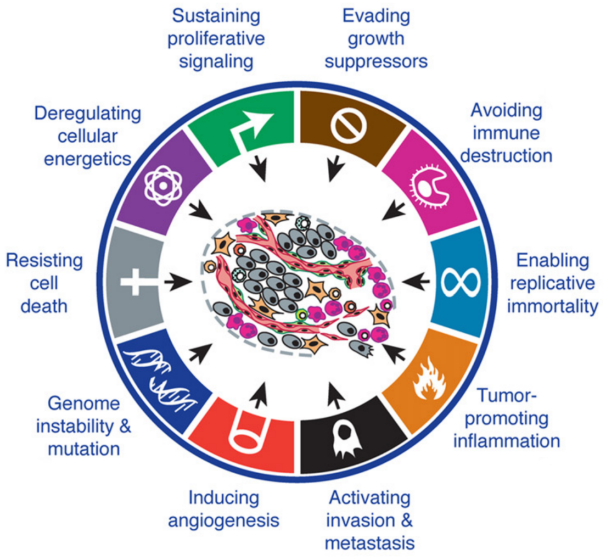


Figure 1.1: The hallmarks of Cancer defined by Hanahan and Weinberg (75).

pre-mRNAs need to be mature to be translated. RNA splicing removes introns (non coding sequences) and exons (coding sequences) are joined together. Several mature mRNAs (isoforms) leading to several proteins can be produced based on a single gene. This is achieved by selecting the right exons coding for the right protein, in a process called alternative splicing. Different kinds of alternative events can occur (176). The most common events in mammalian is exon skipping (or cassette exon) in which an exon is spliced out or retained of the primary transcript (Figure 1.2). Alternative 5' or 3' splice sites change the 3' (respectively 5') boundary of the upstream (respectively downstream) exon. In the mutually exclusive setting, transcripts are formed with one of two exons, but not both. Finally, when an intron is retained in the mature mRNA, it is called intron retention. The splicing pattern of specific isoforms of numerous genes is altered during the oncogenic process, which promotes the emergence of cancer hallmarks. Mature mRNAs are then translated into proteins, following the central dogma of molecular biology introduced by Crick (43). If a mutation occurs in a coding region of the gene, the protein produced may not be fully functional leading to abnormal cellular behavior.

Critical regulatory genes have been linked to the development or progression of cancer (44). Oncogene refers to a category of genes whose expression promotes the occurrence of cancers. Their alterations can introduce a novel gene function or make it insensitive to the regulatory signals. Typically, they encode for proteins, which control cell proliferation or apoptosis. Unlike oncogenes that become hyperactive in cancer cells, tumor suppressor genes lose their function. The most famous tumor suppressor is TP53 called *the guardian of the genome*. Its protein plays a key-role in the integrity of the genome during the cell cycle and managing DNA repair. Many mutations are present in a cancer

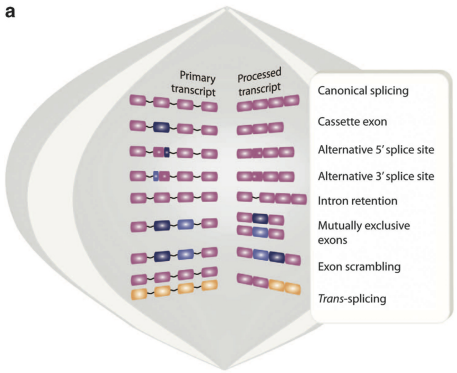


Figure 1.2: Modes of pre-mRNA splicing from Sveen et al. (176).

genome due to the deficiency in DNA repair. Not all mutations contribute equally to the tumor development. Thanks to the redundancy of the genetic code, it is most likely that most do not have any impact. Mutations that actively contribute to oncogenesis are called *drivers*, while *passengers* mutations denote changes in the sequence with no functional impact on the cell (173).

1.1.4 Intratumor heterogeneity

Many mutations are needed to transform a normal cell into a malignant cell. Indeed, if one mutation was enough to convert a healthy cell into a cancer cell, we would not be viable organisms. Cancer cells have to face limits at each step of tumor evolution. Cells with mutations giving a selective advantage at a given step are growing preferentially compared to other cells. Cancer cells compete for growth advantage subject to a positive or negative selection during the tumor lifetime. Cells in a tumor are then not homogeneous but rather organized in different clones (82, 118). These clones carry mutations that have emerged at the beginning of the tumorogenesis and are common to all cells in the tumor (ancestral mutations). They also carry specific alterations giving new traits with potential benefits.
The multiple clonal subpopulations of cancer cells are particularly studied to understand drug resistance. Indeed, the elimination of a dominant clone sensitive to the drug could allow the competitive emergence of a resistant subclone.

1.2 Treating cancer

Cancer treatments are generally classified into local therapies and systemic therapies. Local therapy is what is done locally on the tumor. Typical example is the removal of the tumor and surrounding tissue during surgery. Local treatment can also use high-energy radiation to shrink tumors and kill cancer cells. Systemic therapy affects the whole body. They are designed to kill cancer cells that have reached and affected other parts of the

body. They are organized as cytotoxic agents (chemotherapies) and targeted agents.

1.2.1 Cytotoxic agents

From a therapeutic point of view, cancer is a disease of cellular communication between the malignant cells and their environment. We have seen that cancer cells ignore external signals that stop the cellular division, giving them the ability to proliferate. Cell growth is intrinsically associated to the cell cycle. Cell cycle is the set of steps leading to the duplication of the DNA of a cell and its division into two daughter cells. In a simplified view, the cell cycle starts with the G1-phase where the cell increases in size and prepares the DNA synthesis. The DNA is replicated during the S-phase. In G2-phase, the cell continues to grow to prepare for its division into two daughter cells during the mitosis (Figure 1.3 A). The duration of the cell cycle is the same for malignant and non-malignant cells. However, the proportion of cells that are in the cell cycle is greater in tumors than in healthy tissue. This *growth fraction* is characteristic of tumors and promotes their development. Cytotoxic agents use this characteristic and target cells that proliferate.

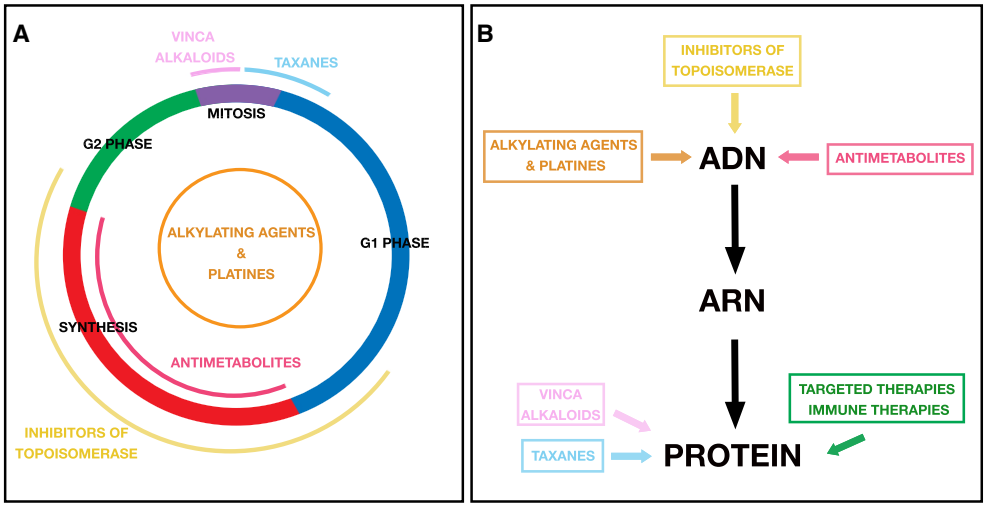


Figure 1.3: Actions of anti-neoplastic agents. A: The cell cycle with the time of actions of the different cytotoxic agents. B. Cytotoxic drugs targets the DNA except spindle poisons that target microtubules. Targeted and immune therapies are designed molecules to targets specific proteins of the cancer cell. Fig inspired from G. Vassal "Bases Experimentales de la chimiotherapie" - XXXème Cours de chimiotherapies, Gustave Roussy

Different kinds of chemotherapies exist. Alkylating agents generate lesions in the DNA strand. An alkyl group is attached to the guanine base of DNA conducting the formation of a covalent bond. These bonds prevent the strands of the double helix from linking, as they should, cause breakage of the DNA strands. When the lesions are not repaired, they accumulate until they lead to cell death. Often described as alkylating-like agents, platinums do not have an alkyl group, nevertheless damage DNA.

Inhibitors of topoisomerase act on enzymes (topoisomerase I and II) that unravel the DNA before synthesis and replication. By inhibiting the phase of religation of the topoisomerase, the DNA is physically blocked leading to the appearance of DNA lesions. Inhibitors of topoisomerase II are also known as anthracyclines. The antimetabolites are another class of chemotherapy. They work as decoys. Their structure is close to the endogenous substances necessary for the synthesis of nucleic acids. They act as competitive inhibitors of enzymes involved in nucleic acids synthesis, which once incorporated interrupts their synthesis.

Spindle poisons disrupt cell division by blocking the microtubules that move the chromosomes to the poles of the cell during the mitosis. Vinca alkaloides inhibit the polymerisation of microtubules, blocking their action. On the contrary, taxanes inhibits the depolymerisation of microtubules. They reach their maximum size and cannot move. These drugs are particularly efficient during the mitosis. They differ from other cytotoxic agents in that they target proteins (the microtubules) instead of the DNA strand directly (Figure 1.3 B).

Inhibitors of topoisomerase and antimetabolites act almost exclusively during the synthesis phase of DNA. Only S-phase cells, at the time of drug exposure, will be sensitive to treatment. Similarly, spindle poisons are efficient during the mitosis (Figure 1.3 A). In this case, a longer or continuous exposure to the drug is preferred. By increasing the exposure of the tumor to the drug, the probability that each of the cells that go into the cycle is exposed to the drug increases. Platinium and alkylating agents being active at any time of the cell cycle, they will preferentially be delivered spaced in time.

1.2.2 Molecular targeted therapies

Since few years, a novel class of antineoplastic agents emerges, called targeted therapies. The novelty is in the way those molecules have been developed. Indeed, we have seen that cytotoxic agents target DNA, and even the surgeon does a targeted therapy on the tumor during its removal. Molecular targeted therapies have been developed to target one of the oncogenesis mechanisms defined by Hanahan and Weinberg(74, 75). Most cytotoxic agents have been developed following a pragmatic development: once an agent has been identified to be efficient to kill cancer cells, its mechanism of action is determined a posteriori. Today, molecular targeted therapies follow a rationale development. The molecular alteration associated to oncogenesis is first identified, specific agents targeting this alteration are designed and then, their activity is demonstrated. Our understanding of targets is essential to give the right treatment to the right patient. Few biomarkers of response to targeted agents have been validated today. The over-expression of HER2 associated to the sensitivity to anti-Her2 agents (115), and the mutation BRAF V600E associated to BRAF inhibitors in melanomas (34) are well-known examples. However, biomarkers for most of the compound are still unknown (149).

Our understanding of action mechanisms of many targeted agents remains unclear. If anti-bodies are specific molecules acting only on their targets, inhibitors of tyrosine kinase are selective and can inhibit several targets, increasing the toxicity of these molecules. Moreover, one alteration is not always associated to the same drug response. Beautiful stories exist, like trastuzumab (anti-HER2) that is efficient to treat breast and colon cancer that over-expressed HER2 (15), but the biological context may lead to unexpected behaviours. The association BRAF V600E mutation and vemurafenib has demonstrated its efficiency to treat melanomas. However, vemurafenib on BRAF V600E mutated colon cancer, increases EGFR activity, causing the tumor to proliferate (141). The idea would be then, to combine different targeted molecules, for example here inhibitor of BRAF and inhibitor of EGFR, to treat those cancers (191).

1.3 Pre-clinical models for cancer research

"All models are wrong but some are useful."

— George Box, 1979(28)

In this statement, Box explains that simple models are required to facilitate our understanding of more complex systems. It is almost impossible for a single model to describe perfectly a real phenomenon, but it can help us explain, predict and understand phenomena that surround us.
This statement was originally related to statistical models, that are a description of a phenomenon using mathematical concepts, but it is also true in biology. Biological models are needed to study mechanisms of tumorogenesis and to find new therapeutic targets. Furthermore, if clinical trials are the only real way to assess drug efficacy and toxicity, they are inadequate for testing the hundreds of drugs currently being developed(181).

1.3.1 *In vivo* models

Different kind of models has been developed to study the different aspects of cancer biology that cannot be explored in people. They are classified as *in vivo* and *in vitro* models. Derived mouse models are the most common *in vivo* models used in oncology. Genetically Modified Mice (GEMs) are mice in which the function of a cancer gene is modified to cause the onset of a given cancer. These models are well suited for studying tumor initiation and progression. As an alternative, tumor cells can be directly transplanted into immunodeficient mice. The so-called *xenograft tumor model*, frequently retain the cell differentiation, morphology, architecture, vasculature, peripheral growth and molecular features of the original tumor from the patient (112). They provide unparalleled opportunities for testing drug sensitivity. However, mice models remains expensive and time-consuming, and engraftment success rates depend on the type of tumor.

1.3.2 *In vitro* models

Cancer cell lines are isolated cells from patient's tumor that are cultured in petri dishes with nutrients. Cellular division enables to collect almost indefinitely replicates of cells. Most of the cell lines are commercially available allowing scientists to work on the same cells everywhere in the world. This *in vitro* model is the most widely used for oncology research because it is easy to produce and maintain. Cancer cell lines demonstrated similar drug response characteristics to those of the primary tumor of origin, such as the BRAF-V600E mutated melanomas sensitive to vemurafenib or HER2 amplification/overexpression conferring sensitivity to lapatinib (16, 34, 63, 97). They have also demonstrated their reliability to predict drug response in several clinical trials (56, 65).

However, even if some studies demonstrated similarity between cell lines and primary tumors (110), others highlighted their genomic divergences compare to tumor samples. For example, Domcke et al (52) analyzed the DNA sequences and transcriptomic profiles of commonly used ovarian cancer cell lines and high-grade serous ovarian cancer tumor samples. They distinguish between 'the good, the bad and the ugly' cell line models according to the number of significant differences with tumor samples. Further, the environment of a petri dish is very different from that of a living organism. The tumor microenvironment is completely absent in cancer cell lines (75), and cells do not have any interactions with other cells (like immune cells). Because these two elements play a key role in the way the tumor responds to antineoplastic agents (119), this is viewed as a limitation. Finally, poor correlations between pharmacological data established on cancer cell lines has been reported leading the authors to make a plea for standardization of pharmacological assays (72, 77).

1.4 Breast cancer

1.4.1 Epidemiology

Breast cancer is the first cancer diagnosed in woman worldwide with 1.7 million new cases diagnosed in 2012. It is the leading cause of death by cancer in woman with 521 900 deaths in 2012 (5). In France, it is the most frequent cancer in woman followed by colorectal and lung cancer. The number of new cases that doubled between 1980 and 2000 with 48 763 new cases diagnosed in 2012. 11 886 deaths have been recorded the same year. Conversely, since 1990 mortality has been decreasing steadily and significantly, with a rate of -1.5% per year between 2005 and 2012 (126) due to earlier diagnosis and improvement in therapeutic strategies.

1.4.2 Factors associated with breast cancer risk

There are a number of risk factors for breast cancer although there are still uncertainties about the involvement and weight of many of them. Nevertheless, knowledge of these risk factors allows the implementation of adapted measures and targeted screening.

- Age is the most important risk factor for breast cancer. The incidence increases with age, the disease is rare under 30 years and increases between 45 and 70 years, then decrease gradually.

- Family and genetic factors: In about 5 to 10% of cases there is a genetic predisposition to breast cancer. Two breast cancer predisposition genes have been characterized: BRCA1 and BRCA2. These genes are tumor suppressor involved in maintaining DNA integrity and genomic stability (194). They are essential for cell division and DNA replication error control (187). Those women present a lifetime risk of breast cancer of 45%-65%. A prophylactic (preventive) mastectomy may be proposed, studies demonstrating that it reduces their risk of breast cancer by 90% (5).

- Women who have started menstruating early (before age 12) and or went through menopause later (after age 55) have an increased relative risk (around 1.05 (73)). The breast is fully developed after the first full-term pregnancy. By increasing the age of first childbirths, the period of breast exposure to carcinogens is increased. Finally, breastfeeding could reduce the risk of breast cancer by 4% for every 12 months of activity, according to a review of 47 studies on 30 countries (21).

- Obesity, diet, tobacco and alcohol are other factors that are associated with breast cancer risk.

Except age and genetic, the relative risk associated to these factors are mostly between 1 and 2. They constitute very small risk factors. In comparison, people who smoke cigarettes are 15 to 30 times more likely to get lung cancer or die from lung cancer than people who do not smoke (137). It is reasonable to think that we will not be able to prevent the future 50 000, and certainly more, new cases diagnosed each year. Improving the management and treatments of these new cases appears then as a real public health issue.

1.4.3 Breast anatomy

The breast is an exocrine gland composed of a mass, an areola and a nipple (Figure 1.4 A). The mammary gland consists of 2 cell compartments: the mesenchymal compartment, perfused by blood vessels and nerves, and the epithelial compartment which is articulated around a network of galactophoric ducts and lobules containing the alveoli. Blood and lymphatic vessels circulate in the connective and adipose (fatty) tissues. Drainage by the lymphatic vessels takes place towards the internal mammary chain, the axillary and supraclavicular lymph nodes. The architecture of the mammary gland evolves throughout life, depending on the age and stage of reproductive life and is constructed under the influence of ovarian sex hormones (estrogens and progesterone).

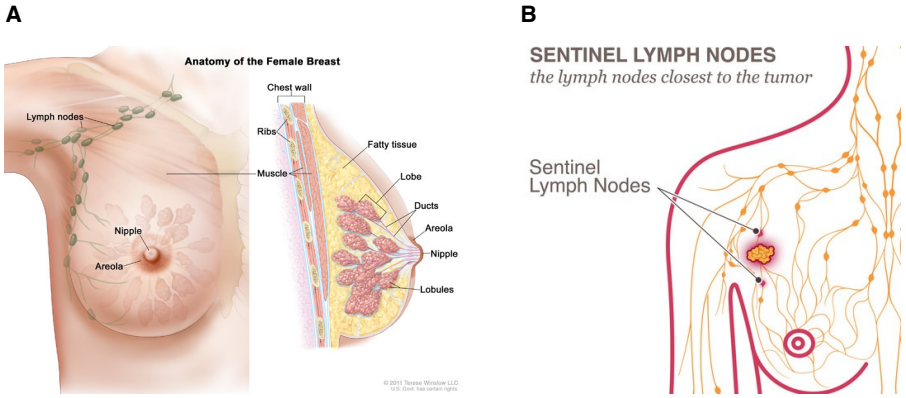


Figure 1.4: Breast anatomy. A) Schematic representation of a breast in section. B) Sentinel lymph node is the first "downstream" node of cancer in the lymphatic circulatory system. It is the entrance door to the spread of cancer cells into the body.

1.4.4 Clinical characteristics

To describe tumors severity and aggressiveness, pathologists commonly classified tumors according to stage, grade and histological type. These information are then used to guide the treatment.

Stage refers to the extent of the cancer and whether it has spread or not. Most staging systems include information about the size of the tumor, whether the lymph nodes are involved (Figure 1.4 B) (and how many lymph nodes are involved), and whether the cancer has spread to other parts of the body. The staging goes from 0 to IV with increasing stage corresponding to increasing tumor size and spread.

The grade quantifies how much the tumor cells are differentiated. It is based on the appearance of cancer cells, the shape of the nucleus and the number of cells in division. Combining these three characteristics, tumors are assigned a grade of 1 to 3. From tumors with well differentiated cells, weakly proliferating, to less differentiated cells with many mitoses.

The first classification of breast cancer is based on histological type. The two most common histological types of breast cancer are ductal and lobular carcinomas. Ductal means that the cancer start in the milk ducts, that carry milk from the milk-producing lobules to the nipple, and spread to the surrounding breast tissues. Lobular means that the cancer began in the milk-producing lobules. Ductal carcinoma in situ (DCIS) accounts for 80-85% while lobular carcinoma in situ constitutes 5% of all in situ tumors. In the invasive case, 75% of breast cancers are invasive ductal carcinoma (IDC) and 15% are lobular (ILC). Others invasive cases are a collection of several histological variants, each of which accounts for no more than 2% of all invasive cases (138).

1.4.5 Molecular subtypes of breast cancer

High-throughput technologies have highlighted the great heterogeneity of breast cancer. Today, breast cancer is considered as different diseases sharing the same anatomical localization. Studies such as those from Perou (136) and Sorlie (169) have shown that on the basis of expression profile, it was possible to identify different subgroups of breast cancers with their own pathological and genomic features, that are different in terms of clinical presentation (lymph nodes invasion, local and regional recurrence, localization of metastases) (Figure 1.5). Since then, many classifications have been proposed (46, 195, 195) but breast cancers are generally classified in four distinct molecular profiles: luminal A, luminal B, HER2+ (Human Epidermal Growth Factor Receptor 2), Basal-like and Normal-like. The majority of breast cancer belongs to the luminals (65-75%), followed by the basal-like (10-15%) and the HER2+ (6-10%) (138).

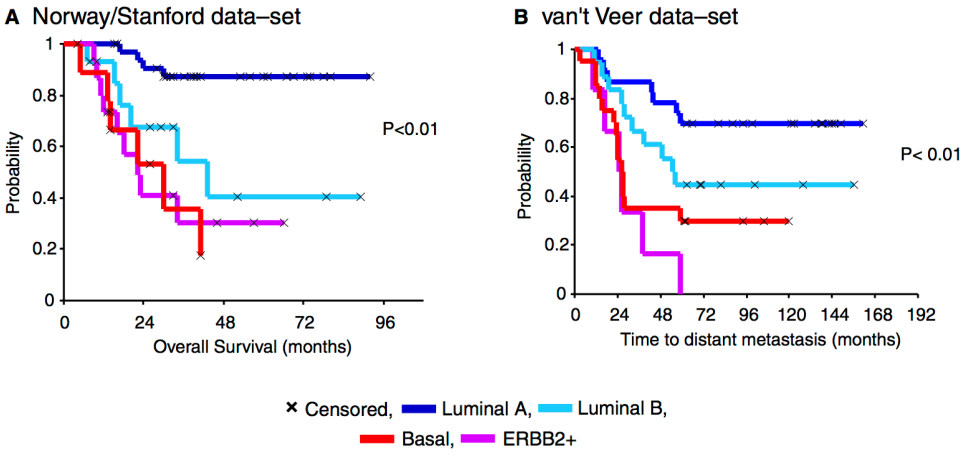


Figure 1.5: Breast cancer subtypes significantly differs in outcome. A) Overall survival for 72 patients with locally advanced breast cancer. (B) Time to development of distant metastasis in the 97 sporadic cases. ERBB2+ denote HER2+ subtype. Figure from Sorlie et al. (168)

Luminals A and B subtypes are defined by an expression of estrogen (ER) and/or progesterone receptor (PR) and no expression for HER2. They are called "luminals" because they share some similarities with luminal cell-lineage gene expression profile. They are the most frequent subtypes (65-75%). Luminals A over expressed cyclin D1 and have PIK3CA mutations. They are characterized by a low proliferation rate (Ki-67 low), are mostly low grade at diagnosis with a relatively good prognosis. Luminals B have a higher expression of proliferative genes and mutations in TP53 and PIK3CA. They are diagnosed with a higher grade compared to luminals A, which is associated with a poorer prognostic. However, luminal tumors are most commonly of low-grade and are associated with an early stage at diagnosis. Both subtypes are particularly sensitive

to targeted hormonal therapy and are associated with a good prognosis (33, 168).

Aggressive tumors are divided in two major subtypes. First are the HER2+ tumors (15-20% of cases) that are characterized by an amplification and overexpression of HER2 tyrosine kinase receptor gene. Their development is very aggressive. HER2+ cancer used to be one of the most lethal cancers. The introduction of the HER2-targeted monoclonal antibody-based treatment (trastuzumab) in early 2000s was considered "not evolutionary but revolutionary" (86). This targeted therapy has reversed the survival curve of these women. Further, new molecules that conjugate antibody and chemotherapy, like T-DM1 (68), dramatically help to control and even cure the disease.

Basal-like are defined as ER negative and HER2 negative. They highly express basal keratins genes and EGFR. Most of them are mutated TP53 with complex genomic rearrangements. Basal-like tumors are most commonly high-grade at diagnosis. They are particularly aggressive with high recurrence rate and a poor prognosis. Today, no targeted therapy exists for this subtype. Basal-like and Triple Negative subtypes are closely related, see section 1.5 for further description.

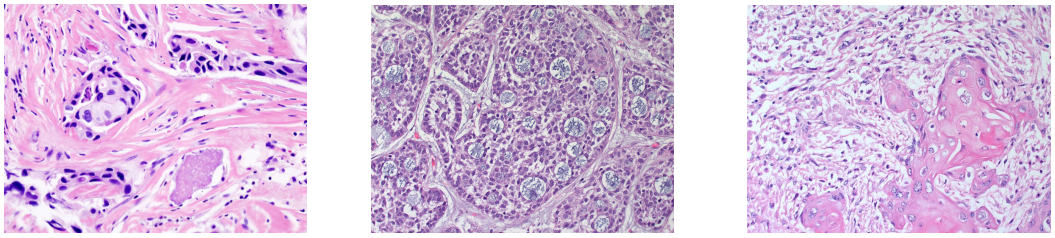
1.5 Triple negative breast cancer

The term "triple negative" breast cancer (TNBC) is an immunohistochemical definition corresponding to the absence of estrogen receptor and progesterone expression and the absence of HER2 receptor expression. In France, the thresholds of negativity retained are less than 10% of cells labeled for hormone receptors. Although TNBCs account for approximately 15% of all cases of breast cancer, there are associated with high recurrence rate and short survival duration (17**?**). Moreover, TNBC tumors are generally larger, are of higher grade, have lymph node involvement at diagnosis, and are biologically more aggressive (71). Beyond the aggressive nature of the disease, this type of tumor presents a heterogeneous clinical behavior that can explain its poor prognosis.

1.5.1 Triple negative breast cancer a heterogeneous disease

Histological classification.
The vast majority of TNBCs are classified histologically as invasive ductal carcinomas. Other subtypes have been described. Among them, the invasive lobular carcinoma represent 1 or 2% of the cases and other rare subtypes are found in less than 1% of all cases of TNBCs (Secretory carcinoma, typical medullary carcinoma, atypical medullary carcinoma, apocrine carcinoma). Some are of specific interest since they are less aggressive and only capable of local recurrence (adenoid cystic carcinoma, spindle- cell metaplastic carcinomas and adenosquamous carcinoma) (Figure 1.6).

Molecular classification.
The first molecular characterization of breast cancer have identified different "intrinsic" subtypes, among them the so-called "basal-like" subtype (BLBC) (136, 169). These



(a) (b) (b)

Figure 1.6: The heterogeneous histological landscape of triple-negative breast cancer. Some examples of TNBC histological subtypes. The most common a) Invasive ductal carcinoma (95%), and two examples of rare subtypes b) Adenoid cystic carcinoma ($< 1\%$) and c) Metaplastic Carcinoma ($< 1\%$)

tumors do not expressed ER, PR and HER2 but express some characteristic genes of normal basal cells such as CK5/6, CK14, CK17 and EGFR (146). BLBC tumors are often assimilated to TNBCs but these two groups are not synonymous (147). Indeed, more than 90% of BLBCs are TNBCs while BLBC represents the most frequent subtype of TNBC (55–81%) (142).

Lehmann et al were among the first to publish a study trying to better dissect the TNBC-specific heterogeneity(100). Through the analysis of 21 public sets of expression data, and 587 TN tumors, this study identified 6 molecular subtypes of TNBC : two basal-like-related subgroups (basal-like 1 (BL1) and 2 (BL2)), two mesenchymal-related sub- groups (mesenchymal (M) and mesenchymal stem-like (MSL)), one immunomodulatory subgroup (IM) and one luminal androgen receptor group (LAR). Each of these subtypes has specific molecular abnormalities. The BL1 and BL2 subgroups are both enriched in proliferation genes. BL1s also express genes involved in DNA repair whereas the BL2 subgroup expresses genes involved in growth signaling pathways. The M subgroup is enriched with genes involved in cell mobility and the epithelial-mesenchymal transition. The MSL subgroup has an expression profile close to the M subgroup and is enriched in genes involved in angiogenesis and in some immune response signaling pathways. The IM subgroup is enriched with genes involved in the immune response and lymphocyte infiltration. Finally, the LAR subgroup that represents about 10% of the TNBCs, shares many common genetically features with luminal-like ER-positive breast cancer. Furthermore, it expresses the androgen receptor (AR) in the presence of a luminal- like expression signature and thus, might be treated with agents that target AR.

The clinical relevance of this classification was evaluated in a retrospective analysis of 130 patients with TNBC treated with chemotherapy before surgery. The authors show that different responses can be observed according to their TNBCs subtypes. Patients with BL1 tumors achieve the highest pathological complete response (pCR, see below for a detail definition) rate (52%) and patients with tumors classified as BL2, LAR and

MSL have the lowest response rates (0%, 10% and 23%, respectively).

Curtis et al published in 2012 a novel breast cancer classification based on the analysis on 2000 breast tumors (METABRIC dataset) (46). They define 10 integrative clusters (IntClust) of breast cancer based on an integrative analysis of genomic and transcriptomic data. The BLBCs are spread among the different subgroups but 80% of them are present in IntClust 4 and 10. IntClust 10 subtype is characterized by a high genomic instability with loss of chromosome 5 and gain in 8q, 10p or 12p. IntClust 4 have extensive lymphocytic infiltration with a strong immune and inflammatory signature and few copy-number aberrations. Patient from IntClust 4 have a better prognosis demonstrating the clinical relevance of this classification.

Despite this new evidence that TNBCs are histologically, genetically and clinically heterogeneous, TNBCs are considered as a single clinical entity and uniformly treated with chemotherapy.

1.5.2 Potentially actionable molecular alterations in TNBC

The emergence of massive parallel sequencing and large-scale project as TCGA (183) or METABRIC (46) allowed establishing the molecular profile of numerous TNBCs (129, 159). TNBC subtype is not a cancer subtype with a high mutation burden. Around 60 somatic mutations are present in each tumor in average (1.68 somatic mutations per Mb of coding regions). Only TP53 mutations are found at a high frequency in TNBC (60-70% of mutations). The next most commonly mutated gene in TNBC is PI3KCA globally mutated in around 10% of tumors (129, 183). All other mutations occur at a low (1–5%) to very low frequency (<1%) in TNBC (MLL3, CDH1, PTEN, RB1, NF1, FOXA1, ERBB2, KRAS, HRAS) (140).

Some TNBC shares similarities with BRCA1 and BRCA2 mutated breast cancers (107, 187). This phenotype is often called “BRCAness” and is characterized by a basal-like phenotype, ER-negativity, EGFR overexpression, MYC amplification, TP53 mutations, extreme genomic instability and sensitivity to DNA-crosslinking agents. If BRCA1 mutations are rare (less than 5%) in breast cancer, it occurs in around 10% of TNBCs (159, 183). Furthermore, germline mutations in BRCA1/2 are associated with pathological high-grade cancers and increase the risk to develop a breast cancer by 5 (12% of women in the general population, 55 to 65% of women with harmful BRCA1 mutation) (87).

Multiple amplifications (PIK3CA, KRAS, BRAF, EGFR, FGFR1, FGFR2, IGFR1, KIT or MET) and deletions (PTEN) have been reported in TNBCs at different frequencies (1-40%) (46, 183). Several pathways are affected by multiple alterations at different levels (DNA repair pathway, PI3K/mTOR pathway, RAS/RAF/MEK pathway, Cell-cycle checkpoints,JAK/STAT pathway, AR pathway Notch pathway, JNK/AP-1 pathway, HIF1-α/ARNT network) (14, 129, 159).

If, no protocol involving these alterations has yet been shown to be clinically effec-

tive, they provide evidence of the substantial clonality and intratumors heterogeneity of TNBC that could lead to the development of resistance to therapies (205).

1.6 Drug-resistance of breast cancer

1.6.1 How to define drug-resistant breast cancers

Evaluation of drug efficiency in breast cancer is a not an easy task. In most cases, oncologists choose the right treatment that is highly effective, at least at first. But sometimes, patient's tumor can stop responding after a variable period. Treatment may also kill most of the cancer cells, leading the patient to go into remission and a few period later cancer can return and no longer responds to treatment. The difficulty in breast cancer comes from this relapse-free period that varies considerably across subtypes. Colleoni et al (39) analyzed 4,105 patients with 24 years of follow-up between 1978 and 1985. They report that most local recurrences occur for all patients within the first 5 years after diagnosis. The annualized hazard is of 10.4% with a peak of 15.2% during the first two years. ER negative patients have a higher risk of relapse during this period compare to ER positive patients. After 5 years, patients with ER positive breast cancer have a significant higher risk of relapse even after 24 years of follow-up, compared to ER negative patients. The authors also denote an improvement in breast cancer relapse-free survival rate and suggest that it is probably related to modern adjuvant treatment. Though the relapse hazard reduces with time (5 to 10 years : 5.4% for ER+ and 3.3% for ER- ; 20 to 24 years: 1.3% for ER+ and 1.4% for ER-), the study highlights the difficulty to clearly define that a tumor has resisted to the treatment or not on such a period. Further, the evolution of antineoplastic treatments since 1985, has probably lead today, to a longer risk period for those patients.

1.6.2 The contribution of neoadjuvant chemotherapy

Neoadjuvant chemotherapy (NAC) consists to administer cytotoxic treatment before surgery (in comparison to adjuvant therapy that is after surgery). It was originally used in locally advanced inoperable disease in order to achieve surgical resection. Today, NAC becomes more and more a standard of care for TNBCs. The objective is to reduce the tumor volume, thus increasing the possibility of breast conservation. At the same time, NAC allows a direct identification of *in vivo* tumor sensitivity to different agents. NAC has demonstrated higher response rate in TNBCs compared to any other breast cancer subtype (41, 199) and studies have shown strong correlation of breast cancer responses to NAC with survival and prognosis (197, 198).

The efficacy of NAC is quantified by the pathological complete response evaluated on surgical specimen. No standardized definition for pCR exists. Some studies have applied the pCR definition to the breast tumor only, whereas others have included the axillary

nodes. The pooled analysis by Cortazar et al.(41) retrieves three major definitions commonly used in a review of 12 international neoadjuvant trials:

- Absence of invasive cancer and in-situ cancer in the breast and axillary nodes
- Absence of invasive cancer in the breast and axillary nodes, irrespective of ductal carcinoma in situ
- Absence of invasive cancer in the breast irrespective of ductal carcinoma in situ or nodal involvement

Although standardization of the definition of pCR is necessary, several studies have shown that patients with pCR tend to have improved disease-free and overall survival compared with patients with residual disease (177).

Approximately 30% of patients with early-stage breast cancer treated with standard neoadjuvant anthracycline and taxane-based chemotherapy achieve a pathologic complete response after treatment. For patients that do not fully respond to NAC the prognosis is really poor with high rates of metastatic recurrence (41, 104). NAC appears then as a very important tool to assess tumor response to a given treatment and speed up decision-making for adjuvant therapy.

Since chemotherapy is done before surgery, NAC makes it possible to obtain tumor material which has undergone treatment without additional constraint for the patient. High-throughput technology evolution enables to identify molecular alterations present in the residual disease. Justin Balko et al (14) molecularly profiled residual disease after NAC of 74 TNBCs. This analysis provides insight on how diverse posttreatment tumors are. Most of the alterations detected were present in less than 5% of the samples. Only three genes were found altered in more than 30% of the samples: mutation in TP53 (89%) and MCL1 (54%) and MYC gene amplification (35%). These results suggest that most of the tumors after NAC do not have the same mutational profile. However, when alterations were categorized into functional pathways, 90% of the tumors contained a genetic alteration potentially treatable with a currently available targeted therapy (Figure 1.7). Several alterations present in the residual disease were associated to patient outcome. Among them, only PTEN alteration was associated to a good overall survival. BRCA1 mutation and gene amplification of JAK2 and MYC demonstrated a trend toward bad overall survival.

Given the heterogeneity of TNBC, personalized treatment represents a hope to treat the 70% of patients with a residual disease.

1.6.3 Mechanisms of resistance

Research over the past decades has uncovered several general mechanisms of drug resistance. Most of them have been recently nicely reviewed (27, 84, 119, 132). In the next subsection we will introduce some of these findings linked with TNBC chemoresistance

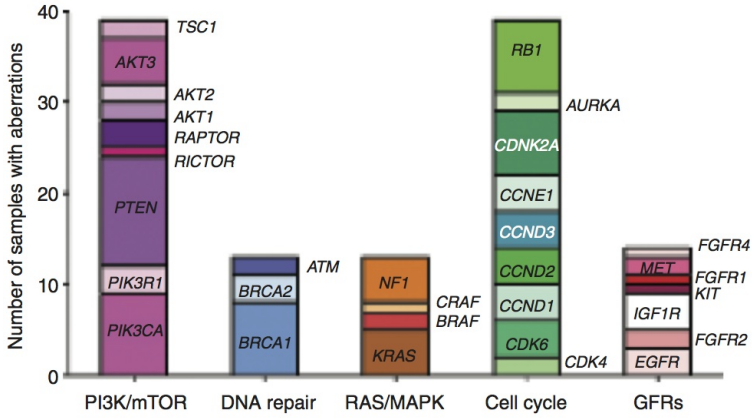


Figure 1.7: Alterations in pathways found in 81 TNBCs residual disease by Balko et al (14)

and summarized in Figure 1.8.

Drug resistance can be mediated by factors that pre-exist before receiving any treatment (intrinsic resistance). But treatments themselves can also act as a selective pressure that drives cancer cells to evolve. Thanks to their genome instability, malignant cells can set up adaptive responses and activate alternative compensatory signaling pathways (109).
One way of preventing the compound from having time to act, is to rapidly move it out of the cell. This process, called drug efflux, reduces the concentration of the drug in the cell and its toxic effect. The cell undergoes fewer lesions on the DNA that cell can repair to survive. Transmembrane proteins such as MDR1, MRP1 or BCRP act as pumps that reject cytotoxic molecules (such as taxanes, topoisomerase inhibitors, and antimetabolites) outside the cell.
β-Tubulin is a structural protein participating together with α-tubulin in the formation of microtubules. It has been demonstrated that when class III β-tubulin is incorporated, microtubule networks are impossible to block leading to chemoresistance. High level of class III β-tubulin expression was linked to taxane based drug-resistance like paclitaxel.
DNA damages normally induce cell cycle arrest either to direct repair the damages, either to activate cell death. Cancer cells may bypass these control mechanisms due to oncogenes or tumor suppressor genes alterations. P53 proteins regulate G1/S cell cycle checkpoint. When TP53 gene is mutated, the cell will not carry out any checks and continue the cycle. MLH1 and MSH2 genes are part of the mismatch repair system that is crucial for maintaining genomic integrity. Alterations of these genes may lead to topoisomerase inhibitors resistance.
Many anti-neoplastic agents exert their cytotoxic effects by inducing apoptosis. Defects in apoptosis mechanisms may confer chemoresistance to tumor cells. The role of BCL-2

protein family in maintaining the balance between cell survival and apoptosis and the one of caspase-3 in the onset of apoptosis have been extensively studied and linked with chemoresistance.
Some antimetabolites are administered in an inactive form and need to be metabolized by the organism to be active. Since the cytotoxicity of these "prodrugs" is low prior to activation, there is a much lower risk for healthy cells, which reduces the associated side effects. Cancer cells may inactivate the metabolic process needed by the prodrug making it inefficient. Cancer cells with overexpression of ALDH1A1 and ALDH3A1 were found less sensitive to alkylating agents.
The tumor microenvironment plays a major role in drug resistance (119). The anarchic vascularization of the microenvironment prevents cytotoxic agents from penetrating tumors. Further, changes in tissue architecture, cell-cell adhesion, integrin expression, and extra-cellular matrix organization have been demonstrated to induce drug-resistance.
Finally, the genome instability of tumor cells creates different subpopulations of cells inside the tumor. Therapeutic pressure may lead to select the resistant subclones while those that are sensitive are eliminated.

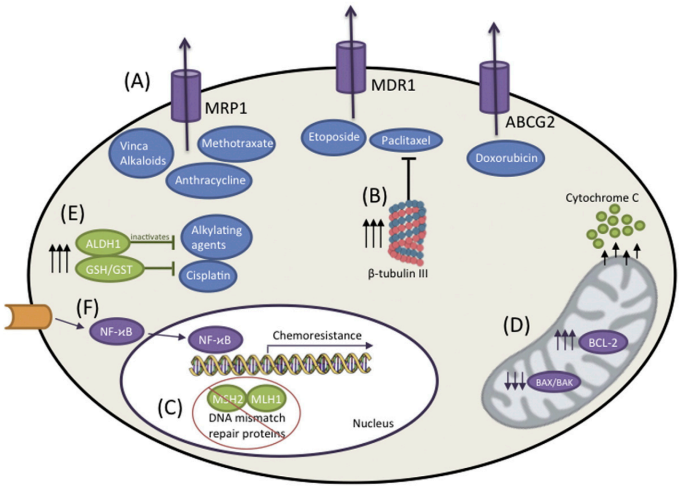


Figure 1.8: Mechanisms of Chemoresistance in TNBC. (A) Drug efflux transports the drug out of the cell (B) Overexpression of *beta*-tubulin III subunit blocks taxanes effect (C) Dysfunction of DNA repair induces several drug resistance (D) Alterations in genes involved in apoptosis prevent chemotherapy-induced apoptosis (E) Alteration of metabolic process required for prodrug activation. Figure from O'Reilly et al. (132)

1.7 Experimental high-throughput technologies for cancer research

Molecular biology has known an unprecedented revolution over the last decades. Emergence of high throughput assays provided huge amount of data that biologist alone cannot analyzed. Technologies are now available to measure not one but thousands data point at a time producing several gigabytes of data. In addition, several levels of complexity can now be analyzed from tumor cells : changes in the DNA sequence, level of gene expression, epigenetic modifications or identity of proteins produced by cancer cells with their quantity associated.

1.7.1 Microarrays

The high-throughput revolution starts with the introduction of microarrays in 1995. They make it possible to quantify the mRNA expression of the whole genome on a simple surface of glass or plastic. They rely on single-strand oligonucleotides called probes chosen to be specific to a transcript. They are amplified and fixed on a miniature solid support. The RNA is extracted from a tumor sample, labeled with a fluorochrome and hybridized on the chip. After the fluorochrome has been stimulated at the appropriate wave length, the signal intensity of the fluorescence light allow quantifying the expression levels of targets which attached to the probe.
One major drawback of this technology is that we quantify only the transcripts designed as probes according to our current knowledge. It is then impossible to discover novel transcripts. Further, its signal level limits fluorescence technology. Genes that are low expressed in a sample could not be distinguished from background chip level, and overexpressed genes may lead to signal saturation limiting their exact quantification.

1.7.2 High-Throughput Sequencing

Next generation sequencing (NGS) is a set of technologies that allow scientists to sequence DNA and RNA much faster and cheaper than Sanger sequencing. We will focus here on the Illumina technology that was used to generate the data analyzed in the thesis.

1.7.2.1 General principles

The DNA extracted from a sample is broken up into small fragments and short sequences of DNA, called adaptors, are attached. These adaptors are used to anchor the fragment on one end of the flow cell (Figure 1.9 A). The DNA attached to the flowcell is then replicated to form small clusters of DNA with the same sequence - forward and reverse strands. When sequenced, each cluster of DNA molecules will emit a signal that is strong enough to be detected by a camera (Figure 1.9 B). To sequence these clusters, several cycles of 4 fluorescently tagged nucleotides compete for addition to the growing

chain. Only one is incorporated based on the sequence of the template. After the addition of each nucleotide, unused reactants are washed away and clusters are excited by a light source. A characteristic signal is emitted, detected by a camera and recorded on a computer. Each of the nucleotide (A, T, C and G) gives off a different color. The fluorescently-labeled terminator group is then removed from the incorporated base and the cycle is repeated until the full DNA is sequenced (Figure 1.9 C). This is what is called sequencing by synthesis. All identical strands are read simultaneously and hundreds of millions of clusters are sequenced in a massively parallel process.

The sequences obtained are small fragments of the genome (reads) that need to be aligned along a reference genome to be analyzed.

1.7.2.2 DNA sequencing

DNA can be sequenced either at the whole genome scale (whole genome sequencing WGS) to determined the complete DNA sequence of an organism's genome, either at the whole exome scale (whole exome sequencing WES). WES consists in the capture, sequencing and analysis of all exons in the human genome. It allows identifying mutations that potentially alter proteins at a much lower cost than WGS since we sequence only regions which are translated into functional proteins.

DNA sequencing allows discovering mutations and polymorphism. After the reads have been mapped along the genome, variant caller software look for positions where the reads sequenced differs from the reference. Genes can mutate in either a somatic or germinal tissue. Germline mutations occur in germ cells that can be transmitted to the next generation of the individual. Somatic or acquired mutations does not affect the reproductive cells but can occur in any other type of cell and cause cancer.

DNA next-generation sequencing also allows quantifying DNA copy number and identifies loss of heterozygosity (LOH). Copy number variation is difficult to perform on whole exome sequencing since the probability that a breakpoint occurs within a gene is low. However recent publication overcome this limitation (161).

1.7.2.3 RNA sequencing

General principles of RNA and DNA sequencing are the same. However, additional steps are needed to convert the RNA to DNA through reverse transcription. The complementary DNA (cDNA) produced is then sequenced as regular DNA. RNA-seq reads are then aligned to a reference and the number of reads that fall into a given gene or exon, quantifies its level of expression. Compared to microarrays, RNA sequencing does not need to design any probes since transcripts are directly sequenced. The gene expression level is estimated by counting data rather than fluorescent signals. The estimation is then much more precise without saturation. RNA-seq enables researchers to perform different kinds of analyses at the gene or transcript level. Comparing the number of

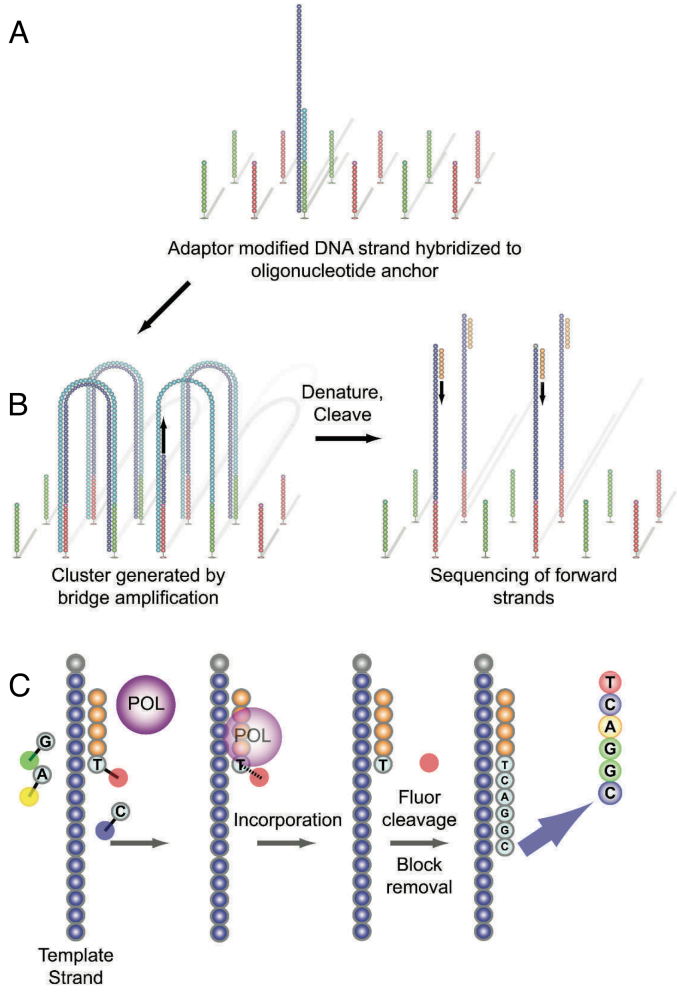


Figure 1.9: Cluster generation and sequencing on the Illumina platform. DNA fragments are digested to small fragments, and adapters are attached. These adapters will then be used to anchor the fragment on one end of the flow cell. Bridge amplification generate cluster of identical sequences being localized in the neighborhood of the first fragment, due to the anchoring. At each cycle, 4 uniquely chromophore compete for addition to the growing chain. Once added, the fluorescence signal of the incorporated base is recorded. The delocking step allows the incorporated nucleotide to bind to another nucleotide and the cycle is repeated until the full DNA molecule is sequenced. Figure from Coonrod et al. (40)

reads that fall into a given exon between two conditions can then identify alternative splicing events.

1.7.3 High-Throughput Screening

High-throughput screening (HTS) is a process of testing numerous drugs against many cancer cell lines (114, 179). It has become a standard method to accelerate the identification of new drugs, by screening large libraries composed of hundreds (or thousands) of drugs candidates. HTS can provide the results in few days or weeks. Researchers and industrials, need to test very large numbers of compounds in a time and cost efficient way, given the high number of drugs currently being developed (181) and all the combination possibilities. Thanks to the miniaturization and automation, HTS reduces reagent use and the time an employee spends doing repetitive tasks. Costs but also human errors are then reduced. Cellular microarrays are solid support (plates) of 96 or 384 wells in which cancer cell lines and compounds are displayed. Response to different compounds or to the same compounds at different concentrations can be analyzed. HTS is a two-step procedure. First, the whole library is tested in primary screens with high dose to identify compounds that are efficient to kill the cancer cell lines treated. Those "hits" are then more precisely tested in a second screening using classically 8 to 10 doses serially diluted. The dose-response curve and the concentration at which the drug inhibited 50% of maximal cell growth (IC_{50}) are classically computed.
In vitro screening of large compound libraries has become an essential tool for drug development. The assessment of drug toxicity or identification of inefficient compounds early in the development process of novel agents, enables to reduce the high attrition rate of these procedures. Further, it is the first step before exposing novel agents to animals and performed human trials.

Large-scale pharmacogenomics studies proposed to couple HTS with high-throughput molecular characterization of cancer cell lines. Statistical methods or machine-learning algorithms are then applied to find biomarkers or companion test associated to drug sensitivity. The identification of these biomarkers is a very difficult task. Collaborative effort organized by the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project have tried to solve this issue. In 2012, they proposed to the scientific community to build models capable of ranking the sensitivity of 18 breast cancer cell lines to 31 compounds (42). Participants had access to exome sequencing, RNA sequencing, methylation and proteomic data to train their model and do their predictions. Teams were scored based on a modified version of the concordance index, scaled between [0,1] (0 means not any predictions were correct, 1 means all predictions were correct). A total of 44 drug sensitivity prediction algorithms were analyzed. The best team reached a prediction score equal to 0.583. Few years later in 2016, they launched the AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge. Here, participants had to predict the response of 85 cancer cell lines (primarily colon, lung, and breast) to combinations of 118 drugs (53). Molecular data (mutations, methylation, copy number, gene expression) as well as monotherapy were available to achieve the challenge. The score used to discriminate participants was a weighted mean of the Pearson correlation across all drug combinations. Among the 90 teams participating, the first ranked team achieved a score of 0.49. Results from the Dream challenges are almost as good

as random which illustrate the difficulties associated to this kind of study, either in the way we used the data, either in the information present in the data.

1.8 Computational biology

High-throughput technologies produce very large matrices of numbers that are hard to interpret. The goal of computational biology is to develop statistical methods and algorithms to give a biological sense to these numbers. The data normalization constitutes the first step of an analysis. It corrects for technical variations introduced during the production of the data and preserves the biological information. This step is crucial so that analyzes are not polluted by signals without biological information that could lead to wrong conclusions. Once data have been cleaned (further quality controls may be needed), samples can be compared.

During my PhD, I have mostly studied transcriptomic data. However, one of my projects described in chapter 3 combines whole exome and RNA sequencing. I was able to apprehend some concepts of variants detection and copy number analysis based on WES. In this section we will introduced some of the methods used to analyze these so-called "-omics" data . We have chosen not to detailed all the existing methods of analysis in order to spare the reader. We preferred to present the methods we used to solve a given problem. We will first present the methods available to study transcriptomic data followed by those dedicated to WES data. We will also discuss the statistical problems associated with the emergence of these high-throughput data.

1.8.1 Unsupervised clustering

Transcriptomic data provides gene expression profile of thousands of genes (or transcripts) for a set of tumor samples. Given the heterogeneity of the tumors, it is interesting to group the tumor samples based on similarity in their overall expression patterns. Unsupervised clustering organizes the samples by calculating similarity of their molecular profile. We expect that samples with similar transcriptomic profiles have similar clinical and biological properties. Hierarchical clustering with Pearson correlation has been widely used to discover novel cancer subtypes and highlight cancer heterogeneity(136, 169). Classically, gene clustering is also performed to identify co-expressed genes. These genes are susceptible to be regulated by the same transcriptional factors or be part of the same signaling pathway.

Before applying any clustering algorithm, we have to filter the data. It is obvious that not all the 22,000 genes of our organism are linked to cancer processes. This space needs to be reduced in order to keep only the most relevant features based on which clusters with biological or therapeutic relevance can be built. When comparing different samples, it is expected that the expression of some genes will differ according to the sample population they belong to. A common practice is then to focus on the genes that

have a large variance across the tumor samples.

Fundamental issue when performing clustering analysis, is how to define the number of clusters. Monti et al (121) have proposed to perform multiple runs of the algorithm by subsampling items (samples or genes) at each iteration. If two items are really similar, they should stay together whatever the subset of items used. The number of time two items are clustered together, adjusted by the number of time they are in the same subsample, is then computed for a given set of cluster k. The number of cluster can then be chosen based on the k giving the highest pairwise consensus values.

1.8.2 Differential gene expression analysis

In comparative transcriptomic, one of the most basic questions is to identify genes that are downregulated or upregulated between several populations. The easiest way to do that is to use classical statistical tests or linear modeling to determine whether the gene expression is statistically different between groups. However, in many transcriptomic analyses we do not have enough observations to reliably estimate the gene variance. It has been proposed to use all other genes whose expression has been quantified, to increase the precision of this estimate. The very popular R package *limma* (150) proposed to "moderate the standard errors across genes" thanks to a Bayesian step. For each gene, the sample variance is computed then adjusted or "shrunk" towards the average variance based on all the genes. According to the authors, "this has the effect of borrowing information from the ensemble of genes to aid with inference about each individual gene".

The advent of the RNA-seq introduced new challenges in the identification of differentially expressed genes (DEG). The use of classical statistical tools relies on the assumptions that continuous fluorescence signal produce by microarray, follows a Student or Gaussian distribution. RNA-seq data are read counts. Consequently, the first differential analysis methods developed, were based on Poisson distribution (204) assuming that the mean is equal to the variance. However, it has been reported that the gene expression variance across multiple biological replicates of in RNA-seq samples, is larger than its mean expression values. This is because the probability that a cDNA fragments originate from a given gene, is not the same between the biological replicates. The negative binomial model has then been proposed (6, 151) enabling additional biological variance in the model. Methods like DESeq (6) or egdgeR (69) used a generalized linear model extensions for negative binomial distributions and likelihood ratio test to assess the significance of the genes.

As an alternative, authors of limma package proposed to transform the data so that they are suitable to fit with standard linear mode (98). The so-called "voom" approach works as follows: 1) A standard linear model is fitted on the log-transformed data for all genes (Figure 1.10 a). 2) A locally weighted regression (LOWESS) is then applied to estimate the relation between the residual standard deviation and the average log-

transformed data for each gene (Figure 1.10 b). 3) The LOWESS estimated function is used to predict the standard deviation of each fitted log-transformed values from 1) (Figure 1.10 c). 4) The inverse squared predicted standard deviation for each observation is added as weight into limma's standard linear model.

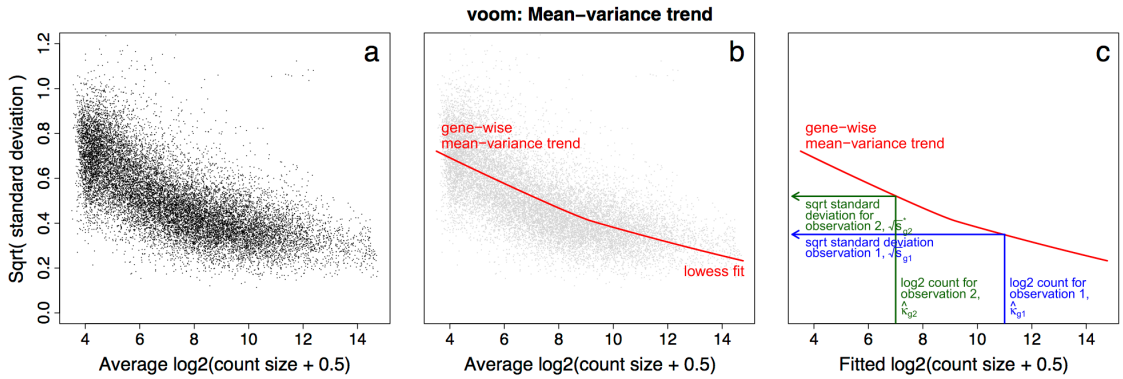


Figure 1.10: Voom mean-variance modeling. (a) Gene-wise square-root residual standard deviations are plotted against average log-count. (b) A functional relation between gene-wise means and variances is given by a robust LOWESS fit to the points. (c) The mean-variance trend enables each observation to map to a square-root standard deviation value using its fitted value for log-count. Figure from Law et al. (98)

1.8.3 Detection of differential alternative splicing

Alternative splicing (AS) has emerged as of particular interest in cancer research (47, 55). Differential gene expression analysis highlights genes with different level of expression between conditions. However, it does not inform whether different transcripts are expressed or not. Further, a gene can be expressed at the same level between conditions but different transcripts can be present, leading to a different phenotype that would not be detected with differential gene expression. Differential splicing analysis appears as a complement to give a more precise understanding of cancer biology. The detection of AS events requires more data than detection of differentially expressed genes. While the gene expression level is estimated by all the reads that map into a given gene, only the reads that include the AS region will count here. This can lead to a substantial number of useless reads. A second major issue is that genes produce many transcripts. Some of them are known, others can be the results of a mutation during the oncogenesis process. Some methods propose to reveal changes in the proportion of each isoform (22, 185). However, accurate reconstruction and quantification of full-length isoforms with the current short read sequencing is particularly difficult (171). In order to avoid isoform reconstruction, one can focus on the distribution of reads in exons and their junctions between conditions. It can then serve as a surrogate to estimate the transcripts present in a sample.

DEXSeq (8) approach proposes to divide the genome into disjoint "counting bins" that are exons or parts of exons. The number of reads falling into each exon inform on the inclusion or "usage" of a specific bin in a sample. The p-value for differential usage of each bin is calculated similarly than DESeq model, by adding an interaction term between the bin and the condition of interest. DEXSeq only considers annotated exons. Results can be biased according to the annotation file used and novel isoform discovery is not possible.
Methods like rMATS (162), propose to use reads that span across exon-exon junctions. These "junction reads" allow identification of novel un-annotated splicing events. Further, it allows distinguishing between cassette exon, alternative 5' and 3' splice site, intron retention and mutually exclusive exons. By combining reads from the inclusion isoform (I) and reads from the skipping isoform (S) an exon inclusion level ψ (or PSI for percentage spliced in) can be estimated (Figure 1.11). The number of unique isoform-specific read positions of the inclusion (l_I) and skipping (l_S) isoform are used to adjust the estimation of ψ as follows :

$$\hat{\psi} = \frac{(I/l_I)}{I/l_I + S/l_S}$$

A likelihood-ratio test is then used to "test whether the difference of the group mean between the two sample groups exceeds a user-defined threshold c, against the null hypothesis $|\Delta\psi_i| = |\psi_{i1} - \psi_{i2}| \leq c$", where i is the AS event and j is the conditions.

Alternative splicing are regulated by post-transcriptional processing induced by RNA-binding proteins (RBPs). These RBPs bind at specific sequences of nucleotides (or motifs) along the genome and can either positively or negatively regulate exon inclusion. Several studies have demonstrated the key role of the binding site position of splicing regulators. The position-dependent analysis of RBPs binding may reveal the different mechanisms that regulates AS events. Authors of rMATS have proposed an approach (133) that will be detailed and extended in chapter 4.

1.8.4 Assessing significance in high-throughput experiments

1.8.4.1 The curse of big data

The area of high-throughput experiments, have seen the development of novel statistical challenges. One of the first was the control of type I error (false positive rate) when multiple inferences are conducted. When a differential gene expression analysis is conducted, the difference between populations is assessed for each gene of the genome. Let us take an example with a *t-test* for simplicity. Suppose that we want to test if a given gene is differentially expressed between cancerous and normal cells. We compute the probability that the observed *t-statistic* is obtained by chance, called p-value. Formally, a p-value is the probability of getting the observed value of the test statistic, or a value with even greater evidence under the null hypothesis H_0 that the gene is not differentially

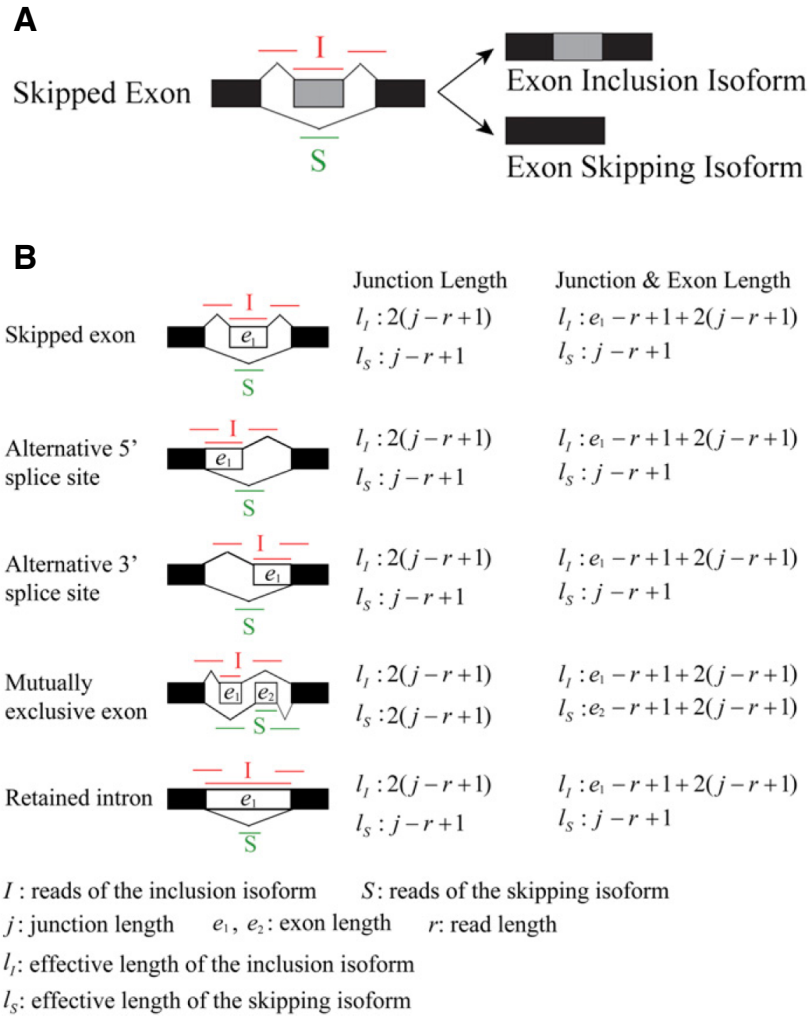


Figure 1.11: The schematic diagrams illustrating how the PSI can be estimated. A) The schematic diagram of an exon skipping event. I represents the exon inclusion reads from the upstream splice junction, the alternative exon itself, and the downstream splice junction. The exon skipping reads S are the reads from the skipping splice junction that directly connects the upstream exon to the downstream exon. B) The schematic diagrams of alternative 5' splice sites, alternative 3' splice sites, mutually exclusive exons and retained introns. Figure from Shen et al. (162).

expressed. We reject the null hypothesis if the p-value is smaller than a given threshold, for example 5%. We have then 5% chance that we called a gene differentially expressed when it was not. Now, suppose we want to find all the genes that are differentially expressed between our two populations. If we consider 10,000 genes, 10,000 t-test would be performed at a significant level of 5%, and we would expect to get 500 (i.e. 5%) false

positives by chance alone. We thus need to control the level at which significance is assessed to get reliable and replicable results.

1.8.4.2 Multiple testing correction

Assume that we want to perform m tests at the level α. The expected number of false positives is then equal to $m\alpha$. The Bonferroni procedure proposes to control the Family-Wise Error Rate (FWER), that is, the probability to do at lest one type I error. To do this, all null hypothesis whose p-values are below α/m are rejected. Then, controlling the FWER generally leads to very few discoveries.

The False Discovery Rate (FDR) is the expected proportion of false positives (FDP) among the rejected hypothesis. The FDR is a less stringent criteria compared to the FWER as it allows a user defined proportion of false positives. Benjamini and Hochberg have introduced a procedure to control the FDR (BH procedure) when the tests are independent or under some kind of positive dependencies (19, 20). Let us sort the m hypotheses $H_1, H_2, \ldots, H_m$ based on their respective p-values $p_1, p_2, \ldots, p_m$ in ascending order $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$. Each individual Benjamini-Hochberg critical values (or q-values) are then computed with the formula

$$\frac{i}{m}\alpha$$

where i is the individual p-value's rank, m is the total number of tests and α is the percentage of false discovery rate allowed. The BH procedure proposes to reject all null hypothesis for all $H_{(i)}$ in $i = 1, \ldots, k$ such as k is the largest index satisfying

$$p_{(k)} \leq \frac{k}{m}\alpha$$

Benjamini and Hochberg procedure has been extensively used during the last twenty years in many scientific fields including astronomy, biology, genetics, medicine, neuroimaging making it today the most cited paper outside its field of publication.

1.8.4.3 Post-hoc inference

Differential analysis can be considered as an unsupervised method in which no prior knowledge are used to detect the differentially expressed genes. One could be interested to find differentially expressed genes among a set of genes belonging to a given biological pathway. Classical multiple testing correction do not correctly control the number of false positives in such rejections sets because of the selection effect, as illustrated by Figure 1.12 from Blanchard, Neuvial & Roquain 2017 (24). Let us assume that y-axis represents the difference of gene expression between two conditions and x-axis the index. Assume that Figure 1.12 A represents genes from a given biological pathway. Seeking for differentially expressed genes here may lead to the identification of 5 significant genes. However, this is only due to the selection effect. Figure 1.12 B represent the whole genes

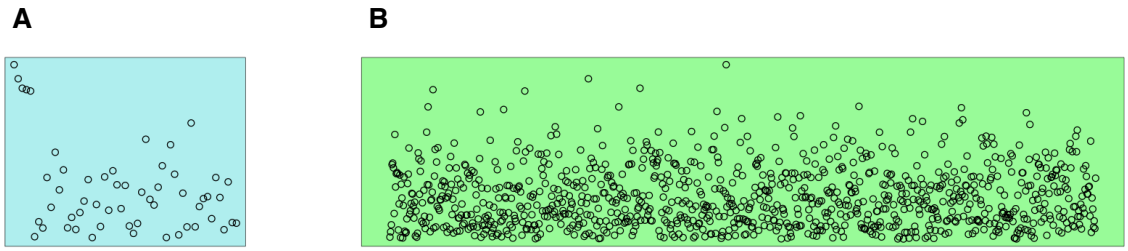


Figure 1.12: Illustration of the post hoc selection effect. Panel A is a random selection of 55 measurements selected from the dataset in panel B (100 measurements). Measures have been generated as i.i.d. absolute values of N (0, 1). Figure from Blanchard, Neuvial & Roquain (24).

were these 5 genes do not longer appear as significantly different. They were just the 5 outliers from previous selection, which is pure noise.

The so-called "post-hoc inference" proposes to build a statistical guarantee on any selected set of genes by considering the overall size of the data. The method gives to the user the possibility to look at the data multiple times before deciding what rejections to make. We will also demonstrate that we can use post-hoc inference to give a statistical guarantee when performing motif enrichment 4.

1.8.5 Whole exome sequencing analysis

1.8.5.1 Somatic variant detection

Whole-exome sequencing focuses on the gene coding sequences of the genome. The first step in the exploitation of the data is the reads alignment or mapping. The goal is to retrieve the genomic position from which the short sequenced reads have been generated. Those reads are aligned on a reference. Presence of SNPs and sequencing errors complicate the task and mismatches between the reference and the read sequence have to be taken into account. The read's position is then chosen, according to the region with the highest probability to fit with the read. The "variant calling" step allows identifying all the mutations present in a sample. Variant callers look for positions where the reads mapped and the reference differ. The ratio between the number of reads carrying the alteration over the total number of reads mapped at this position, the "variant allele frequency" (VAF) is then calculated. In order to distinguish between somatic and germline mutations, somatic variant callers first compare the reads mapped and the reference at a given position. If a difference is detected, the position is compared with the matched germline sample to see if the same variation can be found. Additional information such as variant's position within the gene (ie exonic, intronic), variant classification (ie, synonymous, nonsynonymous, missense, indel), presence of the variant in specific databases (ie dbSNP (163), 1000 Genome Project (171)) and prediction of the functional effect of the variant on the protein (SIFT (163), PolyPhen (2)) can be obtained with variant

annotation softwares like ANNOVAR (203) or snpEff (37). Once the variants have been identified and annotated, we have to filter sequencing errors and low impact variants in a list of around 20000 variants (40) to conduct downstream analysis with unpolluted data. This step requires hypotheses on the variants one would like to keep. Classical approaches begins to use a minimum threshold of variant detection based on the VAF or on the number of reads that carry the variant. Variants from dbSNP or the 10000 Genomes Project should represent only neutral polymorphisms found across human populations. One can assume that deleterious variants should not be found in those databases. Finally, one could limit the analysis to variants with strong impact on the protein like nonsense, missense, or frameshift variants as those predicted deleterious by SIFT and PolyPhen.

The development of precision medicine inevitably involves the clear and precise identification of a patient's mutations. Today, there is no standardization of WGS or WES sequencing, from library preparation to variant filtering. The sequencing depth has a key role in variant detection giving the middle to low purity of tumor samples contaminated by normal or stromal cells. Further, each variant caller is good to detect some kind of variants and bad to identify others. Adding to this the variations in the hypotheses to keep relevant variants across laboratory, discrepancies across studies are today a major challenge. The International Cancer Genome Consortium (ICGC (90)) has conducted a benchmark of complete sequencing pipelines across 5 sites worldwide. A single tumor-blood DNA pair from medulloblastoma was sent to each center then sequenced and analyzed with their own pipeline. According to the authors, "less than 20% of the total number of called variants were called by the 5 centers" and "one third of the mutations were unique to only one center" (Figure 1.13 A). These results are particularly worrying. Indeed, a common practice in WGS/WES is to compare the mutational profile of the in-house cohort to bigger cohort like TCGA or previous studies focusing on the same population. These variations in variant detection make the comparison really challenging and could lead to wrong conclusions. Indeed, ICGC benchmark looked at the impact of these discrepancies in the definition of mutational signatures from Alexandrov and al (4), leading to the establishment of very different signature profile (Figure 1.13 B).

Despite undoubtedly contributions of WGS/WES in biomedical research by reduction in costs and acquisition times, several challenges remain before it can safely become a routine part of clinical care.

1.8.5.2 Copy number detection

Whole exome sequencing was originally designed to identify mutations that are found in protein coding regions. Since the coding regions represent around 1% of the whole genome, researcher can sequence at much higher coverage by WES compared to WGS with the same cost. Several studies have proposed to use WES data for identification of copy number variations (CNVs) (25, 156). However, WES data is much noisier than

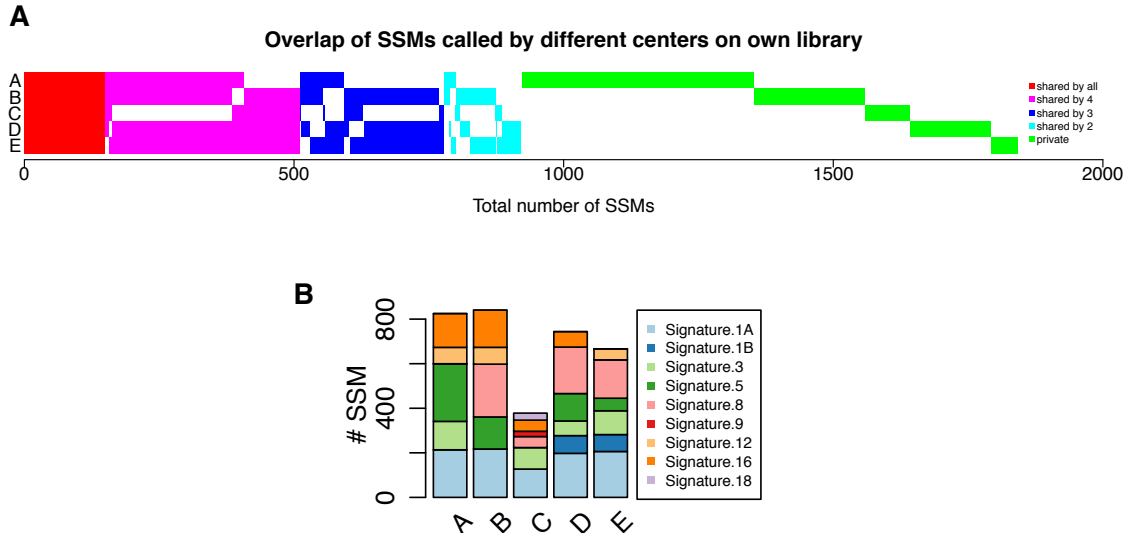


Figure 1.13: Comparison of pipeline to identify simple somatic mutations (SSM). A) Overlap of SSMs called by different centers. B) Mutational signatures for SSMs as defined by Alexandrov et al. (4). The calls from each center were used to fit into the predefined signatures. Figures from Buchhalter et al (30)

WGS due to the exon capture and amplification.
Shen and Seshan have proposed a new framework called FACETS to estimate allele-specific copy number (ASCN) and clonal heterogeneity for NGS data. While classical ASCN methods use only heterozygous SNP, because allelic imbalance can only be measured at these sites, FACETS proposes to use both homozygous and heterozygous SNPs to gain information on total copy number. The total copy number log-ratio (logR) is then computed for all SNPs with minimal depth of coverage in the normal separated by a given intervals to select independent events. The logR is defined as the total read count in the tumor divided by the total read count in the matched normal. In FACETS, it is modeled as a function of the parental copy number in the tumor that is a combination of normal and aberrant copy number, the estimated proportion of cells in the sample with the aberrant copy number (cellular fraction) and terms that correct for bias and to have absolute copy number. The frequency of an allele at a particular locus, the so-called B-allele frequency (BAF), is defined as the log odds ratio (logOR) of the variant -allele count in the tumor versus normal. This definition allow to correct for the reference allele that has frequently higher mapping rates compared to those at the variant allele. Several studies have demonstrated that the analysis of both dimensions (logR and logOR) of the signal jointly, improves precision in the identification of change points in the genome (139). To do so, FACETS use a bivariate Hotelling statistic to detect a breakpoints between two consecutive SNPs combining total and allele-specific read counts. Fianlly, FACETS explicitly consider clonal heterogeneity in the inference of ACSNs giving meaningful insight about the clonal and subclonal structure of the sample.

Copy number analysis from WES data remain a major challenge. Recent comparative studies have demonstrated that no methods perform perfectly and wet-lab experiments or microarrays platforms should be used to confirm the final call (125). However, FACETS proposes a nice unified framework that corrects for sequencing bias, performs ASCN analysis associated with clonal estimation of the tumor in a very efficient running time, making it a tool of choice for this type of analysis.

2

New insight for pharmacogenomic studies from the transcriptional analysis of two large-scale cancer cell line panels

To improve our understanding of the issues and challenges associated with pharmacogenomic data, we have started to investigate two recently published large scale pharmacogenomic datasets. This study allowed us to identify the strengths and limitations of high-hroughput screenings before analyzing a drug screening that we generated. This study have been accepted for publication in Scientific Report.
At the end of the study, a visualization tool that provides a wide range of graphics to help us access the CCLE and GDSC data was realized by a trainee Julie Setbon. During the first two months of her internship, Julie integrated the scripts I developed to explore gene expression, mutations and drug response data from both large-scale pharmacogenomic datasets. The tool allows cross-analysis of heterogeneous data of both datasets using simple descriptive graphs (scatterplot, boxplot, barplot, heatmap-clustering).

2.1 Introduction

One of the most challenging problems in the development of new anticancer drugs is the very high attrition rate. Less than 5% of the drugs entering phase I trials eventually obtain marketing authorization(122). Clinical trials are the only real way to assess drug efficacy and toxicity, but this approach is inadequate for testing the hundreds of drugs currently being developed(181). Scientists need to test hundreds of drugs on numerous tumor models therefore frequently make use of tumor-derived cell lines(16, 63, 209). Such

studies aim to identify genomic biomarkers for predicting the responses of individual patients to the drug and, ultimately, for identifying the best drug for each patient.

In 2012, the first large-scale pharmacogenomics studies provided an unprecedented wealth to the scientific community. The Broad Institute-Cancer Cell Line Encyclopedia (CCLE) provided a collection of 1,036 human cancer cell lines from 36 tumor types, tested for 24 anticancer drugs. The Genomics of Drug Sensitivity in Cancer (GDSC) assessed the sensitivity of 727 cell lines, from 29 tissue types, to 138 drugs. Both datasets contain genome-wide gene expression and sequencing data for a subset of genes. These studies have provided unprecedented amounts of information about molecular profiles and drug sensitivity and have validated several known genetic biomarkers, such as the BRAF-V600E mutation sensitizing melanomas to vemurafenib (34) or ERBB2 amplification/overexpression conferring sensitivity to lapatinib (97)

Previous studies assessed drug sensitivity by pooling all the cell lines or by controlling for tissue source. However, with improvements in our knowledge about tumors, it has become clear that genomic, epigenomic, transcriptional, and proteomic analyses of a given cancer can reveal subtypes differing in pathway activity, progression or treatment response (81, 199). Conversely, the recent success of basket studies(88, 186) have demonstrated that treatment choices can be based on abnormalities shared by tumors originating from different tissue types.

We present here a comprehensive reanalysis of these two recently published large-scale pharmacogenomics resources. We propose an alternative approach in which cell lines are grouped by transcriptomic profile, based on a biological network-driven gene selection process. This molecular classification of cancer cell lines appeared robust across CCLE and GDSC. We further demonstrated the relevance of this novel classification through the drug response We validate our approach by robustly found in CCLE and GDSC as in two external dataset the significant associations between cell line clusters and drug responses.

2.2 Results

2.2.1 A biologically driven approach identifies four robust gene modules

Gene expression profiles were recovered for 471 cell lines, from 24 different tissues, tested in both CCLE and GDSC. Data were curated and annotated with the pipeline of Haibe-Kains *et al.* (72) We developed a three-step biological network-driven process based on transcriptomic data for identifying robust clusters of genes. This process was applied in parallel for each dataset. We first selected the most variant genes from the set of 12,153 genes common to GDSC and CCLE, by the inflexion point method. We then performed hierarchical consensus clustering(121) to identify robust gene modules. Finally, we used String© database software(178) to analyze our gene selection. The goal was to decrease the heterogeneity of each gene cluster. We retained the genes from our initial selection that had (1) high String© database gene connection indices (greater

than 0.7), and (2) similar patterns of expression to other genes within the same biological network (correlation coefficient of at least 0.5)(Figure 2.1 step A). This selection process identified four stable clusters in GDSC (n=183 genes) and five in CCLE (n=210 genes), including a subset of 170 genes common to the two datasets. Distinct functional gene ontologies were associated to each gene modules based on a gene ontology analysis: (Supplementary Figure A.1) Gene Cluster – Extracellular Matrix (GC-ECM; n_ccle=48, n_GDSC=36), Gene Cluster - Migration (GC-Migration; n_ccle=56, n_GDSC=75), Gene Cluster - Immunity-Interferon (GC-Immunity; n_ccle=22, n_GDSC=14) and Gene Cluster - Epithelial Phenotype (GC-Epithelial; n_ccle=63, n_GDSC=58). A set of 21 genes enriched in development processes (GC-Development) was found exclusively in the CCLE dataset.

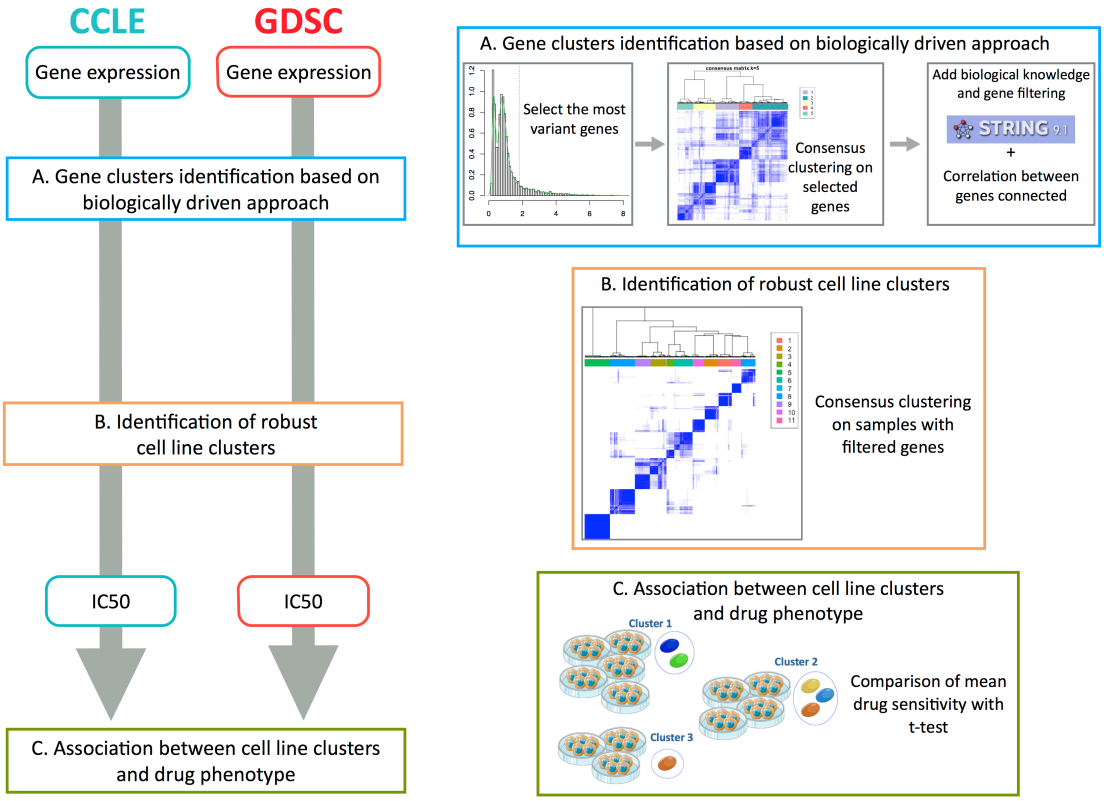


Figure 2.1: Flow chart of the analysis. We apply the same pipeline of analysis independently to CCLE and GDSC. (A) Biologically driven gene selection was performed to build robust clusters of genes. (B) Robust clusters of cell lines were then built using the selected genes. (C) Cell lines clusters have been associated to distinct drug response.

2.2.2 Biologically driven gene selection identifies eleven reproducible cell line clusters

We performed a consensus clustering with the previously selected genes, for each dataset separately, to identify global differences in gene expression between cancer cell lines (Figure 2.1 step B). We obtained eleven stable clusters of cell lines in CCLE and GDSC (Figure 2.2 A and Supplementary Figure 2.G.1).

Previous studies reported strong correlations between the expression profiles of identical cell lines (72). We therefore investigated the closeness of the cell line clusters obtained. We defined the similarity between any two cell lines as the number of datasets in which they clustered together (0=none, 1=CCLE or GDSC, 2=CCLE and GDSC). We assessed the consistency between the clustering patterns obtained with CCLE and GDSC data, using a heatmap clustering of the similarity matrix as a visualization tool. The heatmap shows the number of times that two samples are clustered together across datasets (Figure 2.3A). Groups of cell lines that frequently cluster with each other are shown in darker shades of blue. The heatmap revealed a well defined 11-block, corresponding to the 11 clusters previously identified. A high degree of consistency between the 11 clusters was observed, with 90% accuracy. As the cell line clusters were highly similar, we use the term "cluster" to denote the same group of cell lines from CCLE and GDSC, unless the dataset is specified.

2.2.3 Tissue-of-origin or transcriptomic features dominate cell line clusters

Our eleven clusters can be organized in three major patterns: (i) four clusters of cell lines were derived mostly from tumors from the same tissue of origin. These cell line clusters were named after the organ or cancer subtypes from which most of the cell lines were derived: hematopoietic and lymphoid tissues (HAL), small cell lung cancer (SCLC), skin (SKCM) and breast (BRCA) clusters; (ii) four clusters of cell lines were derived from tissues from the same organ system or had a common embryonic origin: gastrointestinal tract (GI), aerodigestive tract (ADG), glioma and sarcoma (GLSR) and endodermal origin tumors (EDOT) clusters; (iii) three clusters contained cell lines from different tissues of origin. These clusters were named Mixed 1, Mixed 2 and Mixed 3 (Figure 2.3 B and C. Details provided in supplementary data A.21 and A.22).

Clusters of cell lines with common presumptive tissues of origin

Four cell line clusters appeared very homogeneous in terms of tissue lineage: HAL, SCLC, SKCM and BRCA. These lineages accounted for 84%, on average, of the cells of their respective clusters.

The HAL cluster grouped together all the cell lines originating from hematopoietic and lymphoid tissues. This clear clustering pattern can be accounted for by the hematopoietic

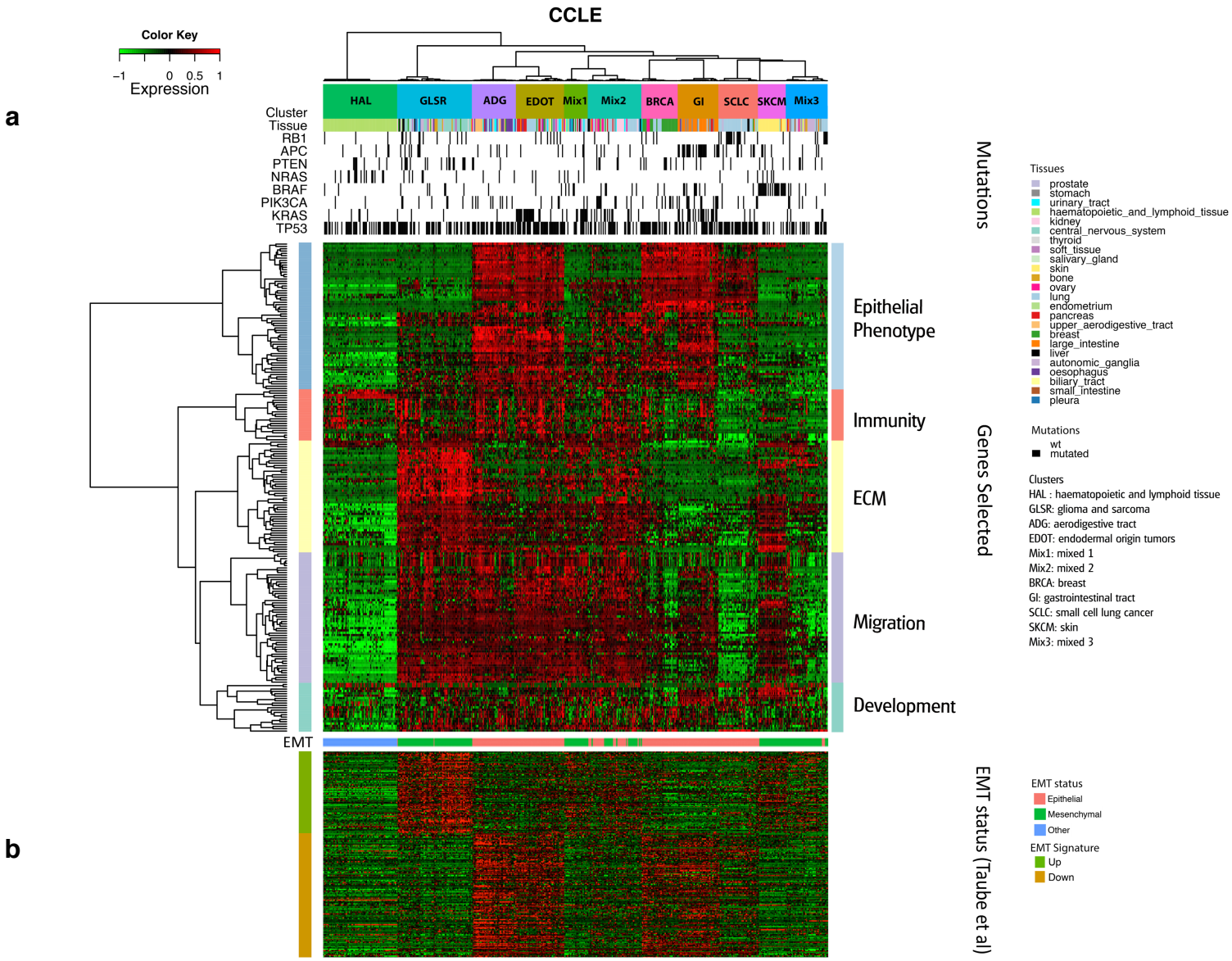


Figure 2.2: Cell line clustering with CCLE data. (A) Heatmap clustering with 471 cell lines (in columns) and 210 selected genes (in rows) for the CCLE data (B) EMT status of the cell lines.

phenotype of this type of tumor. The SKCM cluster was the second most homogeneous cell line cluster in terms of tissue type (92% of the cell lines in this group originated from melanomas). Breast cancer is a heterogeneous disease with a growing number of recognized biological subtypes, including ER+Her2-, Her2+ and triple-negative breast cancer

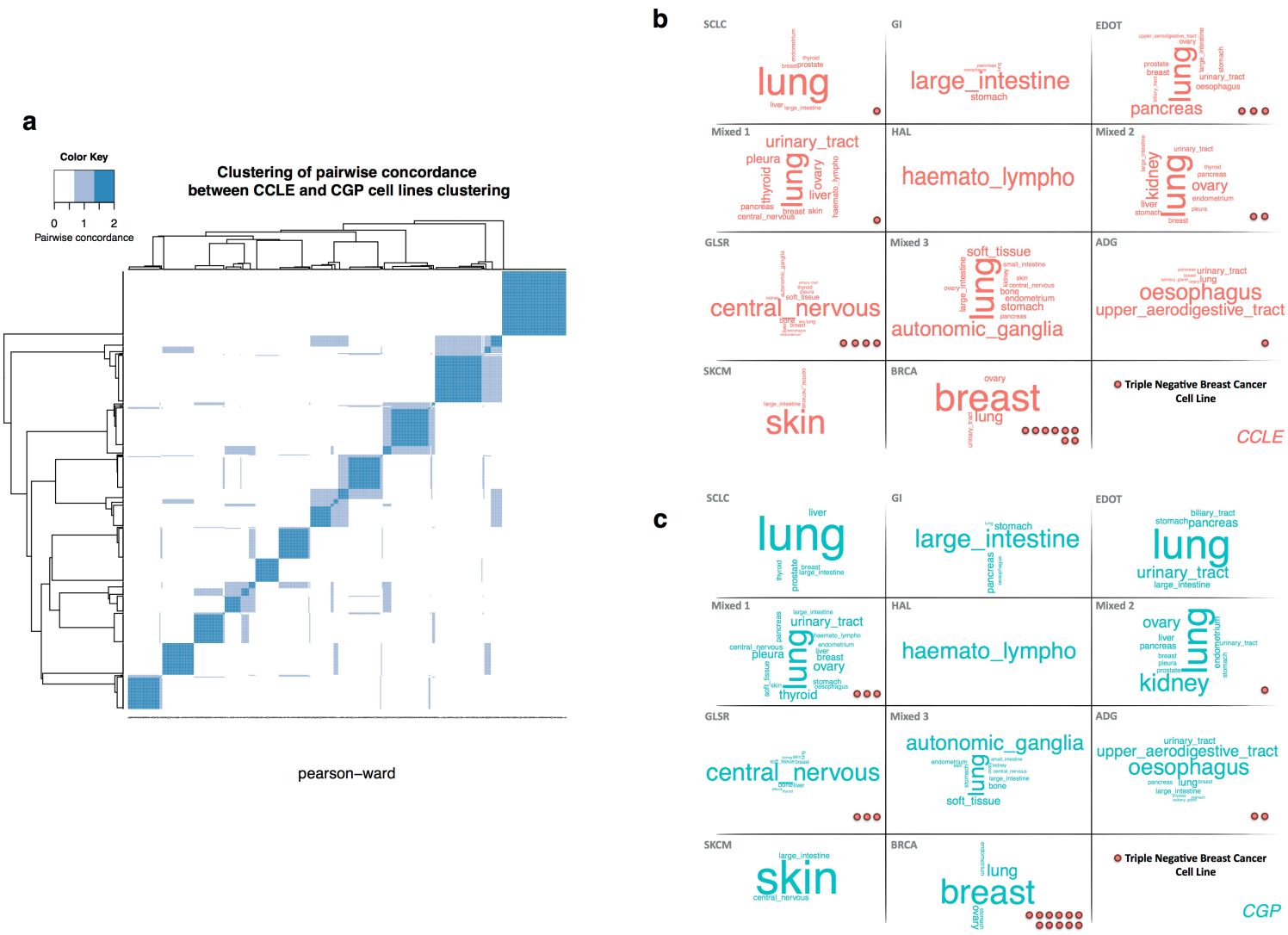


Figure 2.3: Clustering similarity. (A) Color-coded heatmap for similarity between CCLE and GDSC clustering; Tag Cloud represents the tissue composition of cell lines cluster, in CCLE (B) and GDSC (C). The importance of each tissue is indicated by font size. The TNBC cell lines belonging to each cluster are indicated by red dots.

(TNBC), which is the most aggressive subtype. BRCA cluster contained all the breast cancer cell lines defined as ER+Her2- (7/7) and Her2+ (7/7). However, only about half the cell lines defined as triple-negative belonged to this cluster (11/20 in GDSC, 8/20 in

CCLE). The remaining triple-negative breast cancer cell lines were found in six different clusters of cell lines (SCLC, EDOT, Mixed 1, Mixed 2, GLSR and ADG) (Figure 2.3 B and C). SCLC cluster contained 28% of the lung cancer cell lines and 45% of the small-cell lung carcinoma cell lines.
We performed a Gene Set Enrichment Analysis(175) (GSEA) based on our previously defined gene modules to characterized the transcriptomic profile of cell line clusters (Supplementary Figure 2.4). The immunity gene module was strongly expressed in the cell lines of the HAL cluster. Leukemia affects both the bone marrow and lymphocytes, potentially accounting for the detection of immunity gene expression in cell lines derived from a tumor system with no stromal environment. In the SKCM cell line cluster, the epithelial phenotype gene module was downregulated. Furthermore, the activation of the ECM and migration gene modules in this cluster is suggestive of aggressive cancer. In the BRCA and SCLC cell line clusters, the epithelial gene module was expressed, whereas the migration and ECM gene modules were not.

Clusters of cell lines from tissues of the same organ system or common embryonic origin

Some clusters could not be defined on the basis of origin from a single tissue type. However, with a more systemic vision, a consistent organization was obtained for four clusters: GI, ADG, GLSR and EDOT.
Cell lines derived from tumors of the digestive system belonged to two clusters. The ADG cell line cluster consisted mostly of tumors from the esophagus, upper aerodigestive tract, salivary and also urinary glands, whereas the GI cluster grouped together tumors derived from large intestine, stomach and pancreas cancers. About 70% of the cell lines of the GLSR cluster were derived from tumors of the central nervous system, bone, autonomic ganglia and soft tissue. Finally, the EDOT cell line cluster grouped together cell lines derived from tumors of different tissues (e.g. lung, pancreas, urinary tract) arising from the same germ layer (endoderm). The relevance of the EDOT cluster is supported by studies suggesting that oncogenesis may be initiated by the activation of a common pathway in an endodermal progenitor (144).

The ADG, GI and EDOT clusters all displayed strong expression of the genes of the epithelial phenotype module and weak expression of the ECM gene module. According to GSEA, the migration gene module was less strongly expressed in GI cells. For the EDOT cluster, inconsistencies between the CCLE and GDSC datasets were observed concerning the activation or inhibition of migration gene expression at the transcriptomic level only. The GLSR cluster displayed low levels of expression for the epithelial gene module, and high levels of expression for the ECM and migration modules.

Clusters of cell lines from tumors with heterogeneous tissues of origin

Three clusters displayed no particular prevalence of cell lines corresponding to any particular tissue or organ system. They contained cell lines from tumors of 11 to 16 different tissues. We named these clusters Mixed 1, Mixed 2 and Mixed 3. All three of these clusters displayed low levels of epithelial phenotype genes, suggesting that the

cell lines they contained were probably mesenchymal. These clusters also displayed an upregulation of ECM genes. Mixed 1 and 2 displayed an upregulation of migration gene expression. These results suggest that some of the cell lines may have been metastatic in origin or subject to drift, from the characteristics of the tissue of origin to a less differentiated state. In this case, transcriptomic profile is more relevant than tissue of origin.

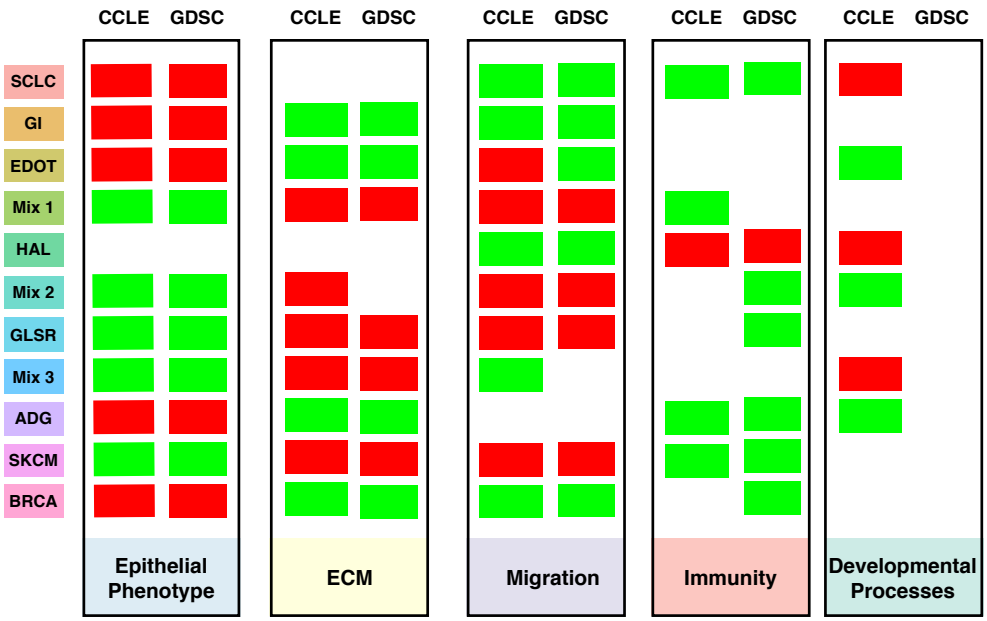


Figure 2.4: GSEA enrichment results. Green indicates gene modules significantly down-regulated, whereas red indicates gene modules significantly upregulated. Results for CCLE are displayed in the left column; results for GDSC are displayed in the right column.

2.2.4 EMT discriminates between cell line clusters

The identification of an epithelial phenotype gene module led us to investigate the epithelial-mesenchymal status of each cell line. A previous study (96) showed that epithelial/mesenchymal transition (EMT)-associated differences in gene expression were a major determinant of the stratification of cancer cell lines based on transcriptomic profiles. Indeed, we found a significant overlap between our gene selections and a published EMT-derived gene signature consisting of 249 genes(180) ($P < 0.0001$, two-tailed Fisher's exact test). We superimposed epithelial/mesenchymal cell line classifications

over our gene expression clusters and found a strong association (Figure 2.2 B and Cb and Supplementary Figure 2.4 b). According to the EMT signature, five cell line clusters (SCLC, GI, EDOT, ADG and BRCA) contained mostly epithelial cell lines, whereas the Mixed 1, Mixed 3, GLSR and SKCM cell line clusters contained mostly mesenchymal cell lines. The Mixed 2 cell line cluster appeared to contain mostly mesenchymal cell lines in GDSC but almost half the cell lines assigned to this cluster in CCLE were epithelial. The HAL cell lines were not concerned by this stratification. Finally, the epithelial/mesenchymal classification was consistent with that obtained with the epithelial phenotype gene module.

2.2.5 Cell line clusters are enriched in somatic mutations

We investigated a common set of 64 genes for the presence of mutations in CCLE and GDSC datasets. However, many inconsistencies between both datasets led us to focus on a set of eight genes (TP53, KRAS, NRAS, APC, PIK3CA, BRAF, PTEN and RB1) for which at least 5% of identical cell lines display mutations in both datasets (Supplementary Information). The mutational profile of cell line clusters was then described based on these genes. Mutation profiles clearly distinguished four clusters (Figure 2.2 A). The SCLC cluster was enriched in RB1 mutations. The GI cluster was rich in APC and KRAS mutations; NRAS mutations were overrepresented in the HAL cluster and the SKCM cluster was enriched in BRAF mutations. Finally, KRAS mutations were particularly abundant in the EDOT clusters. No significant enrichment in mutations was observed for the GLSR, ADG, BRCA and Mixed 3 cell line clusters (Supplementary Tables A.11 and A.12). These clusters have fewer mean mutation rates than the other clusters (GDSC: 13% vs. 19%, t-test p-value = 0.01; CCLE: 17% vs. 22%, t-test p-value = 0.08).

2.2.6 Transcriptomic clustering is more consistent than clustering on the basis of tissue of origin in terms of drug responses

The large-scale drug screening programs of the Broad and Sanger Institutes have provided to the scientific community an unprecedented wealth of publicly available data. Molecular data have been systematically collected for each cell line, but far less information is available for drug screening (Supplementary Information). Moreover, in many cases (25% in CCLE and 45% in GDSC) it was not possible to extract the IC_{50} from the dose-response curve. In order to overcome these issues, both study also report the AUC (area under the dose response curve) that can always be calculated.

We evaluated whether our clustering was more discriminant than the tissue of origin of the cell lines, in terms of drug response. We calculated a pseudo F-statistic separately for IC_{50} and AUC values for each of the 15 drugs common to CCLE and GDSC. This measurement should capture consistency between the clustering and screening data. It is calculated as the ratio of between-group variance in drug response to the corresponding within-group variance(32). High pseudo F values indicate well-separated, compact

clusters. We then compared the pseudo F values calculated with our clustering method with those obtained for 'tissue partitioning' for a given drug (i.e. each tissue being to correspond to a cluster of cell lines).

Twelve of the fifteen drugs had a higher ratio in CCLE and GDSC for our clustering than for clustering based on tissue of origin with the IC50 (Figure 2.5) and ten out of fifteen with the AUC (Supplementary Figure 2.C.1). This trend was confirmed by a t-test comparing the pseudo F values for our clustering with those for 'tissue partitioning' (IC50: CCLE t.test p-value = 0.041, GDSC t.test p-value=0.032, AUC: CCLE t.test p-value = 0.011, GDSC t.test p-value=0.043). PLX4720 (Raf kinase B inhibitor) and PD0325901 (MEK1 and MEK2 inhibitors) were drugs with the largest pseudo F values in both dataset. Paclitaxel was the only molecule in the panel with a higher pseudo F value for tissue partitioning in CCLE and GDSC. As the drug sensitivity results were not used to determine the clustering of the cell lines, these findings provide independent evidence for a major role of mRNA levels in drug sensitivity.

2.2.7 Robust identification of drug response across datasets

Subgroups of patients or cell lines defined on the basis of transcriptomic data have been shown to be associated with differences in drug sensitivity(81, 199). We sought to identify associations between clusters of cell lines and "sensitive" or "resistant" drug phenotypes, for the 15 drugs tested in both CCLE and GDSC. For each dataset and each drug separately, we investigated whether the mean IC_{50} of a given cell line cluster differed significantly from those for the other cell line clusters (see Figure 2.1 step C and Materials and Methods). Six molecules were found to be significantly associated with six different clusters in both CCLE and GDSC (Table 2.1 and Supplementary Table A.13, Supplementary Figure A.2). The SKCM and GI cell line clusters were both significantly more sensitive than the other cell lines to PD0325901 (MEK 1 and MEK 2 inhibitors) (Figure 2.6 A). The association of melanoma and PLX4720 (Raf kinase B inhibitor) is already well established and was confirmed by our analysis. Moreover, an inhibitor of MEK 1 and MEK 2, AZD6244, displayed significantly higher levels of activity in cell lines from the SKCM cell line cluster. Both EGFR inhibitors, erlotinib (Figure 2.6B) and lapatinib, appeared to be significantly more effective against ADG cell lines than against other cell lines. Hematopoietic and lymphoid tissue cells were sensitive to the CDK4/6 inhibitor PD033991. By contrast, SLCL cell lines appeared to be resistant to lapatinib (EGFR and HER2 inhibitor) and the CDK4/6 inhibitor PD033991 was found inefficient to kill GI cell lines. Finally, AZD6244 (inhibitor of MEK1 and MEK2) appeared ineffective to treat BRCA cell. In addition to variation between drug sensitivity and cell lines, previous studies report variations across the different metrics used to report the drug efficacy(57, 72). We then performed similar analysis using AUC. More than half of the associations between cell lines clusters and drug sensitivity were found still significant with AUC (Table 2.1 and Supplementary Information).

We further evaluated the relevance of our clustering regarding the drug sensitivity using two external public datasets. We first study the 118 cell lines tested in common

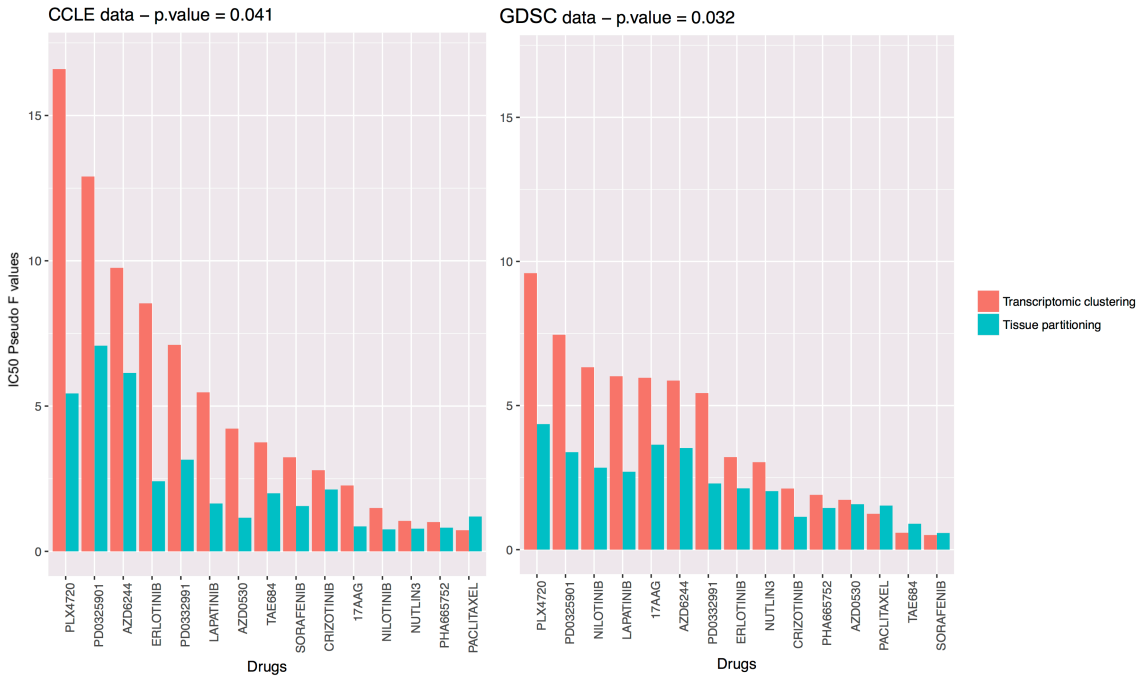


Figure 2.5: Pseudo F value for the 15 drugs common to CCLE and GDSC. The pseudo F index have been computed from the IC50 values for each drug. The pseudo F statistic is the ratio of between-cluster variance to within-cluster variance. Large values of pseudo F indicate well-separated, tight clusters. Drugs are listed in descending order of pseudo F values for clustering.

between the CCLE and the GlaxoSmithKline cell line collection (GSK)(70) on lapatinib and paclitaxel (GDSC was excluded due to small sample size, see Supplementary Information). We found that lapatinib was significantly inactive to kill cells from clusters SKCM and Mixed 1 in both CCLE and GSK (Table 2.1).
Since the set of common cell lines and drugs was small between CCLE, GDSC and GSK (Supplementary Table A.14), we consider the Genentech Cell Line Screening Initiative (gCSI)(78). A panel of 244 unique cell lines and 5 drugs overlap between CCLE, GDSC and gCSI. Instead of AUC, the gCSI reported the mean viability statistic to measure drug efficacy in addition to the IC_{50}. Eight associations between cell lines clusters and drug sensitivity were found significant using the IC_{50} and nine with the mean viability statistic. Among them, the sensitivity of ADG to erlotinib and lapatinib as well as the

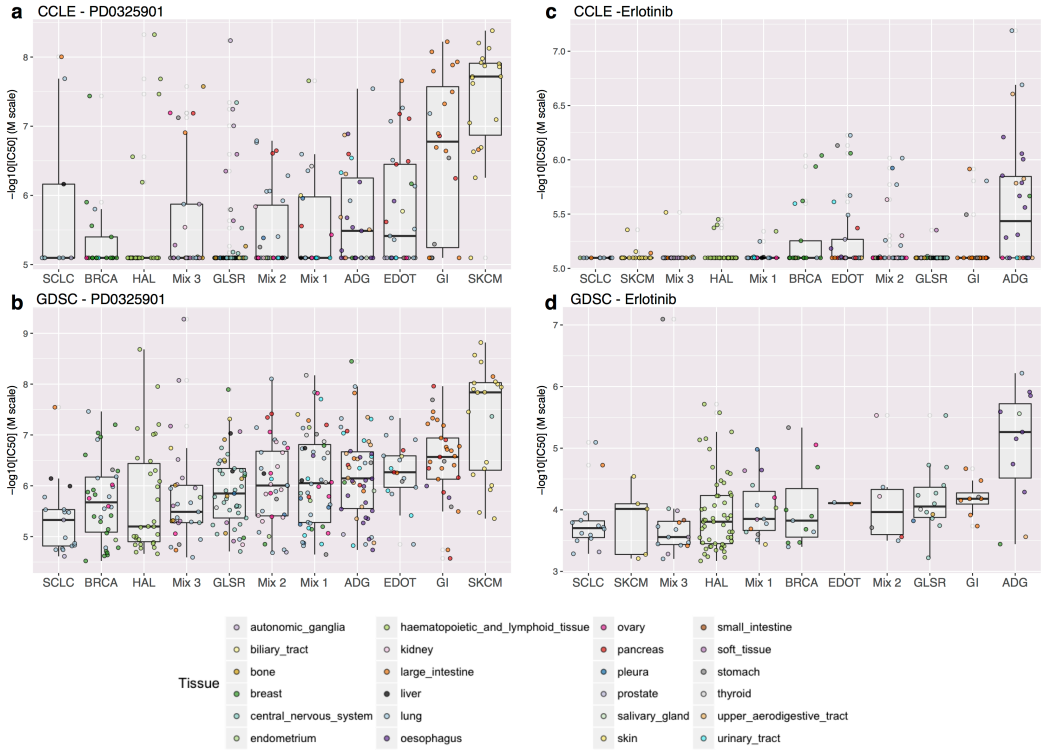


Figure 2.6: Distribution of IC50 values for each in CCLE and GDSC. Ordered according to mean IC50 for the cluster. From resistant (left) to sensitive (right).

efficacy of PD0325901 to kill cells from SKCM cluster were common to CCLE, GDSC and gCSI (Table 2.1 and Supplementary Information).
Our results suggest that our cell line clustering is able to find significant associations with drugs efficacy robustly in four different dataset, despite the large variations across pharmacological data and drug response measures.

2.2.8 Distinct drug profiles were associated with the various cell line clusters

We applied the same procedure to all the drugs tested in the CCLE (24 molecules) and GDSC (129 molecules) studies. For each dataset and each of the 153 drugs separately, we determined whether the mean IC_{50} of a given cell line cluster was significantly different those of the other cell line clusters (Supplementary Tables A.15 and A.16). Overall, the most striking result was the very small number of drugs associated with a sensitive profile (88 associations, including 71 unique drugs) compared to drugs associated with a resistant profile (163 associations, including 92 unique drugs) (Supplementary Information). It was particularly interesting to observed that Mixed 2 and Mixed 3 clusters were

each sensitive to only one drug: respectively midostaurin and vorinostat. Both drugs are targeted agents (PI3K/mTOR inhibitor and HDAC inhibitor). These clusters are made of several cells from different tissue of origin. However, we were able to identify targeted therapies active to kill those cells. These results provide further evidences that our clustering can identify relevant groups of cell sharing unknown features associated to targeted drugs.

Overall, these results suggest that cancer cell lines can be classified, on the basis of their transcriptomic profile, into 11 clusters that may or may not be specific to the tissue of origin. We demonstrated that transcriptomic clustering was more consistent than clustering on the basis of tissue of origin in terms of drug response whatever the drug sensitivity metric considered. We were also able to find several significant associations between clusters of cell lines and "sensitive" or "resistant" drug phenotypes. Many of these associations were robustly found across four different datasets with three different drug response metrics. As the drug sensitivity results were not used to determine the clustering of the cell lines, these findings provide independent evidence about the relevance of this new classification. Furthermore, we show that when we are trying to associate a group of genes from a consistent biological pathway with a group of cell lines, rather than a single gene with a single drug, robust associations can be established across several pharmacologic datasets.

2.3 Discussion

Despite the progress in the development of *in vivo* models, cancer cell lines remain a key tool in cancer research. Patients are usually treated with combination therapy. However, it is important to better understand the mechanisms involved with monotherapies before moving forward to study combination therapies. Here, we introduce a new cell line classification constructed from 471 cell lines derived from tumors from 24 different tissues. A biological network analysis for the most variant genes identified 11 clusters of cell lines. These clusters appeared robust in two large-scale cell line panels. This biologically driven gene selection process, which is probably less sensitive to sample fluctuations than other methods, made it possible to capture strong biological signals that might be concealed by the noise present in microarray data. Several studies have reported that the incorporation of network information improves the stability of gene selection and the biological interpretability of biomarker signatures for a given prediction accuracy (45, 155, 167)

In this new classification, a clear distinction was established between non-epithelial cancer cell lines (GLSR, SKCM, Mixed 3) and epithelial cell lines (EDOT, BRCA, GI). This suggests that EMT-associated differences in gene expression are major determinants of the gene expression-based stratification of cancer cell lines. This new molecular clustering system classified more than 65% of the cell lines differently from the currently used tissue-of-origin cell line classification system. Only four clusters consisted mostly of cell lines originating from a single tissue. Furthermore, three clusters include cells with expression profiles stronger than that of the original tissue (Mixed clusters). Thus, 25%

of the cells lines displayed no link to any tissue of origin or related organ system.

One of the most interesting cases was the triple-negative breast cancer (TNBC). We focused on this subtype, as it is the only subtype of breast cancer without any targeted therapy associated. TNBC were found to be highly heterogeneous, falling into six different clusters. This divergence shows the relevance of studying cell lines from various tumor types. Drug response was dependent on cluster membership, with the EDOT cluster sensitive to chemotherapy, whereas the BRCA cluster was resistant. The widely dispersed TNBC cell lines were mostly mesenchymal, whereas the cell lines of the BRCA cluster were exclusively epithelial. TNBC is increasingly emerging as a heterogeneous disease(79, 100), with tumors differing in histological features, gene expression profiles, clinical behavior, overall prognosis(188) and sensitivity to systemic treatment (41, 104**?**). These findings provide strong evidence to suggest that TNBC heterogeneity is reflected at the cell line level. Our results suggest also that particular attention should be paid to the selection of cell lines for studies of particular types or subtypes of cancer.

By analyzing several large-scale public data sets, we demonstrated that drug efficacy is significantly associated to transcriptomic profile. A comparative analysis recently showed that the gene-expression profiles of the 471 cell lines shared by CCLE and GDSC were highly concordant whereas the reported cell-line drug sensitivities for the 15 drugs tested in both studies were highly inconsistent(72). The authors put forward several hypotheses to explain these discrepancies, including differences in experimental protocols, the viability assay and procedures for summarizing dose response and non-observed IC_{50} (the half maximal inhibitory concentration). Despite discrepancies between the drug sensitivity data retrieved from different databases, we were able to find some robust combinations. Well-known drug associations were found, such as the sensitivity of SKCM lines to vemurafenib. We also found that cancers with BRAF mutations, such as melanoma (202) and cancers with KRAS/BRAF mutations, such as colorectal cancer (213), were more sensitive to MEK inhibitors. Furthermore, CDK4/6 inhibition-induced cell death has been noted in cell lines and xenografts derived from patients with T-cell leukemia (157). SCLC cell lines have been shown to be resistant to lapatinib, but combination with a cytotoxic agent may yield promising results (120).

The decline in the number of new treatments approved in recent years is a major challenge for the pharmaceutical industry. One of the reasons for this decline is the lack of systematic evaluation of therapeutic indications for a drug that is either in advanced development phase or has already obtained a marketing authorization. The so-called "drug repositioning process" proposed to find new therapeutic indications to already approved drugs with faster development times and reduced risks. Furthermore, it allows patients to have access to earlier therapeutic advances(11). Several robust associations were found. Targeted drugs were found efficient to treat clusters of cell lines constituted of cell from different tissues. These drugs are known to be active in one or several tissues that constitute theses clusters. It would be of particular interest to test specifically these drugs on the other tissues represented in these cell lines clusters. For example, cluster ADG is mostly constituted of upper-aerodigestive, oesophagus and urinary tract cancer

cell lines. ADG cluster was particularly sensitivity to the anti EGFR - erlotinib. If EGFR is a validated target for upper-aerodigestive cancer(38, 189, 196) the therapeutic potential of erlotinib has already been highlighted for bladder cancer(143) and showed promising results in phase II for oesophagus cancer(101, 201).

Different types of drugs have been used in the panels. Around 10% of the 153 drugs screened in CCLE and GDSC, and only 1 out of the 15 drugs in common to both studies, are cytotoxic agents. These drugs are expected to be broadly active among the cell line panel since they are not specific molecules. On the contrary, targeted agents are expected to be active only in a subset of cell lines, at least, those carrying the given target. Furthermore, the recent study published by Rees et al(149) demonstrated that target's expression and drug sensitivity were correlated in only 31% of the cases. Grouping cell lines on the basis of their transcriptomic profiles makes it possible to identify subsets of cells with common off-target features. It is then more relevant to compare the drug sensitivity between cell lines of these groups rather than examined the correlation of response of each cell line to a particular drug reported by one dataset with the response of the same cell line to the same drug reported by another dataset. These results suggest that when robust clusters of cell lines based on biologically network-driven approach are considered, consistency between drug responses can be achieved.

In conclusion, our cell line classification provides novel insight for pharmacogenomics studies. As cell lines remain the most widely used models for the preclinical evaluation of candidate cancer drugs, further investigation should be made to use this classification in the development of cancer treatments with the aim of reducing the attrition rate.

2.4 Materials and Methods

Pharmacogenomics data

We collected data from the Broad and Sanger Institutes. The CCLE profiled 24 anti-cancer drugs on 1,036 cell lines. The GDSC screened 138 drugs on 727 cell lines. Both datasets contain genome-wide gene expression and massive parallel sequencing data. All data were recovered, curated and annotated with the pipeline developed by Haibe-Kains *et al.*(72) (the GDSC was referred to the Cancer Genome Project [CGP] in Haibe-Kains et al). We used this pipeline as described in the original article, but with a different method for the normalization of gene expression. Haibe-Kains *et al.* normalized gene expression data by frozen robust multiarray analysis, fRMA(117). This method was designed to combine several datasets and overcome multiple batch issues. This strategy is relevant when trying to ensure assay reproducibility. Even though this approach would be unlikely to have a major effect on gene expression values, we chose to normalize the gene expression data separately with RMA (92), to ensure that the two datasets were perfectly independent. Our analysis focused on 471 cell lines and 15 drugs for which we have transcriptomic and drug sensitivity data available in both the CCLE and GDSC studies.

We collected two large datasets to validate our classification. Data from the Glax-

oSmithKline cell line collection were retrieved from Haibe-Kains *et al.* (72) . The Genentech Cell Line Screening Initiative data were available from compareDrugScreens R package published by Haverty et al.(78)

Gene expression data

Transcriptomic data were restricted to the 12,153 genes common to the two technologies used by GDSC and CCLE (Affymetrix GeneChipHG-U133A and HG-U133PLUS2, respectively). The Jetset method (103) was used to select a unique probe set for given genes. The same probe set was used in both datasets for 83% of the genes.

Drug sensitivity data

The micromolar concentration (μM) at which the drug inhibited 50% of maximal cell growth was used to assess drug sensitivity as well as the area under the dose response curve (AUC). We also consider the mean viability statistic when comparing with gCSI. These measurements were converted to a common scale ($-log10(M)$) for IC_{50}, [0,1] for AUC and 1-mean viability for mean viability), such that high values would be correspond to cell lines sensitive to drugs.

Gene selection by the inflexion point method

We selected the most variant genes, based on the inflexion point of the interquartile range (IQR) distribution for gene expression. This method is more data-driven than a fixed threshold for defining the proportion of genes displaying the highest level of variation. The full procedure is described below. For each gene, we: (1) calculated the IQR for all cell lines, (2) sorted the IQR values of the genes in ascending order, to generate an ordered distribution, (3) estimated the major inflection point of the IQR curve as the point on the curve furthest away from a line drawn between the start and end points of the distribution, and (4) retained genes with an IQR higher than the inflection point.

Gene expression-based identification of cell line clusters

We developed a biological network-driven process based on transcriptomic data, to identify robust clusters of genes and cell lines. This process can be broken down into two parts: A) identification of robust clusters of genes, used for B) identification of robust clusters of cell lines.

A) The gene selection process is a three-step procedure. 1) We selected the most variant genes from among the 12,153 genes common to GDSC and CCLE, by the inflexion point method. 2) We performed hierarchical consensus clustering (ConsensusClusterPlus R Package) to identify robust gene modules. The consensus-clustering step, based on Pearson distance and Ward linkage, identified robust clusters of genes. It involved hierarchical clustering by resampling (1,000 iterations) randomly selected genes. 3) We

identified known biological networks, for each gene cluster separately, using String© database software version 9.1 (http:// string-db.org/). We then applied a two-step selection process: (1) we selected strong biological networks by retaining only genes for which connection scores of at least 0.7 were obtained with String© database software, (2) within each biological network, we selected groups of genes for which expression levels were correlated, with a correlation coefficient of at least 0.5. We used the R package clusterProfiler(214) for comparing and visualizing gene ontologies profiles among gene modules.
B) We applied a consensus-clustering with hierarchical clustering to the cell line gene expression profiles, using the selected genes to visualize the optimal number of stable cell line clusters.

Characterization of cell line clusters at the transcriptomic and mutational levels

Gene set enrichment analysis (GSEA) was performed on genes modules built in step A) of the biological network-driven process described above. We identified up-regulated or down-regulated gene modules, associated with each cell line cluster. An analysis was first performed to identify genes differentially expressed between a particular cluster and all the other cell lines, based on a linear model. For a given cluster k, cell lines were partitioned into two groups $j = \{$*Cluster-k, non-Cluster-k*$\}$. We then performed a differential analysis by comparing the mean gene expression of each group in a linear model (limma R package (150)). The analysis was performed separately for each dataset. The results were used to rank genes in order of significance and to search for overrepresented gene modules, by pre-ranked gene set enrichment analysis (GSEA).

Genes with significantly higher frequencies of mutation in a given cluster were identified by one-tailed Fisher's exact tests. We compared the occurrence of any given mutation in each cell line clusters with that in all the remaining clusters combined.

Identification of cell line clusters common to different studies

We studied the likeness between the clusterings for CCLE and GDSC, by clustering the cell lines with a similarity matrix (hierarchical clustering with Pearson's metric and the Ward agglomerative method). The similarity matrix contains the number of times two cell lines are clustered together in each dataset (0 = never, 1 = only in one classification, 2 = in both classifications). This similarity matrix constitutes a natural visualization tool for assessing the consistency between two clustering patterns. In particular, if we associate a color gradient to the 0–2 range of real numbers, such that white corresponds to 0, and dark blue corresponds to 2, and if we assume that the matrix is arranged so that items belonging to the same cluster are adjacent to each other (with the same item order used to index both the rows and the columns of the matrix), a matrix corresponding to a perfect consensus will be displayed as a color-coded heatmap characterized by blue blocks along the diagonal, on a white background.

The accuracy was calculated as the number of times two cell lines clustered together divided by the number of possible combinations

EMT cell line classification

The "epithelial" or "mesenchymal" status of each cell line was defined with the signature identified by Taube *et al.* (180). This epithelial-to-mesenchymal transition signature consists of 159 downregulated genes and 90 upregulated genes. We performed a hierarchical clustering of cell lines based on these 249 genes and labeled clusters of cell lines according to the overexpression of known epithelial marker genes, known mesenchymal marker genes or neither.

Definition of breast cancer subtypes

Breast cancer subtypes were defined with a bimodal mixture of two Gaussian distributions for ESR1, PGR and ERBB2 gene expression. Triple-negative (TN) breast cancer cell lines were defined by an absence of estrogen and progesterone receptor expression and a lack of ERBB2 overexpression/amplification ($n = 31$). We subsequently defined breast cancer cell lines overexpressing ESR1 but with a lower level of ERBB2 expression as the ER+Her2- subtype (n=7), with cell lines overexpressing the ERBB2 gene defined as the Her2+ subtype (n=7).

Impact of cell line clustering on drug sensitivity

We investigated the relevance of our clustering for drug sensitivity, by comparing the results obtained for this method with those for 'tissue partitioning' (i.e. each tissue of origin being considered to correspond to a cluster of cell lines). We calculated the pseudo F index computed from any drug sensitivity statistic (IC_{50}, AUC, mean viability) for each drug. The pseudo F statistic is the ratio of between-cluster variance to within-cluster variance (32). It is defined as *[Between-cluster variance / (N-K)] / [Within-cluster variance / (K-1)]*, where N is the number of observations (N=471) and K is the number of clusters (K=11 or K=24). Large values of pseudo F indicate well-separated, tight clusters.

The sensitivity and resistant phenotypes of each cell line for a given drug were defined by comparing the drug sensitivity measure between cell lines from any given cluster and the cell lines in all remaining clusters combined. We focus on IC_{50} for clarity. For a given cluster k, cell lines were partitioned into two groups $j = \{$*Cluster-k, non-Cluster-k*$\}$. We then compared the mean IC_{50} values of the two groups in a t test. The sign of the statistical test was used to define the phenotype as sensitive ($t > 0$) or resistant ($t < 0$). We accounted for multiple testing, by calculating the FDR-adjusted p-value for each drug. An FDR-adjusted p-value <0.05 was considered significant.

CCLE vs GDSC			CCLE vs GDSC		
IC50			AUC		
Drug	Cluster	Response	Drug	Cluster	Response
Erlotinib	**ADG**	**Sensitive**	**Erlotinib**	**ADG**	**Sensitive**
AZD6244	**SKCM**	**Sensitive**	**AZD6244**	**SKCM**	**Sensitive**
AZD6244	**BRCA**	**Resistant**	**AZD6244**	**BRCA**	**Resistant**
Lapatinib	SCLC	Resistant	Lapatinib	HAL	Resistant
Lapatinib	ADG	Sensitive	Crizotinib	SKCM	Resistant
PD0332991	GI	Resistant	AZD0530	SKCM	Resistant
PD0332991	HAL	Sensitive	PLX4720	SKCM	Sensitive
PLX4720	SKCM	Sensitive	**PD0325901**	**SKCM**	**Sensitive**
PD0325901	GI	Sensitive			
PD0325901	**SKCM**	**Sensitive**			
CCLE vs gCSI			GDSC vs gCSI		
IC50			IC50		
Drug	Cluster	Response	Drug	Cluster	Response
Erlotinib	**ADG**	**Sensitive**	**PD0325901**	**SKCM**	**Sensitive**
Erlotinib	**Mixed 1**	**Resistant**			
Erlotinib	**GLSR**	**Resistant**			
Erlotinib	**SKCM**	**Resistant**			
Lapatinib	**Mixed 1**	**Resistant**			
Lapatinib	ADG	Sensitive			
PD0325901	BRCA	Resistant			
PD0325901	**SKCM**	**Sensitive**			
CCLE vs gCSI			GDSC vs gCSI		
Mean Viability			Mean Viability		
Drug	Cluster	Response	Drug	Cluster	Response
Erlotinib	**ADG**	**Sensitive**	**PD0325901**	**SKCM**	**Sensitive***
Erlotinib	**Mixed 1**	**Resistant**			
Erlotinib	**GLSR**	**Resistant**			
Erlotinib	**SKCM**	**Resistant**			
Erlotinib	HAL	Resistant			
Erlotinib	SCLC	Resistant			
PD0325901	BRCA	Resistant			
PD0325901	**SKCM**	**Sensitive**			
CCLE vs GSK					
IC50					
Drug	Cluster	Response			
Lapatinib	**Mixed 1**	**Resistant**			
Lapatinib	SKCM	Resistant			

Table 2.1: Significant associations found between CCLE, GDSC, GSK and GCSI. In bold associations found significant in at least three datasets. The association between PD0325901 and SKCM had an adjusted p-values of 0.058 (marked with *).

Supplementary Information

2.A Discrepancies in mutational data between CCLE and GDSC

We investigated a set of 64 genes for the presence of mutations in both the CCLE and GDSC datasets. We investigated the consistency of mutation patterns. Thirteen genes were reported to have no mutations in GDSC, in any of the 471 cell lines studied (AKT2, CCND1, CCND2, CCND3, CDK4, CDK6, EP300, FGFR2, MDM2, MLLT3, MYCL1, MYCN and SMARCB1). We focused on the remaining 51 genes. The most frequently mutated gene was TP53, with a frequency over 60%, mutation frequencies remaining below 20% for the other genes (Supplementary Figure A.3A). A previous study reported high levels of agreement concerning the reported presence of mutations in identical cell lines (72). However, for cell lines displaying mutations of at least one of these genes, the mean proportion of cell lines with mutations of the same genes was about 20% (Supplementary Figure A.3C) and the mean proportion of cell lines displaying identical mutations within the genes was 44% (Supplementary Figure A.3D). Only for eight genes (TP53, KRAS, NRAS, APC, PIK3CA, BRAF, PTEN and RB1) did at least 5% of identical cell lines display mutations in both datasets (Supplementary Figure A.3B). Furthermore, the proportion of identical cell lines mutated was 64%, and the mutations observed were identical in 84% of cases. By contrast, 12% of identical cell lines had mutations for the remaining genes, with 36% of the mutations observed identical (Supplementary Figure A.3 C-D).

2.B Drug screening data

The large-scale drug screening programs of the Broad and Sanger Institutes have provided to the scientific community an unprecedented wealth of publicly available data. Molecular data have been systematically collected for each cell line, but far less informa-

tion is available for drug screening. Considering the 471 cell lines common to CCLE and GDSC, only 47% of the cell lines from CCLE and 90% of those from GDSC have been tested for at least one drug (Supplementary Figure 2.B.1). In CCLE and GDSC, 28% and 22% of the cell lines, respectively, were tested for all compounds. Moreover, in many cases (25% in CCLE and 45% in GDSC) it was not possible to extract the IC50 from the dose-response curve. In these cases, the IC50 was set to the maximum concentration used for screening in the CCLE study, whereas a mathematical extrapolation was applied in the GDSC study. In order to overcome these issues, both study also report the AUC (area under the dose response curve) that can always be calculated.

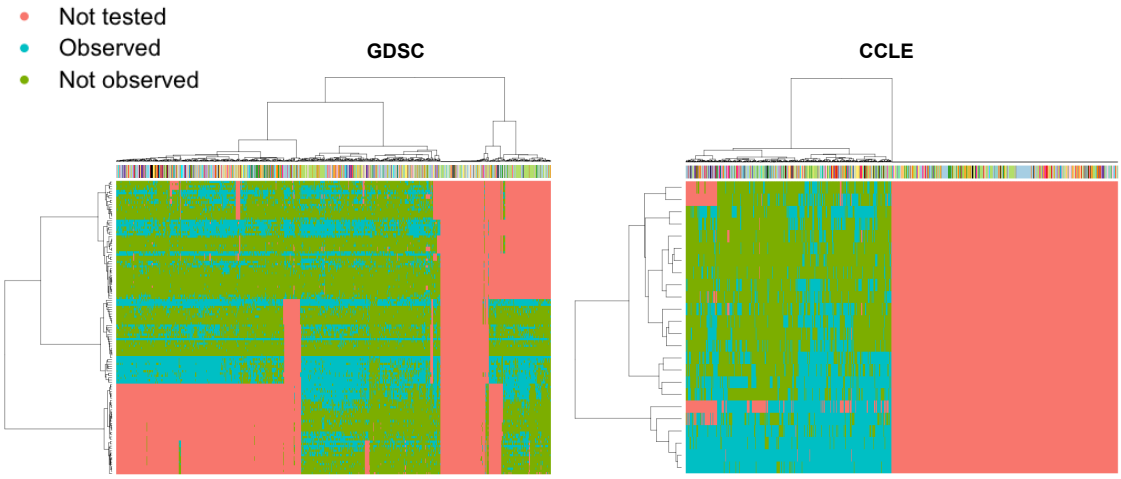


Figure 2.B.1: Number of drugs tested for each cell line. Many drugs were not tested for a large set of cell lines (in red). Even when a test was performed, in many cases IC50 could not be extracted and have been estimated as explained in the main text (green). Values in blue are observed IC50 values.

2.C Comparison between CCLE and GDSC based on AUC

To evaluate the relevance of our method to the variation of drug sensitivity metric, we looked at the area under the dose response curve (AUC). We evaluate the robustness of our cell line classification with the 15 drugs shared between CCLE and GDSC.

We first evaluated whether our clustering was more discriminant than the tissue of origin of the cell lines, in terms of drug response. We calculated a pseudo F-statistic using the AUC values for each of the 15 drugs common to CCLE and GDSC. Ten out of

fifteen drugs had a higher ratio in CCLE and GDSC for our clustering than for clustering based on tissue source (Supplementary Figure 2.C.1). This trend was confirmed by a t-test comparing the pseudo F values for our clustering with those for 'tissue partitioning' (CCLE t.test p-value = 0.011, GDSC t.test p-value=0.043).

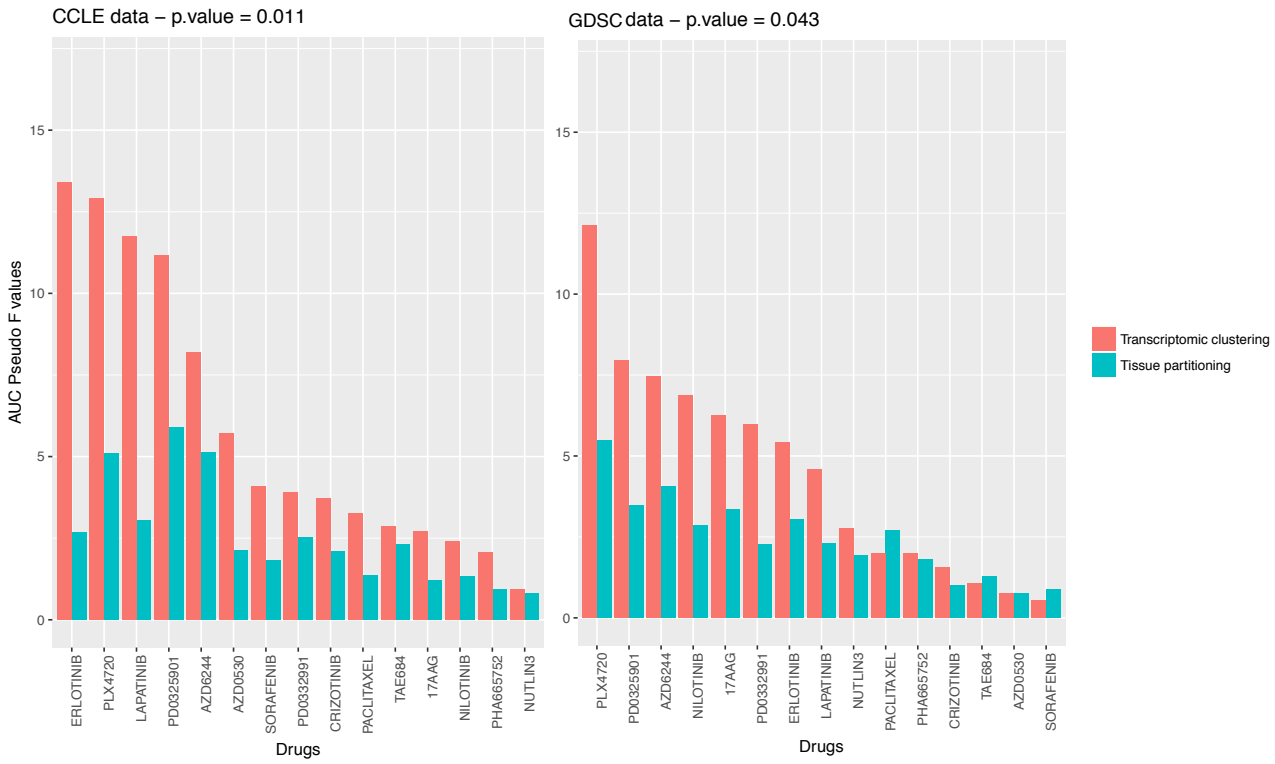


Figure 2.C.1: Pseudo F value for the 15 drugs common to CCLE and GDSC. The pseudo F index have been computed from the AUC values for each drug. The pseudo F statistic is the ratio of between-cluster variance to within-cluster variance. Large values of pseudo F indicate well-separated, tight clusters. Drugs are listed in descending order of pseudo F values for clustering

We then performed the same analyses as those performed with the IC50, to identify association between cell line clusters and drug response but based on the AUC. Four out of the seven cell lines clusters-drugs associations were robustly found using the IC50 and the AUC (Supplementary Table A.17 and ure A.18): ADG cluster being sensitive to erlotinib, SKCM being sensitive to the Raf kinase B inhibitor PLX4720 and the MEK inhibitor AZD6244 whereas the latter is not active in BRCA cell line cluster. In addition, SKCM cluster appears resistant to crizotinib (ALK inhibitor) and AZD0530 (Src

and Abl inhibitor) that is concordant with the wild-type mutational status of SCKM cells for ALK, Abl1/2 and Src (16) . Lapatinib appears inefficient in HAL cluster that is consistent as no EGFR or ERBB2 mutation have been reported in HAL cell lines in both CCLE and GDSC.

Many cell lines never achieved 50% inhibition for many drugs. In this case, data have been either truncated by the maximum dose tested (as in CCLE), either they have been extrapolated (as in GDSC) (Supplementary Figure A.2). On the contrary the AUC can always be calculated. These variabilities can explain the variation of results. However more than half of the cell lines clusters-drug associations were found using both IC50 and AUC.

2.D Comparison between CCLE, GDSC and GSK

We explored the relevance of our classification with two outside datasets. For this, we kept our robust clustering of cell lines define with the CCLE and GDSC. For each dataset and each drug separately, we investigated whether the mean drug sensitivity measure available of a given cell line cluster differed significantly from those for the other cell line clusters (see Materials and Methods). We then investigate if significant cell line clusters-drug associations can be found in these different datasets. First we consider the third dataset used by Haibe-Kains et al1, the GlaxoSmithKline cell line collection (GSK) (70). This panel is composed of 319 cell lines that have been screened on 19 drugs. Only lapatinib and paclitaxel have been tested by GSK among the 15 drugs shared between CCLE and GDSC. A set of 194 cell lines was common to the three studies. However the actual number of cell lines for which we have a sensitivity measure for lapatinib and paclitaxel is smaller especially in GDSC. Supplementary Table A.14 shows the number of cell lines with a sensitivity measure for each drug in each dataset among the 194 cell lines shared between CCLE, GDSC and GSK. Only 70 cell lines have a sensitivity measure for paclitaxel and lapatinib in GDSC. Moreover, the number of cluster with less than 3 cell lines goes up to 4 for paclitaxel and 5 for lapatinib. Given the too small number of cell lines available, we therefore compared only the results between CCLE and GSK. Comparisons were based on IC50, the unique measure available in GSK. Two associations were found in both CCLE and GSK suggesting that lapatinib is inactive in SKCM and in Mixed 1 clusters (Supplementary Table A.18, Supplementary Figure A.5 and 10). Prickett et al4 found some evidence that lapatinib may be more active in melanoma cell lines with ErbB4 mutations than wild-type melanoma cells. However, no

ErbB4 mutations have been called in melanoma cells from cluster SKCM (16).

2.E Comparison between CCLE, GDSC and gCSI

Since the set of common cell lines and drugs was small between CCLE, GDSC and GSK (Supplementary Table A.14), we repeated the analysis with the panel introduced by the Genentech Cell Line Screening Initiative (gCSI) (78). Data were recovered from the R package compareDrugScreens. A panel of 244 unique cell lines and 5 drugs overlap between CCLE, GDSC and gCSI. Once again the overlap of cell lines between GDSC and gCSI is relatively small (Supplementary Table A.14). Drug sensitivity measure was available for less than 3 cell lines in 4 clusters for erlotinib and lapatinib; in 3 cell line clusters for crizotinib and paclitaxel. For these reasons, we focused only on PD0325901 when comparing GDSC and gCSI. Two measures of sensitivity are available in compareDrugScreens. In addition to the classical IC50, they introduced the mean viability statistic that is the arithmetic average of the fitted viabilities at each tested dose. This metric is closely related to the AUC. We used the mean viability to evaluate the robustness of our cell line classification to the drug sensitivity metric variation.
Height associations were found significant in both CCLE and gCSI using the IC50 (Supplementary Table A.19, Supplementary Figure A.7,A.8). Among them three were also found when comparing CCLE to GDSC: ADG cluster sensitive to erlotinib and lapatinib; PD0325901 actives in SKCM. In addition, the resistance of Mixed 1 cluster to lapatinib found when comparing CCLE to GSK appeared once again significant here. Additional associations were found specifically. Erlotinib were inactive to treat cells from Mixed 1, GLSR and SKCM. BCRA cluster was found resistant to PD0325901. Regarding the comparison with GDSC, a strong trend could be observed between SKCM and PD0325901 (gCSI Effect = 0.39, FDR-adjusted p-value =3.05E-10, GDSC Effect = 0.13, FDR-adjusted p-value =0.058).
Considering the mean viability statistic (Supplementary Table A.20, Supplementary Figure A.9 and A.10), 2 associations among those identified when comparing CCLE to GDSC were found again: ADG cluster sensitive to erlotinib and SKCM sensitive to PD0325901. In addition, 8 more clusters were found resistant to erlotinib: SCLC, Mixed 1, HAL, GLSR and SKCM. These findings were supported by several phase II studies reporting the inactivity of erlotinib in patients with gliomas (145, 190) and melanomas8 as well as the lack of EGFR mutation in small-cell lung cancers (164). BRCA appeared resistant to the MEK inhibitor PD0325901. It has been shown that PD0325901 is more active in basal-like breast cancer lines than in luminal and Her2+ lines (83). Indeed,

only 3 out of the 10 breast cancer cell lines from BRCA cluster that have been tested for PD0325901 are TNBC. For GDSC, similar results than those found with the IC50 were found regarding the sensitivity of SKCM to PD0325901.

2.F Distinct drug profiles were associated with the various cell line clusters

The GI cell line cluster was sensitive to drugs targeting the ERK signaling pathway (MEK1/2, Hsp90): 17AAG, AZD6244, PD0325901 (Supplementary Figure 2.F.1). The targeting of this pathway was also found to be effective against the cell lines of the ADG cluster, but only for drugs targeting EGFR and ERBB2. Furthermore, these cell lines seems to be particularly sensitive to drugs acting on the cytoskeleton. SKCM cell lines were sensitive to drugs inhibiting the ERK signaling pathway, regardless of the protein from this pathway targeted (BRAF, MEK1/2, FLT3, JAK2, NTRK1, RET). Hematopoietic cells were sensitive to molecules targeting apoptosis, the cell cycle and the cytoskeleton, and to inhibitors of PARP, topoisomerase I and the ERK signaling pathway. AKT-INHIBITOR-VIII and MK-2206, two AKT1/1 inhibitors, were found efficient on BRCA cells. Indeed, the PI3K/mTOR pathway is commonly deregulated in breast cancer (186). The 15 drugs to which cell lines of the GLSR were sensitive included seven PIK3/mTor inhibitors. Kinase inhibitors and chemotherapy agents targeting the mitotic spindle were also identified as potentially effective drugs against GLSR cells. Chemotherapy agents (gemcitabine, bleomycin, vinblastine) were the most active drugs for killing cells from the Mixed 1 cell line cluster. The other two heterogeneous clusters were each sensitive to only one drug: midostaurin-a PI3K/mTOR inhibitor for Mixed 2, and vorinostat, a HDAC inhibitor, for Mixed 3. Finally, the cell lines of the EDOT cluster were sensitive only to the chemotherapy agent doxorubicin.

Resistance profiles were identified for 163 associations between cell line clusters and drugs. More than 30% of these associations involved inhibitors of the ERK and PI3K/mTOR signaling pathways. All but one of the clusters appeared to be resistant to at least three drugs targeting this pathway. The exception was the Mixed 2 cluster, which was resistant only to the MEK1/2 inhibitor AZD6244). The cell lines of the various clusters were otherwise resistant to a broad range of diverse drugs. The cell lines of the SCLC cluster were resistant to the largest number of molecules (22% of all the drugs tested), including, drugs targeting the apoptosis, NOTCH or Wnt signaling pathways. Drugs targeting the cytoskeleton, mitosis and replication appeared to be the

least effective for killing the cell lines of the GI cluster. By contrast, the cell lines of the ADG cluster were found to be resistant to a drug targeting the cytoskeleton, a drug targeting the Wnt pathway and a drug targeting G-protein–coupled receptors. Topoisomerase I inhibitors, such as topotecan and camptothecin, were found ineffective to kill BRCA cells. Surprisingly, gemcitabine and methotrexate were also found to be inefficient against these cells, despite their widespread use in clinical practice. Mixed clusters presented only a small number of resistant associations.

Figure 2.F.1: Drugs displaying significantly high (A) or low (B) levels of activity in a given pair of cell lines (i.e. in CCLE and GDSC) relative to the remaining cell lines. The 138 drugs from CCLE and GDSC were used. Colors correspond to the drug target family. (C) Distribution, by target family, of the 138 drugs. (D) The drugs significant for each cell line cluster. Drugs and clusters are ordered by number of associations.

2.G Supplementary Figures

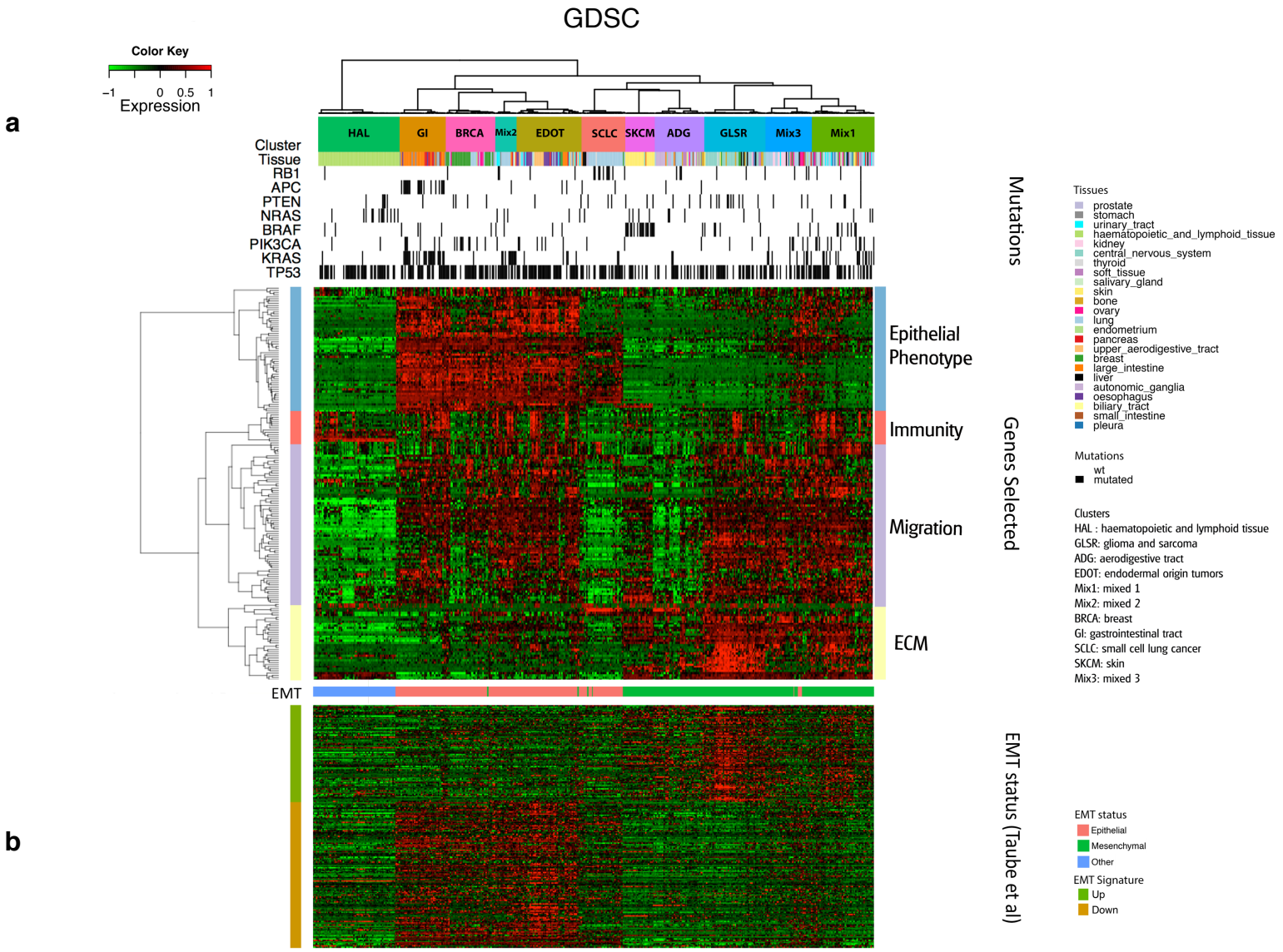


Figure 2.G.1: Cell line clustering with GDSC data. (A) Heatmap clustering with 471 cell lines (in columns) and selected genes (in rows), with GDSC data. (B) EMT status of the cell lines.

3

The genomic and transcriptomic landscape of neoadjuvant-resistant triple-negative breast cancer

Our team generated high-throughput sequencing data to study the resistance to neoadjuvant therapy of triple-negative breast cancers. This project was realized in collaboration with Cécile Laurent and Judith Abecassis. In this study, I supported all RNAseq data analysis and WES results. Cécile Laurent and Judith Abécassis contributed to generating the results based on WES data and Cécile helped to analyze them. A brief review of the literature comparing the tools used to perform alternative splicing analyzes was conducted by Julie Setbon during her internship.

3.1 Introduction

Triple-negative breast cancers (TNBC) are defined by the absence of estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor-2 (HER2) expression. They account for about 15 % of invasive breast cancers and are responsible for a disproportionate number of breast cancer deaths and breast cancer cases among young women (60). TNBCs also have a higher incidence of recurrence and disease progression, with a peak risk of recurrence in the first three years after treatment (130). They represent an important clinical challenge, as no major improvement in treatment

has occurred recently. To date, no single targeted therapy has been approved for the treatment of TNBC, and cytotoxic chemotherapy remains the standard-of-care.
The neoadjuvant setting (i.e chemotherapy before surgery) represents an opportunity to study and monitor *in vivo* treatment-sensitivity of the tumor. A pathological complete response (pCR) obtained after neoadjuvant chemotherapy (NAC) is a strong indicator of the benefit of chemotherapy since patients have a better prognosis than those with residual disease (RD). Despite their relative chemosensitivity, 30%–40% of TNBC patients treated with routine NAC achieved pCR (60). Patients with RD following neoadjuvant chemotherapy have a worse prognosis and overall survival (104, 177).
Previous studies have attempted to identify chemosensitivity markers by comparing patients who either achieved pCR or had extensive residual disease after NAC. No individual genes were identified. Nonetheless, effective pathways were potentially targetable. Indeed, Jiang et al. (93) showed that TNBC carrying mutations in AR and FOXA1 pathways seems to be more sensitive to chemotherapy. Mutational load has also been widely studied and a greater number of mutations per tumor and a higher level of immune activation may be associated with chemosensitivity. Molecular profiling of residual TNBC after NAC demonstrated aggressive pathway activation. For example, resistant TNBCs present more frequent co-amplification of MYC and MCL1 suggesting a potential role of MEK inhibitors or alterations potentially targeted by PI3K/AKT inhibitors or CDK 4/6 inhibitors (13).
However, there is a lack of comprehensive study including matched pre- and post-treatment samples to identify markers of resistance in TNBC after NAC. In this study, we examined pre-therapeutic core biopsy samples and corresponding residual diseases on a total of 15 patients at the DNA and RNA levels. The main goal of this study was to identify genomic alteration and perturbed pathways induced by NAC.

3.2 Clinical characteristics

A total of 15 patients were included in the cohort (Table 3.1). The median age was 47 years (range:35-75 years), 86.7% (n=13) were premenopausal and 53.3% were overweight or obese (BMI > 25). Clinically, patients present large tumor size (mean size = 50mm). Most patients were classified as having stage T1-2 (66.7%, n=10), node-positive breast cancer (73.3%, n = 11) and grade 3 tumors (93.3%, n =14). All patients received neoadjuvant treatment, 66.6% (n=10) were treated by sequential epirubicin/cyclophosphamide and 26.6 % (n=4) were treated by sequential 5-fluorouracil, epirubicin and cyclophosphamide (patients 27,29,32,50). Anthracycline based regimen was followed by sequential

docetaxel for all patients except patient 7. Surgery was performed 4–6 weeks after the end of chemotherapy. At time of surgery 66.7% of the tumors were RCB III (residual cancer burden). The average mitotic index and cellularity of the tumor were similar before and after NAC. Pretreatment core needle biopsies and post-NAC surgical specimens were available for all patients. Lymph nodes were recovered at time of surgery and only 6 patients had enough material to be sequenced.

Variables		Overall
		15
Age	<40 y	2 (13.3%)
	40-55 y	11 (73.3%)
	>55 y	2 (13.3%)
BMI	<19	2 (13.3%)
	19-25	5 (33.3%)
	>25	8 (53.3%)
Menopausal status	post	2 (13.3%)
	pre	13 (86.7%)
BRCA status*	mutated	2 (12.5%)
	non mutated	14 (87.5%)
Clinical tumor size		**50 [20 - 140]**
Clinical T stage	T1-T2	10 (66.7%)
	T3	5 (33.3%)
Clinical N stage	No	5 (33.3%)
	N1-N2-N3	**10 (66.7%)**
Elston-Ellis grade	I-II	1 (6.7%)
	III	**14 (93.3%)**
Mammary surgery	lumpectomy	7 (46.7%)
	Mastectomy	8 (53.3%)
Axillary surgery	sentinel node biopsy	1 (6.7%)
	axillary dissection	14 (93.3%)
Histological tumor size (median [range])		23 [10 - 60]
Number of positive lymph nodes	0	5 (33.3%)
	1-3	**5 (33.3%)**
	>=4	**5 (33.3%)**
Residual Cancer Burden	II	5 (33.3%)
	III	**10 (66.7%)**
Lymphovascular invasion after NAC	no	9 (64.3%)
	yes	5 (35.7%)

*According to Whole Exome Sequencing data for BRCA1 (n=1) and BRCA2 (n=1)
Qualitative variables: number (percentage)
Quantitative variables : median [range]

Variables		Primary Tumor	Residual Disease
		15	15
Mitotic index		**54 [4 - 100]**	**53 [3 - 250]**
Mitotic index cat	=< 10	1 (6.7)	1 (6.7)
	11-22	3 (20.0)	1 (6.7)
	>22	11 (73.3)	13 (86.7)
Tumor cellularity		**70 [30 - 90]**	**70 [10 - 95]**
DCIS, n (%)	yes	3 (20.0)	10 (66.7)
	no	12 (80.0)	5 (33.3)
IT TILs		**5 [0 - 40]**	**5 [0 - 30]**
Str TILs		**20 [5- 70]**	**5 [5- 60]**

Qualitative variables: number (percentage)
Quantitative variables : median [range]

Figure 3.1: Patients and tumors characteristics in the neoadjuvant cohort.

3.3 Transcriptomic analysis of neoadjuvant-resistant triple-negative breast cancer

3.3.1 Gene expression profiles before and after NAC

We performed unsupervised two-way hierarchical clustering to characterize our tumor samples based on common transcriptomic profile. The clustering was based on the 2,426 most variant genes selected by the inflection point method (see material and methods).

Remarkably, all samples for each patient were clustered together with no prominent changes in overall gene expression. Only the primary tumor (PT) of patient 1 was separated from its residual disease (Figure 3.2). These results suggest that the main source of variation between samples was not associated with NAC. Patients clustering showed four groups (clusters I-IV) characterized by the high expression of distinct gene modules. Gene ontologies enrichment analysis suggests that cluster I was driven by extra-cellular matrix genes. Patients from cluster II were characterized by a high expression of ECM and T lymphocytes related genes (Immune-module). Cluster III has shown a high expression level for genes involved in the nervous system processes and metabolism. Finally, the genes expressed primarily by cluster IV were enriched in metabolism. These groups of patients express almost exclusively a given group of genes, suggesting that heterogeneous groups of TNBC may have been selected.

Our lab has recently published a six-metagenes signature of 167 genes to classify the TNBCs (see chapter 5.1 on page 119). We computed the six metagenes in each sample and compared their expression before and after NAC for each patient. We defined a score as the sum of the absolute differences in the expression levels of each metagene between the PT and RD samples (Figure 3.3 A & B) to quantify the differences between both conditions. The signature highlights patients for whom NAC induced no or very small changes in their transcriptome and patients with high differences before and after treatment. Three patients presented changes in expression for the androgen receptor (AR) metagene (patients 27, 35, 45). Changes in the expression levels of Matrix/Invasion metagene were observed in six patients (patients 1, 7, 9, 36, 40, 43) and six patients showed modifications in Immunity metagenes levels (patients 7, 9, 29, 40, 43, 50). We noted that these changes were not unidirectional (not all PTs appeared down-regulated compared to RDs for a given metagene and reciprocally).
Several changes have been identified in the genes involved in immunity processes. We investigated lymphocyte infiltration based on histologic microbiopsy specimens where the presence of stromal tumor-infiltrating lymphocytes (TILs) was evaluated. Patients 1, 9, 29, 35, 37 and 43 has lowed percentage of stromal TILs after NAC than before, while patients 7, 32, 36, 45, 46 were more infiltrated after treatment. However, the percentage of stromal TILS was low for 14 patients ($<$ 60 (50)) (Supplementary Figure 3.A.2 on page 94). We further investigated tumor infiltration by immune cell populations of our samples using the MCP counter algorithm (18). Individually, patients displayed distinct immune profiles across conditions (Figure 3.3 C and Supplementary Figure 3.A.1 on page 93). Patients 7, 9, 40, 43 and 50 exhibited strong changes in the abundance of the

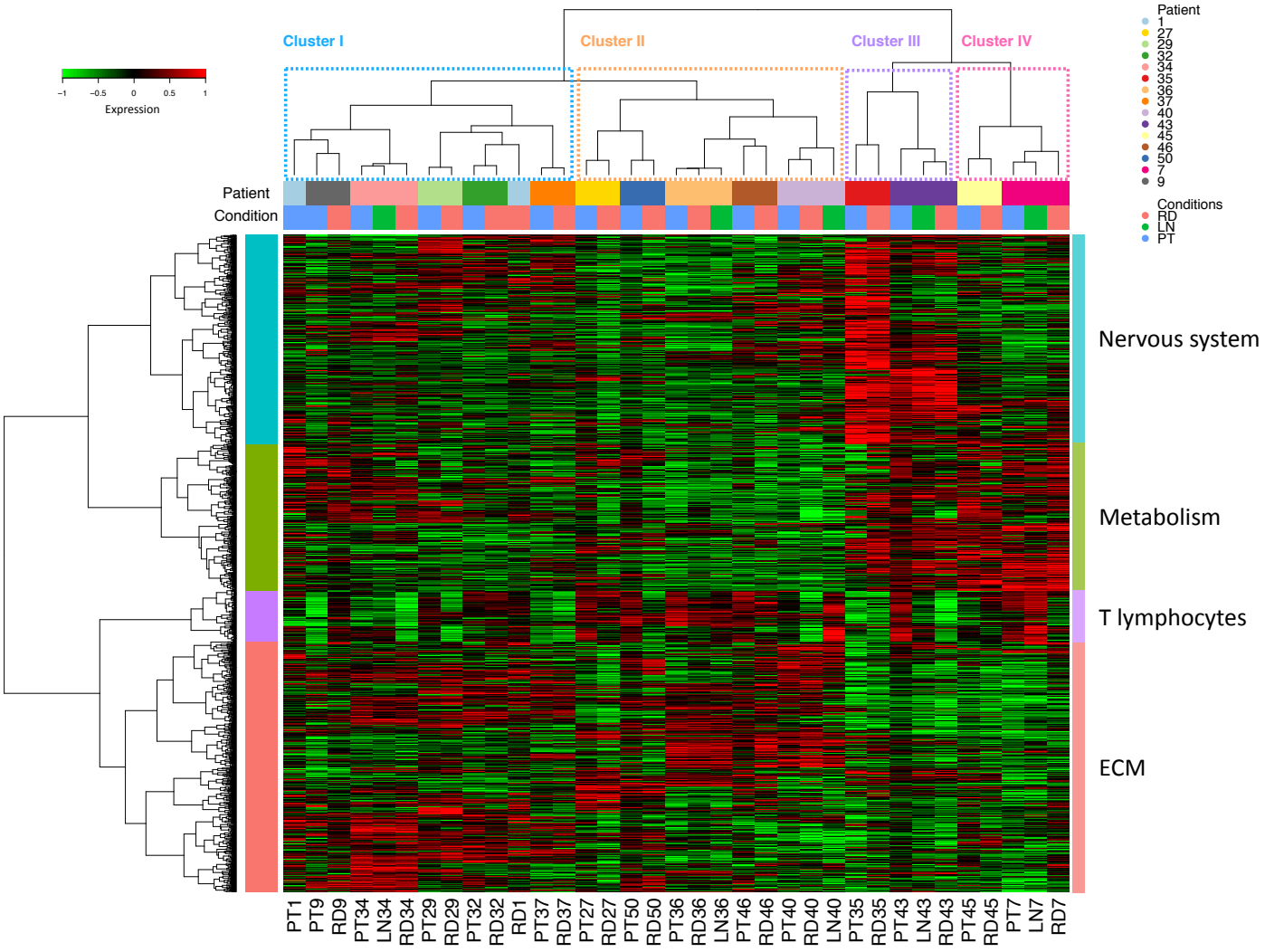
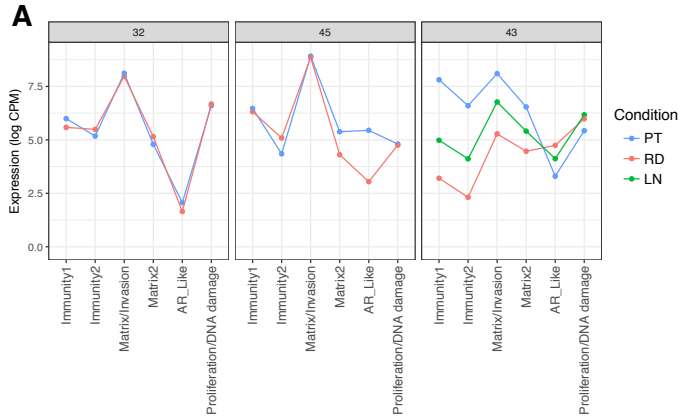


Figure 3.2: Gene expression clustering with RNAseq data based on the 2,426 most variant genes. Data have been rlog transformed.

immune cell population whereas patients 1, 32, 35, 36, 37, 45 showed almost no change. Based on all samples, PTs have a significant lower expression of B cells compared to RDs and lymph nodes (LN) samples (Figure 3.3 D, p-values<0.005).

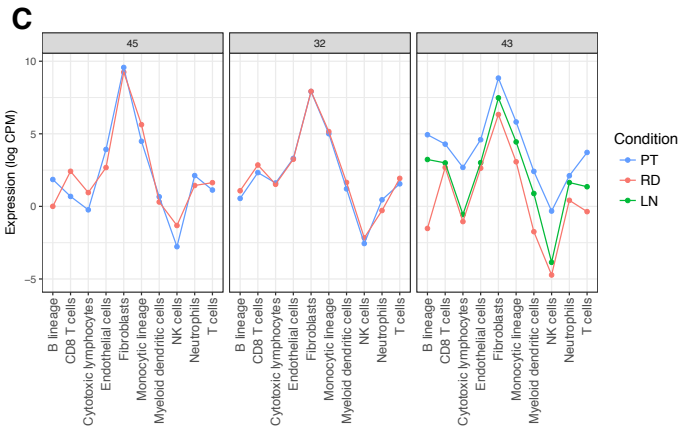
3.3.2 Differential gene expression related to NAC

We assessed the transcriptomic modifications that occurred during NAC using differential gene expression analysis. A small set of 58 genes was found significantly differentially



B

Patient	Bonsang et al	MCP Counter
1	6.09	6.38
7	5.98	**17.12**
9	**10.10**	**21.95**
27	5.73	13.52
29	6.77	14.59
32	1.72	3.31
34	2.77	12.35
35	**7.02**	9.86
36	6.11	7.20
37	2.64	6.99
40	6.12	**15.88**
43	**15.77**	**33.34**
45	4.48	10.50
46	3.91	10.97
50	**7.06**	**20.11**



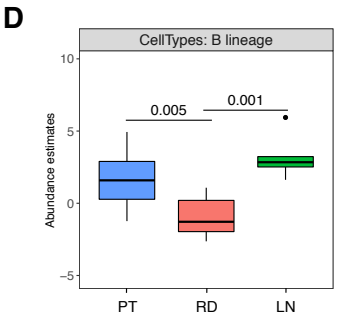


Figure 3.3: Dynamic changes in transcriptomic profiles between conditions. A) Dynamic changes according to TNBC metagenes of Bonsang et al. 2015. Examples of patient without modification, and with changes in AR-like and Immune metagenes. B) A Score that estimates the changes between primary tumors (PT) and residual diseases (RD) according to metagenes from Bonsang et al. 2015 and Becht et al.2016. Patients with the highest scores were highlighted in bold. C) Dynamic changes according to MCP counter metagenes from Becht et al. 2016. D) Boxplot of the B cell lineage in each condition. LN= Lymph nodes.

expressed between PTs and RDs (FDR-adjusted p-value $< 5\%$, Supplementary Figure 3.A.3 on page 95). Among them some known oncogenes and tumor suppressor genes were found (DUSP1, FOS, EGR1, NR4A1, PAX5). Similarly, the comparison between LN and PT identified 46 differentially expressed genes (DEG) Supplementary Figure 3.A.4 on page 96) and no genes were found differentially expressed with a 5% FDR-adjusted p-value between LNs and RDs. These results suggest that gene expression profiles are closer between post-NAC samples (lymph nodes and residual diseases samples) compared

to primary tumor samples. No genes were found in common to the three comparisons. The DEGs of the PTvsRD and PTvsLN comparisons were enriched in processes inducing response to drug, organophosphorus compounds and purine-containing compounds, consistent with exposure of the samples to epirubicine and cyclophosphamide. In addition, genes involved in the activity of RNA polymerase II have been overexpressed in post-NAC samples, suggesting a high transcriptional activity in response to drugs. Several genes involved in proliferation and cancer progression have been identified (FOS, DUSP1, PTGS2, EGR1, CYR61, C7).

3.3.3 Roles of alternative splicing regulation to NAC-resistance

Alternative splicing (AS) is a major source of protein diversity in humans. Differences in the relative expression of isoforms allow the cells to adapt to their environment. Therefore, a gene may have the same level of expression but produce different isoforms and different proteins. The small number of genes with significant variations after exposure to NAC was surprising and led us to assess AS events. The exon inclusion level of each detected event was computed and assessed for significance using the rMATS algorithm (162). We identified 295 skipping exons (SE), 149 mutually exclusive exons (MXE), 36 alternative 5' splice site (A5SS), 41 alternative 3' splice site (A3SS) and 11 retention of intron (RI) differentially regulated between PT and RD samples. In total, 360 AS events were identified in 425 unique genes with the majority (>68%) occurring in 16 or more samples. Some of these genes were known to be associated with tumorogenesis (AGRN, AKT1, ASPSCR1, BCAR4, CDKN1A, CHL1, EIF3E, IKZF1, LAPTM4B, MDM4, PTK2B, RUNX1, SBSN, STIL, STRA6) or TNBCs and breast cancer (ABCC11, ADA, AURKB, CARS, CAST, CD14, CPMEIF3E, NCOR2, PGC, PTS, RAD52, SMARCA4, TOP2A, TPO). The 314 multi-isoform genes were enriched in protein serine/threonine kinase inhibitor activity, histone demethylase, integral and intrinsic component of mitochondrial outer membrane and many pathways involved in NOTCH1 signaling.
Interestingly, a perfect separation between PTs and RDs can be observed based on the exon inclusion levels of the 181 significant splicing events that occurred in all samples (Figure 3.4 A). The correlation between the exon inclusion level and gene expression was low for 98% of the genes (Pearson correlation < 0.4), indicating that expression level was probably not the cause of the altered splicing patterns observed. Except the tumor suppressor PAX5, no splicing events occur in DEGs.

The number of significant differential splicing events (SDSE) between PTs and RDs

with LNs was very large. In total, 784 SDSE in 665 genes were found between PTs and LNs and 1146 SDSE in 922 genes were identified between RDs and LNs. While cassette exons represented more than half of the splicing events between pre- and post-NAC samples (51% PTvsRD and 55% PTvsLN), it represented 44% of all significant events between RDs and LNs (Figure 3.4 B). The correlation between the exon inclusion level and gene expression was low for most of the genes (78%) and only NR4A1 was differentially spliced and expressed. Significant multi-isoform genes between PTs and LNs were enriched in transmembrane transporter, PI3K/AKT/FGFR cascade and its downstream signaling activity (PI-3K cascade: FGFR1/FGFR2/FGFR3/FGFR4, PI3K events in ERBB2 signaling, PI3K/AKT activation). Several studies have shown that PI3K/ AKT pathway protects against anthracycline-induced apoptosis and was associated with resistance to microtubule-targeting drugs (66, 108, 182). The gene differentially spliced between RD and LN display a variety of gene ontologies that were heavily enriched in cilium morphogenesis, assembly and organization, component of Golgi membrane and integrin complex/cell adhesion.

Previous studies report that predominant splicing events detected between cancerous and normal breast samples are SE and RI (55). The majority of the significant AS events detected in our study were SE that accounted for more than half of our detected events but RI accounted for only 3% of the spliced events detected in each comparison. The second most predominant splicing event in our sample was MXE (PTvsRD:33%, PTvsLN:24%, RDvsLN:38%). Only 30 mutli-isoform genes were identified in all three comparisons. This set represented around 1% of all identified mutli-isoform genes. The proportion of each kind of AS event was very similar in each comparison, suggesting that the difference in splicing between each condition is due to the selection of the target genes for splicing rather than the general predominance of a particular general splicing mechanism (Figure3.4 B). Two members of the superfamily of ATP-binding cassette (ABC) transporters, ABCC11 and ABCC12 were commonly spliced between all conditions. Two more members of this family were specifically identified when we compared PT against LN (ABCB4) and RD against LN (ABCC3). The ABC transporters family is known to be important mediators of chemoresistance. Park et al (134) have studied the role of the ABC transporters in patients that received neoadjuvant therapy (FEC plus paclitaxel). The authors have demonstrated that patients with residual disease overexpressed genes of the ABC transporter family compared to patients with pathological complete response.

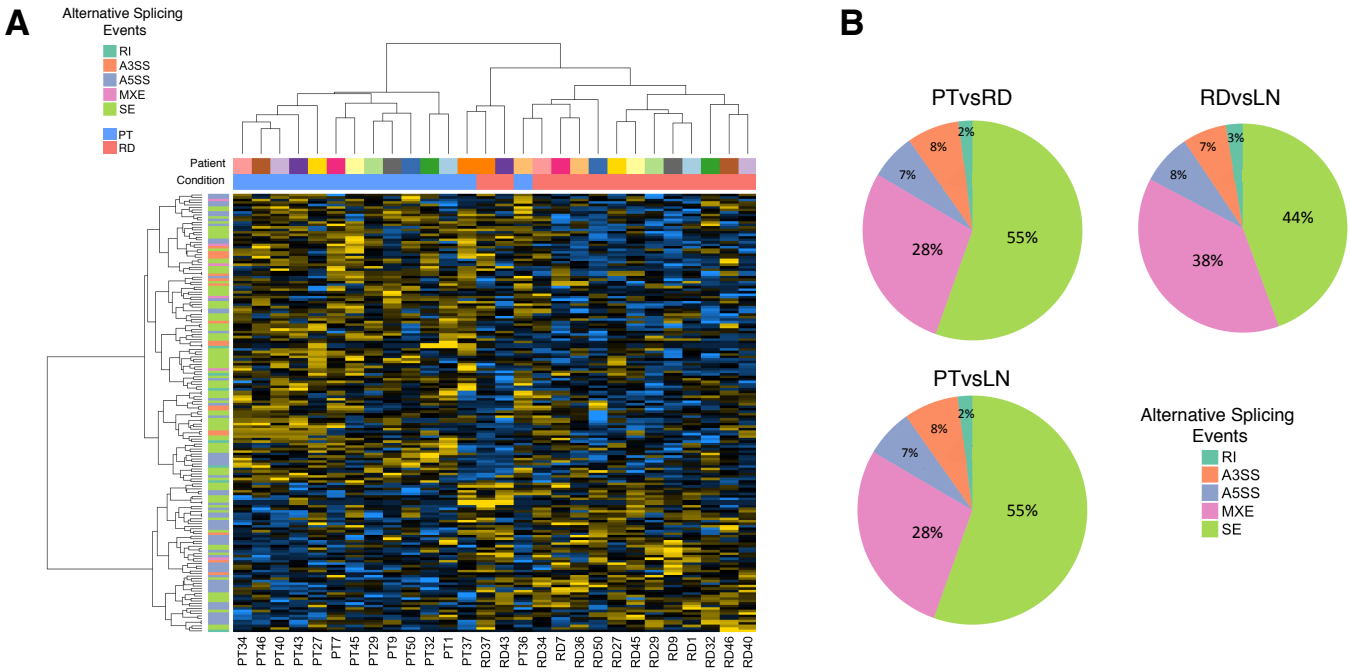


Figure 3.4: Alternative splicing profile of neoadjuvant-resistant triple-negative breast cancers. A) Clustering based on the significant spliced events detected between primary tumors and residual diseases. B) Distribution of differentially spliced events between primary tumors and residual diseases (PTvsRD), residual diseases and lymph nodes (RDvsLN), primary tumors and lymph nodes (PTvsLN). RI=Intron Retention, A3SS=Alternative 3' splice site, A5SS=Alternative 5' splice site, MXE=Mutually exclusive exon, SE = Skipping Exon.

3.3.4 The regulation of AS during NAC is induced by numerous RNA binding proteins

The regulation of AS involves specific splicing factors that can regulate splicing positively or negatively (35). Given the very small set of genes regulated by AS in common under all conditions, we hypothesized that different splicing factors were active in each condition. The identification of the splicing factors responsible for these changes may provide potential targets for reversing chemo-resistance. RNA binding proteins (RBPs) are splicing factors that influence AS patterns by binding to pre-mRNAs in the exons or introns flanking alternative exons. We performed RBP-binding motifs enrichment analysis near alternative exons with an improved version of rMAPS (133), that reliably

correct for multiple testing and allow to precisely identified the binding position of the RBP (see material and method). This method is detailed in chapter 4.

We analyzed a collection of 112 well-characterized RBPs from the literature (9, 51, 148). Motif enrichment analysis was performed separately for each comparison. A total of 74 unique RBPs were found to be significantly enriched around all regulated exons (PTvsRD=47, PTvsLN=27, RDvsLN=25) (Supplementary Figure B.2, B.3, B.4 on page 245). Many of these splicing factors have been previously associated with aberrant splicing in large cohort of breast cancers (166, 211). A set of 6 RBPs were found enriched in all comparisons (QKI, RBM28, RBM8, SF2/ASF, SRSF7 and YBX1). The genes that encode these RBPs were among the top 10% most expressed genes in the cohort (only RBM28 was slightly less expressed). Distinct RBPs were reported in each comparison. For example, members of the RBFOX classes of splicing factors were involved in the regulation of spliced exonbetween PTs and RDs. Former studies report strong an association of these splicing factors with the epithelial-mesenchymal transition in breast cancer (59, 160) leading to a more aggressive and metastatic disease.

Given the number of genes regulated by splicing events with respect to the number of genes with differential expression and the number of relevant RBPs implicated, alternative splicing appears to be a prevalent mechanism induced by the neoadjuvant chemotherapy.

3.4 Copy number profiling of neoadjuvant-resistant triple-negative breast cancer

3.4.1 Heterogeneity of copy number profiles

We analyzed the copy number profiles of 23 samples for which normal tissues were available. These samples correspond to 10 patients with primary tumors and residual diseases samples; 3 of these patients also had lymph node samples. The copy number profiles were determined with whole exome sequencing data using the R package FACETS, which also provided purity and ploidy estimates (161). We observed a total of 1,635 breakpoints and a median of 70 breakpoints per sample. The median ploidy was equal to 2.74, the patient 34 was tetraploid and 3 patients were triploid (Supplementary Figure B.5 on page 248). Post-NAC samples had more rearrangements than primary tumors (average number of breakpoints PTs = 62, RDs =80.8, LNs=69, significant paired Wilcoxon test between PTs and RDs; p-value = 0.002). Altered segment size could be long in

individual samples (average size ∼33Mb). However, when we compared among multiple patients, the boundaries of copy number variations (CNV) were tightened making the average altered CNV segment of ∼ 2Mb. 362 segments were defined as amplifications (over 5 copies) and 68 as deletions (0 copy) with average segment size of ∼1.7 Mb and 1.4 Mb respectively. However, we observed more losses (8145 with 1 copy and an average segment size of ∼3Mb) than gains (4633 with an average segment size of ∼2Mb). On average, an individual sample had 15.7 amplified segments, 201.4 segments gained, 2.9 deleted segments and 354.1 segments lost. Figure 3.5 displays a frequency plot summarizing the distribution of DNA copy number aberrations by condition. Frequent DNA copy number alterations were found in all conditions. We observed recurring gains (over 30%) at 1q, 8q and 20q and recurrent losses (over 50%) at 3p, 5q, almost all chromosome 6, 8p, almost all chromosome 9, 10q, end of 11q, 12q, 15q, 17p and 21q . These results are consistent with previous large scale studies (183, 210).

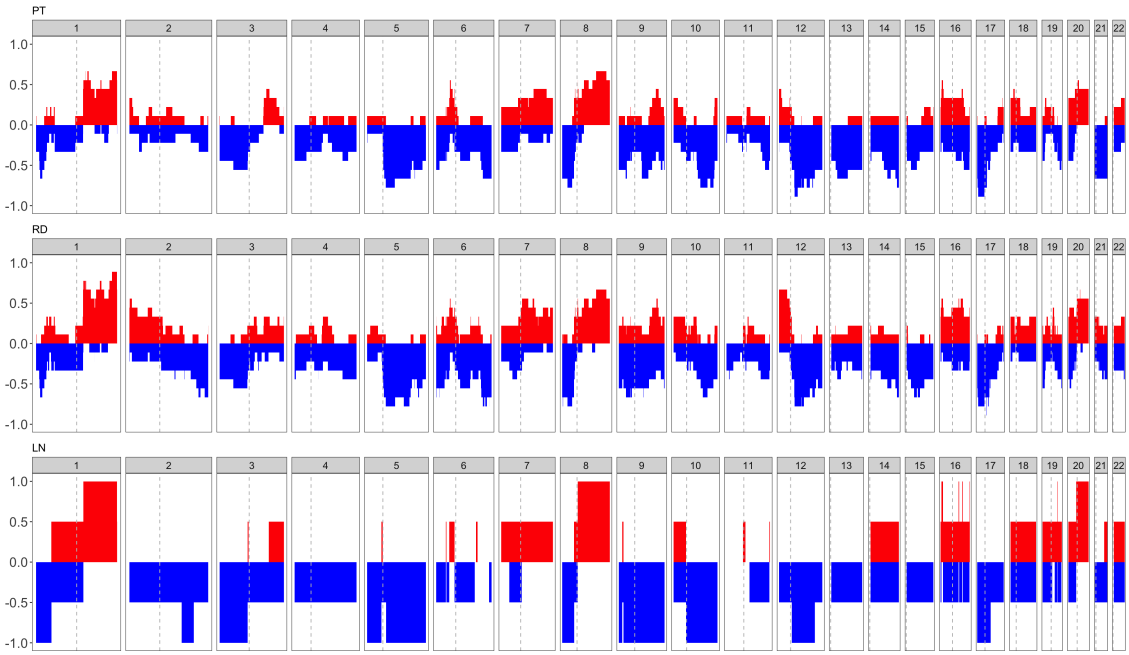


Figure 3.5: Frequency plot of copy number aberrations. Copy number changes of in 10 patients with primary tumors and residual diseases samples, including 3 lymph nodes. The gains and amplifications are represented in red, the losses and deletions in blue.

Several focal amplifications (segment smaller than 10 Mb and over 5 copies) were found in 12 samples, mostly in samples from the same four patients (Figure 3.6 A).

Focal amplifications occurred in 898 genes including 47 known oncogenes. Among them, we can cite CDK6, CXCL1, CXCL2, FOXA1, MDM2, MLLT11, MYC, PPM1D, TBX2 and WT1, which have already been linked to breast cancer. However, no recurrent focal amplifications were identified between patients. Homozygous deletions were observed in 8 samples from 6 patients (Figure 3.6 B). The most frequent homozygous deletions relate to the cyclin-dependent kinase inhibitor genes : CDKN2A, CDKN2A-AS1, CDKN2B (3 patients, 30%). These genes are known tumor suppressor genes and regulators of the cell cycle. They are frequently mutated or deleted in a wide variety of tumors and associated with a high proliferative state. Homozygous deletions were detected for 471 genes including 81 tumor suppressor genes. We can note that residual disease of patient 56 and lymph nodes from patient 34, shared homozygous deletions of 31 tumor suppressor genes located at chromosome 8p.

Multiple amplifications and deletions have been reported in TNBCs at different frequencies (46, 183). MYC plays a role in multiple hallmarks of cancer including those involved in cell proliferation, metabolism, cell death, and cell survival and has been reported as amplified in TNBCs (85). MYC was amplified in 43% of our samples (Figure 3.6 C). Gains of PIK3CA and EGFR were detected in 39% of the samples but were predominantly found in samples from the same patients. Finally, PTEN was lost in more than 50% of the patients, which is more frequent than the 25–30% of cases previously reported (183). Some genes, like KIT or FGFR2 have been previously reported as frequently gained in TNBC, they were predominantly lost in our cohort.

3.4.2 Matched residual/primary tumor CNV analysis

The variation of the number of copies of the residual disease with respect to the corresponding primary tumor showed patients with very different profiles (Figure 3.7). We quantify the similarity between the genomic profiles using the similarity measure introduced by Bollet et al (26). This measure corresponds to the number of common breakpoints, divided by the mean number of breakpoints in either a primary tumor or a residual disease. The mean similarity between PTs and RDs was of 0.46 suggesting that many modifications occurred between the two time points. Patients 34 and 35 demonstrated the highest similarity score (over than 0.8) and patients 29 and 32 the lowest (less than 0.25). While similarity between pre- and post-NAC samples was relatively low, clustering based on this measure groups all samples from the same patients together (Figure 3.6 D).

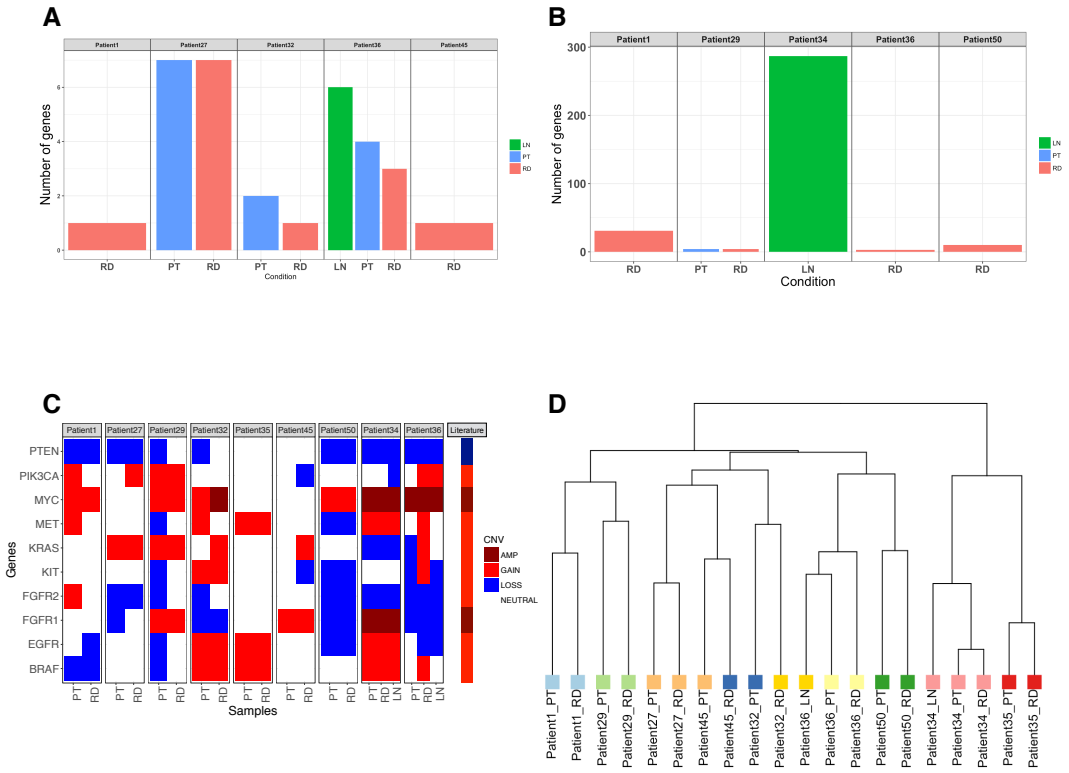


Figure 3.6: A) Number of focal amplifications per sample. B) Number of homozygous deletions per sample. C) Copy number profile of genes previously reported altered in TNBCs. MYC and FGFR1 were reported as amplified, PIK3CA, KRAS, BRAF, EGFR, FGFR2, IGFR1, KIT and MET as gained and PTEN was reported as deleted. D) Sample clustering based on the similarity measure introduced by Bollet et al 2008.

No recurrent segment amplified or deleted were found in common to all patients. 123 amplifications were observed in PTs, 131 in RDs and 108 in LNs. However, these amplifications were mostly found in the same six patients before and after NAC (patients 27, 32, 34, 36, 45, 56). These observations suggest that some tumors are more prone to amplifications than others. A small number of deletions were found only in the primary tumor of patient 29. We detected deletions in 6 residual diseases and 2 lymph nodes specimens. Patient 56 had a huge number of deletions compared to the others RDs (n=56 vs 4.4 in average for the remaining samples). Deletion appeared here as a mechanism specific to post-NAC samples.

Patient 35 displayed an almost full genomic profile with one copy in the primary

tumor and residual disease. Only chromosomes 7, 19p and 20 had two heterozygous copies. The two-copy state LOH was detected at 1q. Further, chromosome 15p was almost amplified (copy number = 5). Patient 35 appeared clearly as a very rare case of tumor. While her clinical profile was consistent with the average cohort characteristics , her gene expression profile appeared as outlier when performing a principal components analysis based on gene expression profiles (data not shown). We were also surprised by the very high tumor purity estimated by FACETS for both samples (0.96 PT, 0.86 RD). As an alternative, allele-specific copy number profile of samples from patient 35 were retrieved with R package sequenza (58). Similar copy number, ploidy and purity were found (data not shown). Whether these findings are real or due to artifacts in the analysis pipeline remain to be investigated.

Amplification of ERBB2 was detected in all three samples from patient 56 (ERBB2 copy number: PT=17, RD=16, LN=15). However, this patient was defined ER-/PR-/ERBB2- by immunohistochemistry (IHC) and FISH before treatment. These results were verified by a novel IHC analysis confirming that ERBB2 was not amplified in the primary tumor (IHC score = 0) but shows an equivocal amplification in the residual disease (IHC score = 2+). We are currently waiting for results using FISH analysis. If these results are confirmed they could have guided the adjuvant treatment of this patient.

3.4.3 CNVs reflect sub-clonal selection under NAC

The evolution of the copy number profile between the two time points can help to estimate the clonal selection of the tumor during chemotherapy. Indeed, even though tumors are prone to genome instability, some events had almost no probability of occurring in the same cell and can reasonably be considered coming as derived from sub-clonal selection. An example of this type of "impossible events" is a segment that is deleted in the primary tumor but is gained or amplified in the residual disease. We then count for each patient the number of impossible events between PTs and RDs (see material and methods). A total of 197 impossible events occurred in 9 patients with an average size of 1.4 Mb (min= 3 bases, max= 13Mb). They mainly occurred on chromosomes 1, 6, 11, 16 and were absent from chromosome 13 and 18. No such events were observed in patient 35, only one event was observed in patient 34 and 4 in patient 27. The average number of impossible events in the remaining patients was 27.4. We observed the greatest number of this kind of alteration in patients 32, 29 and 45 (n=40, 38 and 31 respectively). 95% of the impossible events were unique. The remaining segments concerned do not appear in

more than 2 patients. We may hypothesize that early sub-clonal selection events would allow a longer divergence time between PT and RD and would result in a large number of CNV differences. These differences in tumor progression timelines may be critical to better understand the mechanisms of chemoresistance.

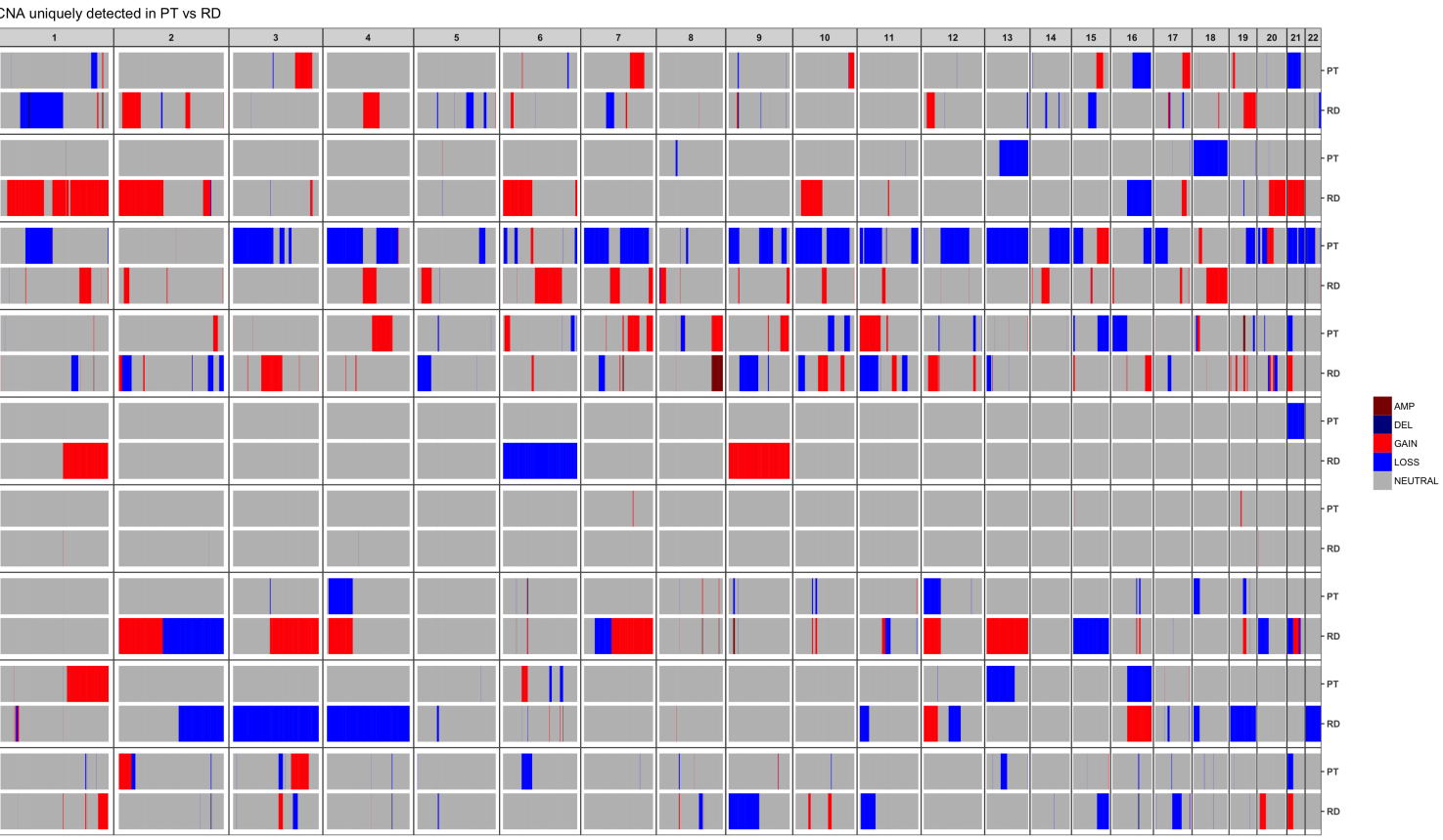


Figure 3.7: The variation of the number of copies of the residual disease with respect to the corresponding primary tumor. Only segments that were not similar between each matched patient sample were shown.

Overall these results show that no recurrent copy number alterations were present in more than 50% of the patients or common to a specific condition. The copy number profiles were very patient-specific reflecting the heterogeneity of the triple-negative tumors present in our cohort. In addition, major modifications of the copy number profile occurred between primary tumors and residual diseases suggesting sub-clonal selection induced by the chemotherapy.

3.5 Mutational profiles of neoadjuvant-resistant triple-negative breast cancer

3.5.1 Overall detection of genetic alterations

The mutational profile was retrieved for the 20 samples for which normal tissue were available. After filtering on possible false positive somatic variants and filtering on non-somatic polymorphisms, the average number of mutations per megabase was 1.09 (range 0.08-2.5). Primary tumors appeared to be less mutated than residual disease and lymph node (0.93, 1.19, 1.37, respectively) (Figure 3.8 A). These results confirm to the prevalence of somatic mutations in breast cancer published by Alexandrov et al. (4).

We investigated the mutational spectrum of six transitions (Ti) and transversions (Tv) in our collection. We found more transversions C>A compared to previous study (94), followed by C>T and C>G (Figure 3.8 B). We observed an average of 131 somatic nucleic variants (SNV) per sample (range 42-297), with an average of 114 mutations in primary tumors, 139 in residual disease and 167 in lymph nodes. Half of these SNVs were exonic, the remainder being intronic or intergenic mutations.
We detected large variations in the number of mutations per sample in the cohort. Patient 1 was the least mutated patient with about 50 somatic SNVs (with 66% represent exonic SNVs), while the patient 35 was the most mutated with more than 270 mutations (around 50% corresponded to exonic SNVs)(Figure 3.8 C).

We assessed the status of known germline mutations that have been described previously in breast and other cancers. Only two patients with germline mutations in BRCA1 and BRCA2 were identified (Patient1 : BRCA1- non-synonymous exonic g.41245027G>A and Patient46: BRCA2 - non-synonymous exonic g.32912361G>A). Two patients (patients 34 and 35) were mutated in ATM (Ataxia telangiectasia mutated), a DNA damage response gene. Three patients were mutated in BAP1 but these mutations were exonic synonymous or intronic mutations (Patient 9, 37 and 36).
A total of 1030 genes containing somatic mutations were detected in primary tumors and 1254 in residual diseases. 334 genes were found mutated in the 2 nodes. The most mutated gene of the cohort was TP53 (7 out of 9 patients , 80% of the samples). All of the following genes with somatic mutations detected were found in only 2 or 3 patients (Figure 3.8 D). Somatic mutations in BRCA2 were found only in patient 1 (intronic SNV in PT and RD). No systematic mutational pattern was observed between primary tumors and residual diseases.

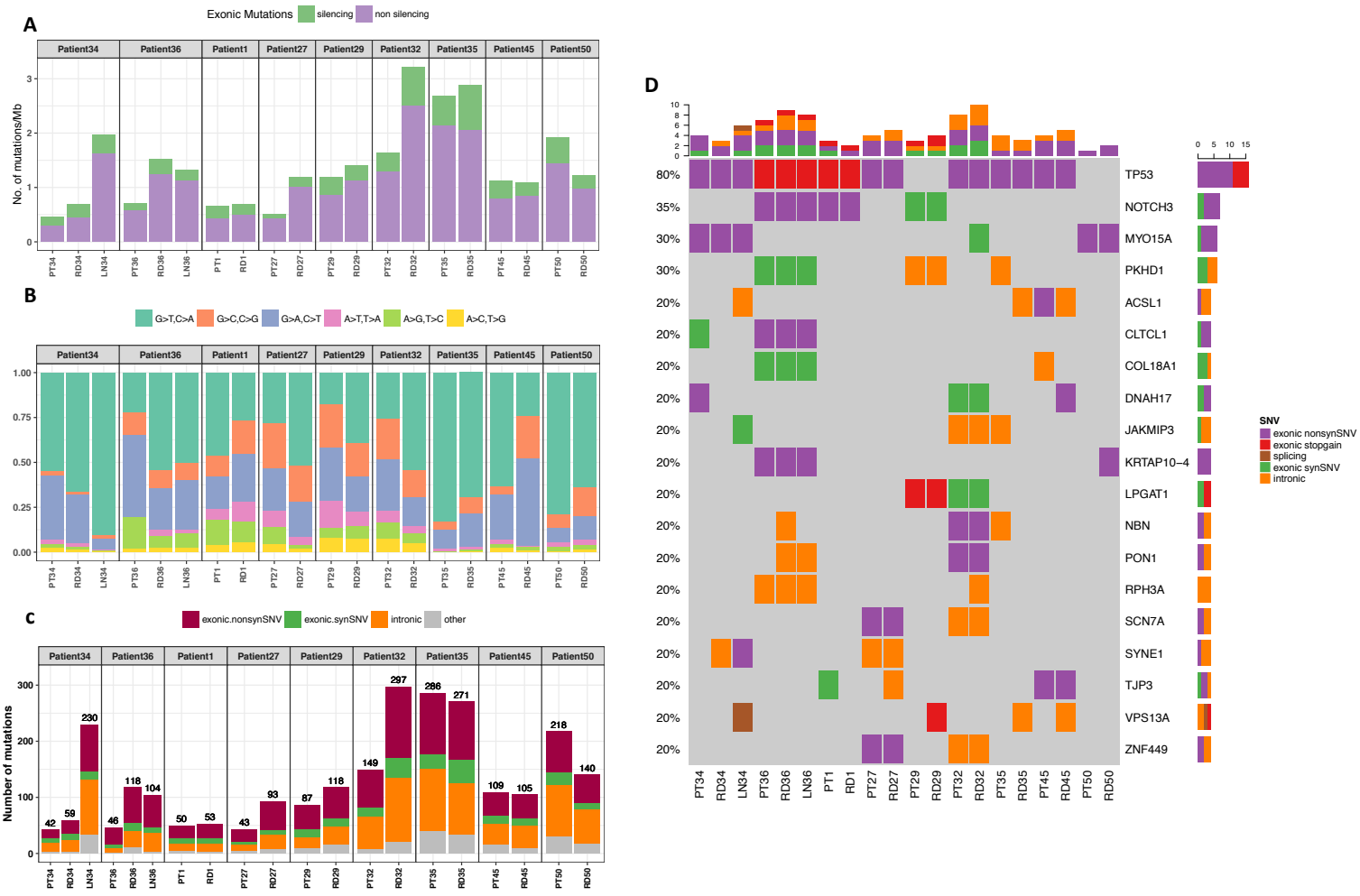


Figure 3.8: Mutation pattern and spectrum in pre- and post-neoadjuvant from 9 resistant triple-negative breast cancer cases. A) Distribution of somatic mutations per megabase. Non-silencing mutations have been defined as an exonic stopgain, stoploss or non synonymous. B) The mutational spectrum of transition and transversion C) Number of mutations per sample stratified by type of SNV. D) Genes with somatic mutation frequently detected (in at least 3 samples).

Somatic mutations have already been reported in large cohorts of breast cancers (110, 129). Based on a list of 128 genes known to breast cancer, only 65 mutations were found in our cohort mostly in a single sample or patient (Supplementary Figure 3.A.6 on page 98).

We next investigated whether known pathways were differentially mutated across conditions using the methodology proposed by Lips et al. (106). Using REACTOME

pathways, we defined a mutated pathway if at least one of its gene members was mutated. We then performed the Fisher's exact test to identify pathways potentially more or less mutated in each condition. Of the 634 REACTOME pathways identified with at least one mutation in at least one sample, only one pathway was significant (p-value= 0.043) and found in 5 RDs : JNK c-JUN kinases phosphorylation and activation mediated by activated human TAK1. However, the association did not remain statistically significant after correcting for multiple testing. 22 pathways were found specifically in RDs, including 9 that were found in 3 to 4 samples. They correspond to pathways involved in JNK and NFκB activation mediated by activated human TAK1 or related pathways (TAK1 activates NFκB by phosphorylation and activation of IKKS complex, CA dependent events, SOS mediated signaling, RAF/MAP kinase cascade, signal regulatory protein SIRP family interactions, RIP-mediated NFκB activation via DAI ,PKA mediated phosphorylation of CREB, SMAD2 SMAD3 SMAD4 heterotrimer regulates transcription).

We observed different types of mutational profiles per sample (Supplementary Figure 3.A.5 on page 97). Residual diseases of patients 29, 32 and 36 shared a large number of mutations with their matched primary tumors (70% for patient 29, 75% for patient 32 and 93 % for patient 36). However, the large number of mutations specific to residual diseases suggests that tumors remained almost unchanged but have continued to accumulate alterations. These results may reflect an increased genomic instability under treatment. Approximately 50% of the mutations were found in the PT of patients 1 and 27 was found in the RD. However, more than 50% of the mutations found in the RD were specific. Finally, clear differences between primary tumor and residual disease were observed for patients 34, 35, 45, 50, with less than 35% of mutations in common. These findings may indicate sub-clonal selection under treatment.

3.5.2 Mutational Signatures in NAC resistant TNBCs

Mutational signatures have been recently introduced as being representative of specific mutational processes. We used the R package deconstructSigs (153) to extract known mutational signatures (4) that might contribute to the specific mutation profiles of pre- and post-treatment SNVs. Applied to our 20 samples, 17 signatures were revealed, including 10 previously observed (4, 128, 129) (Figure 3.9 and Supplementary Figure B.6 on page 249). These signature have been associated with age at diagnosis, activity of the AID/APOBEC family, failure of DNA double-strand break-repair by homologous recombination, defective DNA mismatch repair, and altered activity of the polymerase

POLE. Other detected signatures have been formerly reported only in a small number of cancers. Signature 12 has been reported only in liver cancer and signature 14 has been found in gliomas low grade and uterine carcinomas. Both signatures have unknown aetiologies. Signature 7 was associated with exposure to ultraviolet light and detected in head and neck cancers, oral gingivo-buccal squamous carcinomas and melanomas. Finally, signature 4 was detected in a large panel of cancers and associated with tobacco mutagens. Alexandrov et al. could not validate the signatures R2, R3 and U1. However, the signature R2 was found in 35% of our samples.
The patterns of mutational signatures were almost the same between primary tumors and corresponding residual diseases. However, we could note the presence of 1 or 2 signatures specific to each matched samples. The same signatures were found in patients 34 and 50 before and after NAC, at the opposite of patient 32 with 4 or 3 signatures specific to each condition. Interestingly, the signatures observed in LN have always been found in their matched PT and RD.

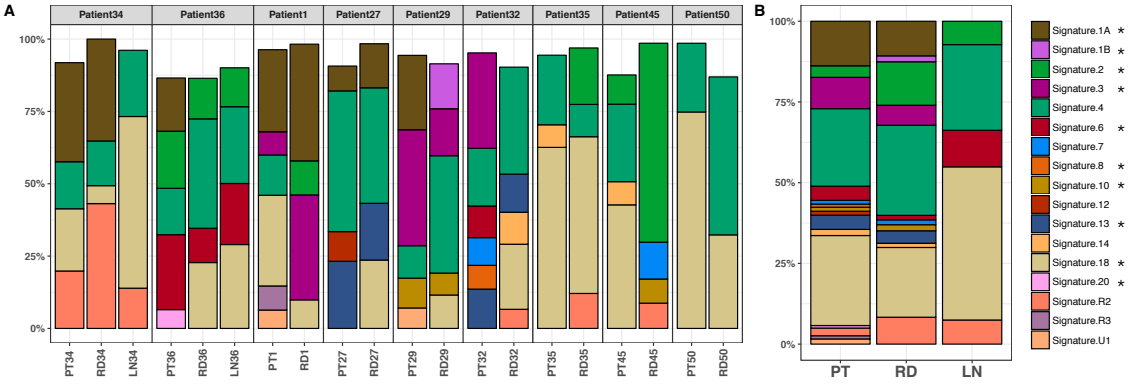


Figure 3.9: Mutational signatures in pre and post-NAC TNBCs. The signatures marked by a "*" have been previously associated with breast cancer.

3.5.3 Functional pathways altered in drug-resistant TNBC

Several key pathways have already been reported altered in post-treatment triple-negative breast cancer samples. Based on the functional groups used by Balko et al. (14), we have observed three alterations in three patients that were present in all their pre- and post-NAC samples (patient 32 amplification CDK6, patient 34 amplification FGFR1, patient 36 amplification CCND3) (Figure 3.10 A).
Different definitions may exist on altered cancer pathways. Therefore, we extended

our analysis by examining 11 pathways reported by Lorio et al. (110). We identified 8 pathways altered in our cohort (cell cycle, DNA repair, PI3K/AKT/mTOR, RAS/RAF/MEK/ERK, RTK signaling, cell death, TP53 signaling and transcription signaling) (Figure 3.10 B). All samples have alterations in at least one of these pathways. TP53 signaling was the most altered pathways in PTs and RDs and we observed a major increase of alterations in cell cycle signaling in RDs. The number of alterations remains stable in most processes, but the loss of alterations in the post-treatment samples in the two associated Ras-ERK and PI3K-mTOR pathways can be noted.

3.5.4 Clonal mutational heterogeneity

To characterize the clonal architecture of pre- and post-neoadjuvant triple negative breast cancers, we performed sub-clonality analysis using QuantumClone (48). Most of the patients showed important changes in clonal composition after therapy. We observed many loss of mutations or increased frequency in the post-treatment samples of clones that were uncommon in the pre-treatment sample (Supplementary Figure B.7 on page 251). Although these changes could be accounted for by sampling effects, they may provide evidence of significant sub-clonal selection. The mean number of subclones estimated per sample was 4.9 (range 3-9). The patient 35 had the largest number of estimated clones suggesting numerous tumor changes. We did not observe a significant association between the predicted number of clones and the number of mutations per patient across the cohort (Spearman's rho=0.24, p-value=0.47). We searched for mutations that could confer selective advantage in each sub-clones. Putative driver mutations were determined as being predicted deleterious by at least three prediction tools (FATHMM (165), SIFT (127), PolyPhen (2) and SNPeff (37)). 30 clones out of 46 harbored driver mutations (average of 3.1, range:1-10). We observed more putative driver mutations in clones belonging to RDs (mean=4.3, range:1-8, 13 clones in 9 patients) compared to those from PTs (mean=2.7, range:1-10, 11 clones in 9 patients). However, the difference was not significant.

3.6 Personalized medicine

At present, the standard of care in the post-NAC setting for TNBC patients is observation. Patients with TNBC who resist to neoadjuvant chemotherapy are less likely to respond to conventional antineoplasic treatment as the tumor or future metastases have been exposed to chemotherapy in the neoadjuvant setting. Screening for actionable genes with somatic mutations in resistant sub-clones may reveal great potential for adjuvant

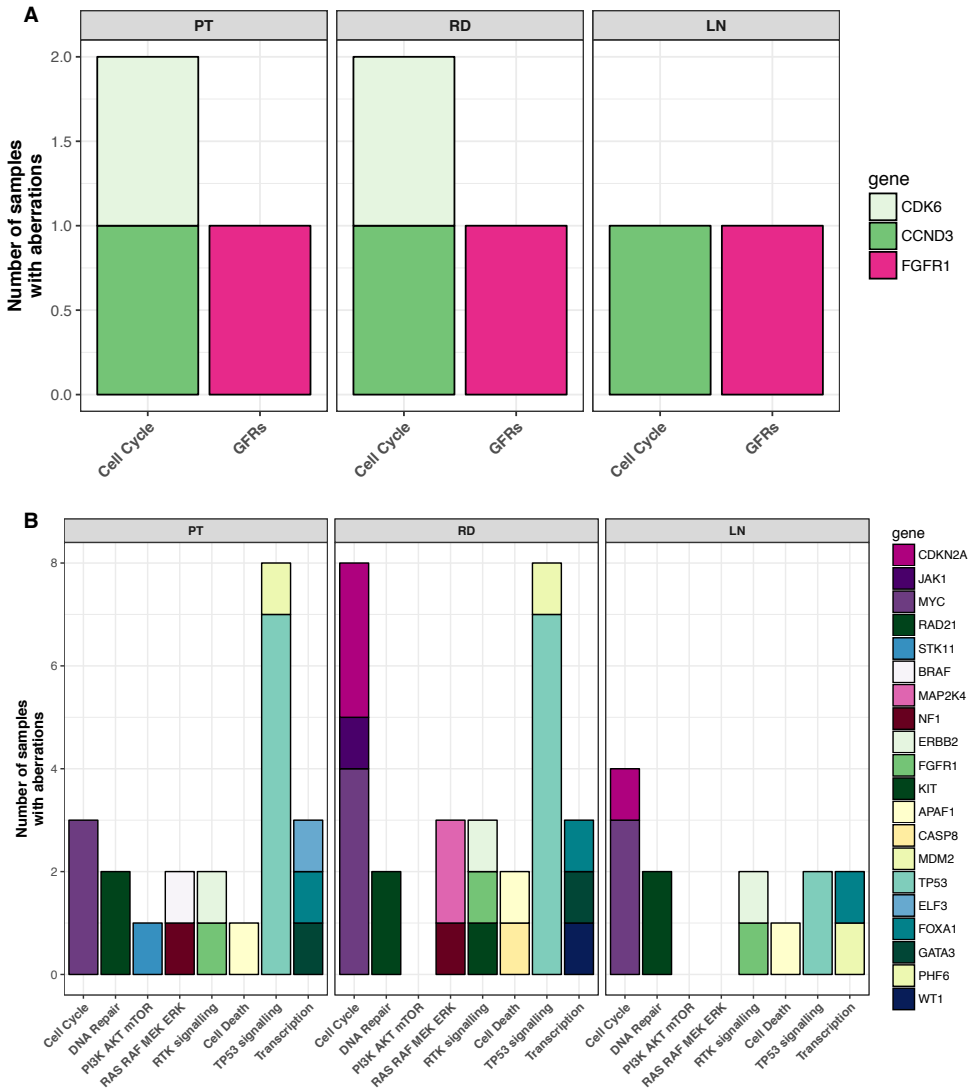


Figure 3.10: Targetable pathways in TNBCs. A) Pathways associated with cancer reported by Balko et al. (14). B) Pathways associated with cancer reported by Lorio et al. (110).

trials. We evaluated the druggability of genes with non-silencing somatic aberrations for each clones by recovering gene-drug interactions from curated FDA approved, preclinical and clinical databases through rDGIdb (184, 200). A total of 663 current or prospective molecules interact with 91 genes with somatic non silent mutations found in the cohort. More than 80% of the estimated sub-clones had mutations in genes possibly actionable (Supplementary Figure B.7 on page 251). Importantly, distinct tailored drugs could

Table 3.1: Patients with specific alterations that can benefit from targeted therapies. Drug-genes interaction were retrieved from FDA approved and currated databases (rDGIdb (184, 200)), and two clinical trials SAFIR01 (10) and SHIVA (99).

Genes	Alteration type	PT	RD	PT and RD	Drug
FGFR1	Amplification	0	0	Patient 34	E_3810, Everolimus, GDC_0980, BGJ398
FRS2	Amplification	0	0	Patient 27	Sorafenib
MDM2	Amplification	0	0	Patient 27	RO5503781
PDGFRB	Mutation	0	0	Patient 32	Imatinib,Sorafenib
KIT	Mutation	0	Patient 35	0	Imatinib, Midostaurin
RET	Mutation	0	Patient 36	0	Imatinib
STK11	Mutation	Patient 1	0	0	Everolimus
BRAF	Mutation	Patient 50	0	0	Vemurafenib, Cobimetinib ,Trametinib

be assigned to primary or residual cancer. Among the 153 drugs found in more than two patients, 58 were inhibitors of proteins or tyrosine kinases, MEK inhibitors or other targeted therapies (Figure 3.11).

We then restricted our analysis to gene-drug interactions based only on drugs that were FDA-approved or used in two recent major prospective trials SAFIR01 (10) and SHIVA (99) (Table 3.1 and Supplementary Figure3.A.7 on page 99). The goal of these two studies was to assess the efficacy of several molecularly targeted agents outside their indications on the basis of tumor molecular profiling. 8 patients of our cohort were eligible for targeted therapies according to predefined study algorithms. Interestingly, three patients (27, 32, and 34) have alterations found in their primary tumor that persist after treatment (FGFR1, FRS2 and MDM2 amplification and PDGFRB mutation). Similarly, patients 35 and 36 have alterations that emerged in their residual diseases (ERBB2 amplification, KIT and RET mutation). A dedicated targeted therapy using imatinib or trastuzumab could have been proposed as an adjuvant treatment to these patients.

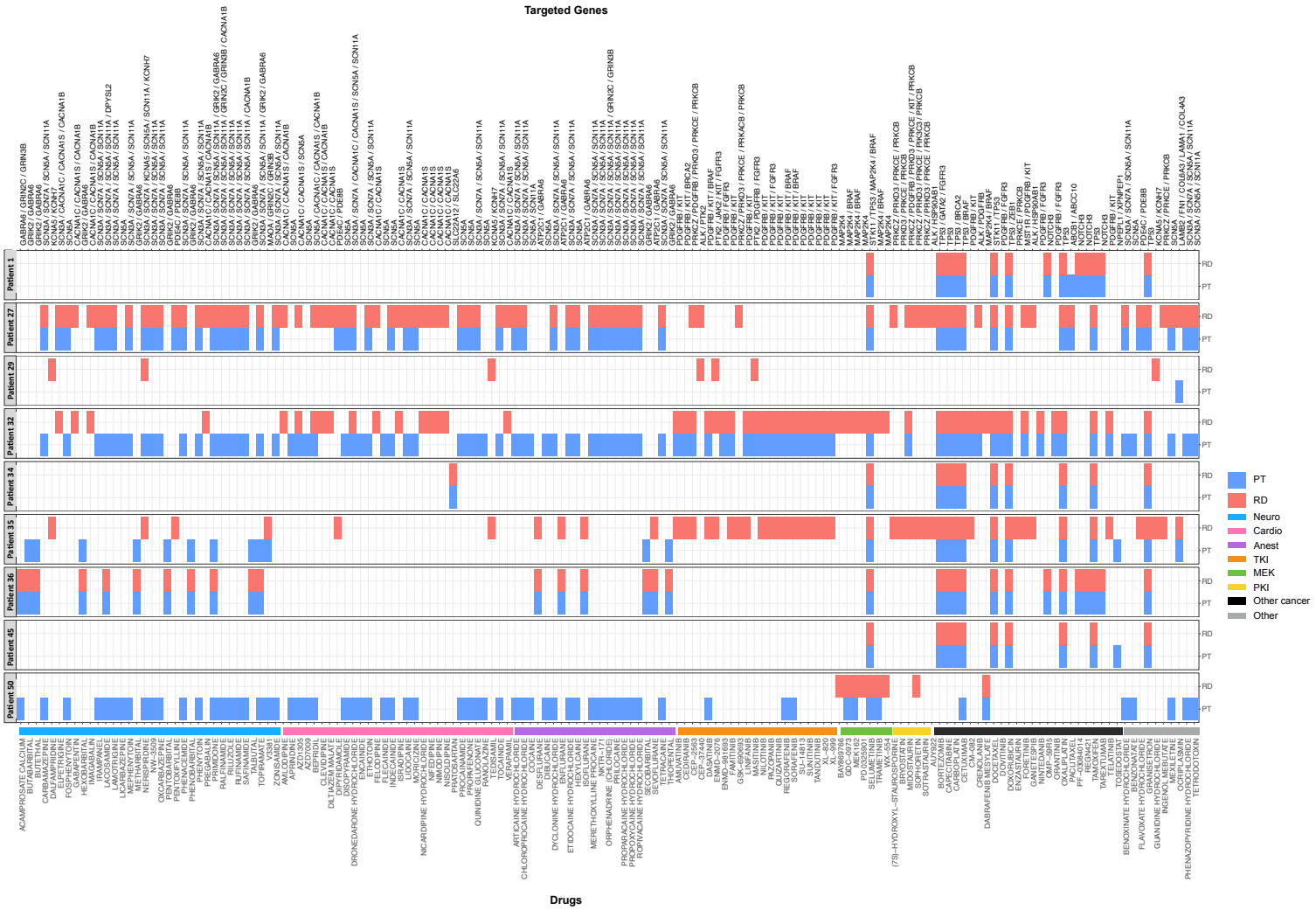


Figure 3.11: Mutations in genes with potential druggable implications. Drugs associated with 51 genes with non-silencing mutation in the cohort. We report the 153 drugs found in at least two patients. Drug names are display at the bottom of the panel. The top of the panel present the mutated genes detected in our cohort associated with the drug. PT: Primary tumor, RD: Residual disase, Neuro: Neurology, Cardio: Cardiovascular drug, Anest: Anesthetic, TKI: Tyrosine kinase inhibitor, PKI: Protein kinase inhibitor, MEK: MEK inhibitor.

3.7 Discussion

We report the transcriptional landscape and genomic profile of a cohort of 16 patients with neoadjuvant-resistant triple-negative breast cancer. The selected patients were chosen to represent tumors highly resistant to standard-of-care anthracyclines-taxanes chemotherapy. To our knowledge, this is the first study dedicated to analyse the resistance of TNBC to neo-adjuvant chemotherapy with pre- and post-treatment matched samples based on WES and RNAseq data.

Studies based on large cohorts of TNBC tumors have shown extensive genomic and transcriptomic tumor heterogeneity (46, 100, 183). Although our cohort is small with 16 patients, we also observed large heterogeneity among our patients. We first identified four groups of patients according to their gene expression profiles. At the individual level, some patients had changes in gene expression for AR and immune modules, while others had no change for 5 TNBC-related processes. In addition, we observed a greater inter-tumor heterogeneity than intra-tumor heterogeneity suggesting that individual variations are greater than those induced by NAC. The heterogeneous nature of TNBC could be one of the reasons why we found only a very small set of differentially expressed genes between conditions and why no mutation or copy number alteration was associated with a given condition. Similar results have previously been reported on TNBC (13, 93, 106) suggesting that hundreds of different genomic alterations may be associated with resistance to standard neoadjuvant chemotherapy.

Evidence for clonal selection under treatment was observed at both CNV and SNV level. One of our initial hypotheses was that a clonal selection was made under treatment, leaving one or more clones resistant to chemotherapy. We have identified several sub-clones emerging in the residual disease, but it remains unclear how much they are related to NAC resistance. Although genome transcriptome concordance remains unclear (3, 124), it was surprising to detect almost no modification in gene expression. However, we identified many alternative isoforms differentially expressed between conditions. Among the genes differentially spliced between PTs and RDs we can cite TOP2A, NOTCH1, ABCC11 and MDM4. TOP2A encodes the DNA topoisomerase II involved in both DNA replication and transcription. The anthracyclines block the topoisomerase II that causes DNA lesions. Many isoforms are produced by TOP2A. They are organized into cytoplasmic isoforms that are functionally inactive, as opposed to nuclear isoforms. These regulate the transport of proteins between the cytoplasmic and nuclear

compartments, which can reduce the concentration of the drug in the cell (49, 123). The NOTCH1 gene is a member of the NOTCH family that regulates the proliferation of cancer cells. NOTCH1 has been associated with multi-drug resistant genes such as ABCC11 and MDM4 (MRP1) in triple-negative breast cancer (23, 36, 215). In addition, our study suggests that many RNA binding factors regulated AS events. The deregulation of RNA binding factors may plays an important role in NAC resistance of TNBCs. The exact cause of deregulation of RBP remains unknown, as are their individual and combinatorial impacts on TNBC resistance. Nevertheless, the over-representation of some RBP binding sites and the distinct profiles of alternative transcripts could be biomarkers potentially useful in the management of TNBCs.

The analysis presented here has some limitations inherent to the small number of patients studied. The identification of recurrent genomic alterations and altered biological processes was then limited.

Another limitation is the lack of tumor-matched normal DNA for 6 tumors

However, our small cohort have highlighted two rare cases of TNBCs. 1) One patient with an almost full genomic profile with one copy in the primary tumor and residual disease, 2) one patient with ERBB2 amplification. While we observed ERBB2 amplifications in pre- and post-NAC samples, the amplification have been so far confirmed only in the residual disease. IHC and sequencing analyses were performed on different core biopsy specimens, which prevent us from making definitive conclusions. Indeed, spatial tumor heterogeneity may be responsible of this inconsistency. However, it has been shown that in some cases the hormonal receptor of the tumor and the ERBB2 status change over time. Lindstrom et al. (105) have studied biopsy results from more than 1,000 women who were diagnosed with early-stage breast cancer and developed a locally or metastatic advanced breast cancer. All patients received adjuvant therapy. About 15% of the later biopsies showed HER2 status had changed from the original biopsy (Primary positive/relapse negative = 8.7%, Primary negative/relapse positive = 36 5.8%). Assuming that these processes can be influenced by adjuvant therapies it may be the same with neo-adjuvant treatment. Treatment decisions are usually based on the IHC status from the original breast cancer biopsy. The case of this patient emphasizes that the IHC should be redone on surgical samples to define adjuvant treatment.

The prognosis of these patients remains very bleak. We have sought to identify new therapies that could have been used specifically for each patient. We have shown that a wide range of drugs interact with a small set of mutated genes in our cohort. In addition,

8 out of 10 patients were eligible for FDA-approved cancer treatments that have been tested in two recent "basket" studies. Many of the alterations identified in the primary tumor persist in residual disease. These results suggest that many tumors were chemo-resistant prior to NAC. We strongly encourage the development of studies with larger cohorts of matched primary tumors and residual diseases to identify potential targeted therapies that could be used in neoadjuvant trials.

3.8 Conclusion

The results presented here have yet to be consolidated and clarified. So far, we have mostly looked for global changes that could explain the resistance to treatment. Given the molecular heterogeneity and the small size of our cohort, the task has proven difficult. We must now carefully analyze each patient and try to identify his own mechanism of resistance.

3.9 Material and methods

Patients

We analyzed a cohort of patient with invasive breast cancer all treated by NAC (NEOREP Cohort, CNIL declaration number 1547270) treated at Institut Curie, Paris, between 2005 and 2012. We included patients with NAC-resistant triple-negative breast cancer for which frozen core needle biopsies and post-NAC surgical specimens were available. Cases were considered estrogen receptor (ER) or progesterone receptor (PR) negative (-) if at less than 10% of the tumor cells expressed estrogen and/or progesterone receptors (ER/PR), in accordance with guidelines used in France(76). HER2 expression was determined by immunohistochemistry and scoring was performed according to the American Society of Clinical Oncology (ASCO)/College of American Pathologists (CAP) guidelines(212). Scores 3+ were reported as positive, score 1+/0 as negative (-). Tumors with scores 2+ were further tested by FISH. Resistance to NAC was evaluated based on the pCR. A pCR was defined as the absence of residual invasive cancer cells in the breast and axillary lymph nodes (ypT0/is +/ypN0). Lymphovascular invasion was defined as the presence of carcinoma cells within a definite endothelial-lined space (either lymphatic or blood vessels). The presence of lymphovascular invasion was evaluated on standard formalin fixed paraffin embedded post-NAC surgical pathologic specimens without additional staining. The lymph nodes were sequenced when the material was available (6 patients).

Whole Exome Sequencing and Variant Calling

Genomic DNA (500 ηg - 1 μg) from the biopsies and the corresponding surgical specimens were extracted. The Agilent SureSelect solution capture exome array (SureSelect Human All Exon v5) was used to capture exomes from tumor samples, using the procedure supplied from the manufacturer's instructions. The library was sequenced on an Illumina Hi-Seq2500 platform in paired-end 100 bp at Institut Curie sequencing platform with an average read depth of coverage of 100x.
Reads were aligned on the human genome reference hg19/GRCh37 by Burrows-Wheeler Aligner v0.7.5a. PCR duplicates were removed using the MarkDuplicates (Picard) algorithm and reads were filtered on exome target using BedTools. The Genome Analysis Toolkit (GATK v3.5) was used for local realignement and quality score recalibration.
Haplotype caller (GATK suite), Mutect2 (GATK suite) and Varscan (v2.4.1) were used to call somatic single nucleotide variants (SNV). Mutation calling was assessed only on the 10 pairs primary tumor / residual disease for which the corresponding normal samples were available because false positive rate remained too high for the other samples. Somatic variants were annotated with snpEff and Annovar (RefGene 102105 and snp138 database). Base coverage was assessed by depthOfCoverage tool from Genome Analysis Toolkit.
Variants found by a single caller over 3, with a low mappability or repeated region, with a mutant allelic fraction supported by <5 reads, or with a total coverage <11 reads or in over mutated exon or in several patients were disregarded. They constituted the first level of filtering and were used to retrieve the mutational signature. Mutations in the Single Nucleotide Polymorphism Database (dbSNP build 138), the 1000 Genomes Project, and Exome Aggregation Consortium dataset (ExAC) with a minor allele frequency of <1% and not found in COSMIC database were excluded as putative germline sequence alterations. They were the second level of filtering and were used to characterize each sample and perform clonal reconstruction.

RNA sequencing

RNA-seq was performed on biopsies and corresponding surgical specimens using Illumina Hi-Seq2500 leading to paired-ends 100 x 100 bp with 100x expected coverage. Alignments were performed on human reference sequences using TopHat v.2.0.621. Reads with mapping quality ¡20 and reads marked as duplicates by Picard v.1.97 were excluded from further analysis.
Gene- level read counts were obtained using HTSeq-count (7) and RefSeq hg19/GRCh37

gene annotation.

Selection of the most variant genes We selected the most variant genes, based on the inflection point of the interquartile range (IQR) distribution for gene expression. The gene expression was previously rlog transformed with DESeq2 package (111). This method is more data-driven than a fixed threshold to define the proportion of genes with the highest level of variation. The full procedure is described below. For each gene, we: (1) calculated the IQR for all cell lines, (2) sorted the IQR values of the genes in ascending order, to generate an ordered distribution, (3) estimated the major inflection point of the IQR curve as the point on the curve furthest away from a line drawn between the start and end points of the distribution, and (4) retained genes with an IQR higher than the inflection point.

Gene expression analysis Differential expression analyzes on RNA-seq data were performed using the limma-voom package (98). Each comparison was adjusted for paired samples. Patient 35 was excluded from the analysis because she appeared as an outlier in the PCA analysis (data not shown). The histogram of p-values estimated by limma-voom between PTs and LNs, showed a uniform distribution. We apply the method fdrtool (174) to estimate the variance of the null distribution and correct the p-values. We accounted for multiple testing, by calculating the FDR-adjusted p-value for each drug. An FDR-adjusted p-value <0.05 was considered significant.

Alternative splicing analysis We used rMATS (v 3.0.9) for paired study design (162) to determine the differential alternative splicing events between the three conditions. Four types of alternative splicing events (skipped exon (SE), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), and retained intron (RI)) based on annotated RefSeq genes were tested for inclusion-level difference $\geq 5\%$, and the events with FDR-adjusted p-value < 0.05 and absolute inclusion level difference > 0.1 were determined as differential alternative splicing events.

RBP motif enrichment analysis We have used an improved version of rMAPS (133) to find splicing factors involved in the regulation of spliced exons. Our method is presented in the chapter 4. For each of the RBP motif, we investigated their enrichment in the 250-bp of flanking sequences in the upstream and downstream intronic sequence of SE events ($\Delta\psi \geq 0.05$, FDR-adjusted p-value < 0.05), excluding splice sites, and we compared the motif enrichment score to a set of non-regulated splicing events ($minPSI > 0.15$ and $max\psi < 0.85$ and FDR-adjusted p-value $> 50\%$) of the rMATS analysis. A *post-hoc* FDR was computed to identify the RBPs significantly enriched in

the sequences of regulated exons. Here we used a *post-hoc* FDR < 0.25.

Copy number analysis

Copy number analyses with FACETS (161) which also provided purity and ploidy estimates was performed. Amplifications were called at segments with ≥ 6 copies and homozygous deletions at 0 copies.
"Impossible events" used to estimate sub-clonal selection based on copy number were determined if they satisfy one of the following criteria: 1) amplified or gain in PT and deleted or lost in RD, 2) deleted or lost in PT and amplified or gain in RD.

Supplementary Information

3.A Supplementary Figures

Figure 3.A.1: Individual changes in patient gene expression. A) According to six TNBC-related metagenes defined by our team (Bonsang et al.). B) According to gene expression modules corresponding to 10 immune cell populations defined by Becht et al.

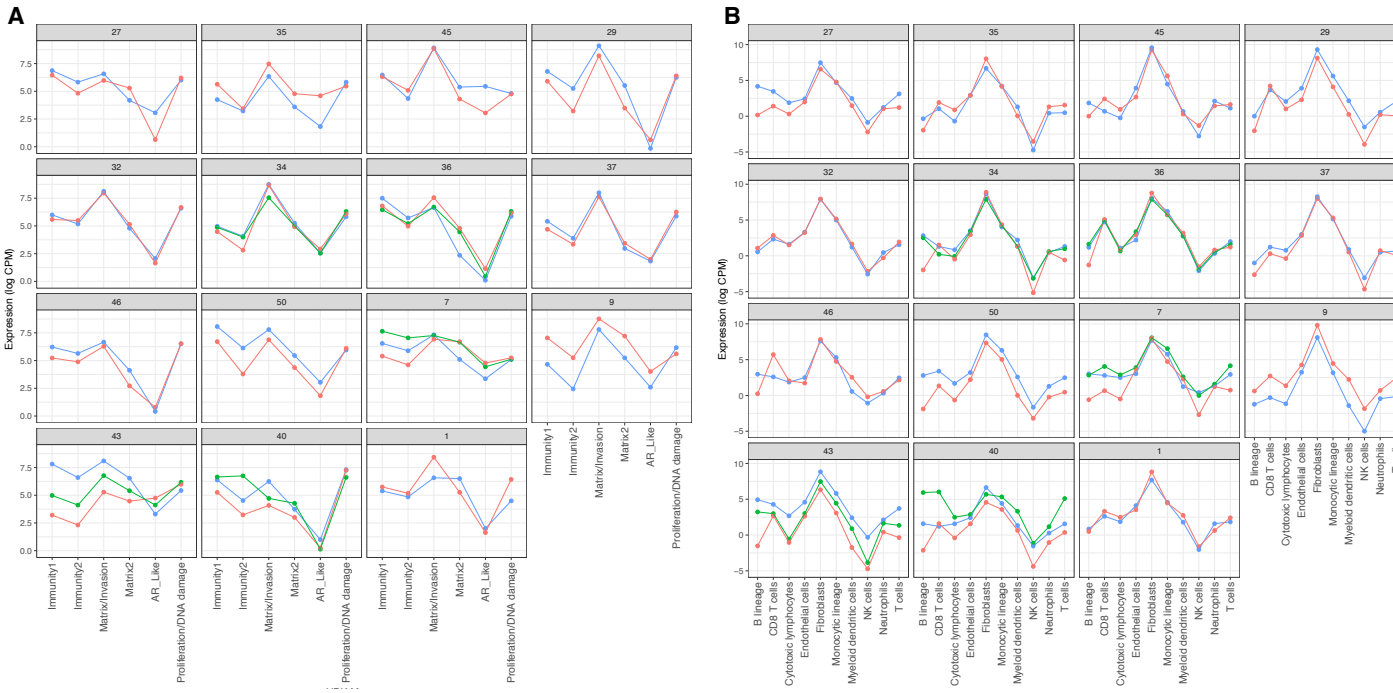


Figure 3.A.2: Stromal tumor- infiltrating lymphocytes before and after NAC. Patients in red have high expression for A) the Immune modules defined by Bonsang et al. B) the T lymphocytes immune module defined with the most variant genes.

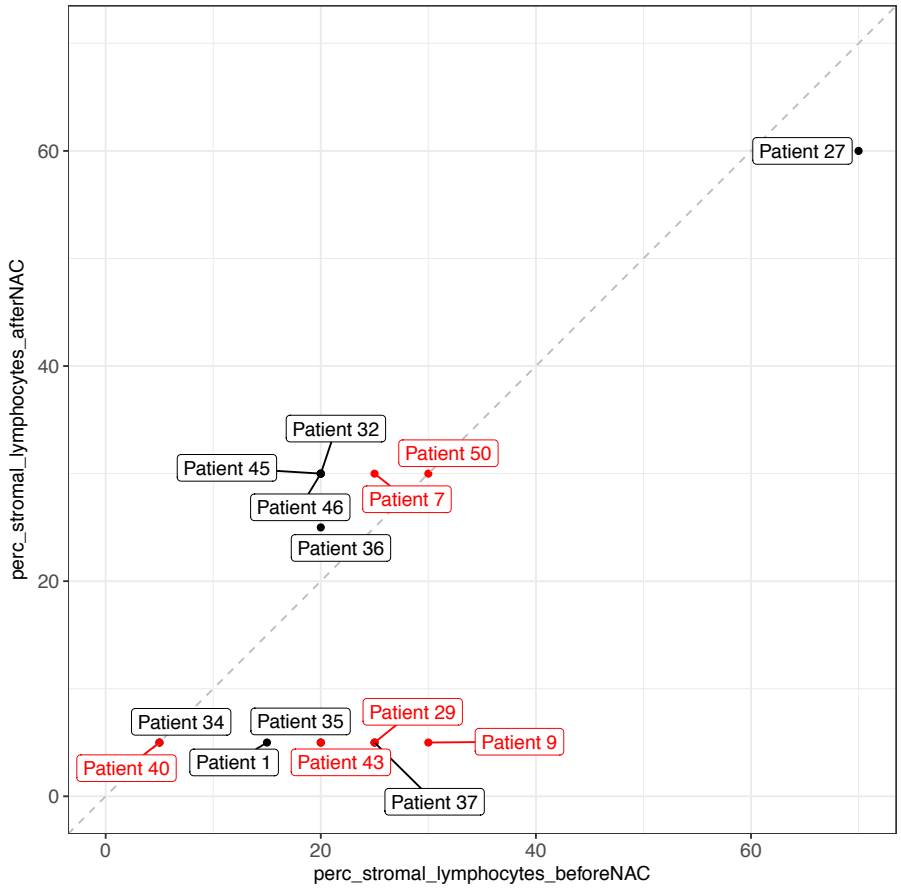


Figure 3.A.3: Clustering based on genes differentially expressed between PT and RD

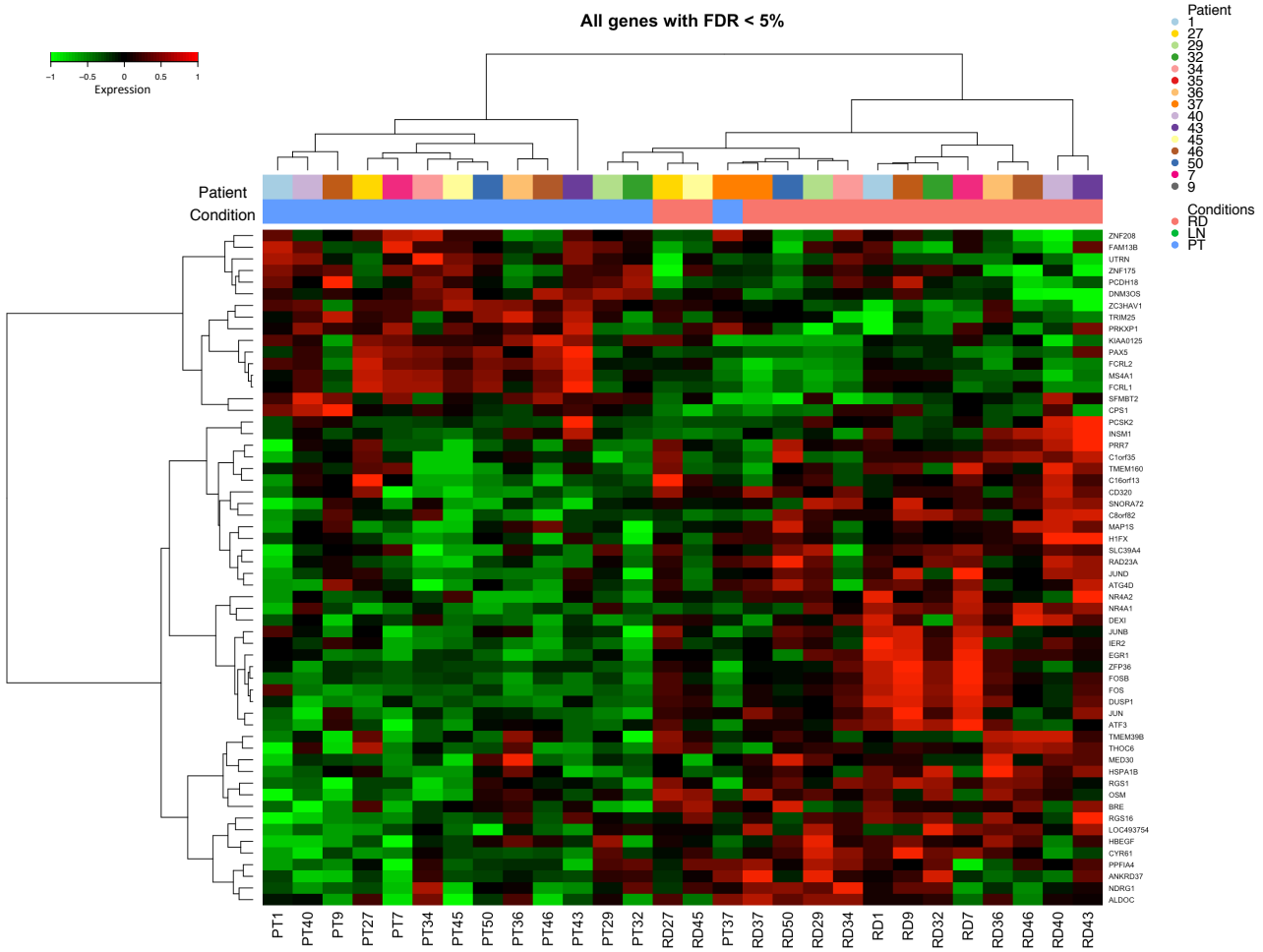


Figure 3.A.4: Clustering based on genes differentially expressed between PT and LN

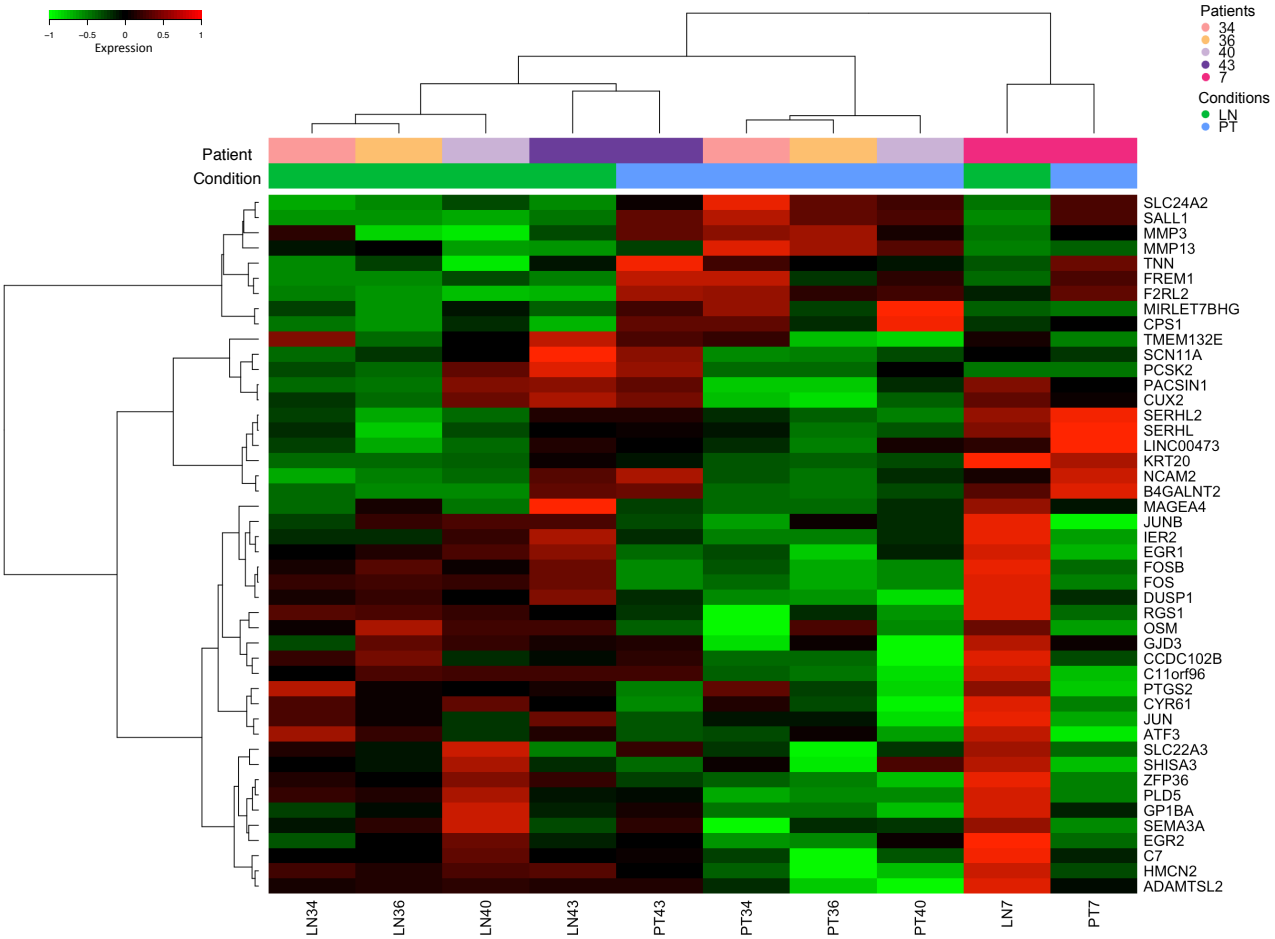


Figure 3.A.5: Number of mutations shared between paired samples

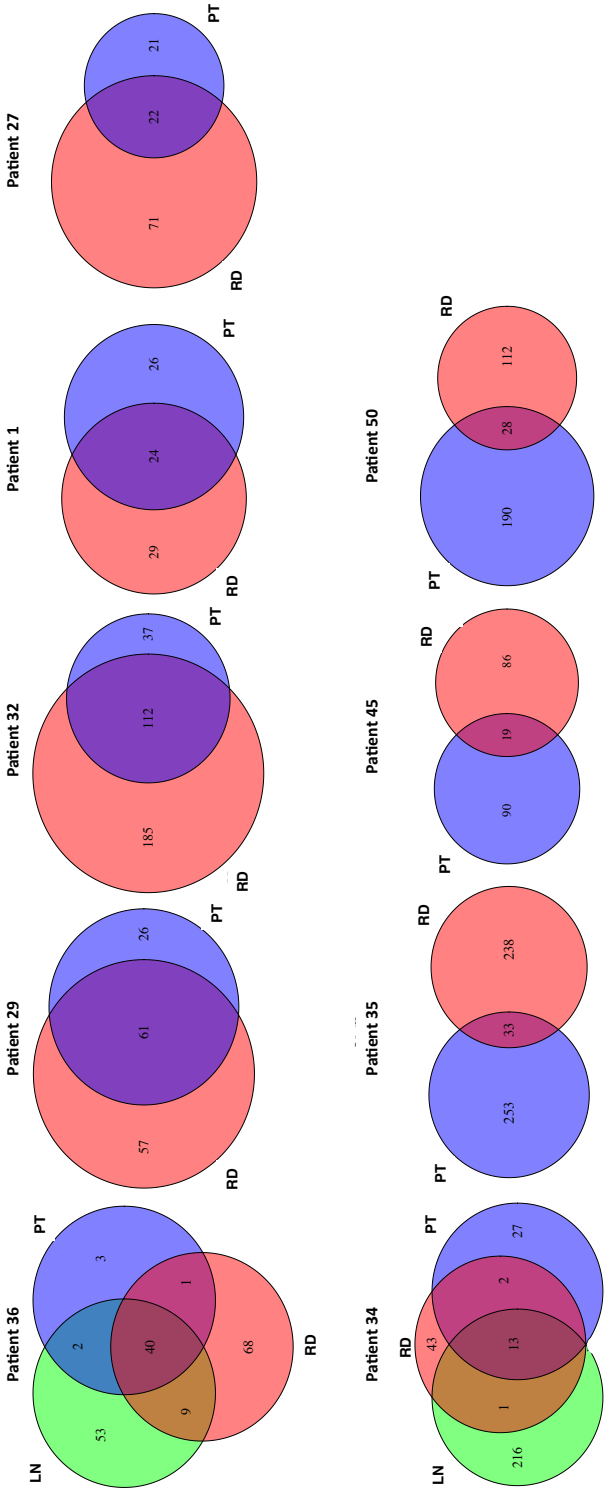


Figure 3.A.6: Number of somatic mutations in genes known to breast cancer. Only 34 genes out of 128 that have been previously found in large TNBC cohort were found mutated in our patients.

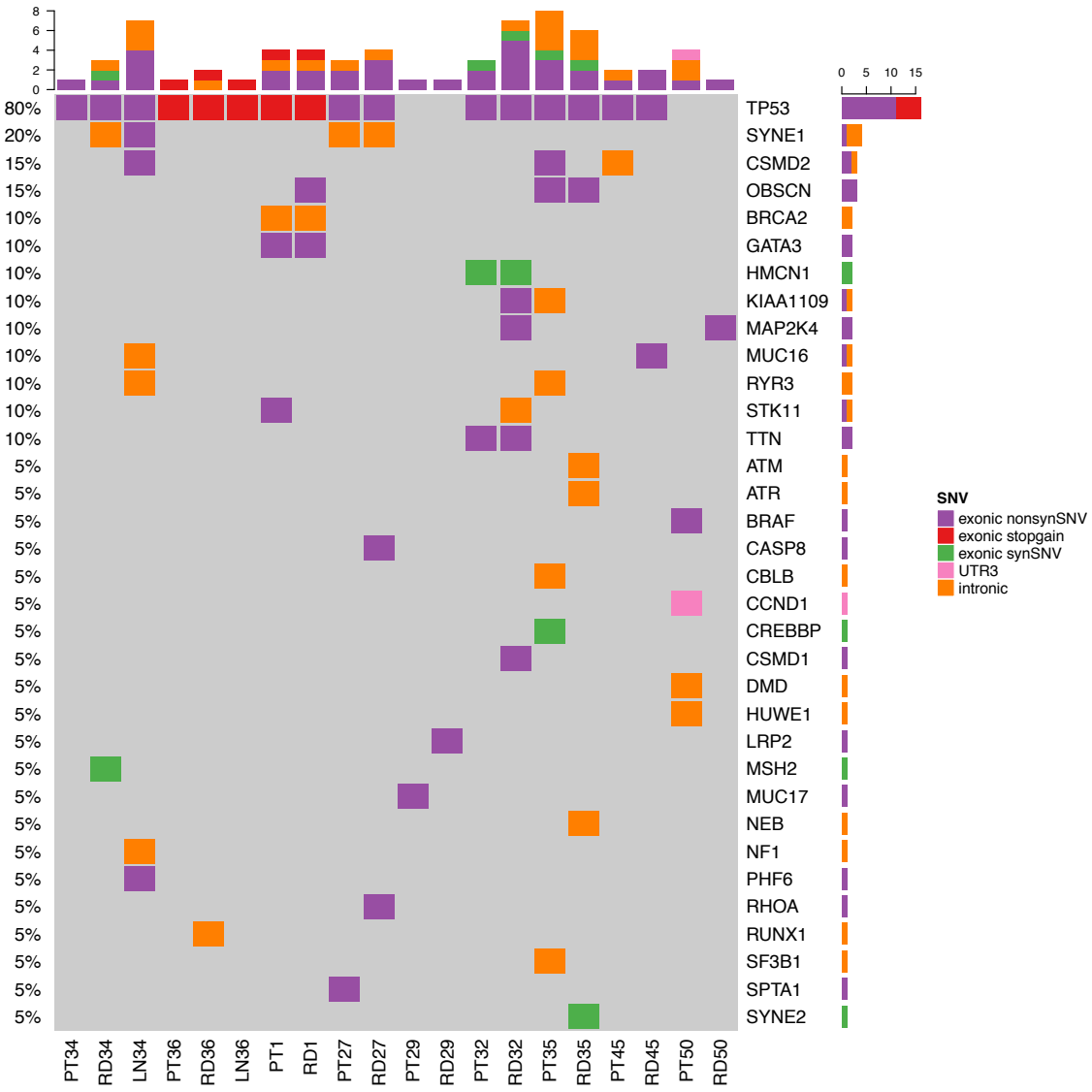


Figure 3.A.7: Treatment algorithm in SHIVA A) and SAFIR01 B)

A

Table 1. Treatment algorithm in the experimental arm

Targets	Molecular abnormalities	Molecularly targeted agents
KIT, ABL1/2, RET	Activating mutation or amplification[a]	Imatinib 400 mg qd PO
PI3KCA, AKT1	Activating mutation or amplification	Everolimus 10 mg qd PO
AKT2,3, mTOR, RAPTOR, RICTOR	Amplification	Everolimus 10 mg qd PO
PTEN	Homozygous deletion or heterozygous deletion + inactivating mutation or heterozygous deletion + IHC confirmation	Everolimus 10 mg qd PO
STK11	Homozygous deletion or heterozygous deletion + inactivating mutation	Everolimus 10 mg qd PO
INPP4B	Homozygous deletion	Everolimus 10 mg qd PO
BRAF	Activating mutation or amplification	Vemurafenib 960 mg bid PO
PDGFRA/B, FLT3	Activating mutation or amplification	Sorafenib 400 mg bid PO
EGFR	Activating mutation or amplification	Erlotinib 150 mg qd PO
ERBB2/HER2	Activating mutation or amplification	Lapatinib 1000 mg qd PO + Trastuzumab 8 mg kg^{-1} IV followed by 6 mg kg^{-1} IV q3w
SRC	Activating mutation or amplification	Dasatinib 70 mg bid PO
EPHA2, LCK, YES1	Amplification	Dasatinib 70 mg bid PO
ER, PR	Protein expression >10%	Tamoxifen 20 mg qd PO (or letrozole 2.5 mg qd PO if contra-indication)
AR	Protein expression >10%	Abiraterone 1000 mg qd PO

Abbreviations: AR = androgen receptor; bid = twice a day; ER = oestrogen receptor; IHC = immunohistochemistry; IV = intravenously; LOH = loss of heterozygosity; mTOR = mammalian Target Of Rapamycin; PO = orally; PR = progesterone receptor; qd = daily; q3w = every 3 weeks.
[a]Druggable focal amplification was defined as gene copy number ⩾6 for diploid tumours and ⩾7 for tetraploid tumours and an amplicon size of ⩽1 Mb or ⩽10 Mb if protein overexpression confirmed by IHC.

B

	Number of patients (assessable for efficacy)	Number of patients treated in phase 1 or 2 trials	Number of patients with antitumour activity (%)*
All patients	48 (43)	28	13 (30%)
FGF4 amplification, treated with FGFR inhibitor E-3810	2 (2)	2	1 (50%)
EGFR amplification, treated with EGFR inhibitors erlotinib and cetuximab-temsirolimus	2 (2)	1	1 (50%)
EGFR amplification or AKT1 or PIK3CA mutation, treated with AKT or mTOR inhibitors (everolimus and GDC-0980)	2 (1)	1	1 (100%)
FGFR1 amplification, treated with FGFR inhibitors E-3810 (n=3) or BGJ398 (n=6)	9 (8)	9	2 (25%)
FGFR1 amplification or PIK3CA mutation, treated with FGFR inhibitor BGJ398	2 (1)	2	0
FGFR1 amplification or PIK3CA mutation, treated with mTOR inhibitor everolimus	1 (1)	0	0
FGFR2 amplification or PIK3CA mutation, treated with FGFR inhibitor BGJ398	1 (1)	1	0
IFG1R amplification or PIK3CA mutation, treated with mTOR inhibitor CCI-223	1 (1)	1	1 (100%)
MET gain, treated with MET inhibitor onartuzumab	1 (1)	1	0
FRS2 amplification, treated with Raf inhibitor sorafenib	1 (1)	0	0
AKT1 mutation or *AKT2* amplification, treated with AKT1 and/or mTOR inhibitor everolimus (n=4) or ridaforolimus plus MK2206 or 0752 (n=2) or plus CC223 (n=1)	7 (6)	3	3 (50%)
PIK3CA mutation or amplification†, treated with PI3K, AKT, or mTOR inhibitors (everolimus n=9, GDC-0980 n=2,[17] GDC-0068 n=1,[18] or cetuximab-temsirolimus n=1; associated with chemotherapy in one patient)	13 (12)	4	4 (33%)
RPTOR amplification, treated with mTOR inhibitor everolimus or axitinib-everolimus	2 (2)	1	0
CCND1 amplification, treated with CDK4 inhibitor BAY 1000394	1 (1)	1	0
AR amplification, treated with AR inhibitor bicalutamide	1 (1)	0	0
MDM2 amplification, treated with MDM2 inhibitor RO5503781	1 (1)	1	0
MGMT amplification, treated with alkylating agent temozolomide	1 (1)	0	0

*Objective response or stable disease for >16 weeks, seen overall in n=28. †*PIK3CA* amplification seen in only one patient.

***Table 3*: Genomic targets and matched drugs**

4

Assessing significance in motif enrichment analysis

When I started studying alternative splicing events and RNA binding proteins (RBPs) enrichment, I had the opportunity to meet Olivier Saulnier, a PhD student in Olivier's Delattre laboratory. Olivier studies the splicing landscape of Ewing's sarcoma. From this meeting a collaboration was born to improve a tool called rMAPS developed by the authors of rMATS, to identify the binding positions of the RBPs around the exons. This web tool allows very few customizations and we have encountered many server errors when using it. These errors have already been reported on the google group of the method with no satisfactory answers. In addition, we observed that no multiple testing correction was considered. We decided to implement an internal version of the method and add some improvements. This chapter is composed of a first theoretical part in which we recall the biological context, what the existing method realizes, its limits and what we propose to improve it. In a second part we present two applications. First, on our neoadjuvant-resistant triple-negative breast cancer samples, then, on Ewing's sarcoma cell lines in collaboration with Olivier Saulnier.

4.1 Background

RNA-binding proteins (RBPs) are proteins that bind to specific DNA sequences (called motifs) to regulate post-transcription of mRNAs. Motif enrichment analysis aims at identifying the RBPs involved in the regulation of a given set of exons. For this, one wants to identify motifs that are found significantly more often in a set of given nucleotide sequences than expected by chance. Further, it has been shown that the binding position

of the RBP (before or after the exon) has a key role in splicing regulation (216).
The authors of rMATS have introduced a method called rMAPS (133) for identifying the RBPs' binding positions around skipped exons. The goal of rMAPS is to identify known motifs of RBPs that are significantly enriched in differentially regulated exons between two sample groups as compared to control (background) events. rMAPS takes a set of 6-mer/7-mer degenerate RBP's motifs and a rMATS output file from which significant and background exons are selected. The number of times each motif matches each sequence in a local region is computed as a motif enrichment score. This score is calculated in the upstream exon, upstream flanking intron, target exon, downstream flanking intron and downstream exon separately based on a sliding window (Fig 4.1) . The positional distribution of the score is plotted as a map, from which the user can visually identify enriched binding regions (Figure 4.7 A). To assist in the decision, local comparative enrichment is assessed by Wilcoxon's rank sum test between included (respectively excluded) exons versus background exons in each sliding window. The minimal p-values are returned separately for exons or their upstream/downstream flanking introns to identify specific binding position of the RBPs (Fig 4.2 A).
rMAPS provides an efficient strategy to identify binding positions of RBPs. However, we believe that the statistical framework can be improved to better assess the presence of RBP' motifs in a sequence with a more precise identification of their location. First, rMAPS does not correct for multiple testing. The sliding window process constructs a large set of p-values that have to be adjusted to reliably control the false discovery proportion. In addition, considering only the minimal p-values of the upstream/downstream flanking introns does not allow the user to precisely identify the RNA binding site according to a statistical metric.

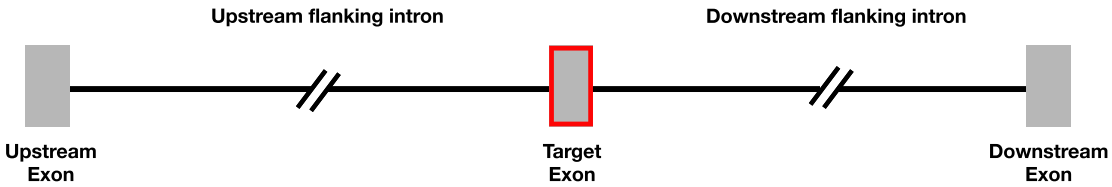


Figure 4.1: Schematic representation of an exon and its upstream/downstream exons.

In this work we propose to extend rMAPS method by computing a false discovery proportion (FDP) corrected for multiple testing for any set of successive p-values along the sequence. Binding regions of RBPs can then be defined as sets of p-values with FDP

controlled at a given level α.

4.2 The rMAPS algorithm

In order to properly assess the exact role of each RBP, rMAPS proposes to distinguish between included, excluded and background exons as follows. We denote by $\Delta\psi = \psi_{i1} - \psi_{i2}$ the difference in exon inclusion level (ψ) between the two conditions, and by q the FDR-adjusted p-value provided by rMATS. Included exons are defined as events with $q < \alpha$ and $\Delta\psi > c$. Conversely, excluded exons are defined as events with $q < \alpha$ and $\Delta\psi < -c$. In rMAPS, α and c are both set to 5%. Let μ_ψ^{i1} and μ_ψ^{i2} be the mean inclusion level ψ of exon i across all replicates of condition 1 and 2, respectively. Background exons are defined as non-significant events ($q > 0.5$) without inclusion level changes ($\min(\mu_\psi^{i1}, \mu_\psi^{i2}) < 0.85$ and $\max(\mu_\psi^{i1}, \mu_\psi^{i2}) > 0.15$). Those selected exons are filtered, in order to keep unique exons for included, excluded and background sets separately. Further, all exons common to different sets are discarded. The filtering step guarantees that motif occurrence are counted once for sequence around target exon and avoids confusion that presence of similar exons in different sets might introduce.
Up- and downstream flanking intronic sequences (Fig 4.2 B) of 300 nt are then retrieved. The consensus splice sites that drive exon recognition is excluded by removing the 20 nt sequence within the 3' splice site and the 6nt sequence within the 5' splice. Each motif is analyzed as follows. For each set of exons (included, excluded, background), this sequence data is summarized as an enrichment score, which is defined as the number of times the motif is present in a sliding window of 50nt (Fig 4.2 B), for the 250 successive window positions. An unilateral Wilcoxon's sum rank test is performed for each window, to assess the local enrichment of the motif in included or excluded exons compared to background exons. For each motif and each up- and downstream flanking intronic sequences, we obtain a set of 250 p-values that are spatially organized along the sequence. The smallest (raw) p-value is used to identify significant" enrichment of the RBP's motif in a given flanking intronic region. The mean enrichment score of each set of exons is plotted on a so-called RNA binding map, which is used to visually identify specific binding location of the RBP.

While rMAPS provides a very interesting strategy, we have identified some limits. First, each p-value only informs about the sliding window it comes from. However, rMAPS use it as an information concerning the entire sequence of 300 nt. Second, no

multiple correction is applied to the 250 p-values, which is prone to false discoveries as discussed in section 1.8.4.

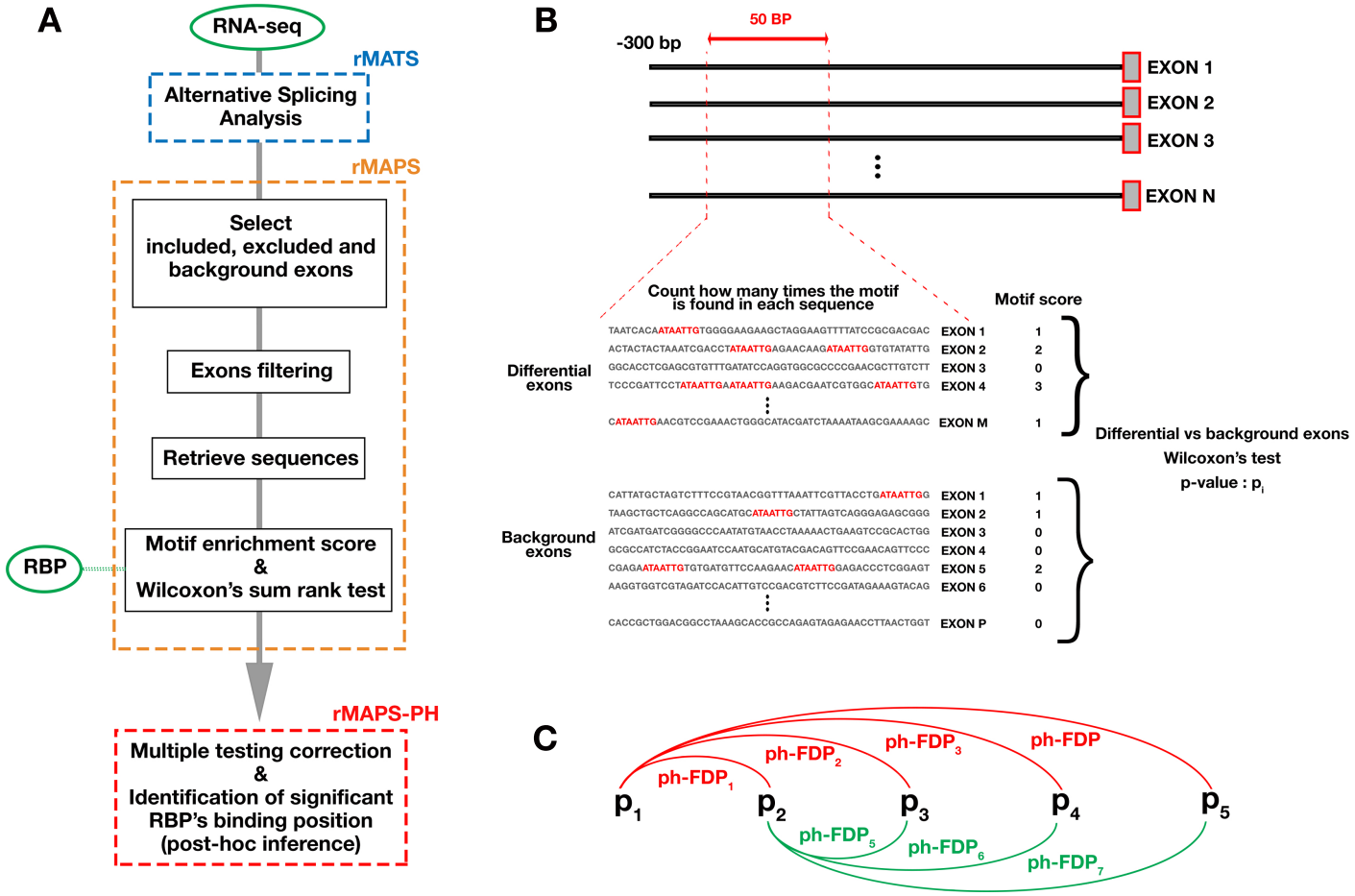


Figure 4.2: Extension of rMAPS algorithm. A) Overall workflow of rMAPS-PH. B) Computation of the motif enrichment score in each sliding window of 50 nt and the Wilcoxon's test associated. C) Computation of the post-hoc false discovery proportion. The post-hoc FDP is computed for all successive sets of p-values along the sequence.

4.3 Mathematical framework for post-hoc inference

4.3.1 Objective of the method

rMAPS algorithm computes a p-value for each sliding window along the sequence. One way to statistically identify RBP's binding position, is to identify sets of successive p-values along the sequence that are statistically significant. If we can define the false discovery proportion for each set of successive p-values, based on a user defined FDP threshold, we can identify regions where the motif enrichment score for included (respectively excluded) exons is significantly higher compared to background exons. In addition, the user has a statistical guarantee on the identified region. In this context, two issues are associated with the multiple testing correction.

1. Multiple sets of successive p-values have to be evaluated. Further, most of these sets are nested.
   As a consequence, one would like to build a statistical guarantee simultaneously on all subsets of p-values. Thus, the FDP can be controlled for each subset of p-values.

2. When computing the FDP on a subset of p-values from a larger set, we have to take care of the *selection effect*. When choosing a subset of items, one must ensure that the observed signal is real and not noise whose effect is amplified by the selection. That is why the overall size of the data set should be considered to compute the FDP on each selected subset.

Post-hoc inference, as introduced by Goeman and Solari (67), proposes to compute an upper bound on the number of false positives on any selected set of tested hypotheses. Classical multiple testing procedures propose to control the error rate set by the user, to determine significant testing hypotheses. In the *post-hoc* setting, the user provides a set of testing hypotheses to reject, and the post-hoc multiple testing procedure computes the error rate.

4.3.2 Post-hoc inference by closed testing procedure

The procedure of Goeman and Solari relies on the concept of *intersection hypothesis* and *closed testing* principle to control the FWER. Let us consider H a set of m null hypotheses to be tested and Ω the collection of all non empty subset of the index $\{1 \leqslant i \leqslant m\}$: $H_\Omega = \cap_{i\in\Omega} H_i$ for $\Omega \subseteq \{1 \leqslant i \leqslant m\}$. H_Ω is called an intersection hypothesis. Suppose m=3 with the hypotheses H_A, H_B, H_C. The collection Ω of intersect hypotheses is

composed of H_A, H_B, H_C, $H_{AB} = H_A \cap H_B$, $H_{AC} = H_A \cap H_C$, $H_{BC} = H_B \cap H_C$and $H_{ABC} = H_A \cap H_B \cap H_C$ (Fig 4.3 A). The intersection hypothesis H_{ABC} is true if and only if all hypotheses that compose it H_A, H_B, H_C are true.

We call *local test*, a valid α level test for every intersection hypothesis of H_Ω. The local tests produce a set of rejected (significant) hypotheses that are not corrected for multiple testing. In Fig 4.3 B, the hypotheses rejected by the local tests are the ones marked with a red cross. The closed testing approach rejects a hypothesis at level α if and only if it and every hypothesis that includes it in an intersection is rejected at level α. On the example Fig 4.3 C, H_A is rejected by the closed testing procedure because the four hypotheses $H_A, H_A \cap H_B, H_A \cap H_C$ and $H_A \cap H_B \cap H_C$ are all rejected by their local test. The closed testing procedure does not reject H_B because $H_B \cap H_C$ is not rejected by the local test. (113) proved that if the local test is a valid α-level test, then the closed testing procedure controls FWER at level α . The collection of rejected hypotheses by the closed testing procedure will be denoted by χ.

For any set of hypotheses R, Goeman and Solari define $\tau_\alpha(R)$ as the largest subset of R for which the corresponding intersection hypothesis is not rejected by the closed testing procedure:

$$\tau_\alpha(R) = \max\{\sharp\Omega : \Omega \in R, H_C \notin \chi\}$$

where Ω is the collection of intersection hypotheses from R and χ the rejected hypotheses by the closed testing procedure. The authors demonstrates that with probability $1 - \alpha$, the number of false positives in any such R is upper-bounded by $\tau_\alpha(R)$. This bound can therefore be used for post-hoc inference.

The approach of Goeman and Solari requires to perform a huge number of tests that is all $2^m - 1$ possible interactions between the m hypotheses. In practice, this is computationally inefficient when m becomes large. The authors have developed so-called "shortcuts", which yield computationally efficient procedures at the price of increased conservativeness and narrower applicability due to restrictions on the form of the test. In particular, when the p-values are independent or satisfy a particular form of positive dependence, called PRDS for positive regression dependence (see (20)), a post hoc procedure can be derived using Simes' local tests. However, our approach for motif enrichment analysis requires to look at a huge number of p-values sets (more than 30,000) and the implementation of Goeman and Solari approach in the R package "cherry" does not suit for this type of case. Blanchard, Neuvial & Roquain (24) have introduced a

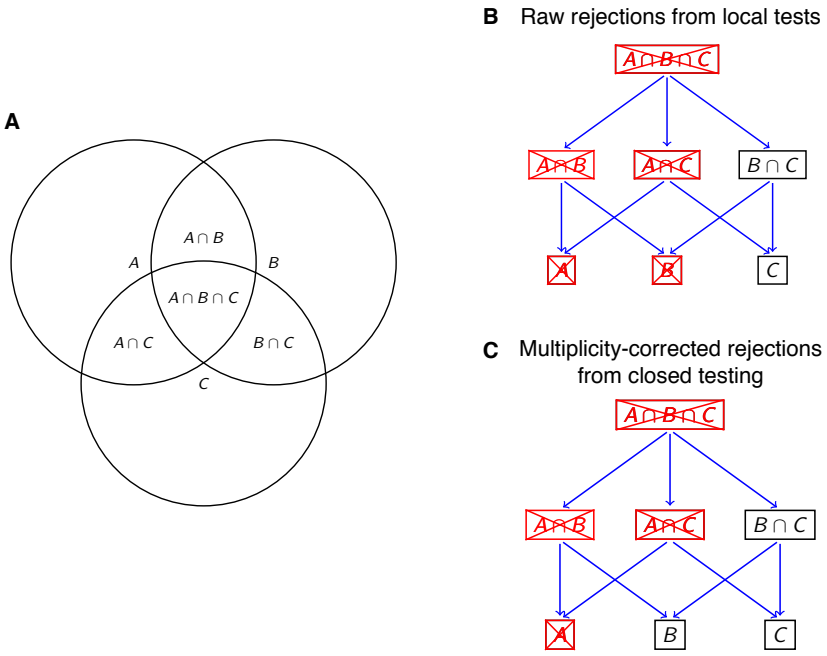


Figure 4.3: Intersection hypothesis and closed testing principle. A) Generation of all intersection hypotheses from a set of three hypotheses A,B and C. B) Rejected intersection hypotheses according to local tests. Hypotheses marked with a red cross are the rejected hypotheses before multiple testing correction. C) Rejected intersection hypotheses according to closed testing procedure. After multiple testing correction, only $H_{ABC}, H_{AB}, H_{AC}, H_A$ and H_B are rejected. Figures from Jelle Goeman talk at GDR Statistique et Santé, 2013/06/24

procedure based on a novel risk measure called the *Joint Risk* (JR) that gives the same bound and that can be computed more easily.

4.3.3 Post-hoc inference by controlling the Joint Risk

Let us consider H_0, the unobserved subset of H corresponding to the true null hypotheses (ie the non-significant tests) and m_0 the unobserved number of true null hypotheses. Let us denote by $(p_i)_{1\leqslant i\leqslant m}$ the p-values associated to each hypothesis of H and $(q_i)_{1\leqslant i\leqslant m_0}$ the p-values of each hypothesis of H_0. We denote by $(p_{(i)})_{1\leqslant i\leqslant m}$ and $(q_{(i)})_{1\leqslant i\leqslant m_0}$, the ordered p-values.

The FWER is defined as the probability to at least one false rejection. Romano and Wolf (152) have proposed a generalization called the k-FWER that controls the proba-

bility to do k or more false rejections : $\mathbb{P}(|R \cap H_0| > k-1)$. Now suppose that we want to construct k collections of rejection hypotheses (R_k) for k lists of hypotheses nested in H. These procedures must be adapted because of the selection effect. The JR control introduced by Blanchard, Neuvial & Roquain (24), proposes to simultaneously control the k-FWER for all k (Equation 4.1).

Let $(R_k)_{1\leqslant k\leqslant m}$ be a nested list of k candidate hypotheses. $(R_k)_{1\leqslant k\leqslant m}$ is said to control the JR at level $\alpha \in [0,1]$ if

$$\mathbb{P}(\forall k \in \{1,\dots,m\}, |R_k \cap H_0| > k-1) \leq \alpha \tag{4.1}$$

The JR control ensures that for each k, the probability that R_k contains more than $k-1$ false rejections is less or smaller than α. The authors demonstrated that by controlling the JR, we can control the number of false positives in any number of arbitrary sets of selected hypotheses $(R_k)_{1\leqslant k\leqslant m}$ simultaneously.

Given a JR controlling family $(R_k)_{1\leqslant k\leqslant m}$, the upper bound on the number of false positives with probability smaller than α, for any hypotheses sets R is given by

$$|R| \wedge \min_{i\in 1,\dots,|R|} \{|R \cap (R_k)^c| + k - 1\} \tag{4.2}$$

In order to construct the rejection sets, a thresholding-based rejection approach using the Simes' inequality can be defined.

Simes' Inequality : If the p-values $(p_i)_{1\leqslant i\leqslant m}$ are independent or under specific positive dependence (PRDS) then,

$$\mathbb{P}(\exists k \in \{1,\dots,m_0\} : q_{(k)} \leqslant \alpha k/m_0) \leq \alpha \tag{4.3}$$

Equation 4.3 guaranties that under PRDS, the set $R_k = \{1 \leqslant i \leqslant m : p_i \leqslant \alpha k/m\}, 1 \leqslant k \leqslant m$ satisfies the JR control at level α.

Finally, for any set of hypotheses $R \subset \{1,\dots,m\}$, the post-hoc bound at level α on the maximum number of false positives is given by

$$t_\alpha^{Simes}(R) = |R| \wedge \min_{k\in 1\leq\dots\leq|R|} \{\sum_{i\in R} \mathbb{1}_{\{p_i > \alpha k/m\}} + k - 1\} \tag{4.4}$$

Equation 4.4 give the same bound than the one introduced by Goeman and Solari (24) that is simpler to implement.

4.3.4 Application to motif enrichment analysis.

Recall that for each RBP and each flanking intronic sequence, we have a collection of 250 p-values, from which we want to identify subsets of successive p-values along the sequence, with a desired false discovery proportion. These sets correspond to regions where the RBP's motif can be observed significantly more in the included (respectively excluded) exons compared to the background exons. To do this, we just have to compute the post-hoc bound defined in 4.4 for all successive sets of p-values along the sequence, from which we can derive post-hoc false discovery proportion (ph-FDP, Fig 4.2 C). Based on this statistical measure, RBP binding location can be defined according to a given level α desired by the user.

4.4 Implementation

The current version of rMAPS is a web-server from which only the RBP binding maps and the smallest p-values can be retrieved. In this work, we introduce an improved version of rMAPS, called rMAPS-PH. rMAPS-PH includes a novel implementation of rMAPS implemented in R that gives more flexibility to the user, along with the post-hoc procedure to identify significant RBPs and their binding sites. We re-encode the construction of the post-hoc bound proposed by BNR2017 (24) in C++, from the R implementation available in the SansSouci package (`https://github.com/pneuvial/sanssouci`). The whole post-hoc procedure was implemented in C++ and integrated in R with RCPP package (54). Parallel computing is possible to speed up the computing execution. Today, the complete method consists of two scripts that will be further released and made available online as an R package.

The algorithm start with a rMATS output file from which inclusion, exclusion and background exons are selected. The flanking intronic sequences of these exons are then retrieved with SAMtools (102) based on a reference fasta file (ie Hg19). The rMATS FDR p-value level and the difference in exon percentage of inclusion ($\Delta\psi$) are set to 5%. We look for motif enrichment in sequence of 300 nucleotides in the flanking intronic sequence. Those values are set by default, but they can be parameterized by users. Compared to rMAPS, our implementation gives more flexibility. The RBP investigated are those provided by rMAPS, but users can input additional motifs. Using a sliding window of 50 nt, motif enrichment scores of each collection of exons are computed separately, in their upstream and downstream introns. Similarly to rMAPS, we slide the window by 1 nt at a time. The mean motif enrichment score of included, excluded

and background exons are recorded in each sliding window and used to plot the RNA binding map (Fig 4.4 A).

To identify significant locations of binding site, we first perform a unilateral Wilcoxon's rank sum test for each sliding window that compares motif enrichment scores between included (respectively excluded) versus background exons. We return a map of the distribution of p-values along the sequence (Fig 4.4 B). Then, the ph-FDP is computed for all sets of successive p-values. By default, all sets with ph-FDP $< \alpha = 0.25$ are called significant. As a consequence, many nested set of p-values are evaluated and can pass the user-defined threshold α. We choose to report the largest set of p-values that pass the post-hoc threshold with the largest average Wilcoxon's statistic test. The predicted binding site is reported on both RNA binding map score and p-values (Fig 4.4 C and D). We called the complete procedure "rMAPS-PH"

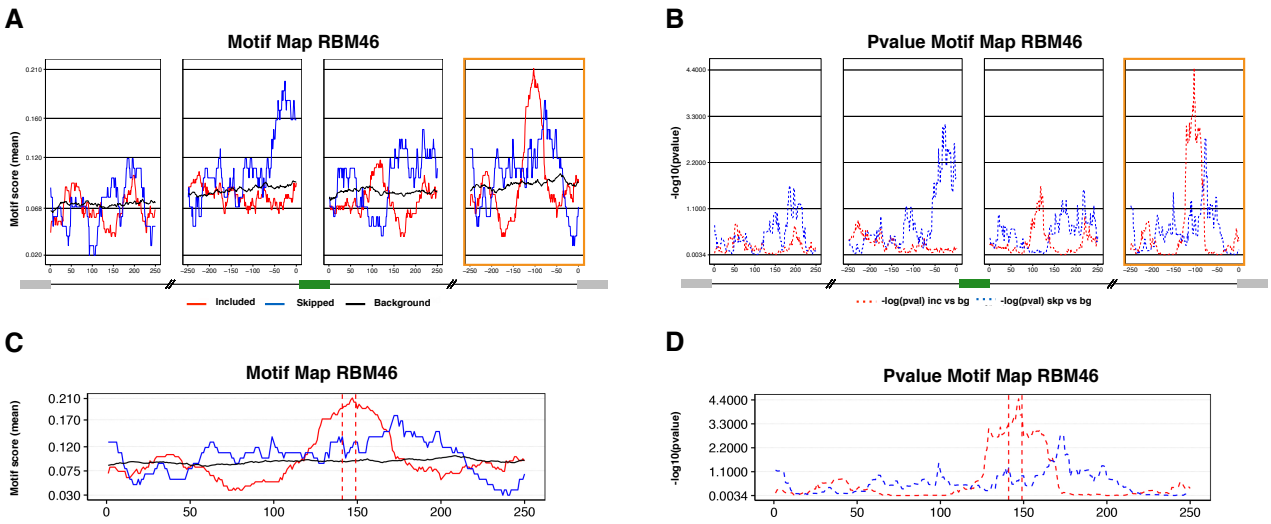


Figure 4.4: RNA binding map for RBP RBM46. A) The motif enrichment score distribution along the sequence for including (red), excluding (blue) and background (black) exons. B) The Wilcoxon's p-values distribution along the sequence for comparison between included (red) or excluded (blue) exons versus background exons. Y-axis is the $-\log_{10}(p$-value). C) and D) correspond to the 250 nucleotides before the downstream flanking exon (frame highlighted in orange). Vertical dashed red lines show the significant binding region identified by the post-hoc inference that contains less than 25% of false positives.

4.5 Results

4.5.1 Application to neoadjuvant-resistant triple-negative breast cancer samples

We compared the classical rMAPS and our rMAPS-PH method to identify RBPs involved in the regulation of the many exons detected between pre- and post-neoadjuvant triple negative breast cancer tumors (see Chapter 3). In the RBP file provided by rMAPS, several RBPs share the same motif and some RBPs may have multiple motifs. We do not know how classical rMAPS deals with these cases. In our approach, we grouped the RBPs with the same motif and distinguished when a RBP had several motifs. While classical rMAPS gives the smallest p-value for the whole upstream or downstream intronic sequences studied, we report a result for each set of 250 nt sequences near flanking and spliced exons. This approach makes it possible to more precisely identify the RBP binding site.

We first compared the recovered exons by each method considered to be significantly included, excluded and background. Both methods recovered the same exons included (n=256) and excluded (n=101). In terms of background exons, our method recovered 7 additional exons (rMAPS = 8706, rMAPS-PH = 8713). The visual inspection of the motif maps obtained showed very similar results between the two methods (Supplementary Figure 4.7). However, we can note that the scores and the p-values differ.
The classic version of rMAPS identified 106 unique RBPs significantly enriched, with an uncorrected p-value at 5%, in upstream and downstream introns. rMAPS-PH identified 47 unique RBPs with a ph-FDP of 25%. These RBPs have also been detected by classical rMAPS. However, the calling (ie whether the RBP includes or excludes) was different between classical rMAPS and rMAPS-PH for 29 out of the 47 RBPs identified in common (Figure 4.5). We observed 4 RBPs that are significantly enriched in different regions (upstream or downstream), 9 RBPs with different calling (same region but different behavior : include/exclude) and 16 RBPs with different regions and calling. Indeed, classical rMAPS has detected many more significantly enriched regions than rMAPS-PH. The greater number of RBPs and regions identified by rMAPS is not surprising because no multiple correction is performed, leading to an increase in false discoveries. In addition, the large number of similar regions referred to as included and excluded by rMAPS (in purple on Figure 4.5) also suggest poor control of the false positive rate. It is unlikely that so many RBPs have two different functions when they bind to the same position.

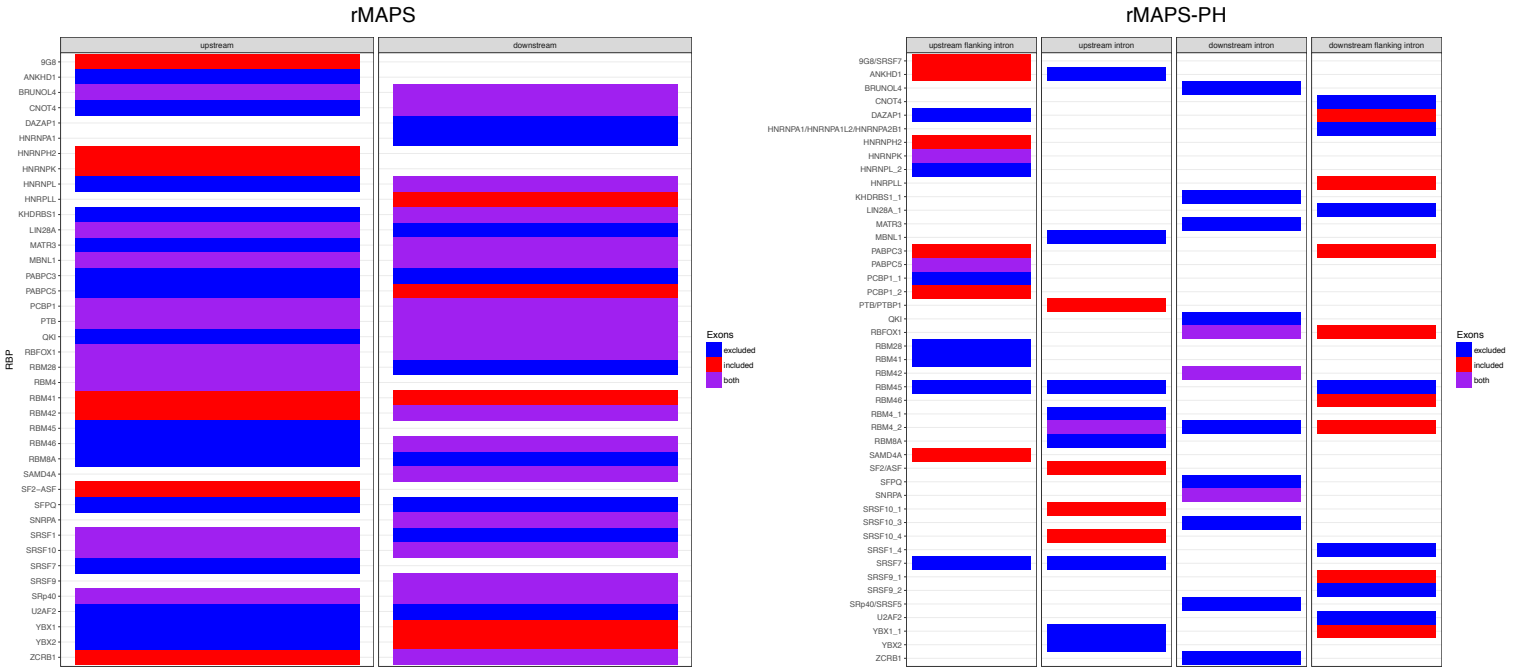


Figure 4.5: Calling for the 47 RBPs detected in common by classical rMAPS and rMAPS-PH. Motif enrichment detected in upstream or downstream intronic sequences. The color indicates the role of the RBP on the regulated exons when binding to the given location. Blue for exclusion, red for inclusion and purple for both.

4.5.2 Application to Ewing's sarcoma cell lines

In this section we introduced the results of our collaboration with Olivier Saulnier on Ewing's sarcoma cell lines.

Ewing's sarcoma is an aggressive cancer of bone and soft tissues affecting children and young adults. It is driven, in 85% of cases, by a chromosomal translocation, which generates the chimeric transcription factor EWS-FLI1. Previous studies have indicated that in addition to be a transcriptional regulator, EWS-FLI1 also impacts splicing. To further explore the function of EWS-FLI1 in splicing regulation, Saulnier et al performed transcriptome-wide splicing events analysis on multiple Ewing's sarcoma cell lines following EWS-FLI1 knock-down. We used rMAPS-PH to identified RBPs that regulate alternative splicing in these cells. This work has not yet been published, therefore the results were anonymised for confidentiality issues.

Saulnier et al have shown that nearly 300 genes were identified significantly altered at exon level, depending on EWS-FLI1 expression. They assume that EWS-FLI1 has an indirect function via partners (RBPs). In order to investigate how EWS-FLI1 modulates alternative slicing, we conducted RBPs motif enrichment analysis with rMAPS-PH. Some RPBs' motif appear to be significantly enriched in intronic sequences flanking alternatively spliced exons upon EWS-FLI1 status (Fig 4.6 A and B). They studied the impact of one of the significant RBP identified, called here $RBP\alpha$, on the splicing mediated by EWS-FLI1, using RT-PCR. They selected a set of exons from the 300 identified with significant splicing shift and compared the isoform abundance of these exons upon EWS-FLI1 or $RBP\alpha$ inhibition. It appears that EWS-FLI1 regulate inclusion of these exons while $RBP\alpha$ regulates skipping of this exon (Fig 4.6 C). Previous studies on myoblasts (normal muscle cells) have demonstrated that $RBP\alpha$ includes the exon when it binds in the upstream intronic sequences and skip the exon when it binds in the downstream intronic sequences (Fig 4.6 D). Our method detects two regions in downstream intronic sequences enriched in $RBP\alpha$ motif (Fig 4.6 E). One of the region detected is concordant with the one previously reported that suggests $RBP\alpha$ includes the regulated exons when it binds in the downstream sequences (red peak). The second region detected is located just after the regulated exons. At this position, $RBP\alpha$ skip the regulated exons suggesting that EWS-FLI1 inhibits the normal function of $RBP\alpha$ in Ewing's sarcoma cells. This study suggests that EWS-FLI1 may interact with $RBP\alpha$ in order to inhibit its function on alternative splicing in Ewing cells.

4.6 Discussion

The classical rMAPS integrates RNA-seq results of differential alternative exon regulation, with information of RBP motif occurrence to understand how these differential AS events are regulated. The main objective of rMAPS is to identify RBPs with significant binding site enrichment and potential position-dependent regulatory roles. However, we believe that the current proposed methodology does not reliably address the problem. First on the correction of multiple tests, and on the other hand in the identification of binding sites.

We have extended the statistical framework of rMAPS with post-hoc inference to correct for multiple testing. Our approach allows the user to identify RBPs that regulate

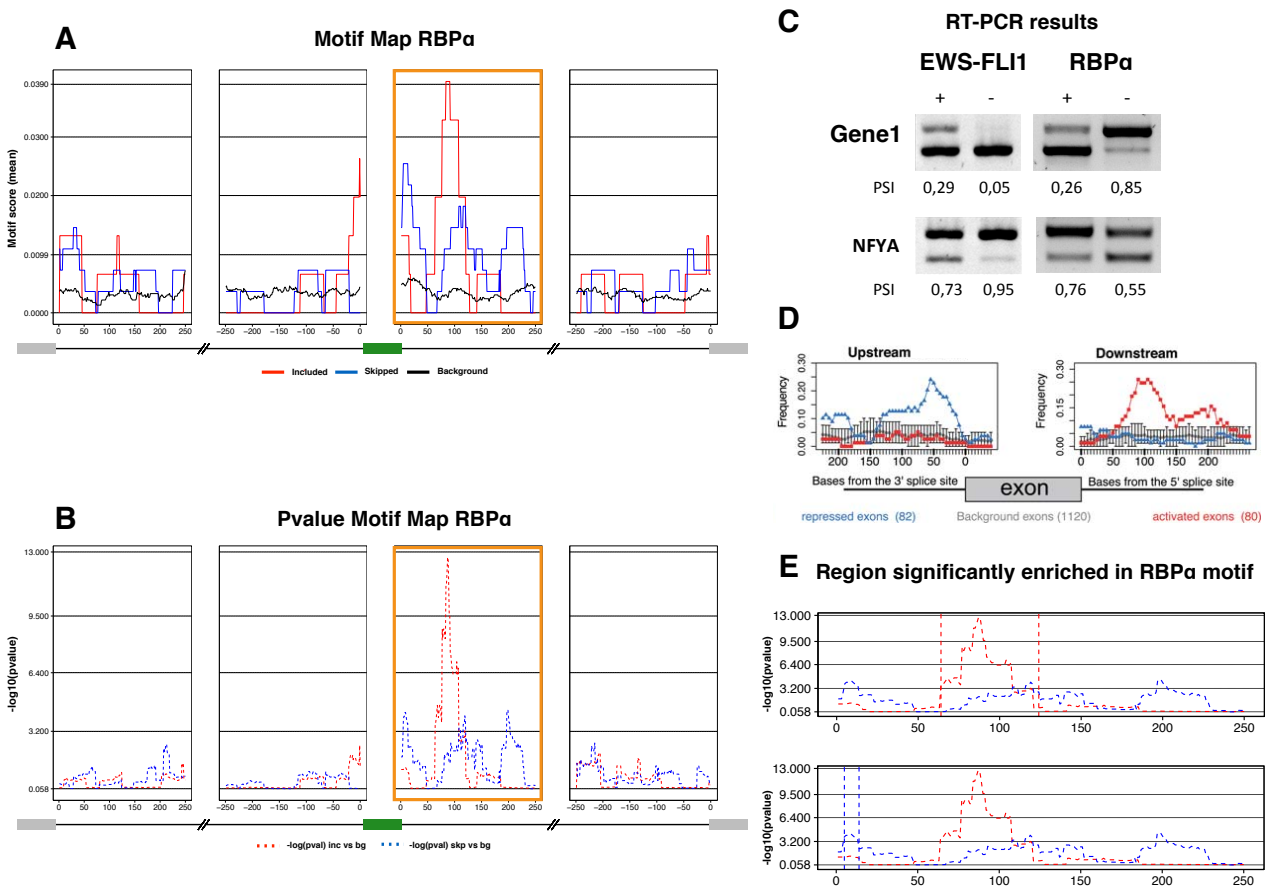


Figure 4.6: The roles of $RBP\alpha$ in Ewing's sarcoma A) Motif map of scores for $RBP\alpha$, B) Motif map of p-values for $RBP\alpha$. C) RT-PCR results for EWS-FLI1 and $RBP\alpha$. D) Physiological behavior of $RBP\alpha$ in myoblasts cells. E) Significantly enriched regions found with rMAPS-PH (frame highlighted in orange). The vertical dotted lines delineate the identified binding site in the upstream sequence.

skipped exons against a set of background exons. Our implementation in R allows more flexibility for the user than classical rMAPS web server. Furthermore, our approach reliably correct for multiple testing and allows the identification of the precise binding location of these RBPs. We have demonstrated that we have identified reliable RBPs that have been validated biologically. Further investigations have to be done to validate the binding sites.

The rMAPS part was implemented according to some information that we were able to recover by exchanging emails with the authors. We can see that we have little dis-

crepancy with the classical rMAPS and further research will be done to estimate their impact from a biological point of view.

The current implementation only works with an output file of rMATS as input. However, the method can be easily derived to fit other common methods of alternative splicing analysis such as DEXSeq or Voom. The actual implementation of rMAPS-PH has limits especially with regard to computing time. For example, based on the 9070 exons selected in Section 4.3.2, it took about 50 minutes to complete the entire process (motif extraction and motif enrichment) on 10 cores with CentOS 7, with Intel Xeon at 2,20GHz. More accurate tests are still needed to identify the sources of the slow-down. Preliminary studies have shown that computational time depends on the number of exons selected. The greater the number of exons for which the intronic sequences are to be recovered, the longer the execution time. We should consider whether similar results can be obtained by using a smaller number of randomly selected controls. The available implementation today tests all successive sets of p-values along the sequences. This leads to calculating the ph-FDP for more than 30,000 sets of p-values. Since all sets are nested, it may not be necessary to calculate the ph-FDP for each set. Future research will attempt to address this issue.

In our analysis, we used a 25% significance threshold for the post-hoc inference. This threshold is an upper bound on the number of false positives present in each set of p-values called significant. We expect this value to only affect the length of the region called significant rather than the call itself. Our current approach uses the Simes' local test to compute the post-hoc bound of false positive. To use this approach we must assume that our p-values satisfy the PRDS assumption from (20). This assumption is quite commonly made in genomics, as it is the assumption under which the widely-used Benjamini-Hochberg method controls FDR. However, the Joint Risk control has been introduced in a general framework in order to develop computationally feasible post hoc procedures that are adaptive to unknown dependence. The authors have demonstrated that Simes' local test can be improved using permutation tests. The permutation test could bypass the possible conservativeness of the JR control provided by Simes' inequality. We have already implemented the JR control using a permutation test adapted from the permutation algorithm for step-down minP adjusted p-values introduced by Ge, Speed and Dudoit (64). In the future, we intend to integrate this approach that could give more power to the method, which would yield similar results with a higher level of confidence.

4.7 Conclusion

Cancer specific splice variants of genes that control the cell proliferation and DNA damage, invasion and apoptosis have recently been identified (193). While the identification of such events is essential to better understand the disease, the ultimate goal is to develop therapeutic approaches that target these specific variants. Differential splicing analysis in comparative RNA-Seq experiments, identifies hundreds or even thousands of differential AS events. Rather than targeting the altered proteins, one could target the upstream mechanisms responsible for the occurence of these variants. The identification proteins involved in these mechanisms therefore seem crucial for developing a new therapeutic strategies.

Here we provide a tool to precisely identifying the RNA binding proteins involved in the regulation of a given set of differentially spliced exons between two conditions. In addition, the identification of precise binding sites makes it possible to design specific CLIP-seq experiments to validated the results.

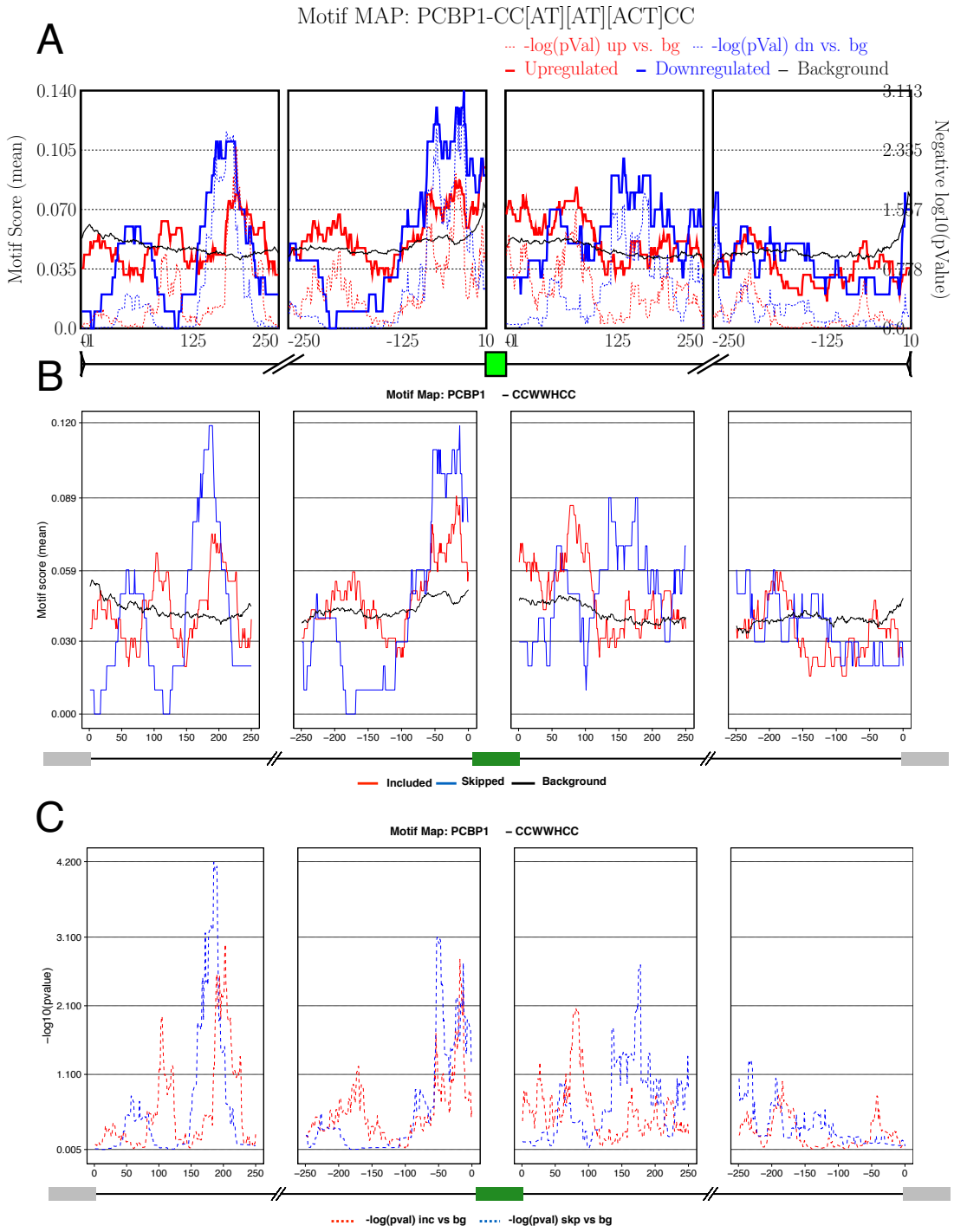


Figure 4.7: Comparison of classical rMAPS and rMAPS-PH for PCBP1. A) Results from rMAPS. Scores are in plain lines, p-values are in dashed lines. B) Motif map of scores obtains with rMAPS-PH. C) Motif map of raw-p-values obtains with rMAPS-PH. "CCWWHCC is the RBP's binding motif in IUPAC nomenclature.

5

Collaborations within RT2 Lab

During my PhD, I had the opportunity to collaborate on many projects. It was a great opportunity to work further with my colleges on subjects directly related to my PhD and to improve our exchanges to share our individual backgrounds. Other outside collaborations were launched after very good coffee breaks or lunch, with people from different teams or laboratories. These very enriching and successful collaborations are presented in the following two chapters.

In the following two studies, I have assisted in the curation, clean-up and normalization of data, participated in the development of the methodological and statistical framework and contributed to the interpretation of the results.

5.1 Biological network-driven gene selection identifies a stromal immune module as a key determinant of triple-negative breast carcinoma prognosis

This work is based on the study of Lehmann et al: 'Identification of human triple-negative breast cancer subtypes and pre-clinical models for selection of targeted therapies'. In this work, they developed a classification of TNBCs in which a signature of the 2188 gene was used to classify the tumors. They further suggest that this classification can be used to classify xenografts and cell lines. This study provides a significant insight into TNBC biology and was among the first to demonstrate the heterogeneity of TNBCs. However, in this study we found some key concerns. Authors have normalized the 21 sets of breast cancer data from different microarray platforms in a single process. Several studies report that many discrepancies and artifacts can be introduced by

pre-processing different array platforms together (167). The very large number of genes used to classify the samples can cause instability due to the noise introduced and lead to a lack of reproducibility (1, 167, 206). We also thought it would be unwise to use a tumor-based gene signature to classify *in vitro* and *in vivo* models. Indeed, the stromal environment is very different between tumors and xenografts and most of stromal genes are not expressed in cancer cell lines. The Lehman classification applied to pre-clinical models can then lead to misleading results.
We decided to refine this analysis. From the same 21 datasets, we have distinguished microarray platforms in a training (262 HGU-133A chips) and a validation (295 HGU-133Plus2 chips) set. From the training set, we constructed a gene signature based on biological networks to decrease the intrinsic instability of molecular classification methods. It has been shown that incorporation of biological knowledge improves the stability of gene selection and the biological interpretation of the signature(45, 62, 155). We developed a two-step biological network-driven gene selection process: 1) identification of the most variable genes displaying highly correlated patterns of expression, 2) gene filtering within known biological networks. The final signature composed of 167 genes, is then less sensitive to fluctuations compared to a broader signature. This six-metagene signature is enriched in different gene ontologies: two clusters were enriched in immunity genes, one in proliferation/DNA damage genes, one in AR pathway genes, and two in matrix/invasion genes. Hierarchical clustering was performed on the validation set and two external datasets (Ignatiadis (89) (n=314) and METABRIC (46) (n=254)). Six reproducible subgroups of TNBCs were independently identified with similar gene expression pattern in the four sets of data generated on different microarray technologies. We investigated the prognostic or predictive value of each metagenes with survival data from METABRIC dataset. Multivariate analysis showed that strong expression of the Immunity2 module was associated with good patient outcome. This result is consistent with recent studies that have demonstrated the key role of the immune component (from tumor-infiltrating lymphocytes to upregulation of cell immune-regulating pathways) in the clinical outcomes of various epithelial cancers (31, 61). We also compared Immunity2 metagene with other published immune signatures and demonstrated that it was strongly correlated with B-cell, T-cell and $CD8C^{+}$ cell signatures.

Today TNBC is the only subgroup of breast cancer for which no targeted treatment is available. TNBCs samples with high expression of Immunity2 metagene may define a subgroup of TNBCs for which the use of immunotherapy is a new therapeutic opportunity.

ORIGINAL RESEARCH

OncoImmunology 5:1, e1061176; January 2016; © Institut Curie

Biological network-driven gene selection identifies a stromal immune module as a key determinant of triple-negative breast carcinoma prognosis

H Bonsang-Kitzis[1,2,3], B Sadacca[1,2,4], AS Hamy-Petit[1,2], M Moarii[5,6], A Pinheiro[1,2], C Laurent[1,2], and F Reyal[1,2,3,*]

[1]Residual Tumor & Response to Treatment Laboratory; RT2Lab; Translational Research Department; Institut Curie; Paris, France; [2]U932 Immunity and Cancer; INSERM; Institut Curie; Paris, France; [3]Department of Surgery; Institut Curie; Paris, France; [4]Laboratoire de Mathématiques et Modélisation d'Evry, Université d'Évry Val d'Essonne; UMR CNRS 8071, ENSIIE, USC INRA, France; [5]Mines Paristech; PSL-Research University; CBIO-Centre for Computational Biology; Mines ParisTech; Fontainebleau, France; [6]U900, INSERM; Institut Curie; Paris, France

Keywords: immune signature, molecular subtypes, prognosis, triple-negative breast cancer

Abbreviations: AR, Androgen receptor; BC, Breast cancer; BC_CL, Breast cancer cell lines; BCSS, Breast cancer-specific survival; CCLE, Cancer Cell Line Encyclopedia; CDF, Consensus distribution function; CGP, Cancer Genome Project; EMT, Epithelial-mesenchymal transition; ER, Estrogen receptor; GE, Gene expression; HER2, Human epidermal growth factor receptor 2; IHC, Immunohistochemistry; IM, Immunomodulatory; LAR, Luminal androgen receptor; M, Mesenchymal; MSL, Mesenchymal stem-like; no CT, No chemotherapy; NPI, Nottingham Prognostic Index; pCR, Pathological complete remission; PR, Progesterone receptor; RMA, Robust multichip average; TILs, Tumor-infiltrating lymphocytes; TNBC, Triple-negative breast cancer; TNBC_CL, Triple-negative breast cancer cell lines.

Triple-negative breast cancer (TNBC) is a heterogeneous group of aggressive breast cancers for which no targeted treatment is available. Robust tools for TNBC classification are required, to improve the prediction of prognosis and to develop novel therapeutic interventions. We analyzed 3,247 primary human breast cancer samples from 21 publicly available datasets, using a five-step method: (1) selection of TNBC samples by bimodal filtering on ER-HER2 and PR, (2) normalization of the selected TNBC samples, (3) selection of the most variant genes, (4) identification of gene clusters and biological gene selection within gene clusters on the basis of String© database connections and gene-expression correlations, (5) summarization of each gene cluster in a metagene. We then assessed the ability of these metagenes to predict prognosis, on an external public dataset (METABRIC). Our analysis of gene expression (GE) in 557 TNBCs from 21 public datasets identified a six-metagene signature (167 genes) in which the metagenes were enriched in different gene ontologies. The gene clusters were named as follows: Immunity1, Immunity2, Proliferation/DNA damage, AR-like, Matrix/Invasion1 and Matrix2 clusters respectively. This signature was particularly robust for the identification of TNBC subtypes across many datasets (n = 1,125 samples), despite technology differences (Affymetrix© A, Plus2 and Illumina©). Weak Immunity two metagene expression was associated with a poor prognosis (disease-specific survival; HR = 2.68 [1.59–4.52], p = 0.0002). The six-metagene signature (167 genes) was validated over 1,125 TNBC samples. The Immunity two metagene had strong prognostic value. These findings open up interesting possibilities for the development of new therapeutic interventions.

Introduction

TNBC, defined by the absence of estrogen and progesterone receptor expression and a lack of *HER2* overexpression/amplification, is an aggressive disease accounting for 15%–20% of breast cancers. It differs from other molecular subtypes [1–3] in displaying axillary lymph node involvement, local and regional recurrence, differences in the time lag to metastasis (distant metastatic events occurring within 5 y of diagnosis), high rates of brain, lung and distant nodal metastasis and in its response to neoadjuvant treatment.

TNBC constitutes a major clinical challenge because there has been no substantial improvement in treatment for this subgroup in the recent past. Even if adjuvant chemotherapy has significantly improved outcome, reducing the risk of death by approximately 30%,[4] but these cancers do not respond to endocrine or targeted therapy. TNBC is, thus, currently the breast cancer subgroup with the worst outcome.[5] Moreover, the shape of the

*Correspondence to: F Reyal; Email: fabien.reyal@curie.fr
Submitted: 03/03/2015; Revised: 06/02/2015; Accepted: 06/08/2015
http://dx.doi.org/10.1080/2162402X.2015.1061176

survival curve for this subgroup differs from that for other BC subtypes: there is a sharp decrease in survival during the first 3–5 y after diagnosis, but distant relapses, occurring after this interval, are much less common.[5]

TNBC is a highly heterogeneous group of tumors differing in terms of their histological features, GE profiles, clinical behavior, overall prognosis [6] and sensitivity to systemic treatment.[7-9]

Robust classifiers are urgently required, to improve our understanding of the molecular basis of TNBC and to define novel therapeutic interventions. Lehmann et al. recently published a classification of six molecular subtypes of TNBC[10] and developed a website (http://cbc.mc.vanderbilt.edu/tnbc/) [11] for the classification of TNBC samples on the basis of their GE profiles. This classification has been shown to be relevant, as it identifies the main biological component and pathways of TNBC. However, the large number of genes defining this TNBC molecular classification (2,188 genes) constituted a potential source of instability.[12,13]

We developed a two-step biological network-driven gene selection process: (1) identification of the most variant genes displaying highly-correlated patterns of expression, (2) direct connection of these genes within known biological networks. This method has been reported to be efficient for the construction of molecular signatures.[14,15] We defined a robust TNBC molecular subtype classification, providing considerable biological insight, with great potential for use in the development of therapeutic interventions. We also identified a stromal immune module GE profile strongly correlated with TNBC prognosis.

Results

TNBC gene expression profiles identify six main gene clusters

GE profiles were obtained from 21 publicly available datasets, containing data for 3,247 primary human breast cancer samples. These profiles were processed according to the flow chart in **Fig. 1**. The training set included samples hybridized on HGU-133A Affymetrix© arrays (12 datasets, $n = 1{,}995$), to eliminate cross-platform discrepancies and to ensure robust normalization. The validation set included samples hybridized on HGU-133Plus2 Affymetrix© arrays (9 datasets, $n = 1{,}014$). We filtered out 42 outlier samples from the training set and 17 from the validation set.

We also collected two large datasets, for the validation of our classification: the Ignatiadis set ($n = 996$) and the METABRIC set ($n = 1{,}992$). The processing of these two datasets has been described elsewhere.[16,17]

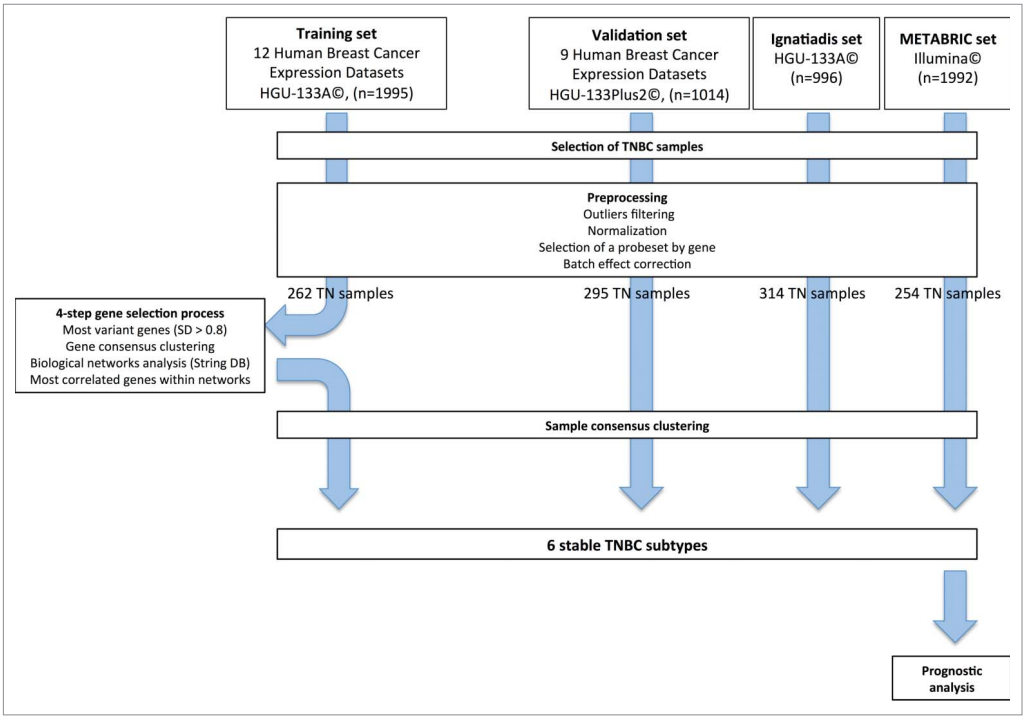


Figure 1. Methodology flow chart.

5.1 Biological network-driven gene selection identifies a stromal immune module as a key determinant of triple-negative breast carcinoma prognosis

Bimodal filtering on ER-PR and HER2 GE identified 262, 295, 314 and 254 TNBC samples in the training set, the validation set, the Ignatiadis set and the METABRIC set, respectively.

We developed a gene selection process based on biological networks, to decrease the intrinsic instability of molecular classification methods.

We identified the 830 most variant genes (SD > 0.8) in the training set ($n = 262$). A consensus clustering method and hierarchical clustering identified four main gene clusters. Further increases in cluster number yielded no significant increase in the consensus distribution function (CDF) area (**Fig. S1** and Materials and Methods).

The various gene clusters were associated with different gene ontologies (**Fig. S2**). The clusters were thus named as follows (**Fig. S3A**): Immunity cluster (145 genes), Proliferation/DNA damage cluster (397 genes), AndrogenReceptor(AR)-like cluster (139 genes) and Matrix/Invasion cluster (149 genes).

The Immunity cluster was the most homogeneous, with strong correlations between the GE profiles of most of the genes within this cluster (**Fig. S3B**).

We used String© database software to analyze our gene selection, with the aim of decreasing the heterogeneity of each main gene cluster. We retained the genes from our initial selection that (1) had high String© database gene connection indexes (greater than 0.7, **Fig. S4**), (2) had similar patterns of expression to other genes within the same biological network (correlation coefficient of at least 0.5). We selected a final set of 167 genes [Immunity cluster (80), Proliferation/ DNA damage (15), AR-like(15), Matrix/Invasion (57)] (**Fig. S5**).

Following biological network-driven gene selection, it became clear that the original Immunity and Matrix/Invasion clusters were more accurately described by splitting them into two subclusters displaying minor differences [Immunity1 (33), Immunity2 (47), Matrix/Invasion1 (43), Matrix2 (14)] (**Fig. S6A**). This approach yielded an increase in the area under the CDF curve (**Fig. S7**).

For each of the six gene clusters identified in this way, we defined a metagene. The Immunity1 and Immunity2 metagenes displayed similar patterns of expression, with a Pearson correlation coefficient of 0.58; the Pearson correlation coefficient for the expression patterns of Matrix/Invasion1 and Matrix2 was 0.48. The Proliferation/DNA damage and Matrix metagenes displayed the strongest inverse correlation (coefficients of −0.43 and −0.60 for Matrix/Invasion1 and Matrix2, respectively) (**Fig. S6B**).

We validated this six-gene cluster classification, by applying hierarchical clustering based on the 167 genes selected to the validation set ($n = 295$). Clustering was highly consistent between the training and validation gene sets (concordance: 93–100%).

The six gene clusters identify six stable TNBC subgroups

Hierarchical clustering was performed on the four TNBC datasets [training set (262), validation set (295), Ignatiadis (314) and METABRIC (254)]. For Affymetrix© arrays, we used the 167 selected genes. For the Illumina© platform, we used 153 common genes. We identified six reproducible subgroups of TNBC, for which GE patterns were similar in the training set and in the three validation sets (total of 1,125 samples). The corresponding heatmaps are shown in **Fig. 2**. The Pearson correlation coefficients for the relationships between each sample subgroup centroid in the three validation sets and the corresponding subgroup centroid in the training set are shown in **Fig. 2**.

We illustrated the dynamic links between genes within a biological network, as defined by the String© database, by showing GE levels for a "prototype sample" (**Fig. S8**).

We compared our sample classification with those reported by Lehmann et al. and Curtis et al. (**Fig. S9**). Our classification appears very different from that of Lehmann at first glance (χ^2 test p value = 0.05), but the samples assigned to Centroids one and six (with high-levels of Matrix/Invasion 1 and Matrix 2 gene expression, respectively) tended to be classified as Mesenchymal (M) or Mesenchymal stem-like (MSL), the samples in Centroid 5 (strong expression of Immunity2 genes) tended to be classified as Immunomodulatory (IM), and the samples in Centroid 4 (strong expression of AR-like genes) tended to be classified as of the Luminal androgen receptor (LAR) subtype (**Fig. S10A and Fig. S10B**). Curtis et al. aimed at defining a new classification across all cancer subtypes, not specific to TNBC subtypes. In this classification, the TNBC samples were mostly classified as IntClust10 or IntClust4, with an even distribution.

Prognostic value of the Immunity2 metagene in TNBC

The prognostic value of the 167-gene TNBC signature was assessed with the METABRIC dataset. The 254 TNBC samples were split into two subgroups: a subgroup treated by chemotherapy (n = 139) and a subgroup not treated by chemotherapy (n = 115). The chemotherapy-naive (noCT) population and the chemotherapy-treated population were significantly different (**Table S1**). The patients in the noCT population were older (mean age of 61.5 y vs. 50.1 y, $p < 1.2^{10-11}$), more likely to be postmenopausal (77% vs. 47%, $p = 5.38^{10-5}$), and their tumors were of lower grade ($p = 0.01$), with less lymph node involvement (81% vs. 17%, $p < 2.2^{10-16}$), a lower Nottingham Prognostic Index (NPI < 3.4, 17% vs. 2%, $p = 2.57^{10-5}$), and less cellularity ($p = 0.03$).

Univariate analysis identified three factors significantly correlated with a poor outcome (distant disease-free survival) in the chemotherapy-treated population: NPI > 5.4 (HR = 2.15 [1.28–3.60], $p = 0.003$); *p53* mutation (HR = 2.42 [1.15–5.09], $p = 0.02$); and weak Immunity2 metagene expression (HR = 2.59 [1.54–4.34], $p = 0.0002$) (**Table 1A**, **Fig. 3A**). We did not include *p53* mutation status in the multivariate model, due to missing data ($n = 79$). A NPI > 5.4 and low-levels of Immunity2 metagene expression were retained in the multivariate model and were significantly associated with a poor outcome (HR = 2.30 [1.36–3.89], $p = 0.002$; HR = 2.68 [1.59–4.52], $p = 0.0002$, respectively) (**Table 1A**). The combined variable, NPI score/Immunity2 metagene expression was found to be of particular interest. In a first model, a NPI score greater than 5.4 was associated with a worse prognosis: HR = 3.98 [2.00–7.92], $p = 8.72^{10-5}$. For patients with NPI scores of 5.4 or below, Immunity2 metagene expression discriminated between two

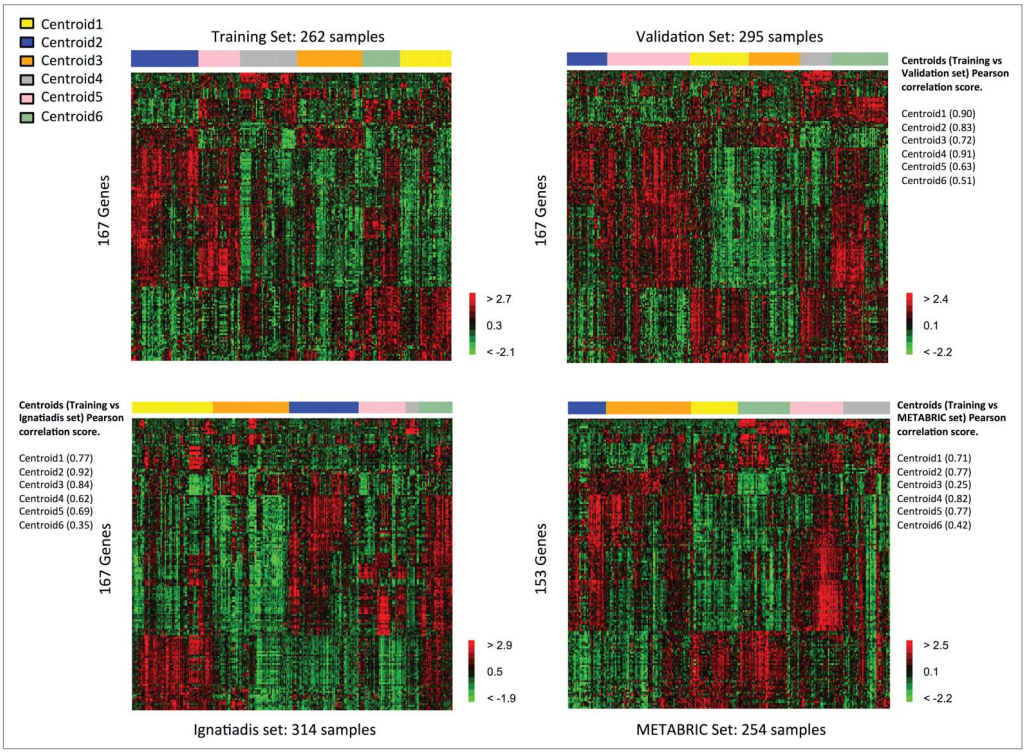


Figure 2. Heatmaps of the selected genes in the TNBC training set (upper left) and the TNBC validation sets (upper right: validation, lower left: Ignatiadis, lower right: METABRIC).

groups of patients with different outcomes (HR = 2.90 [1.51–5.56], $p = 0.001$). In a second model, NPI3 patients can also be split into two groups on the basis of Immunity2 metagene expression. The NPI3 group with high-levels of Immunity2 metagene expression had a prognosis similar to that of the NPI1/2 group with low-levels of Immunity2 metagene expression (**Table 1B**, **Fig. 3A**).

Univariate analysis identified four factors significantly correlated with poor outcome in the noCT population: tumor size >20 mm (HR = 2.36 [1.01–5.48], $p = 0.04$), lymph node-positive status (HR = 3.66 [1.65–8.11], $p = 0.001$), NPI score >5.4 (HR = 10.69 [2.74–41.76], $p = 0.001$) and low-levels of Immunity2 metagene expression (HR = 2.33 [1.09–4.95], $p = 0.03$) (**Table 2A**, **Fig. 3B**). Two of these factors were retained in the multivariate model: NPI score >5.4 (HR = 12.03 [3.05–47.50], $p = 0.0004$) and low-levels of Immunity2 metagene expression (HR = 2.42 [1.13–5.16], $p = 0.02$) (**Table 2A**). As in the chemotherapy-treated subpopulation, the combined variable, NPI score/Immunity2 metagene expression discriminated between two groups of patients with different outcomes in this noCT population (**Table 2B**, **Fig. 3B**). The chemotherapy-naive group contained only seven patients classified as NPI3. Stratification of this subgroup defined on the basis of treatment was therefore not considered methodologically relevant.

We compared the prognostic value of the Immunity2 metagene with that of eight previously published immune signatures,[18-25] using the METABRIC dataset.

We generated a heatmap (**Fig. S11**) of the GE profiles of each of the above prognostic signatures applied to the METABRIC dataset. The samples were ordered according to our classification of low/high Immunity2 metagene expression. Expression patterns were very similar between the Immunity2 GE signature and all the other GE signatures, with the exception of the Bianchini, Karn and Burstein (BLIS) gene-expression signatures.

We first performed a univariate analysis of the prognostic value of the eight-GE signatures, as described in the corresponding original manuscripts. The Rody, Sabatier, Teschendorff, Desmedt, Gu-Trantein Tfh, Gu-Trantien Th1 and Burstein signatures were significantly correlated with the prognosis of TNBC. The Bianchini and Karn GE signatures were not correlated with the prognosis of TNBC (**Fig. S12**, **Table S2**). We then performed a multivariate analysis. We included NPI score, the Immunity2 metagene and each of the Rody, Sabatier, Teschendorff, Desmedt, Gu-Trantein Tfh, Gu-Trantien Th1,

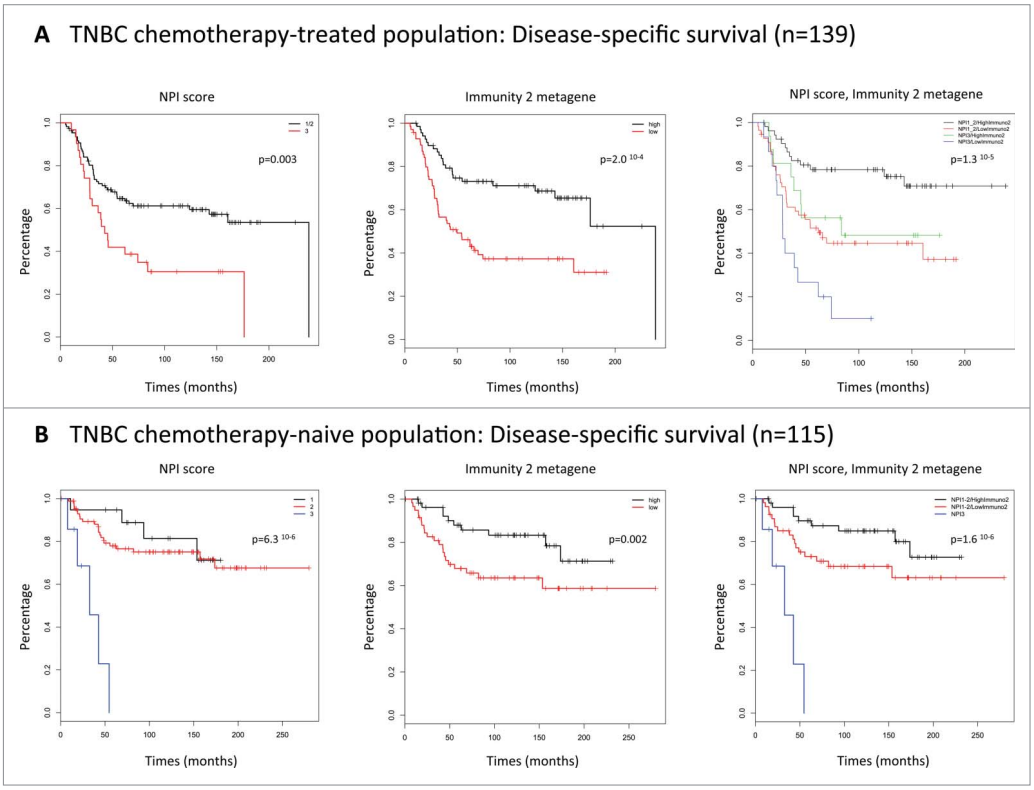


Figure 3. (**A**) Kaplan–Meier plots. Disease-specific survival of the chemotherapy-treated population ($n = 139$). NPI score. Immunity2 metagene. NPI score/Immunity2 metagene. (**B**) Kaplan–Meier plots. Disease-specific survival of the noCT population ($n = 115$). NPI score. Immunity2 metagene. NPI score/Immunity2 metagene.

and Burstein signatures, one-by-one, in the model. In all comparisons the only significant variables remaining in the multivariate model were NPI score and the Immunity2 metagene (**Table S2**).

The Immunity2 metagene corresponds to B-cell and T-cell pathways

String database connections between the Immunity1 or Immunity2 genes and the genes of the eight published prognostic immune signatures [18-25] are provided in **Fig. 4**. The gene intersection was poor, but our immune signature nevertheless appears to be strongly correlated with other published signatures (Supplementary data), suggesting the use of similar immune pathways. The Immunity2 metagene was strongly correlated with the expression metagenes of the above signatures (coefficient greater than 0.8), except for the Bianchini, Karn and BLIS metagenes (**Fig. S13**).

We explored the pathways relating to the immune metagenes in detail, by analyzing the correlation between the expression of the Immunity1 and Immunity2 metagenes and the metagenes defined by Gatza et al.[26] (IFN-α, IFNγ, STAT3, TGF-β, TNF-α) and Palmer et al.[27] (LB, LT, $CD8^+$, GRANS, LYMPHS). This analysis was performed on the METABRIC dataset published by Curtis et al.[17]

We showed that the Immunity2 metagene was highly correlated with the B-cell, T-cell and $CD8^+$ cell metagenes (Pearson correlation scores: 0.93, 0.91, 0.87, respectively) (**Fig. S14**). The Immunity1 metagene was highly correlated with the interferon alpha and gamma pathways (Pearson correlation scores: 0.97, 0.94, respectively).

Furthermore, in cancer cell lines (CCLE and CGP datasets), the Immunity2 metagene displayed very low-levels of expression, similar to those of the $CD8^+$ metagene (**Fig. S15**). This was true for all cell lines and BC_CLs tested.

Moreover, the IFN-γ, IFN-gamma, STAT3, TGF-β, TNF-α, LB, LT, GRANS metagenes were more strongly expressed in TN BC_CLs than in HER2-positive and luminal BC_CLs (**Fig. S16**).

We investigated Immunity2 GE in white blood cell populations (Palmer et al.[27]), by performing a consensus clustering of the Immunity2 genes on Palmer's dataset. This analysis identified four stable clusters of the genes of the Immunity2 signature. Some genes were more strongly expressed in B cells (GZMA, GZMB, CCR7, LY96, MS4A1, CD74 for example), others in T

Table 1 A. Survival analysis (disease-specific survival). Chemotherapy-treated population. Univariate and multivariate analysis.

		Univariate analysis		Multivariate analysis	
139 triple-negative breast cancer patients		**DS-survival HR [95% CI]**	***p* value**	**DS-survival HR [95% CI]**	***p* value**
Menopausal status	Pre	1			
	Post	1.56 [0.95–2.55]	0.08		
Tumor size (mm)	<20 mm	1			
	>20 mm	1.03 [0.58–1.82]	0.92		
Tumor grade	II	1			
	III	1.23 [0.45–3.39]	0.69		
Lymph node status	0	1			
	1	0.84 [0.42–1.65]	0.61		
NPI score	<5.4	1		1	
	>5.4	**2.15 [1.28–3.60]**	**0.003**	**2.30 [1.36–3.89]**	**0.002**
Cellularity	Low	1			
	Moderate	0.57 [0.22–1.46]	0.24		
	High	0.59 [0.25–1.39]	0.23		
P53 status	Wild-type	1			
	Mutant	**2.42 [1.15–5.09]**	**0.02**		
Immunity1 metagene expression	High	1			
	Low	0.97 [0.60–1.58]	0.91		
Immunity2 metagene expression	High	1		1	
	Low	**2.59 [1.54–4.34]**	**0.0002**	**2.68 [1.59–4.52]**	**0.0002**
Proliferation/DNA damage metagene expression	High	1			
	Low	1.13 [0.69–1.84]	0.63		
AR-like metagene expression	High	1			
	Moderate	1.07 [0.59–1.94]	0.82		
	Low	0.98 [0.50–1.94]	0.96		
Matrix/Invasion1 metagene expression	High	1			
	Low	1.23 [0.76–2.01]	0.40		
Matrix2 metagene expression	High	1			
	Low	0.99 [0.61–1.61]	0.96		

Abbreviations: NPI, Nottingham Prognostic Index; AR, androgen receptor; HR, hazard ratio; CI, confidence interval.

cells (CD3D, CCL2, CD14, CD2, LCK, IL7R), and still others in granulocytes (named Pax cells) (CD163, MNDA, NCF2, CSF2RB, FGL2) (**Fig. S17**). These findings suggest that, even if the "Immune2" signal is highly homogeneous within tumor samples (the entire set of genes being coordinately either over- or under-expressed), different subpopulations of cells express different subsets of these genes in the periphery.

The Immunity2 metagene is probably expressed by stromal cells

In TNBC cell lines (TNBC_CL), genes from the Immunity2 module displayed very low medians and narrow ranges of expression, suggesting that they were expressed only in the tumor stromal compartment. A similar trend was observed for all BC_CLs. The Immunity1 module genes had higher median expression levels and a broader range of expression in TNBC_CL and in all BC_CL, suggesting that Immunity1 genes were expressed by the tumor cells (**Figs. 5A** and **B**).

Furthermore, we explored the contributions of stromal and cancer cells to Immunity1 and Immunity2 expression in detail, by comparing our gene lists to the "stromal contribution to global GE evaluated in PDX RNAseq data", as defined by Isella et al.[28] The Immunity2 metagene had a very high stromal fraction, as for the Matrix/Invasion1 and Matrix2 metagenes. The Immunity1 metagene had a very low stromal fraction, like the AR-like and Proliferation/DNA damage metagenes (**Fig. S18**).

The Immunity2 metagene opens up interesting new possibilities for therapeutic interventions

To highlight the new opportunities for therapeutic intervention provided by this study, we represented the existing drugs (with or without US Food and Drug Administration approval) for each metagene (**Fig. S19 and** Supplementary data). Some are undergoing clinical investigation in patients with TNBC.

We explored the links between PD1, PDL1, CTLA4 (and their respective metagenes) and the Immunity2 metagene. We compared the Immunity2 metagene with the TILs signature defined by Schalper et al.,[29] who showed that PD-L1 mRNA synthesis was associated with increases in the expression of TILs and recurrence-free survival. This analysis was performed on the METABRIC dataset. The PD1 and CTLA-4 metagenes were constructed from the genes most strongly correlated with the PD1 and CTLA-4 genes, respectively (Pearson correlation score >0.8). The PDL1 metagene was defined by Sabatier et al.[30]

The Immunity2 metagene was highly correlated with the PD1, PDL1 and CTLA-4 metagenes (Pearson correlation: 0.90, 0.96, 0.91, respectively). The coefficient of correlation between the Immunity2 metagene and the TILs signature was up to 0.90 (**Fig. S20**).

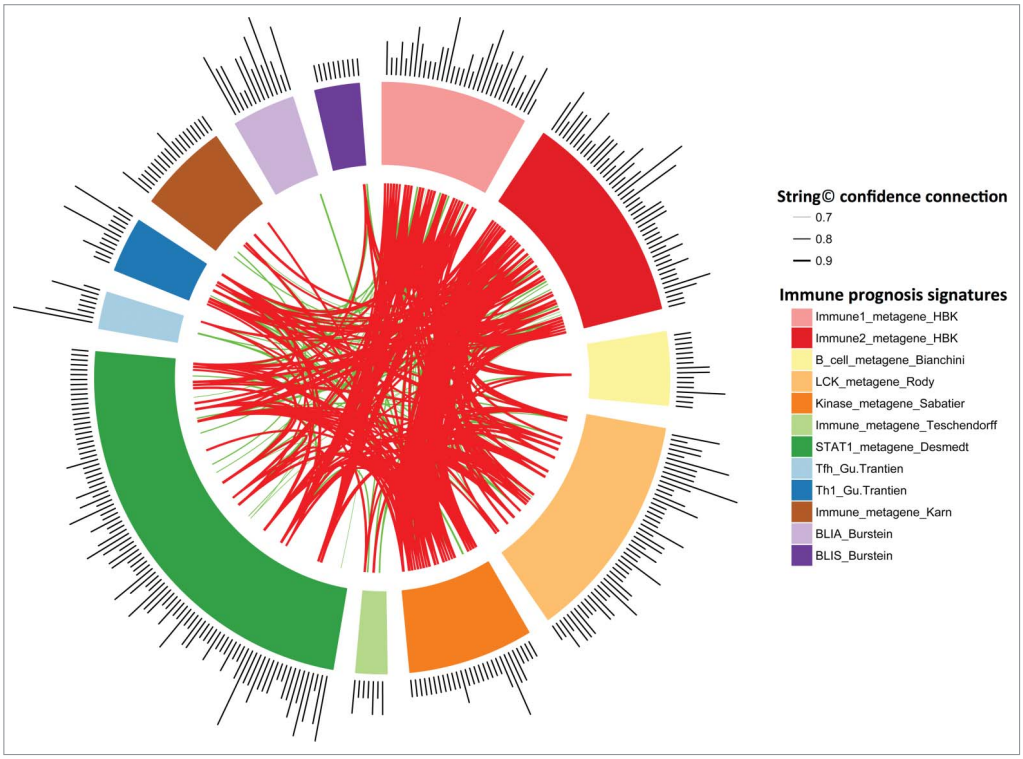


Figure 4. String Software connections between our Immunity1 and Immunity2 genes and the genes of eight previously published prognostic immune signatures. Stronger associations between genes are represented by thicker lines. Associations between genes with a coefficient < 0.9 are shown in green. Associations between genes with a coefficient ≥ 0.9 are shown in red. Associations between genes with a coefficient between 0.4 to 0.7 are not shown.

In cell lines, the PD1, PDL1, CTLA-4 and TILs metagenes were very weakly expressed, like the Immunity2 metagene (**Fig. S21**).

Using the METABRIC dataset, we compared the prognostic value of these metagenes (PD1, PDL1, CTLA4 and TILs) with that of the Immunity2 metagene. In univariate analysis, high-levels of PD1, PDL1, CTLA-4 and TILs metagene expression were associated with a good prognosis (**Fig. S22, Table S3**). In multivariate analysis, we included NPI score, the Immunity2 metagene and each of the PD1, PDL1, CTLA4 and TILs metagenes, one-by-one, in the model. In all comparisons, the only significant variables remaining in the multivariate model were NPI score and the Immunity2 metagene.

Discussion

New tools for classifying TNBCs are urgently required, to improve our understanding of the molecular basis of TNBC and to identify potentially useful novel therapeutic interventions. By analyzing the GE profiles of 1,125 TNBCs, we identified a six-metagene signature (167 genes) in which the various metagenes were enriched in different gene ontologies: two clusters were enriched in immunity genes, one in proliferation/DNA damage genes, one in AR pathway genes, and two in matrix/invasion genes. This signature appeared to be particularly robust for identifying TNBC subtypes across different datasets, independently of the gene chip technology used to generate the data. Furthermore, one metagene (Immunity2) was found to be of strong prognostic value for TNBC samples.

Lehmann et al.[10] recently developed a classification of TNBCs in which a 2,188-gene signature was used to classify tumors. They suggested that this classification could also be used to classify xenografts or cell lines. They also developed a website (http://cbc.mc.vanderbilt.edu/tnbc/) for the classification of TNBC samples.[11] This study provided important biological insight into the molecular drivers of TNBC, but it also raised several key concerns. First, the normalization process involved data from different platforms. Several studies have shown that large discrepancies in signature composition and absences of

Table 1 B. Survival analysis (disease-specific survival). Chemotherapy-treated population. Two univariate models. Combination of NPI score and Immunity2 metagene expression

139 triple-negative breast cancer patients		DS-survival HR [95% CI]	*p* value
NPI score/Immunity2 metagene expression	NPI1-2/HighI2	1	
	NPI1-2/LowI2	**2.90 [1.51–5.56]**	**0.001**
	NPI3	**3.98 [2.00–7.92]**	**8.72^{10-5}**
NPI score/Immunity2 metagene expression	NPI1-2/HighI2	1	
	NPI1-2/LowI2	**2.91 [1.51–5.59]**	**0.001**
	NPI3/HighI2	**2.31 [0.96–5.57]**	**0.06**
	NPI3/LowI2	**6.30 [2.89–13.78]**	**3.87^{10-6}**

Abbreviations: NPI, Nottingham Prognostic Index; I2, Immunity2; HR, hazard ratio; CI, confidence interval.
NPI = [0.2 × S] + N + G.
S: tumor size (cm).
N: number of lymph nodes involved (0 = 1, 1−3 = 2, >3 = 3).
G: tumor grade according to Elston and Ellis (Grade I=1, Grade II = 2, Grade III = 3).

concordance concerning outcome may be due to differences in the array platform and preprocessing method used.[12] Second, Lehmann et al. used a very large number of genes (2,188 genes) to establish their molecular signature, and this may have constituted a source of instability, due to the noise introduced.[12,13] As shown by Weigelt et al.,[31] microarray-based single-sample predictors do not allocate individual samples to a given molecular subtype reproducibly, probably because the use of large numbers of genes leads to instability of the classification when new samples are added. Third, it would be unwise to transpose this classification to various *in vitro* and *in vivo* breast cancer models (primary tumor xenografts, cell lines, cell line-derived xenografts), because the stromal environment and the original tumor are very different.[32,33] We found that genes from the Immunity compartment (Immunity2 module) were highly relevant for the classification of TNBC samples and that these genes were not expressed in

Table 2 A. Survival analysis (disease-specific survival). Chemotherapy-naive population. Univariate and multivariate analysis.

		Univariate analysis		Multivariate analysis	
115 triple-negative breast cancer patients		**DS-survival HR [95% CI]**	***p* value**	**DS-survival HR [95% CI]**	***p* value**
Menopausal status	Pre	1			
	Post	1.31 [0.56-3.06]	0.53		
Tumor size (mm)	<20 mm	1			
	>20 mm	**2.36 [1.01-5.48]**	**0.04**		
Tumor grade	I-II	1			
	III	1.33 [0.51-3.49]	0.56		
Lymph node status	0	1			
	1	**3.66 [1.65-8.11]**	**0.001**		
NPI score	<3.4	1		1	
	3.4-5.4	1.36 [0.47-3.96]	0.57	1.55 [0.53-4.51]	0.43
	>5.4	**10.69 [2.74-41.76]**	**0.001**	**12.03 [3.05-47.50]**	**0.0004**
Cellularity	Low	1			
	Moderate	1.91 [0.54-6.71]	0.31		
	High	1.42 [0.41-4.90]	0.58		
P53 status	Wild-type	1			
	Mutant	0.90 [0.17-4.63]	0.90		
Immunity1 metagene expression	High	1			
	Low	1.56 [0.76-3.19]	0.22		
Immunity2 metagene expression	High	1		1	
	Low	**2.33 [1.09-4.95]**	**0.03**	**2.42 [1.13-5.16]**	**0.02**
Proliferation/DNA damage metagene expression	High	1			
	Low	1.14 [0.56-2.32]	0.72		
AR-like metagene expression	High	1			
	Moderate	0.96 [0.42-2.20]	0.92		
	Low	0.74 [0.28-2.00]	0.56		
Matrix/Invasion1 metagene expression	High	1			
	Low	0.48 [0.23-1.01]	0.06		
Matrix2 metagene expression	High	1			
	Low	1.31 [0.64-2.66]	0.46		

Abbreviations: NPI, Nottingham Prognostic Index; AR, androgen receptor; HR, hazard ratio; CI, confidence interval.

Table 2 B. Survival analysis (disease-specific survival). Chemotherapy-naive population. Univariate analysis. Combination of NPI score and Immunity2 metagene expression.

115 triple-negative breast cancer patients		DS-survival hazard Ratio [95% CI]	*p* value
NPI score/Immunity2 metagene expression	NPI1-2/HighI2	1	
	NPI1-2/LowI2	**2.13 [0.95-4.78]**	**0.07**
	NPI3	**12.89 [4.07-40.82]**	**$1.37^{10^{-5}}$**

Abbreviations: NPI, Nottingham Prognostic Index; I2, Immunity2; HR, hazard ratio; CI, confidence interval.

BC_CLs. The observed lack of reproducibility between classifiers may reflect major differences in the methodology and aims of the studies concerned. Further validation will be required before these models can be used in routine clinical practice.

We developed a strategy for the definition of a GE signature based on the analysis of biological networks for the most variant genes. Within these networks, we then analyzed GE parameters, to select the genes with the most strongly correlated patterns of

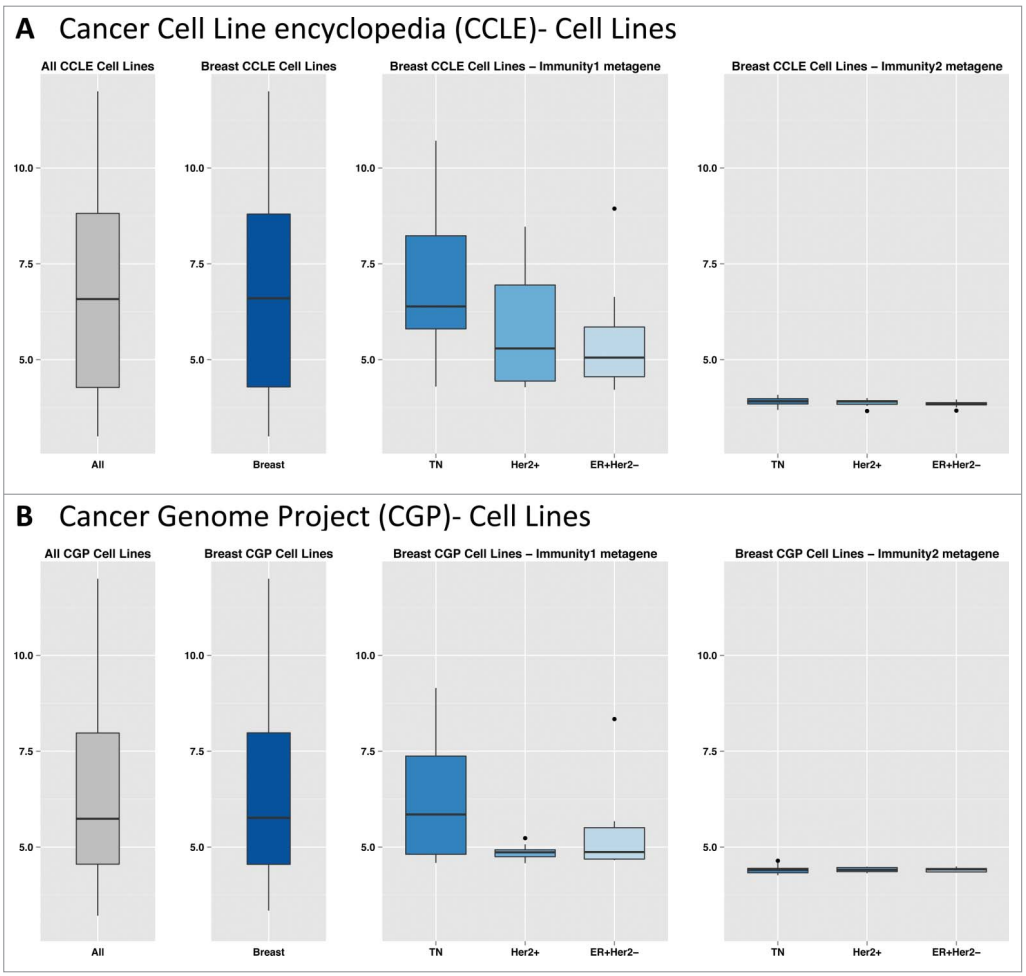


Figure 5. (**A**) Boxplots of gene expression for the Immunity1 and Immunity2 metagenes, in each breast cancer cell line subtype from the CCLE. (**B**) Boxplots of gene expression for the Immunity1 and Immunity2 metagenes in each breast cancer cell line subtype from the CGP.

expression. The validation process showed that our gene matrix identified similar GE patterns across 1,125 TNBC samples. This first step in biological network analysis, which is probably less sensitive to sample fluctuations than other methods, made it possible to capture strong biological signals that might be concealed by the noise present in microarray data. Several studies have reported that the incorporation of network information improves the stability of gene selection and the biological interpretability of biomarker signatures for a given prediction accuracy.[14,15,34]

The Immunity2 module was identified as a strong prognostic factor for disease-specific survival (strong expression of this metagene is correlated with a good outcome), regardless of the characteristics of the tumor (NPI score, tumor size, tumor grade and lymph node status). It clearly suggest the presence of an hemopoietic infiltrate, composed of activated cytotoxic T cells, B cells, myeloid cells, natural killer cells and neutrophils. This module includes adhesion molecule-associated genes (SELL, ITGB2), and genes encoding proteins involved antigen processing and presentation (CD74 or ligand, HLA-DRA), B-lymphocyte cell surface molecules (PTPRC, ITGB2, HLA-DRA), the caspase cascade (CASP1), complement pathway (C1QA, C1QB), CTL-mediated immune responses to target cells (ITGB2, CD3D, GZMB), dendritic cell regulation of Th1 and Th2 development (CD2, IL7R), granzyme-mediated apoptosis (GZMA, GZMB), IL12-mediated signaling events (CD3D, HLA-DRA, GZMA, LCK), the IL2 signaling pathway (LCK), interleukin-3, 5 and GM-CSF signaling (HCK, BLNK, CSF2RB), T-cell surface molecules (PTPRC, CD3D, CD2, ITGB2), and the T-cell receptor signaling pathway (PTPRC, CD3D, HLA-DRA, LCK).

Burstein et al.[25] identified four different TNBC subtypes (LAR, MES, BLIS, BLIA) with the identification of similar pathways and a prognostic value for the BLIA subgroup similar to that for the signature identified in our study. This subgroup displays an upregulation of B-cell, T-cell, and natural killer cell immune-regulating pathways and an activation of STAT transcription factor-mediated pathways. The authors showed that the prognosis was worse for basal-like immune-suppressed tumors than for basal-like immune-activated tumors, for both disease-free survival ($p = 0.04$) and disease-specific survival ($p = 0.039$).

Several recent studies have demonstrated the importance of tumor-infiltrating lymphocytes (TILs) in controlling the clinical progression of various epithelial cancers.[35] In breast cancer, recent advances in GE profiling have revealed an association between immune signatures and favorable outcomes.[29,36] A gene signature enriched in cytotoxic $CD8^+$ T-cell genes and genes associated with natural killer cell activity has been reported.[37] However, the ability of $CD8^+$ T cells to control human breast cancer is probably counteracted by the presence of immunosuppressive cells, $CD4^+$ T-regulatory cells or macrophages: immunohistochemistry (IHC) analysis of tissue microarray data for 179 treatment-naive breast tumors revealed that high-levels of macrophages and $CD4^+$ T cells were correlated with poor overall survival, whereas a combination of high-levels of $CD8^+$ T cells and low-levels of macrophages and $CD4^+$ T cells was correlated with higher overall survival.[38] Intratumoral B cells have also been associated with a favorable prognosis in breast cancer.[39] In ER-negative breast cancers, a STAT1 signaling metagene,[16] and a B-cell metagene[19] were found to be associated with better outcomes. Another group identified an immune response-based prognostic gene module (C1QA, XCL2SPP1, TNFRSF17, LY9, IGLC2, HLA-F) associated with a better prognosis than for other ER-negative breast cancers, regardless of lymph node status and lymphocytic infiltration.[40] According to Bertucci et al.,[41] the IM subtype (overlapping with medullary breast cancers, a rare form of TNBC with a prominent lymphocytic reaction) is associated with a favorable prognosis. The two immune modules identified in this study had many biological connections with other eight immune prognosis signatures published for TNBC.[18-25]

Neoadjuvant chemotherapy is increasingly being used for TNBC, because these tumors have a poor prognosis, are assumed to be chemosensitive and no alternative specific systemic treatment is available. Patients with a complete pathologic response (pCR) after neoadjuvant chemotherapy have a better outcome than those with residual disease, and pCR is a good surrogate for long-term survival and cure in this specific subgroup.[9,42]

The Immunity2 metagene was not found to be predictive of response to neoadjuvant chemotherapy in TNBC (272 fine needle aspirations of TNBC samples for which information about pCR or its absence was available from the eight datasets previously published by Ignatiadis et al. [16]) (data not shown). This lack of relationship may have resulted from the use of fine needle aspiration biopsy samples. The Immunity2 genes, which are largely expressed in the stromal environment, were less strongly expressed in fine needle aspiration samples than in tumor samples (**Fig. S23**).

However, intratumoral immune responses are known to be correlated with clinical outcomes in TNBC. This may reflect the role of immune cells in the activity of cytotoxic chemotherapeutic agents. Some chemotherapeutic drugs, such as anthracyclines, act not only through direct cytotoxic effects, but also by activating $CD8^+$ T-cell responses. Conflicting results have been published on the ability of other immune-based classifiers to predict outcome in TNBC. High-intratumoral levels of $CD8^+$ T cells [43] or TILs [36,44] are associated with better clinical responses to anthracycline-based chemotherapy. West et al. [45] reported that high-levels of lymphocyte GE were associated with a high rate (74%) of complete pathological responses to neoadjuvant anthracycline-based chemotherapy. In 2011, Sabatier et al. [20] showed, by gene-expression profiling, that "Immune High" patients (59%) were more likely to present pCR than "Immune Low" patients (43%), but this difference was not significant ($p = 0.29$). In 2014,[46] they showed that "*PDL1* mRNA expression high" (57%) patients presented higher rates of pCR than "*PDL1* mRNA expression low" (43%) patients ($p < 0.001$). Wimberley et al. [47] showed that PDL1 protein levels in the epithelium and stroma were correlated with pCR only in hormone receptor-positive and HER2-amplified breast cancers. Denkert et al. [44] demonstrated the importance of TIL and immune GE signatures for predicting pCR in breast carcinoma. However no significant difference in pCR rate was detected between lymphocyte-predominant breast cancer (LPBC) and no-LPBC in the anthracycline-taxane subgroup.

However, the results of these studies suggest that clinical outcomes in ER-negative breast cancers, including TNBC in particular, are strongly influenced by tumor immune responses and are, thus, highly responsive to immunotherapies. The possible use of immunotherapy approaches to treat TNBC (tumor vaccine approaches, immune-checkpoint inhibitors, antagonists of immunosuppressive molecules and adoptive cell therapies) should be investigated in detail.[48]

The other metagenes studied had no significant prognostic or predictive value. However, they identified sound biological networks providing opportunities for therapeutic intervention. The Immunity1 metagene included genes involved in the interferon α/β signaling pathway or cytokine signaling (STAT1, IRF7, IRF27, OAS1, OAS2, PMSB8, XAF1, IFIT1, IFITM1, ISG15, IGS20, IF6, MX1), the Toll-like receptor signaling pathway (STAT1, CXCL9, CXCL10, CXCL11, CCL5, IRF7), cell-cycle checkpoints and DNA synthesis (PSMB8, PSMB9). Patients displaying strong expression of this metagene often also had high-levels of Immunity2 and Proliferation/DNA damage metagene expression, suggesting the possible existence of common pathways. The IDO1 (indoleamine 2, 3-dioxygenase 1) gene is a particularly interesting potential target. It encodes a tryptophan-degrading enzyme known to suppress antitumor $CD8^+$ T cells and it contributes to the inhibition of anticancer immune responses.[48] This immunosuppressive enzyme is actually investigated as a promising candidate target in cancer immunotherapy.

A subset of TNBC tumors strongly expresses AR-regulated genes.[49] AR expression has been reported to be lower in triple-negative breast tumor cells than in other types of breast cancer. The overall frequency of AR expression in carcinoma cells varies considerably between studies (0–53%).[50,51] We identified strong expression of AR pathway genes in 25% of our population. The biological role of androgens in TNBC remains a matter of debate. Immunohistochemical studies investigating the presence of AR in tumor cells have reported conflicting results for clinical outcome; some studies have suggested that AR expression is advantageous for survival,[52-54] whereas others found no significant effect.[55] Lehmann et al. found that the LAR subtype of TNBC displayed the lowest frequency of pCR (10%). The presence of AR in a subset of TNBC patients suggests that androgenic pathways in tumor cells could be targeted in at least some TNBC patients. The widespread availability of agents targeting the AR also makes this approach potentially appealing, as it would be straightforward to incorporate such treatment into clinical practice.

The Matrix/Invasion1 metagene included genes associated with β1 integrin cell surface interactions, ECM-receptor interaction or integrin family cell surface interactions (NID1, TGFBI, COL5A1, COL5A2, COL6A3, COL3A1, COL1A1, COL1A2, COL11A1, FN1, FBN1, THBS1, THBS2), the TGF β signaling pathway (DCN, COMP, THBS1, THBS2), the inhibition of matrix metalloproteases (MMP2, TIMP3), and the AP-1 transcription factor network (DCN, COL1A2, MMP2). Metalloproteinases (MMPs) and their tissue inhibitors are involved in several key pathways of tumor growth, invasion and metastasis.[56,57] The expression and activity of MMPs has been linked to advanced stages of breast cancer, greater tumor invasion and the construction of metastatic formations.[58,59,60] Some studies have highlighted the importance of matrix MMP expression by stromal cells as a prognostic factor in the TNBC subtype.[61] These molecules are thus attractive targets for drug development.[62]

The Matrix2 metagene included genes associated with the AP-1 transcription factor network (FOS, EGR1, FABP4, DUSP1), the EGR receptor signaling pathway (FOS, DUSP1, EGR1), the Wnt or ALK signaling pathway (CAV1), the MAPK signaling pathway (FOS, DUSP1) or Trk receptor signaling mediated by the MAPK pathway (FOS, EGR1), the mTOR signaling pathway (IGF1), the PPAR signaling pathway (ADIPOQ, CD36, FABP4), and androgen-mediated signaling (FOS, EGR1). These pathways may contribute to cell motility and tumor cell invasion[63] and play a prominent role in epithelial-mesenchymal transition (EMT) and in stem cells. These metagenes are strongly expressed in mesenchymal cells and metaplastic breast cancers.[4] Metaplastic breast cancers have lineage plasticity, including spindle cell foci, and display osseous or cartilaginous differentiation.[64] Some drugs targeting the pathways relating to the metagenes identified here may be of particular interest for the treatment of TNBC (PI3K/mTOR inhibitor, Wnt/β catenin inhibitor).

Conclusion

In conclusion, our 167-gene TNBC molecular signature, consisting of six metagenes, appears to be particularly robust for the identification of TNBC subtypes. Furthermore, expression of the Immunity2 metagene was strongly correlated with prognosis, and many biological targets have been identified within the corresponding biological network. These findings open up interesting new possibilities for the development of new therapeutic interventions.

Patients and Methods

Data normalization and quality control

We collected 21 publicly available datasets (described in the supplementary data) containing raw GE data from microarray analyses (Affymetrix© Gene Chip Human Genome HG-U133A and HG-U133Plus2) of 3,247 primary human breast cancer samples. The data were normalized by the robust multichip average (RMA) procedure from the EMA R package.[65] The datasets were split into training (HGU-133A Affymetrix© arrays, 12 datasets, $n = 1{,}995$) and validation (HGU-133Plus2 Affymetrix© arrays, (9 datasets, $n = 1{,}014$) sets. We also collected two large datasets, to validate our classification: The Ignatiadis dataset ($n = 996$) and the METABRIC dataset ($n = 1{,}992$). Data processing for these two datasets has been described elsewhere.[16,17]

Determination and preprocessing of triple-negative breast cancer samples

We identified the TNBC samples in each dataset, using a bimodal mixture of two Gaussian distributions for ER and *HER2* gene expression, and the median value for PR expression.

The training, validation and Ignatiadis datasets

Batch effects were eliminated by the median centering of each probe-set across arrays and by a, independent quantile normalization of all arrays for each dataset. We controlled for outliers with the Array Quality Metrics R package.

The METABRIC set

We fitted a linear model (limma R package) to remove the batch effect and probes were filtered according to three criteria: probe quality,[66] GC content and presence in more than 5% of the samples. We centered expression values, using the R function *scale()*.

Gene selection process

Consensus clustering was applied to the training set, to determine the optimal number of robust gene clusters for the most variant genes (standard deviation > 0.8). We investigated the enrichment of each gene cluster in particular types of genes. We then identified known biological networks, for each gene cluster separately, using String© database software version 9.1 (http://string-db.org/).[67]

We then applied a two-step selection process: (1) we selected strong biological networks by retaining only genes for which connection scores of at least 0.7 were obtained with String© database software, (2) within each biological network, we selected groups of genes with for which expression levels were correlated, with a correlation coefficient of at least 0.5.

For each dataset (the training, validation, Ignatiadis and METABRIC sets), we applied a hierarchical clustering procedure to the TNBC GE profiles, using the selected genes to visualize the optimal number of stable TNBC subtypes.

Prognostic analysis

Prognostic analysis was performed on the METABRIC set published by Curtis et al.[17]

Expression data were summarized by a metagene for each gene cluster (details in the supplementary material). The clinical and pathological variables available for each dataset are described in the supplementary data. Qualitative variables were compared in χ^2 tests or Fisher's exact tests, as appropriate. Quantitative variables were analyzed in Student's *t*-tests. Survival analyses were performed separately for patients with and without chemotherapy. Survival analyses were performed, with the Kaplan–Meier estimate of the survival function. The endpoint of these analyses was breast cancer-specific survival (BCSS). Survival curves were compared in log rank tests. Hazard ratios were estimated with Cox's proportional hazard model.

Expression of the gene signature in human triple-negative breast cancer cell lines

We downloaded the GE profiles of the human cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE)[68] of Novartis/the Broad Institute and the Cancer Genome Project (CGP)[69] of the Sanger Institute. We normalized all the cell lines from different tissues together.

All statistical analyses were performed with R software (www.cran.r-project.org). P -values < 0.05 were considered statistically significant.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

The authors thank Sergio Roman-Roman for the reviewing of the study and the manuscript.

Funding

This work was supported by Institut Curie, INCa (the french National Cancer Institute) Grant INCa-DGOS-4654, ANR-10-IDEX-0001-02 PSL, ANR-11-LABX-0043, CIC IGR Curie 1428 and Fundation ARC (association for Research Against Cancer).

Supplemental Material

Supplemental data for this article can be accessed on the publisher's website.

References

1. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Ggenom 2006; 7:96; PMID:16643655; http://dx.doi.org/10.1186/1471-2164-7-96
2. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001; 98(19):10869-74; PMID:11553815; http://dx.doi.org/10.1073/pnas.191367098
3. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009; 27(8):1160-7; PMID:19204204; http://dx.doi.org/10.1200/JCO.2008.18.1370
4. Hennessy BT, Gonzalez-Angulo A-M, Stemke-Hale K, Gilcrease MZ, Krishnamurthy S, Lee JS, Fridlyand J, Sahin A, Agarwal R, Joy C et al. Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. Cancer Res 2009; 69(10):4116-24; PMID:19435916; http://dx.doi.org/10.1158/0008-5472.CAN-08-3441
5. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. N Engl J Med 2010; 363(20):1938-8; PMID:21067385; http://dx.doi.org/10.1056/NEJMra1001389
6. Turner NC, Reis-Filho JS. Tackling the diversity of triple-negative breast cancer. Clin Cancer Res 2013; 19 (23):6380-8; PMID:24298068; http://dx.doi.org/10.1158/1078-0432.CCR-13-0915
7. Von Minckwitz G, Blohmer JU, Costa SD, Denkert C, Eidtmann H, Eiermann W, Gerber B, Hanusch C, Hilfrich J, Huober J et al. Response-guided neoadjuvant chemotherapy for breast cancer. J Clin Oncol 2013; 31 (29):3623-30; PMID:24002511; http://dx.doi.org/10.1200/JCO.2012.45.0940
8. Liedtke C, Mazouni C, Hess KR, André F, Tordai A, Mejia JA, Symmans WF, Gonzalez-Angulo AM, Hennessy B, Green M et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. J Clin Oncol. 2008; 26 (8):1275-81; PMID:18250347; http://dx.doi.org/10.1200/JCO.2007.14.4147
9. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, Bonnefoi H, Cameron D, Gianni L, Valagussa P et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet 2014; 384 (9938):164-72; PMID:24529560; http://dx.doi.org/10.1016/S0140-6736(13)62422-8
10. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol J. A.. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J

5.1 Biological network-driven gene selection identifies a stromal immune module as a key determinant of triple-negative breast carcinoma prognosis

Clin Invest 2011; 121(7):2750-67; PMID:21633166; http://dx.doi.org/10.1172/JCI45014

11. Chen X, Li J, Gray WH, Lehmann BD, Bauer JA, Shyr Y, Pietenpol JA. TNBCtype: A Subtyping Tool for Triple-Negative Breast Cancer. Cancer Informat 2012; 11:147-56; PMID:22872785; http://dx.doi.org/10.4137/CIN.S9983
12. Sontrop HMJ, Moerland PD, van den Ham R, Reinders MJT, Verhaegh WFJ. A comprehensive sensitivity analysis of microarray breast cancer classification under feature variability. BMC Bioinformatics 2009; 10:389; PMID:19941644; http://dx.doi.org/10.1186/1471-2105-10-389
13. Abraham G, Kowalczyk A, Loi S, Haviv I, Zobel J. Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. BMC Bioinformatics 2010; 11:277; PMID:20500821; http://dx.doi.org/10.1186/1471-2105-11-277
14. Fröhlich H. Network based consensus gene signatures for biomarker discovery in breast cancer. PLoS One 2011; 6(10):e25364; PMID:22046239; http://dx.doi.org/10.1371/journal.pone.0025364
15. Cun Y, Fröhlich HF. Prognostic gene signatures for patient stratification in breast cancer: accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. BMC bioinformatics 2012; 13:69; PMID:22548963; http://dx.doi.org/10.1186/1471-2105-13-69
16. Ignatiadis M, Singhal SK, Desmedt C, Haibe-Kains B, Criscitiello C, Andre F, Loi S, Piccart M, Michiels S, Sotiriou C. Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis. J Clin Oncol 2012; 30(16):1996-2004; PMID:22508827; http://dx.doi.org/10.1200/JCO.2011.39.5624
17. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486(7403):346-52; PMID:22522925; http://dx.doi.org/10.1038/nature10983.
18. Bianchini G, Qi Y, Alvarez RH, Iwamoto T, Coutant C, Ibrahim NK, Valero V, Cristofanilli M, Green MC, Radvanyi L et al. Molecular anatomy of breast cancer stroma and its prognostic value in estrogen receptor-positive and -negative cancers. J Clin Oncol 2010; 28 (28):4316-4323; PMID:20805453; http://dx.doi.org/10.1200/JCO.2009.27.2419
19. Rody A, Karn T, Liedtke C, Pusztai L, Ruckhaeberle E, Hanker L, Gaetje R, Solbach C, Ahr A, Metzler D et al. A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast Cancer Res 2011; 13(5):R97; PMID:21978456; http://dx.doi.org/10.1186/bcr3035
20. Sabatier R, Finetti P, Cervera N, Lambaudie E, Esterni B, Mamessier E, Tallet A, Chabannon C, Extra JM, Jacquemier J et al. A gene expression signature identifies two prognostic subgroups of basal breast cancer. Breast Cancer Res Treatment 2011; 126:407-420; PMID:20490655; http://dx.doi.org/10.1007/s10549-010-0897-9
21. Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol 2007; 8 (8):R157; PMID:17683518; http://dx.doi.org/10.1186/gb-2007-8-8-r157
22. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart M. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res 2008; 14(16):5158-65; PMID:18698033; http://dx.doi.org/10.1158/1078-0432.CCR-07-4756
23. Gu-Trantien C, Loi S, Garaud S, Equeter C, Libin M, de Wind A, Ravoet M, Le Buanec H, Sibille C, Manfouo-Foutsop G et al. CD4+ follicular helper T cell infiltration predicts breast cancer survival. J Clin Invest 2013; 123(7):1-20; PMID:23778140; http://dx.doi.org/; http://dx.doi.org/10.1172/JCI67428
24. Karn T, Pusztai L, Holtrich U, Iwamoto T, Shiang CY, Schmidt M, Müller V, Solbach C, Gaetje R, Hanker L et al. Homogeneous datasets of triple negative breast cancers enable the identification of novel prognostic and predictive signatures. PloS One 2011; 6(12): e28403; PMID:22220191; http://dx.doi.org/10.1371/journal.pone.0028403
25. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua SA, Savage MI, Osborne CK, Hilsenbeck SG, Chang JC et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. Clin Cancer Res 2014; 21(7):1688-98; PMID:25208879; http://dx.doi.org/10.1158/1078-0432.CCR-14-0432.
26. Gatza ML, Lucas JE, Barry WT, Kim JW, Wang Q, Crawford MD, Datto MB, Kelley M, Mathey-Prevot B, Potti A et al. A pathway-based classification of human breast cancer. Proc Natal Acad Sci U S A 2010; 107:6994-9; PMID:20335537; http://dx.doi.org/10.1073/pnas.0912708107
27. Palmer C, Diehn M, Alizadeh AA, Brown PO. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. BMC Genomics 2006; 7:115; PMID:16704732; http://dx.doi.org/10.1186/1471-2164-7-115
28. Isella C, Terrasi A, Bellomo SE, Petti C, Galatola G, Muratore A, Mellano A, Senetta R, Cassenti A, Sonetto C et al. Stromal contribution to the colorectal cancer transcriptome. Nature Genet 2015; 47(4):312-9; PMID:25547594; http://dx.doi.org/10.1038/ng.3224
29. Schalper K A., Velcheti V, Carvajal D, Wimberly H, Brown J, Pusztai L, Rimm DL. In situ tumor PD-L1 mRNA expression is associated with increased tils and better outcome in breast carcinomas. Clin Cancer Res 2014; 20:2773-82; PMID:24647569; http://dx.doi.org/10.1158/1078-0432.CCR-13-2702
30. Sabatier R, Finetti P, Mamessier E, Adelaide J, Chaffanet M, Ali HR, Viens P, Caldas C, Birnbaum D, Bertucci F. Prognostic and predictive value of PDL1 expression in breast cancer. Oncotarget 2015; 6 (7):5449-64; PMID:25669979
31. Weigelt B, Mackay A, A'hern R, Natrajan R, Tan DS, Dowsett M, Ashworth A, Reis-Filho JS. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. The Lancet Oncol 2010; 11 (4):339-49; PMID:20181526; http://dx.doi.org/10.1016/S1470-2045(10)70008-5
32. De Wever O, Mareel M. Role of tissue stroma in cancer cell invasion. J Pathol 2003; 200(4):429-47; PMID:12845611; http://dx.doi.org/10.1002/path.1398
33. Lum DH, Matsen C, Welm AL, Welm BE. Overview of human primary tumorgraft models: comparisons with traditional oncology preclinical models and the clinical relevance and utility of primary tumorgrafts in basic and translational oncology research. Curr Protoc Pharmacol 2012;Chapter 14(801):Unit 14.22; PMID:23258598; http://dx.doi.org/10.1002/0471141755.ph1422s59
34. Sanavia T, Aiolli F, Da San Martino G, Bisognin A, Di Camillo B. Improving biomarker list stability by integration of biological knowledge in the learning process. BMC Bioinformatics 2012; 13 Suppl 4(Suppl 4):S22; PMID:22536969; http://dx.doi.org/10.1186/1471-2105-13-S4-S22
35. Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer 2012; 12(4):298-306; PMID:22419253; http://dx.doi.org/10.1038/nrc3245
36. Denkert C, Loibl S, Noske A, Roller M, Müller BM, Komor M, Budczies J, Darb-Esfahani S, Kronenwett R, Hanusch C et al. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol 2010; 28 (1):105-13; PMID:19917869; http://dx.doi.org/10.1200/JCO.2009.23.7370
37. Finak G, Bertos N, Pepin F, Sadekova S, Souleimanova M, Zhao H, Chen H, Omeroglu G, Meterissian S, Omeroglu A et al. Stromal gene expression predicts clinical outcome in breast cancer. Nat Med 2008; 14 (5):518-27; PMID:18438415; http://dx.doi.org/10.1038/nm1764
38. DeNardo DG, Brennan DJ, Rexhepaj E, Ruffell B, Shiao SL, Madden SF, Gallagher WM, Wadhwani N, Keil SD, Junaid SA et al. Leukocyte complexity predicts breast cancer survival and functionally regulates response to chemotherapy. Cancer Discov 2011; 1 (1):54-67; PMID:22039576; http://dx.doi.org/10.1158/2159-8274.CD-10-0028
39. Schmidt M, Böhm D, von Törne C, Steiner E, Puhl A, Pilch H, Lehr HA, Hengstler JG, Kölbl H, Gehrmann M. The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res 2008; 68(13):5405-13; PMID:18593943; http://dx.doi.org/10.1158/0008-5472.CAN-07-5206
40. Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol 2007; 8 (8):R157; PMID:17683518; http://dx.doi.org/10.1186/gb-2007-8-8-r157
41. Bertucci F, Finetti P, Cervera N, Charafe-Jauffret E, Mamessier E, Adélaïde J, Debono S, Houvenaeghel G, Maraninchi D, Viens P et al. Gene expression profiling shows medullary breast cancer is a subgroup of basal breast cancers. Cancer Res 2006; 66(9):4636-44; PMID:16651414; http://dx.doi.org/10.1158/0008-5472.CAN-06-0031
42. Von Minckwitz G, Untch M, Blohmer J-U, Costa SD, Eidtmann H, Fasching PA, Gerber B, Eiermann W, Hilfrich J, Huober J et al. Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. J Clin Oncol 2012; 30(15):1796-804; PMID:22508812; http://dx.doi.org/10.1200/JCO.2011.38.8595
43. Mattarollo SR, Loi S, Duret H, Ma Y, Zitvogel L, Smyth MJ. Pivotal role of innate and adaptive immunity in anthracycline chemotherapy of established tumors. Cancer Res 2011; 71(14):4809-20; PMID:21646474; http://dx.doi.org/10.1158/0008-5472.CAN-11-0753
44. Denkert C, von Minckwitz G, Brase JC, Bruno V. Sinn, Stephan Gade, Ralf Kronenwett, Berit M. Pfitzner, Christoph Salat, Sherene Loi, Wolfgang D. Schmitt et al. Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers. J Clin Oncol 2015; 33(9):983-91; PMID:25534375; http://dx.doi.org/10.1200/JCO.2014.58.1967
45. West NR, Milne K, Truong PT, Macpherson N, Nelson BH, Watson PH. Tumor-infiltrating lymphocytes predict response to anthracycline-based chemotherapy in estrogen receptor-negative breast cancer. Breast Cancer Res 2011; 13(6):R126; PMID:22151962; http://dx.doi.org/10.1186/bcr3072
46. Sabatier R, Finetti P, Guille A, Adelaide J, Chaffanet M, Viens P, Birnbaum D, Bertucci F. Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization. Mol Cancer 2014; 13:228; PMID:25277734; http://dx.doi.org/10.1186/1476-4598-13-228
47. Wimberly H, Brown JR, Schalper K, Haack H, Silver MR, Nixon C, Bossuyt V, Pusztai L, Lannin DR, Rimm DL PD-L1 Expression Correlates with Tumor-Infiltrating Lymphocytes and Response to Neoadjuvant Chemotherapy in Breast Cancer. Cancer Immunol Res 2015; 3(4):326-32; PMID:25527356; http://dx.doi.org/10.1158/2326-6066.CIR-14-0133
48. Stagg J, Allard B. Immunotherapeutic approaches in triple-negative breast cancer: latest research and clinical prospects. Ther Adv Med Oncol 2013; 5(3):169-81; PMID:23634195; http://dx.doi.org/10.1177/1758834012475152
49. Doane AS, Danso M, Lal P, Donaton M, Zhang L, Hudis C, Gerald WL. An estrogen receptor-negative

breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. Oncogene. 2006; 25(28):3994-4008; PMID:16491124; http://dx.doi.org/10.1038/sj.onc.1209415
50. Ogawa Y, Hai E, Matsumoto K, Ikeda K, Tokunaga S, Nagahara H, Sakurai K, Inoue T, Nishiguchi Y. Androgen receptor expression in breast cancer: relationship with clinicopathological factors and biomarkers. Inter J Clin Oncol 2008; 13(5):431-5; PMID:18946753; http://dx.doi.org/10.1007/s10147-008-0770-6
51. He J, Peng R, Yuan Z, Wang S, Peng J, Lin G, Jiang X, Qin T. Prognostic value of androgen receptor expression in operable triple-negative breast cancer: a retrospective analysis based on a tissue microarray. Med Oncol 2012; 29(2):406-10; PMID:21264529; http://dx.doi.org/10.1007/s12032-011-9832-0
52. Luo X, Shi Y-X, Li Z-M, Jiang W-Q. Expression and clinical significance of androgen receptor in triple negative breast cancer. Chin J Cancer 2010; 29(6):585-90; PMID:20507730; http://dx.doi.org/10.5732/cjc.009.10673
53. Rakha E A, El-Sayed ME, Green AR, Lee AHS, Robertson JF, Ellis IO. Prognostic markers in triple-negative breast cancer. Cancer. 2007; 109(1):25-32; PMID:17146782; http://dx.doi.org/10.1002/cncr.22381
54. McNamara KM, Yoda T, Miki Y, Chanplakorn N, Wongwaisayawan S, Incharoen P, Kongdan Y, Wang L, Takagi K, Mayu T et al. Androgenic pathway in triple negative invasive ductal tumors: its correlation with tumor cell proliferation. Cancer Sci 2013; 104(5):639-46; PMID:23373898; http://dx.doi.org/10.1111/cas.12121
55. Peters A A, Buchanan G, Ricciardelli C, Bianco-Miotto T, Centenera MM, Harris JM, Jindal S, Segara D, Jia L, Moore NL et al. Androgen receptor inhibits estrogen receptor-alpha activity and is prognostic in breast cancer. Cancer Res 2009; 69(15):6131-40; PMID:19638585; http://dx.doi.org/10.1158/0008-5472.CAN-09-0452
56. Jiang Y, Goldberg ID, Shi YE. Complex roles of tissue inhibitors of metalloproteinases in cancer. Oncogene 2002; 21(14):2245-52; PMID:11948407; http://dx.doi.org/10.1038/sj/onc/1205291.
57. Deryugina EI, Quigley JP. Matrix metalloproteinases and tumor metastasis. Cancer Metastasis Rev 2006; 25 (1):9-34; PMID:16680569; http://dx.doi.org/10.1007/s10555-006-7886-9
58. Vizoso FJ, González LO, Corte MD, Rodríguez JC, Vázquez J, Lamelas ML, Junquera S, Merino AM, García-Muñiz JL. Study of matrix metalloproteinases and their inhibitors in breast cancer. Br J Cancer2007; 96(6):903-11; PMID:17342087; http://dx.doi.org/10.1038/sj.bjc.6603666
59. McGowan PM, Duffy MJ. Matrix metalloproteinase expression and outcome in patients with breast cancer: analysis of a published database. Annals Oncol 2008; 19(9):1566-72; PMID:18503039; http://dx.doi.org/10.1093/annonc/mdn180
60. Figueira RCS, Gomes LR, Neto JS, Silva FC, Silva IDCG, Sogayar MC. Correlation between MMPs and their inhibitors in breast cancer tumor tissue specimens and in cell lines with different metastatic potential. BMC Cancer 2009; 9:20; PMID:19144199; http://dx.doi.org/10.1186/1471-2407-9-20
61. González LO, Corte MD, Junquera S, González-Fernández R, del Casar JM, García C, Andicoechea A, Vázquez J, Pérez-Fernández R, Vizoso FJ. Expression and prognostic significance of metalloproteases and their inhibitors in luminal A and basal-like phenotypes of breast carcinoma. Human Pathol 2009; 40(9):1224-33; PMID:19439346; http://dx.doi.org/10.1016/j.humpath.2008.12.022
62. Fingleton B. Matrix metalloproteinases as valid clinical targets. Curr Pharmaceut Design 2007; 13:333-46; PMID:17313364; http://dx.doi.org/10.2174/138161207779313551
63. Shin S-Y, Rath O, Zebisch A, Choo S-M, Kolch W, Cho K-H. Functional roles of multiple feedback loops in extracellular signal-regulated kinase and Wnt signaling pathways that regulate epithelial-mesenchymal transition. Cancer Res 2010; 70(17):6715-24; PMID:20736375; http://dx.doi.org/10.1158/0008-5472.CAN-10-1377
64. Gibson GR, Qian D, Ku JK, Lai LL. Metaplastic breast cancer: clinical features and outcomes. Am Surg 2005; 71:725-30; PMID:16468506
65. Servant N, Gravier E, Gestraud P, Laurent C, Paccard C, Biton A, Brito I, Mandel J, Asselain B, Barillot E et al. EMA - A R package for Easy Microarray data analysis. BMC Res Notes 2010; 3(1):277; PMID:21047405; http://dx.doi.org/10.1186/1756-0500-3-277
66. Barbosa-Morais NL, Dunning MJ, Samarajiwa S A, Darot JF, Ritchie ME, Lynch AG, Tavaré S. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res 2010; 38(3):e17; PMID:19923232; http://dx.doi.org/10.1093/nar/gkp942
67. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M et al. STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 2009; 37(Database issue):D412-D416; PMID:18940858; http://dx.doi.org/10.1093/nar/gkn760
68. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483(7391):603-607; PMID:22460905; http://dx.doi.org/10.1038/nature11003
69. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J et al. Europe PMC Funders Group Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012; 483(7391):570-5; PMID:22460902; http://dx.doi.org/10.1038/nature11005

5.2 A Stromal Immune Module Correlated with the Response to Neoadjuvant Chemotherapy, Prognosis and Lymphocyte Infiltration in HER2-Positive Breast Carcinoma Is Inversely Correlated with Hormonal Pathways

The study of Bonsang-Kitsis et al 2015 (79) introduced a novel signature to classify TNBCs. We decided to conduct similar analysis on HER2-positive breast cancer, given the interesting results presented in our previous analysis.

HER2+ breast cancer tumors are characterized by an amplification and high expression level of HER2 tyrosine kinase receptor gene. They are very aggressive tumors with a high rate of early distant metastatis. Targeted therapy with a monoclonal antibody targeting the HER2 receptor (trastuzumab) has considerably changed the patient outcomes. However, a fraction of HER2+ tumors appears to be resistant to trastuzumab. Robust classifiers are then required to find novel therapeutic strategies for these women. Based on a collection of 21 publicly available datasets we applied the same biological network-driven gene selection process used in Bonsang-Kitsis et al 2015. The gene signature was build on a training and validated on a validation set. The final gene signature was composed of 138 genes that form six clusters of genes enriched in different gene ontologies: Immunity, Interferon, Signal transduction, Hormonal/survival, Tumor suppressors/Proliferation and Matrix. We further validated the six-metagene signature with Ignatidis(89) and METABRIC (46) dataset.

We have demonstrated strong inverse interactions between immunity metagene and estrogen, progesterone and androgene receptor (ER, PR, AR) hormonal pathways. The ER, PR and AR genes are more expressed in samples with low Immunity metagene expression level than in the "Immunity high" subgroup in the four datasets.
A multivariate analysis was conducted to assess whether the 138-gene HER2+ signature associated with clinical characteristics was predictive of response to neoadjuvant chemotherapy (NAC). It was performed on 82 HER2-positive samples from the Ignatiadis dataset. Samples with ER positive status and high expression of Immunity metagene were associated with pathological complete response (pCR). The "Immunity2" metagene described in Bonsang-Kitsis et al 2015 was not predictive of pCR in TNBCs although both immune metagene (from HER2+ samples and TNBC samples) were strongly cor-

related in the four sets of data (correlation coefficients greater than 0.94 in the training and validation sets, Ignatidis and METABRIC). We applied both signatures to the whole population for the Ignatiadis dataset, and we analyzed the pCR rates as a function of breast cancer subtype and Immunity metagene status. The results showed that the high expression of the two metagenes was significantly associated with higher pCR rates in ER+/HER2- and HER2+. For TNBC only the Immunity metagene was significant. The prognostic value of the signature was assessed on 248 HER2+ samples from the METABRIC dataset. High expression level of the Immunity metagene was significantly associated with good prognosis only in ER- samples.
To complete our analysis associated with the immune process, we investigated the correlation between Immunity metagene expression and lymphocyte infiltration. We analyzed an independent set of HER2+ tumors for which both histology and gene expression data were available (n = 27). We demonstrated that the high expression level of the Immunity metagene was significantly correlated with the amount of both intratumoral and stromal lymphocyte infiltration.

Our work reports one of the first immune signatures identified as both predictive and prognostic, reflecting histological immune infiltration in HER2+ breast cancers. Given the key role played by the immune processes, immunotherapies may represent a therapy with a high impact for such cancers.

PLOS ONE

RESEARCH ARTICLE

A Stromal Immune Module Correlated with the Response to Neoadjuvant Chemotherapy, Prognosis and Lymphocyte Infiltration in *HER2*-Positive Breast Carcinoma Is Inversely Correlated with Hormonal Pathways

Anne-Sophie Hamy[1☯], Hélène Bonsang-Kitzis[1,2☯], Marick Lae[3], Matahi Moarii[4,5], Benjamin Sadacca[1,6], Alice Pinheiro[1], Marion Galliot[1], Judith Abecassis[1,4,5], Cecile Laurent[1], Fabien Reyal[1,2]*

1 Institut Curie, PSL Research University, Translational Research Department, INSERM, U932 Immunity and Cancer, Residual Tumor & Response to Treatment Laboratory (RT2Lab), Paris, France, **2** Department of Surgery, Institut Curie, Paris, France, **3** Department of Tumor Biology, Institut Curie, Paris, France, **4** Mines Paristech, PSL-Research University, CBIO-Centre for Computational Biology, Mines ParisTech, Fontainebleau, France, **5** U900, INSERM, Institut Curie, Paris, France, **6** Laboratoire de Mathématiques et Modélisation d'Evry, Université d'Évry Val d'Essonne, Evry, France

☯ These authors contributed equally to this work.
* fabien.reyal@curie.fr

OPEN ACCESS

Citation: Hamy A-S, Bonsang-Kitzis H, Lae M, Moarii M, Sadacca B, Pinheiro A, et al. (2016) A Stromal Immune Module Correlated with the Response to Neoadjuvant Chemotherapy, Prognosis and Lymphocyte Infiltration in *HER2*-Positive Breast Carcinoma Is Inversely Correlated with Hormonal Pathways. PLoS ONE 11(12): e0167397. doi:10.1371/journal.pone.0167397

Editor: William B. Coleman, University of North Carolina at Chapel Hill School of Medicine, UNITED STATES

Received: September 19, 2016

Accepted: November 14, 2016

Published: December 22, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: A-S Hamy-Petit was supported by an ITMO-INSERM-AVIESAN cancer translational research grant. Funding was also obtained from the Site de Recherche Intégrée en Cancérologie/ INstitut National du Cancer (InCa-DGOS-4654). The funders had no role in study design, data

Abstract

Introduction

HER2-positive breast cancer (BC) is a heterogeneous group of aggressive breast cancers, the prognosis of which has greatly improved since the introduction of treatments targeting *HER2*. However, these tumors may display intrinsic or acquired resistance to treatment, and classifiers of *HER2*-positive tumors are required to improve the prediction of prognosis and to develop novel therapeutic interventions.

Methods

We analyzed 2893 primary human breast cancer samples from 21 publicly available datasets and developed a six-metagene signature on a training set of 448 *HER2*-positive BC. We then used external public datasets to assess the ability of these metagenes to predict the response to chemotherapy (Ignatiadis dataset), and prognosis (METABRIC dataset).

Results

We identified a six-metagene signature (138 genes) containing metagenes enriched in different gene ontologies. The gene clusters were named as follows: Immunity, Tumor suppressors/proliferation, Interferon, Signal transduction, Hormone/survival and Matrix clusters. In all datasets, the Immunity metagene was less strongly expressed in ER-positive than in ER-negative tumors, and was inversely correlated with the Hormonal/survival metagene. Within the signature, multivariate analyses showed that strong expression of the

collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: BC, breast cancer; BCSS, breast cancer-specific survival; BC_CL, breast cancer cell lines; CCLE, Cancer Cell Line Encyclopedia; CGP, Cancer Genome Project; ER, estrogen receptor; GE, gene expression; *HER2*-positive BC, *HER2*-positive breast cancer; *HER2*, human epidermal growth factor receptor 2; PR, progesterone receptor; pCR, pathological complete response; RMA, robust multichip average; TILs, tumor-infiltrating lymphocytes.

"Immunity" metagene was associated with higher pCR rates after NAC (OR = 3.71[1.28–11.91], $p = 0.019$) than weak expression, and with a better prognosis in *HER2*-positive/ER-negative breast cancers (HR = 0.58 [0.36–0.94], $p = 0.026$). Immunity metagene expression was associated with the presence of tumor-infiltrating lymphocytes (TILs).

Conclusion

The identification of a predictive and prognostic immune module in *HER2*-positive BC confirms the need for clinical testing for immune checkpoint modulators and vaccines for this specific subtype. The inverse correlation between Immunity and hormone pathways opens research perspectives and deserves further investigation.

Introduction

HER2-positive breast carcinomas (BCs) are defined by amplification and overexpression of the *HER2* tyrosine kinase receptor gene (17q12). The tumors of this subgroup have aggressive pathological features and a high rate of early distant metastatic events. They are routinely treated with a combination of docetaxel plus a monoclonal antibody targeting the *HER2* receptor (trastuzumab). Other drugs also appear to be of major interest and will probably be made available for routine treatment in the near future (lapatinib, pertuzumab and T-DM1).

HER2-positive BCs constitute a heterogeneous group of tumors differing in histological features, gene expression profiles, clinical behavior, overall prognosis, and response to conventional systemic cytotoxic therapy. Trastuzumab-based treatments have been used for the last decade and have substantially improved outcomes in patients with early or metastatic *HER2*-positive BC. However, some *HER2*-positive tumors display intrinsic or acquired resistance to trastuzumab. Robust classifiers are required, both to improve our understanding of the molecular basis of *HER2*-positive BC and to develop novel therapeutic interventions.

We developed a two-step biological network-driven gene selection process: 1) identification of the most variable genes displaying highly correlated patterns of expression, 2) direct connection of these genes within known biological networks. This method has been shown to construct molecular signatures efficiently [1–3]. We defined a *HER2*-positive molecular subtype classification and identified a stromal immune module gene expression profile strongly correlated with predicted response to chemotherapy, prognosis and lymphocytic infiltration. This classification provides considerable biological insight, and has potential for use in the development of therapeutic interventions, such as novel immunotherapies in particular.

Material and methods

Data normalization and quality control

Training, validation and Ignatiadis datasets. We collected 21 publicly available datasets (described in the S1 File) containing raw gene expression data for 2893 primary human breast cancer samples. The data were normalized by the robust multichip average (RMA) procedure from the EMA R package [4]. The datasets were split into training (HGU-133A Affymetrix* arrays, 12 datasets, $n = 1921$) and validation (HGU-133Plus2 Affymetrix* arrays (9 datasets, $n = 972$) sets. Batch effects were eliminated by the median centering of each probe-set across arrays and by an independent quantile normalization of all arrays for each dataset. We controlled for outliers with the Array Quality Metrics R package. We also collected two large

datasets to validate our classification: The Ignatiadis dataset (Affymetrix data $n = 996$) [5] and the METABRIC dataset (Illumina data $n = 1992$) published by Curtis *et al.* [6].

Determination and preprocessing of *HER2*-positive breast cancer samples

We identified the *HER2*-positive samples in the training and validation datasets, on the basis of transformed ERBB2 mRNA expression, as described by Gong *et al.* [7], and using the bimodal distribution of ERBB2 expression for the Ignatiadis and the METABRIC dataset.

Gene selection process

Consensus clustering with the ConsensusClusterPlus R Package was applied to the training set with a ward inner, final linkage and Pearson distance, to determine the optimal number of robust gene clusters for the most variable genes (standard deviation>0.8). We investigated the enrichment of each gene cluster in particular types of genes, and categorized and labeled genes clusters according to the different gene ontologies. We then identified known biological networks, for each gene cluster separately, using String* database software version 9.1 (http://string-db.org/) [8]. We then applied a two-step selection process: 1) we selected strong biological networks by retaining only genes for which connection scores of at least 0.7 were obtained with String* database software, 2) within each biological network, we selected groups of genes with correlated expression patterns and a correlation coefficient of at least 0.5.

For each dataset (the training, validation, Ignatiadis and METABRIC sets), we applied a hierarchical clustering procedure with a ward inner, final linkage and Pearson distance to the *HER2*-positive gene expression (GE) profiles, using the selected genes to visualize the optimal number of stable *HER2*-positive subtypes.

Metagene construction

We defined a metagene as an aggregate patterns of gene expression. Metagene expression was assessed by calculating the median normalized expression values of all probe sets in the respective gene clusters for each sample. The metagene value for each sample was then discretized on the basis of the median value, as "high" or "low".

Association between expression of the Immunity metagene and that of ESR1, PGR, and AR

All the analyses were performed on all four datasets (training, validation, Ignatiadis, METABRIC). The levels of expression of ESR1, PGR and AR were compared between "Immunity low" and "Immunity high" samples, by ANOVA. Levels of Immunity metagene expression were compared between samples positive and negative for ER, PR, and AR, by ANOVA. We also performed ANOVA for each gene of the Immunity metagene as a function of ER status.

Analysis of the predicted response to NAC

We analyzed the predicted response to chemotherapy in the datasets published by Ignatiadis *et al.* [5]. Expression data were summarized by defining a metagene for each gene cluster. The clinical and pathological variables available for each dataset are described in S1 File. Qualitative variables were compared with logistic regression models.

Prognostic analysis

Prognostic analysis was performed on the METABRIC set. Expression data were summarized by defining a metagene for each gene cluster. The clinical and pathological variables available for each dataset are described in S1 File. Survival analyses were performed for the whole population, and separately for ER-positive and ER-negative patients, by calculating Kaplan-Meier estimates of the survival function. The endpoint of these analyses was breast cancer-specific survival (BCSS). Survival curves were compared in log-rank tests. Hazard ratios were estimated with Cox's proportional hazard model. Predictive and prognostic analyses were performed with the R survival package. Variables associated with pCR or BCSS with a *P*-value <0.10 in univariate analysis were included in the multivariate model. Variables with *P*-values <0.05 in multivariate analysis were considered statistically significant.

Correlation with tumor-infiltrating lymphocyte levels

We downloaded the gene expression data from the REMAGUS 02 trial [9] and retrieved 27 samples for which paraffin-embedded tissue sections were available at our institution. All patients enrolled in this study gave their informed written consent. Histologic microbiopsy specimens were evaluated independently for the presence of a lymphocytic infiltrate (intratumoral TILs and stromal TILs by one BC pathologist (ML) and one breast physician (ASH) unaware of the gene expression classification. Percentages of TLs and StrL were compared, as a function of Immunity metagene status, in ANOVA. The correlations between Immunity metagene expression and the percentages of TLs and StrL were assessed by calculating Pearson's correlation coefficient.

Expression of the gene signature in human breast cancer cell lines

We downloaded the gene expression profiles of the human cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) [10] of Novartis/the Broad Institute and the Cancer Genome Project (CGP) [11] of the Sanger Institute. We normalized the data for all the cell lines from different tissues together.

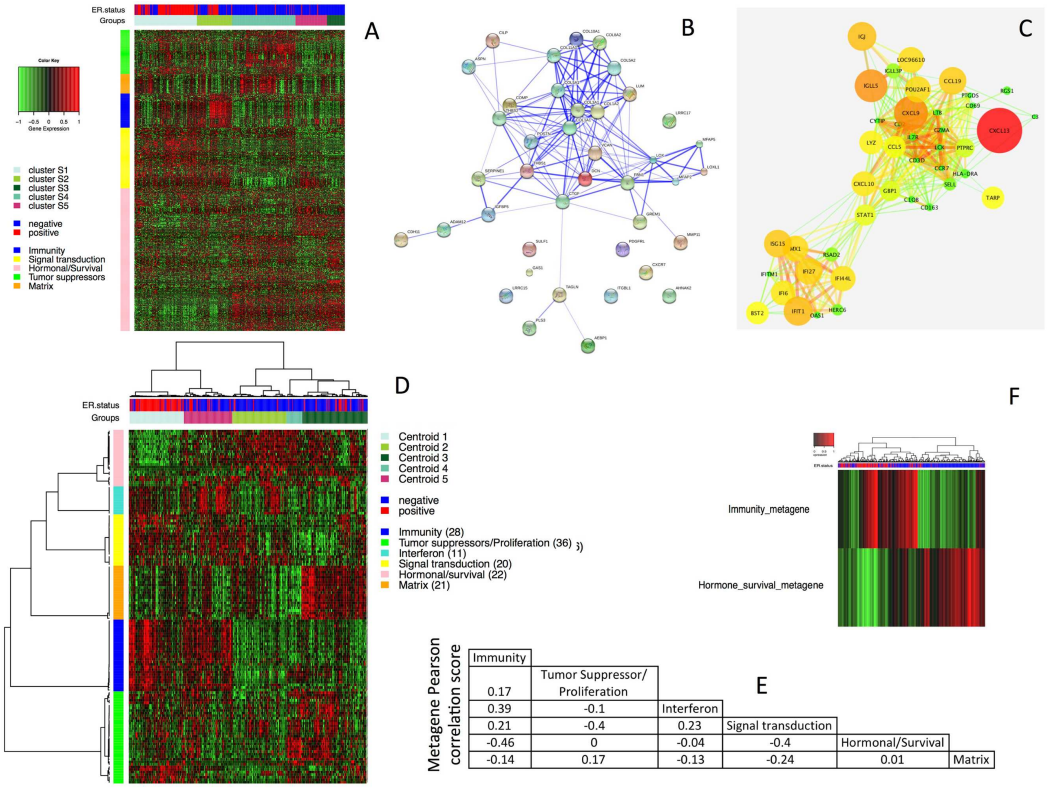
Statistical analysis

Data were processed and statistical analyses were carried out with R software version 3.1.2 [12] (www.cran.r-project.org).

Results

HER2-positive gene expression profiles identify six main gene clusters

HER2-positive BC samples were selected from 21 publicly available datasets (n = 3,247 breast cancer samples) and separated into a training set and a validation set (S1 File and S1 Fig). In the training set, we applied a gene selection process based on biological networks (Fig 1A to 1C), to decrease the instability intrinsic to molecular classification methods (see S1 File), as previously described for triple-negative breast cancers (TNBCs) [3]. We selected a final set of 138 genes (S1 Table), composed of six gene clusters enriched in different gene ontologies: Immunity ($n = 28$), Interferon ($n = 11$), Signal transduction ($n = 20$), Hormonal/survival ($n = 22$), Tumor suppressors/Proliferation ($n = 36$), Matrix ($n = 21$) (Fig 1D). We defined a metagene for each of the six gene clusters identified in this way (S1 File). The Immunity and Interferon metagenes displayed similar patterns of expression. The Immunity and Hormonal/survival metagenes displayed the strongest inverse correlation for expression (coefficient of -0.46) (Fig 1E and 1F). The correlations between the 138 genes and the metagenes are



E

Metagene Pearson correlation score

Immunity	Tumor Suppressor/ Proliferation	Interferon	Signal transduction	Hormonal/Survival	Matrix
0.17					
0.39	-0.1				
0.21	-0.4	0.23			
-0.46	0	-0.04	-0.4		
-0.14	0.17	-0.13	-0.24	0.01	

Fig 1. Gene selection process. A Heatmap showing the 616 most variable genes in the 448 *HER2*-positive samples (training set). **B** String database software confidence view of the Matrix genes cluster. Stronger associations between genes are represented by thicker lines. **C** Cytoscape View for the Immunity gene cluster. GE correlations between genes are indicated by edges (edge color varies from green to red and edge size increases with increasing correlation) and gene expression variance is represented by node color (node color varies from green to red and node size increases with increasing variance). **D** Heatmap showing the relative expression of 138 selected genes in 448 *HER2*-positive samples from the training set. **E** Table of Pearson's correlation coefficient values for the correlations between the 6 metagenes. **F** Heatmap showing the anticorrelation between the Immunity and the Hormone/Survival metagene.

doi:10.1371/journal.pone.0167397.g001

described in more detail in S1 File. For validation, we applied hierarchical clustering methods to three additional independent *HER2*-positive datasets; a validation set (n = 194), the Ignatiadis dataset (n = 82) and the METABRIC dataset (n = 248) (S1 File and S2 Fig).

The expression of the Immunity metagene is strongly associated with ER status, PR, and AR status

Given the inverse correlation between Immunity metagene and the Hormonal/survival metagene expression (Fig 2A) and with the strong correlation of Hormonal/survival metagene expression with ESR1 expression (Pearson correlation coefficient = 0.77), we compared levels of ESR1, PGR and AR expression as a function of Immunity metagene status (Fig 2B). These three genes were consistently more strongly expressed in the "Immunity low" subgroup than in the "Immunity high" subgroup ($p < 10^{-16}$, $p < 10^{-8}$, $p = 0.002$ respectively). Similar results were obtained with the other three datasets, although less consistently for PR and AR (S1 File).

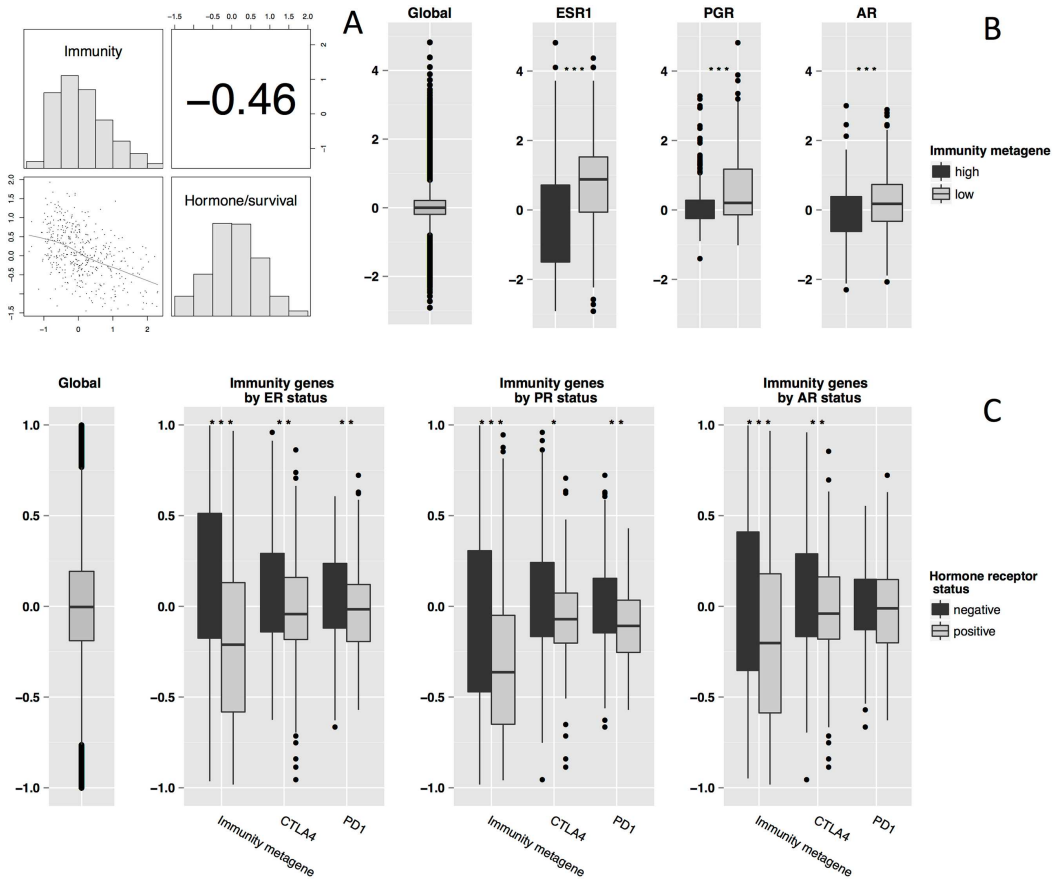


Fig 2. Association between Hormone genes expression and Immunity genes expression. A Correlation of Immunity metagene and Hormone/Survival metagene expression (training set). Pearson's correlation coefficient is -0.46 (95% CI [-52.7–38.0], $p<10^{-16}$). **B** Boxplots of global gene expression and ESR1, PGR and AR expression by Immunity metagene status, "low" versus "high" in the training set (A). P-values for ANOVA are $p = 10^{-16}$, $p = 10^{-6}$ and $p = 0.0002$, respectively. **C** Boxplots of Immunity metagene and immune gene (CTLA4 and PD1) expression levels by ER, PR and AR status in the training set (A). The p values for ANOVA were $p<10^{-16}$, $p = 0.002$ and $p = 0.008$ for the Immunity metagene, CTLA4 and PD1 by ER status, respectively; $p = 0.0001$, $p = 0.05$ and $p = 0.001$ by PR status, respectively; and $p<10^{-6}$, $p = 0.006$ and $p = 0.23$ by AR status, respectively. The statistical significance (p-value) of the difference between gene expression values is indicated by black stars (p-value $\leq$ 0.05: *; p-value$\leq$ 0.01: **; p-value$\leq$ 0.001: ***).

doi:10.1371/journal.pone.0167397.g002

We then compared the levels of expression of our Immunity metagene with those of two other immune genes (CTLA4 and PD1; PDL1 was not available on the HGU133a Chip) as a function of ER, PR, and AR status. The Immunity metagene and CTLA4 were significantly more strongly expressed in the ER-negative, PR-negative, and AR-negative subgroups (Fig 2C). PD1 was significantly more strongly expressed in ER-negative and PR-negative tumors, but the difference in expression levels according to AR status was not significant for this gene. Similar findings were obtained when we compared each of the genes of the Immunity metagene separately as a function of ER status, and across the three other datasets. The results were less consistent for PR and AR (see S1 File). The proportions of tumors in the Immunity

metagene "low" and "high" subgroups as a function of ER status differed significantly in three of the four datasets. ER-positive samples were more likely to be in the Immunity metagene "low" group, whereas ER-negative samples were more likely to be in the Immunity metagene "high" group (S1 File).

These findings suggest that there are strong inverse interactions between immune pathways that are captured by the Immunity metagene and ER, PR, and AR hormonal pathways in *HER2*-positive breast cancer tumors.

Predictive value of the Immunity metagene in *HER2*-positive breast cancers

We assessed the value of the six metagenes for predicting the response to neoadjuvant chemotherapy (NAC) on 82 *HER2*-positive samples from the Ignatiadis dataset. Univariate analysis identified four factors (ER status, tumor grade, and Immunity and Hormone/survival metagene expression) correlated with pathological complete response (pCR) (Table 1). In multivariate analysis, both ER status and the Immunity metagene were significantly associated with pCR (ER-positive: OR = 0.29 [0.09–0.82] versus ER-negative (reference class), $p = 0.02$; Immunity metagene "high" expression: OR = 3.71, 95% CI [1.28–11.91], *versus* "low" expression

Table 1. Association of clinical factors and gene cluster expression with pathological response rates after neoadjuvant chemotherapy in the Ignatiadis dataset, univariate and multivariate analysis.

		n	Univariate analysis			Multivariate analysis		
			OR	IC	*pval*	OR	IC	*pval*
Age	<50 y.o.	39	1					
	> = 50 y.o	43	1.1	[0.42–2.9]	0.84			
ER status	ER negative	38	1			1		
	ER positive	44	0.23	[0.08–0.63]	**0.006**	0.29	[0.09–0.82]	0.023
PR status	PR negative	78	1					
	PR positive	4	NA	NA*	0.99			
Tumoral size	T1 and T2	34	1					
	T3	21	0.34	[0.08–1.14]	0.096			
	T4	27	0.41	[0.12–1.23]	0.122			
Nodal status	N0	12	1					
	N1,N2 or N3	55	1.02	[0.26–5.1]	0.974			
Tumor grade	Grade I or II	24	1					
	Grade III	51	4.16	[1.22–19.26]	**0.037**			
Immunity metagene	low	41	1			1		
expression	high	41	4.57	[1.65–14.2]	**0.005**	3.71	[1.28–11.91]	**0.019**
Tumor suppressor/proliferation	low	41	1					
metagene	high	41	1.61	[0.62–4.3]	0.333			
Interferon metagene	low	41	1					
expression	high	41	0.49	[0.18–1.27]	0.149			
Signal transduction metagene	low	41	1					
expression	high	41	1.27	[0.49–3.33]	0.628			
Hormone/survival metagene	low	41	1					
expression	high	41	0.22	[0.07–0.61]	**0.005**			
Matrix metagene	low	41	1					
expression	high	41	1.27	[0.49–3.33]	0.628			

*: OR not available, no pCR in the PR-positive group

doi:10.1371/journal.pone.0167397.t001

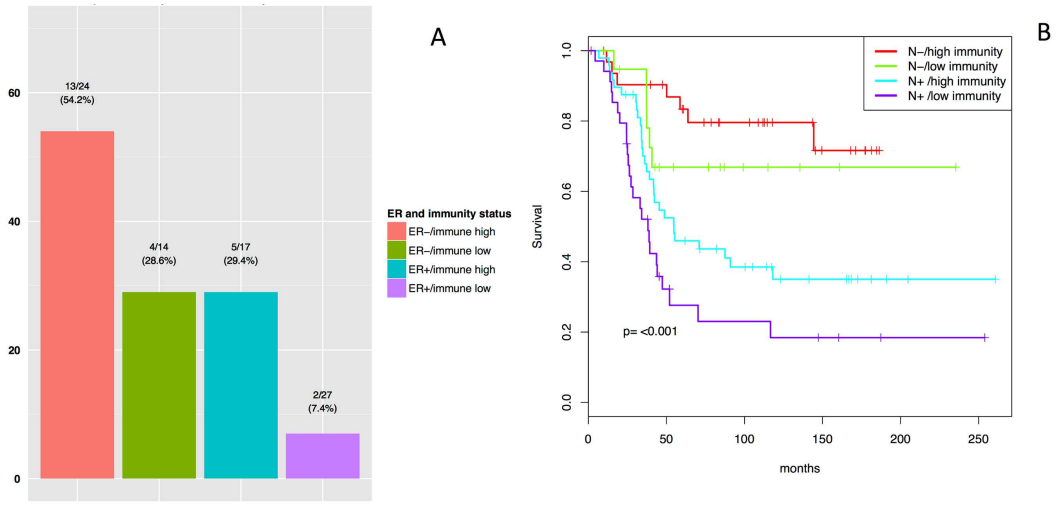


Fig 3. pCR and DSS outcomes in the Ignatiadis and the METABRIC dataset. A: pCR rates by ER and Immunity metagene status (low *versus* high in the Ignatiadis dataset). **B:** Kaplan-Meier plots. Disease-specific survival of the ER-negative population ($n = 138$) according to Immunity metagene expression (low/high) and nodal status in the METABRIC dataset.

doi:10.1371/journal.pone.0167397.g003

(reference class), $p = 0.02$) (Fig 3A). Analyses in the subset of patients that did not receive trastuzumab ($n = 75$) yielded similar results (S1 File).

We compared the predictive value of the Immunity metagene with that of nine immune signatures or metagenes already validated as predictors of the response to chemotherapy for breast cancer, notably in *HER2*-positive BCs [13–18]. In multivariate analysis, the Immunity metagene and six of the other signatures or metagenes tested were identified as predictive of the response to chemotherapy. The smallest *p*-value obtained was that for our Immunity metagene ($p = 0.019$), OR = 3.71, 95% CI [1.28–11.91] (S2 Table).

We then investigated the reasons for which the Immunity metagene (28 genes) was predictive of pCR in *HER2*-positive BCs, whereas the Immunity2 metagene (47 genes) published by Bonsang *et al.* [3] was not in a TNBC population [3], despite the strong correlation between these two signatures in three independent datasets (correlation coefficients: 0.96; 0.94 and 0.96 in the training set, METABRIC and Ignatiadis dataset, respectively). We applied both signatures to the whole population for the Ignatiadis dataset, and analyzed pCR as a function of breast cancer subtype and Immunity metagene status. We found that pCR rates were significantly higher in the "Immunity high" subgroup in *HER2*-negative/ER-positive (16.7% *versus* 8.4%, OR = 2.17, $p = 0.05$), *HER2*-positive (43.6% *versus* 16.7%, OR = 3.84, $p = 0.01$), and TNBC breast cancers (37.3 *versus* 22.6%, OR = 2.08, $p = 0.03$) (S3A Fig). A similar pattern was observed for the Immunity2 metagene (*HER2*-negative-ER positive: 16.5% *versus* 8.1%, OR = 2.22, $p = 0.05$), *HER2*-positive (45.5% *versus* 18.7%, OR = 3.57, $p = 0.01$), and TNBC breast cancers (36.3 *versus* 24.6%, OR = 1.75, $p = 0.08$; S3B Fig), but the difference was not statistically significant ($p = 0.08$) in the TNBC subgroup. Interestingly, Immunity metagene status appeared to have a larger effect on pCR rates in the *HER2*-positive subgroup (OR = 3.84 and 3.57, respectively) than in the ER-positive (OR = 2.17 and 2.22, respectively) and TNBC (OR = 2.08 and 1.75, respectively) subgroups. The Immunity metagene therefore seems to be

associated with the response to NAC in all breast cancer subtypes, with a marked effect in terms of both the strength and magnitude of the association in the *HER2*-positive subgroup.

Prognostic value of the Immunity metagene in *HER2*-positive breast cancers

The prognostic value of the 138-gene *HER2*-positive signature was assessed with 248 *HER2*-positive samples from the METABRIC dataset. Univariate analysis identified five factors (menopausal status, tumor size, nodal status, Immunity and Signal transduction metagene expression) significantly correlated with a poor outcome (disease-specific survival) (Table 2).

In multivariate analysis, nodal status (node-negative *versus* node-positive) was significantly associated with a poor outcome (HR = 3.29 [2.14–5.06], $p<0.001$), and there was a trend towards association between high levels of Immunity metagene expression and better disease-free survival (DFS; HR = 0.70 [0.48–1.01], $p = 0.054$). In the ER-negative population, the Immunity metagene was found to be of significant prognostic value in multivariate analysis ($n = 138$) (HR = 0.58 [0.36–0.94], $p = 0.026$; Fig 3B), but was not associated with DFS in the ER-positive population ($n = 110$) ($p = 0.43$). We compared the prognostic value of the Immunity metagene with that of nine previously published immune signatures or metagenes known to predict survival in several breast cancer subtypes [14,17–22]. None of the signatures or metagenes described above was significantly associated with prognosis (S2F Table).

The Immunity metagene is correlated with tumor-infiltrating lymphocytes (TILs) in *HER2*-positive breast cancer

We then investigated the correlation between Immunity metagene expression and lymphocyte infiltration. We analyzed an independent set of *HER2*-positive tumors for which both histology and gene expression data were available ($n = 27$). Intratumoral TILs (TLs) and stromal TILs (StrL) were evaluated separately. Intratumoral TIL percentages were significantly higher in patients with strong Immunity metagene expression than in those with weak Immunity metagene expression (24% and 9%, respectively, $p = 0.001$) (Fig 4A). The same pattern was observed for the percentage of stromal TILs (36% versus 16.6%, $p = 0.009$) (Fig 4B). The coefficients of correlation between Immunity metagene expression level on the one hand and the percentage of intratumoral TILs (Fig 4C) or stromal TILs (Fig 4D) on the other hand were high ($r = 0.60$, $p<0.001$ and $r = 0.69$, $p<0.00001$ respectively). Lymphocyte infiltration is shown for two specimens, one with weak (Fig 5A and 5B), and the other with strong lymphocyte infiltration (Fig 5C and 5D). The Immunity metagene was therefore strongly correlated with the amount of lymphocyte infiltration in both the stromal compartment and the tumor bed.

The Immunity metagene corresponds to the B-cell, T-cell and CD8 cell pathways

The Immunity metagene was strongly correlated with several published immune signatures (S4 Fig and S1 File), suggesting the use of similar immune pathways (see S1 File). We analyzed the correlation between expression of the Immunity and Interferon metagenes and expression of the metagenes defined by Gatza *et al.* [23] (IFN-alpha, IFN-gamma, STAT3, TGF-beta, TNF-alpha) and Palmer *et al.* [24] (LB, LT, CD8, GRANS, LYMPHS). This analysis was performed on the METABRIC dataset. The Immunity metagene was highly correlated with the B-cell, T-cell and CD8 cell metagenes (Pearson correlation coefficients: 0.89, 0.86, and 0.90, respectively; S5 Fig). We also assessed the correlations between the expression of PD1, PDL1,

Table 2. Survival analysis (disease-specific survival) in the METABRIC dataset (univariate and multivariate analysis); whole population and ER-negative population.

		Whole population (n = 248)							ER negative population (n = 138)						
		n	Univariate analysis			Multivariate analysis			n	Univariate analysis			Multivariate analysis		
			HR	IC	*pval*	HR	IC	*pval*		HR	IC	*pval*	HR	IC	*pval*
Age at diagnosis	< = 45 y. o.	52	1		-				25	1					
	45–55	59	0.67	[0.4–1.14]	0.142				19	0.65	[0.36–1.18]	0.153			
	>55	130	0.66	[0.43–1.04]	**0.071**				23	0.62	[0.35–1.1]	0.103			
Menopausal status	Pre	74	1		-				32	1					
	Post	167	0.68	[0.46–1]	**0.051**				33	0.67	[0.41–1.09]	0.11			
Tumoral size	< 20 mm	68	1		-				15	1					
	> = 20 mm	173	1.87	[1.18–2.96]	**0.008**				52	1.51	[0.85–2.69]	0.159			
Tumor grade	I	3	1		-				10	1	-	-			
	II	53	1.66	[0.22–12.19]	0.621				55	0.942	[0.48–1.85]	0.863			
	III	178	1.81	[0.25–13.05]	0.554				10	NA	NA	NA			
ER status	negative	135	1		-										
	positive	108	0.74	[0.51–1.07]	0.108										
PR status	negative	193	1		-				65	1					
	positive	50	0.84	[0.53–1.34]	0.46				2	2.3	[0.56–9.49]	0.25			
Nodal status	N-	105	1		-	1			13	1			1		
	N+	138	3.26	[2.13–5.01]	**<0.001**	3.29	[2.14–5.06]	**<0.001**	54	3.55	[1.93–6.51]	**<0.001**	3.57	[1.94–6.55]	**<0.001**
NPI	GP	38	1		-				6	1					
	IP	155	1.26	[0.71–2.25]	0.433				35	1.01	[0.42–2.4]	0.988			
	PP	50	3.32	[1.78–6.19]	**<0.001**				26	2.81	[1.15–6.84]	**0.023**			
Metagene expression															
Immunity	low	122	1		-	1			54	1					
	high	121	0.71	[0.49–1.03]	**0.073**	0.70	[0.48–1.01]	**0.054**	81	0.58	[0.36–0.94]	**0.028**	0.58	[0.36–0.94]	**0.026**
TS /proliferation	low	121	1		-				50	1					
	high	122	1.04	[0.72–1.51]	0.828				85	0.84	[0.51–1.38]	0.491			
Interferon	low	122	1		-				78	1					
	high	121	1.23	[0.85–1.78]	0.278				57	1.28	[0.79–2.07]	0.316			
Signal transduction	low	121	1		-				72	1					
	high	122	1.48	[1.02–2.14]	**0.04**				63	1.34	[0.83–2.17]	0.232			
Hormone/ survival	low	122	1		-				114	1					
	high	121	0.94	[0.65–1.36]	0.751				21	1.35	[0.72–2.52]	0.351			
Matrix	low	121	1		-				69	1					
	high	122	1.05	[0.73–1.52]	0.785				66	1.03	[0.64–1.67]	0.889			

Abbreviations: GP: good prognosis, IP: intermediate prognosis, PP: poor prognosis; TS: tumor suppressor

doi:10.1371/journal.pone.0167397.t002

CTLA4, and that of their respective metagenes. The PD1 and CTLA-4 metagenes were constructed from the genes most strongly correlated with the PD1 and CTLA-4 genes, respectively (Pearson's correlation coefficient > 0.8). The PDL1 metagene was defined by Sabatier *et al.* [25]. Pearson's correlation coefficients for the relationships between the Immunity metagene and each individual gene were strong for PD1 and CTLA-4 (Pearson's correlation coefficient: 0.75 and 0.84, respectively), and weaker for PDL1 (0.36), but the expression of all three metagenes was strongly correlated with that of the Immunity metagene (Pearson's correlation

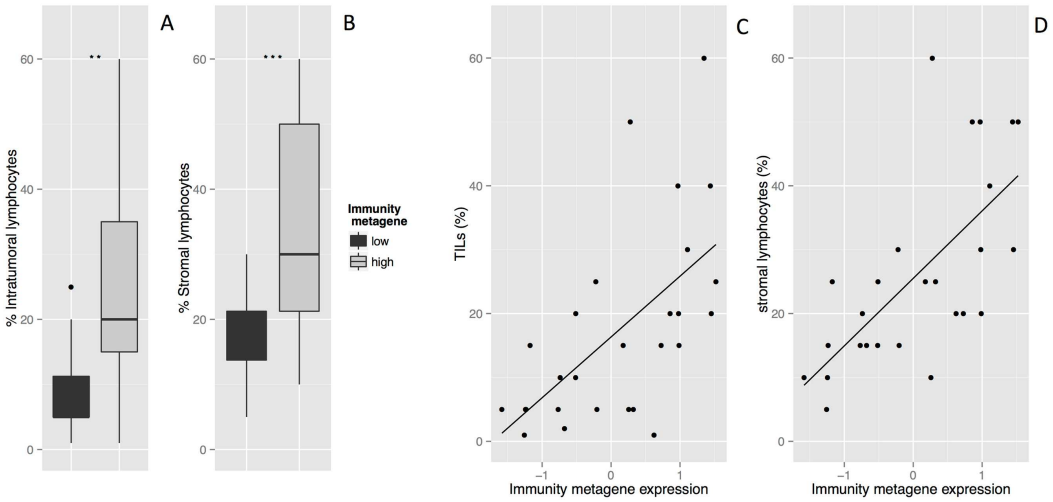


Fig 4. Association between tumor-infiltrating lymphocyte levels and Immunity metagene expression in the REMAGUS dataset. A: Percentage of intratumoral TILs according to Immunity metagene status (low *versus* high). **B** Percentage of stromal TILs according to Immunity metagene status (low *versus* high). **C**: Correlation between metagene expression and the percentages of intratumoral TILs. **D**: Correlation between metagene expression and the percentage of stromal TILs.

doi:10.1371/journal.pone.0167397.g004

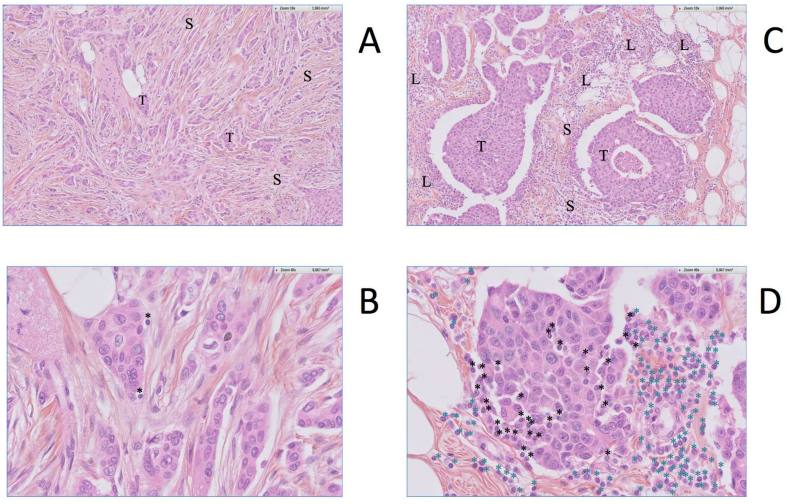


Fig 5. Lymphocytic infiltration in breast tumors. A and **B:** Tumor specimen with weak lymphocytic infiltration (A: zoom x10 B: zoom x 40). Abbreviations: S = stroma, T = tumor, L = lymphocytes. Intratumoral TILs are indicated by a black star. **C** and **D**: Tumor specimen with prominent lymphocytic infiltration. (C: zoom x10 D: zoom x 40). Abbreviations: S = stroma, T = tumor, L = lymphocytes. Intratumoral TILs are indicated by a black star; stromal TILs are indicated by a blue star.

doi:10.1371/journal.pone.0167397.g005

coefficient: PD1: 0.89, PDL1: 0.95, CTLA-4: 0.93), opening up new possibilities for therapeutic intervention.

The Immunity metagene is probably expressed by stromal cells

In breast cancer cell lines (CCLE and CGP datasets), the Immunity metagene displayed very low levels of expression, similar to those of the CD8 metagene (S6A and S6B Fig), consistent with expression only in the tumor stromal compartment. This pattern was observed for all cell lines and breast cancer cell lines tested. The Interferon module genes had higher median expression levels and a broader range of expression than those of the Immunity metagene in breast cancer cell lines, consistent with their expression by tumor cells. We also explored the contributions of stromal and cancer cells to the expression of the Immunity and Interferon metagenes in detail, by comparing our gene lists with the "stromal contribution to global gene expression evaluated in PDX RNAseq data", as defined by Isella *et al.* [26]. The stromal fraction of the Immunity metagene was high, although lower than those of the Matrix and the Tumor suppressor/proliferation metagenes. The Interferon metagene had a low stromal fraction, like the Hormone/survival and Signal transduction metagenes (S6C Fig). Although these data relate to the colon cancer PDX model, they provide support for the stromal expression of the Immunity metagene.

Discussion

By analyzing the gene expression profiles of 448 *HER2*-positive breast cancers, we identified a six-metagene signature (138 genes) in which each of the various metagenes was enriched in a different gene ontology. Within these metagenes, we identified an immune stromal module inversely correlated with the ER and hormonal pathways and strongly associated with the predicted response to chemotherapy, prognosis, and tumor lymphocyte infiltration. We report here one of the first immune signatures identified as both predictive and prognostic, reflecting histological immune infiltration in *HER2*-positive breast cancers. We also provide a relevant analysis by HR status.

We previously developed a strategy for defining gene expression signatures based on the analysis of biological networks for the most variable genes [3]. Since the early 2000s, a molecular classification of breast cancers has emerged that is continually being refined. Several authors have proposed TNBC subclassifications [3,27,28] but, to our knowledge, only one classifier has been published, but was not subsequently validated in *HER2*-positive BC [18]. The various metagenes in our signature were enriched in different gene ontologies: two clusters were enriched in immunity genes, one in signal transduction genes, one in hormonal/survival genes, one in tumor suppressor/proliferation genes and one in matrix genes. Unlike several other teams [29–31], we did not identify a subgroup to tumors overexpressing androgen receptor pathways in *HER2*-positive BCs by our biology-driven approach. The expression of the Immunity and Hormone/survival metagenes accurately predicted the response to NAC, but the expression of the Hormone/survival metagene had no significant effect in multivariate analysis, because the information it provided largely overlapped with ER status. Moreover, only the Immunity metagene was found to be of significant prognostic value.

Several authors have previously identified immunity patterns in *HER2*-positive BC. The Immunity module identified in our study had many biological connections with other predictive or prognostic immune signatures published for *HER2*-positive breast cancers [13–21], but it outperformed previous classifiers. This module includes genes encoding chemokines for T cells (CXCL10, CXCL9, CCL5), B cells (CXCL13), both B and T cells (CCL19) or other immune cells (CXCL13, CCL5); chemokine receptors (CCR7); cytokines (LTB); adhesion

molecule-associated genes (SELL), and genes encoding proteins involved antigen processing and presentation (HLA-DRA), B-lymphocyte cell surface molecules (PTPRC, HLA-DRA), complement pathway proteins (C1QB), and proteins involved in CTL-mediated immune responses to target cells (CD3D), dendritic cell regulation of Th1 and Th2 development (CD2, IL7R), granzyme-mediated apoptosis (GZMA), IL12-mediated signaling events (CD3D, HLA-DRA, GZMA, LCK), the IL2 signaling pathway (LCK), T-cell surface molecules (PTPRC, CD3D, CD2), and molecules of the T-cell receptor signaling pathway (PTPRC, CD3D, HLA-DRA, LCK). It was also strongly correlated with the B-cell, T-cell and CD8 cell pathways.

There was a marked significant inverse association between ESR1 expression and that of the Immunity metagene. Similar inverse associations were found between PGR, AR and immunity, but these associations were weaker and less consistent. There is growing evidence for sex-based differences in the innate and adaptive immune responses underlying susceptibility to infectious diseases and the prevalence of autoimmune diseases. A higher proportion of men than of women display infectious diseases and their severity is also greater in men than in women [32]. By contrast, many autoimmune diseases predominantly affect women [33]. There are also difference between men in terms of humoral and cellular responses to infection and vaccination, with women often displaying higher response rates and mounting stronger humoral responses [34]. Estrogen receptors are expressed in most of the cells of the innate and adaptive immune system, including T cells, B cells, neutrophils, macrophages, dendritic cells (DC), and natural killer (NK) cells [35]. The effects of major sex steroid hormones were reviewed by Giefing-Kröll [36]. Estradiol and testosterone have opposite effects on the cells of the adaptive and innate immune systems, with estradiol having mostly enhancing and testosterone mostly suppressive effects. Estrogens affect the expression of some chemokine receptors (CCR1 and CCR5) by T cells [37]. They also affect B-cell development [38], decrease the cytotoxicity of NK cells [39] and regulate DC development [40]. TReg-cell frequencies within the $CD4^+$ population change considerably during the ovarian cycle, with potential effects on immunoregulation [41]. Unlike the differences between the sexes in terms of infection and auto-immunity, the relationships between tumor immunology, sex and steroid hormones have remained largely unexplored. In two phase III trials, immunotherapy had a significant beneficial effect on survival only in male patients [42,43]. However, it remains unclear whether there is a true "sex" effect on the efficacy of immunotherapy or whether these findings are purely incidental.

The interaction between the ER, immunity and *HER2* pathways is complex. There is increasing evidence to suggest that interactions between *HER2* and hormone-receptor pathways play an important role in disease progression and that there is extensive, complex, bidirectional, crosstalk between the *HER2* and ER pathways [44]. Immune signatures have been reported to have a predictive or prognostic role mostly in ER-negative breast cancers [45–48]. In *HER2*-positive breast cancer subtypes, Rody found that an immune T-cell metagene was of predictive value in both ER-positive and ER-negative *HER2*-positive BC [49]. The prognostic value of HDDP was demonstrated in both subgroups (11), but its value for predicting the response to NAC was not evaluated as a function of ER status. Conversely, the IRSN-23 [15] was not predictive in the ER-positive subpopulation. However, few authors determined the predictive [18] or prognostic value of their metagene or signature as a function of ER status within *HER2*-positive breast cancers [5,13,14,19–21]. The inverse association observed between ESR1 expression and immunity genes may be an important piece of the puzzle, and merits further investigation.

Consistent with previous reports [13,15,16], we found that the Immunity metagene was predictive of the response to NAC in *HER2*-positive BC. However, despite the similar gene

module identification methods used and the strong correlation between the Immunity metagene and the Immunity2 metagene previously described by our team for TNBC [3], the Immunity metagene was predictive of the response to chemotherapy in *HER2*-positive BC, whereas the Immunity2 metagene was not predictive of the response to chemotherapy in TNBC. This finding was reported in the princeps report by Ignatiadis, in which high immune module scores were strongly and independently associated with a higher probability of pCR probability in *HER2*-positive tumors, whereas this association, although still significant, was weaker in TNBC [5]. ER-positive tumors have long been described as chemoresistant, with low pCR rates after NAC. Taking Immunity metagene expression into account, pCR rates ranged from 7.4 to 29.4%, with the highest rates close to those of ER-negative tumors.

The Immunity metagene was also prognostic in *HER2*-positive ER-negative breast cancer. The impact of immunity on prognosis has been reported before [21](Alexe et al., 2007) [21][14][18,20,21]. Together with our work, these findings suggest that immunity gene expression is highly predictive and of prognostic value in *HER2*-positive breast cancer. Nevertheless, the *HER2*-positive patients of the METABRIC dataset did not receive targeted anti-*HER2* therapies, and our results would probably be influenced by adjuvant trastuzumab treatment.

We also demonstrated a correlation between Immunity metagene expression and stromal and intratumoral lymphocyte infiltration. The significance of TILs has recently become apparent, with advances in tumor immunology and the availability of cancer immunotherapies. TIL levels are strongly correlated with breast cancer subtype, and are higher in *HER2*-positive BCs than in ER-positive BCs, but lower than in TNBCs [50]. TIL levels are consistently higher in ER-negative tumors than in ER-positive tumors [51]. This was also found to be the case when the analysis was limited to *HER2*-positive BC only [52], [50]. The value of TIL levels for predicting pCR after NAC is less clear in *HER2*-positive BC than in TNBC. Stromal TILs and the lymphocyte-predominant breast cancer phenotype (LPBC) were strongly associated with treatment response in the GeparSixto trial [13]. However, this effect was found to be nonlinear in the NeoALTTO trial, and the optimal cutoff value remains unclear [52]. Two large studies in the adjuvant setting gave conflicting results. A positive association between higher levels of TILs and greater benefit from trastuzumab in *HER2*-positive disease was found in a retrospective analysis of the FinHER trial [50], whereas the opposite result was reported in the ALLIANCE N9831 study [53]. No difference in DFS between chemotherapy and chemotherapy plus trastuzumab was found in LPBC, whereas benefits of trastuzumab in addition to chemotherapy were observed only in non-LPBC. Thus, the prognostic impact of TILs on survival remains a matter of debate in *HER2*-positive BC. A few authors have reported a correlation between TIL and stromal lymphocyte levels and gene expression in *HER2*-positive breast cancers [13,15,21]. If this correlation is further validated, TIL levels could be used as a surrogate marker for the Immunity metagene, as TIL assessment is carried out in routine practice and is currently undergoing standardization [54].

Conclusion

Our work opens up a number of exciting therapeutic perspectives in *HER2*-positive breast cancers. Due to the high immunogenicity of *HER2*-positive breast cancers and the considerable predictive and prognostic impact of immunity in this subtype, immunotherapies may soon become part of the therapeutic arsenal for such cancers. Preclinical models have suggested that there is synergy between anti-*HER2* monoclonal antibody and anti-PD-1 [55] or anti-CTLA4 antibodies [56]. The PANACEA phase Ib/II trial is currently investigating the use of pembrolizumab (KEYTRUDA®) in combination with trastuzumab, to determine whether the addition of an anti-PD-1 treatment can overcome trastuzumab resistance in patients with *HER2*-

positive breast cancer whose cancer spread whilst they were on trastuzumab. Future challenges in the field of immunity and *HER2*-positive breast cancers include:

1. The public accessibility of large sets of gene expression data for tumors from patients treated with *HER2*-targeting treatments. As treatments are constantly changing for this breast cancer subtype, it is important for expression data to be shared promptly, to facilitate comprehensive research and the identification of predictive and prognostic markers in patients treated with cutting edge care.
2. Improvements in our understanding of hormone and immunity pathways in *HER2*-positive breast cancers. In particular, it would be very useful to determine whether a subset of patients with *HER2*-positive ER-positive cancers could be effectively treated by a combination of endocrine therapy/immune checkpoint blockade/ targeted therapy, without the need for chemotherapy.
3. Drug positioning strategies in HER2-positive BC, because, by contrast to other breast cancer subtypes, the *HER2*-targeting drug pipeline contains many candidates despite the comparative rarity of this particular disease.
4. The selection criteria for the candidates most likely to benefit from immune checkpoint blockade is a key point. The use of PD-L1 as a surrogate marker of anti-PD-1 efficacy remains controversial, even in cancers for which immunotherapy treatments have proved effective, and few data are available for breast cancer. The standardization and demonstrations of the reproducibility of published immune signatures would be useful, as would improvements in our understanding of the prognostic value of TILs in *HER2*-positive breast cancers. Moreover, it remains to be determined whether and how the immunogenic power of tumors with low expression of immunity genes could be enhanced.

Once these challenges have been overcome, given the outstanding results of immunotherapy for other cancers (e.g. melanoma, lung cancer) and the expected efficacy of such treatment for *HER2*-positive disease, such therapies could revolutionize the course of *HER2*-positive breast cancer in the near future.

Supporting Information

S1 Fig. Methodology flow chart.
(PDF)

S2 Fig. Heatmaps of the selected genes in the *HER2*-positive datasets. Training set (upper left); validation set (upper right), Ignatiadis (lower left), METABRIC (lower right).
(JPG)

S3 Fig. pCR rates by breast cancer subtype and Immunity metagene. A: pCR rates by breast cancer subtype by Immunity metagene status (low *versus* high). B: pCR rates by breast cancer subtype by Immunity2 metagene status (low *versus* high) as previously published by Bonsang et al [3].
(PDF)

S4 Fig. Heatmaps of the gene expression profiles of published immune signatures and connections between all immune genes. Fig A. Heatmap of the gene expression profiles of the nine immune predictive signatures or metagenes previously published, applied to the Ignatiadis dataset. The samples were ordered according to our classification of Low/High 'Immunity' metagene expression. B: Heatmap of the gene expression profiles of the immune prognostic signatures or metagenes previously published, applied to the METABRIC dataset. The samples

were ordered according to our classification of Low/High 'Immunity' metagene expression. C: String Software connections between genes of our Immunity metagenes and the genes of previously published predictive or prognostic immune signatures or metagenes. Stronger associations between genes are represented by thicker lines. Associations between genes with a coefficient < 0.9 are shown in green. Associations between genes with a coefficient ≥ 0.9 are shown in red. Associations between genes with a coefficient between 0.4 to 0.7 are not shown.
(PDF)

S5 Fig. Distribution histograms for our Immune metagenes and immune pathways. Distribution histograms for our Immune metagenes (Immunity and Interferon) and the immune pathway metagenes published by Gatza *et al.* (Interferon alpha, Interferon gamma, STAT3, TGF beta, TNF alpha) and Palmer *et al.* (B Cell, T Cell, CD8 T Cells, Granulocytes, Lymphocytes), Pearson correlation coefficient values and pairwise scatter plots.
(PDF)

S6 Fig. Gene expression for the Immune metagenes and pathway, in cell lines and xenografts. A. Boxplots of gene expression for the Immune metagenes, the immune pathway metagenes (published by Gatza *et al.* and Palmer *et al.*) and the PD1, PDL1, CTLA4 metagenes in breast cancer cell lines from the CCLE (A) and the CGP (B). C: Boxplots of the stromal contribution to global gene expression evaluated with PDX RNAseq data (Isella *et al.*), for each of the gene clusters for our signature.
(PDF)

S1 File. Supplementary methods and results.
(PDF)

S1 Table. 138-gene signature.
(XLS)

S2 Table. Association of published immune signatures or metagenes with response to chemotherapy and prognosis. Response to chemotherapy is assessed in the Ignatiadis dataset (univariate and multivariate analysis) (S2A to S2E Table). The association of published immune signatures or metagenes with prognosis is assessed in the METABRIC dataset (univariate analysis) (S2F Table).
(XLS)

Acknowledgments

The authors thank Vassili Soumelis and Sergio Roman-Roman for reviewing of the study and the manuscript.

Author Contributions

Conceptualization: FR AP MG.

Data curation: BS.

Formal analysis: BS JA.

Investigation: BS CL.

Methodology: CL MM.

Project administration: FR CL.

Resources: ML.

Supervision: FR.

Validation: ML.

Writing – original draft: ASH HBK.

Writing – review & editing: ASH.

References

1. Fröhlich H. Network based consensus gene signatures for biomarker discovery in breast cancer. PLoS One. 2011; 6: e25364. doi: 10.1371/journal.pone.0025364 PMID: 22046239
2. Cun Y, Fröhlich HF. Prognostic gene signatures for patient stratification in breast cancer: accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. BMC Bioinformatics. 2012; 13: 69. doi: 10.1186/1471-2105-13-69 PMID: 22548963
3. Bonsang-Kitzis H, Sadacca B, Hamy-Petit AS, Moarii M, Pinheiro A, Laurent C, et al. Biological network-driven gene selection identifies a stromal immune module as a key determinant of triple-negative breast carcinoma prognosis. Oncoimmunology. 2015;
4. Servant N, Gravier E, Gestraud P, Laurent C, Paccard C, Biton A, et al. EMA—A R package for Easy Microarray data analysis. BMC Res Notes. BioMed Central Ltd; 2010; 3: 277.
5. Ignatiadis M, Singhal SK, Desmedt C, Haibe-Kains B, Criscitiello C, Andre F, et al. Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis. J Clin Oncol. 2012; 30: 1996–2004. doi: 10.1200/JCO.2011.39.5624 PMID: 22508827
6. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486: 346–352. doi: 10.1038/nature10983 PMID: 22522925
7. Gong Y, Yan K, Lin F, Anderson K, Sotiriou C, Andre F, et al. Determination of oestrogen-receptor status and ERBB2 status of breast carcinoma: a gene-expression profiling study. Lancet Oncol. 2007; 8: 203–211. doi: 10.1016/S1470-2045(07)70042-6 PMID: 17329190
8. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2009; 37: D412–6. doi: 10.1093/nar/gkn760 PMID: 18940858
9. De Cremoux P, Valet F, Gentien D, Lehmann-Che J, Scott V, Tran-Perennou C, et al. Importance of pre-analytical steps for transcriptome and RT-qPCR analyses in the context of the phase II randomised multicentre trial REMAGUS02 of neoadjuvant chemotherapy in breast cancer patients. BMC Cancer. 2011; 11: 215. doi: 10.1186/1471-2407-11-215 PMID: 21631949
10. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. NIH Public Access of anticancer drug sensitivity. 2012; 483: 603–607.
11. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Europe PMC Funders Group Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012; 483: 570–575.
12. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Sta- tistical Computing. Vienna, Austria 2009. [Internet]. http://www.R-project.org
13. Denkert C, von Minckwitz G, Brase JC, Sinn BV, Gade S, Kronenwett R, et al. Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without Carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers. J Clin Oncol Off J Am Soc Clin Oncol. 2015; 33: 983–991.
14. Gu-Trantien C, Loi S, Garaud S, Equeter C, Libin M, Wind De A, et al. CD4+ follicular helper T cell infiltration predicts breast cancer survival. J Clin Invest. 2013; 123: 1–20.
15. Sota Y, Naoi Y, Tsunashima R, Kagara N, Shimazu K, Maruyama N, et al. Construction of novel immune-related signature for prediction of pathological complete response to neoadjuvant chemotherapy in human breast cancer. Ann Oncol Off J Eur Soc Med Oncol ESMO. 2014; 25: 100–106.
16. Stoll G, Enot D, Mlecnik B, Galon J, Zitvogel L, Kroemer G. Immune-related gene signatures predict the outcome of neoadjuvant chemotherapy. Oncoimmunology. 2014; 3: e27884. doi: 10.4161/onci.27884 PMID: 24790795
17. Rody A, Karn T, Liedtke C, Pusztai L, Ruckhaeberle E, Hanker L, et al. A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast Cancer Res BCR. 2011; 13: R97. doi: 10.1186/bcr3035 PMID: 21978456

18. Staaf J, Ringnér M, Vallon-Christersson J, Jönsson G, Bendahl P-O, Holm K, et al. Identification of subtypes in human epidermal growth factor receptor 2—positive breast cancer reveals a gene signature prognostic of outcome. J Clin Oncol Off J Am Soc Clin Oncol. 2010; 28: 1813–1820.
19. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, et al. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res. 2008; 14: 5158–5165. doi: 10.1158/1078-0432.CCR-07-4756 PMID: 18698033
20. Finak G, Bertos N, Pepin F, Sadekova S, Souleimanova M, Zhao H, et al. Stromal gene expression predicts clinical outcome in breast cancer. Nat Med. 2008; 14: 518–527. doi: 10.1038/nm1764 PMID: 18438415
21. Alexe G, Dalgin GS, Scanfeld D, Tamayo P, Mesirov JP, DeLisi C, et al. High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates. Cancer Res. 2007; 67: 10669–10676. doi: 10.1158/0008-5472.CAN-07-0539 PMID: 18006808
22. Perez EA, Thompson EA, Ballman KV, Anderson SK, Asmann YW, Kalari KR, et al. Genomic Analysis Reveals That Immune Function Genes Are Strongly Linked to Clinical Outcome in the North Central Cancer Treatment Group N9831 Adjuvant Trastuzumab Trial. J Clin Oncol Off J Am Soc Clin Oncol. 2015;
23. Gatza ML, Lucas JE, Barry WT, Kim JW, Wang Q, Crawford MD, et al. A pathway-based classification of human breast cancer. Proc Natl Acad Sci U S A. 2010; 107: 6994–6999. doi: 10.1073/pnas.0912708107 PMID: 20335537
24. Palmer C, Diehn M, Alizadeh A a, Brown PO. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. BMC Genomics. 2006; 7: 115. doi: 10.1186/1471-2164-7-115 PMID: 16704732
25. Sabatier R, Finetti P, Mamessier E, Adelaide J, Chaffanet M, Ali HR, et al. Prognostic and predictive value of PDL1 expression in breast cancer. Oncotarget. 2015; 6: 5449–5464. doi: 10.18632/oncotarget.3216 PMID: 25669979
26. Isella C, Terrasi A, Bellomo SE, Petti C, Galatola G, Muratore A, et al. Stromal contribution to the colorectal cancer transcriptome. Nat Genet. Nature Publishing Group; 2015; 1–11.
27. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011; 121: 2750–2767. doi: 10.1172/JCI45014 PMID: 21633166
28. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua S, et al. Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-negative Breast Cancer. Clin Cancer Res Off J Am Assoc Cancer Res. 2014;
29. Doane AS, Danso M, Lal P, Donaton M, Zhang L, Hudis C, et al. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. Oncogene. 2006; 25: 3994–4008. doi: 10.1038/sj.onc.1209415 PMID: 16491124
30. Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, et al. Identification of molecular apocrine breast tumours by microarray analysis. Oncogene. 2005; 24: 4660–4671. doi: 10.1038/sj.onc.1208561 PMID: 15897907
31. Lehmann-Che J, Hamy A-S, Porcher R, Barritault M, Bouhidel F, Habuellelah H, et al. Molecular apocrine breast cancers are aggressive estrogen receptor negative tumors overexpressing either HER2 or GCDFP15. Breast Cancer Res. 2013; 15: R37. doi: 10.1186/bcr3421 PMID: 23663520
32. Klein SL. The effects of hormones on sex differences in infection: from genes to behavior. Neurosci Biobehav Rev. 2000; 24: 627–638. PMID: 10940438
33. Lockshin MD. Sex differences in autoimmune disease. Lupus. 2006; 15: 753–756. PMID: 17153846
34. Cook IF. Sexual dimorphism of humoral immunity with human vaccines. Vaccine. 2008; 26: 3551–3555. doi: 10.1016/j.vaccine.2008.04.054 PMID: 18524433
35. Fish EN. The X-files in immunity: sex-based differences predispose immune responses. Nat Rev Immunol. 2008; 8: 737–744. doi: 10.1038/nri2394 PMID: 18728636
36. Giefing-Kröll C, Berger P, Lepperdinger G, Grubeck-Loebenstein B. How sex and age affect immune responses, susceptibility to infections, and response to vaccination. Aging Cell. 2015; 14: 309–321. doi: 10.1111/acel.12326 PMID: 25720438
37. Mo R, Chen J, Grolleau-Julius A, Murphy HS, Richardson BC, Yung RL. Estrogen regulates CCR gene expression and function in T lymphocytes. J Immunol Baltim Md 1950. 2005; 174: 6023–6029.
38. Sakiani S, Olsen NJ, Kovacs WJ. Gonadal steroids and humoral immunity. Nat Rev Endocrinol. 2013; 9: 56–62. doi: 10.1038/nrendo.2012.206 PMID: 23183675
39. Hao S, Zhao J, Zhou J, Zhao S, Hu Y, Hou Y. Modulation of 17beta-estradiol on the number and cytotoxicity of NK cells in vivo related to MCM and activating receptors. Int Immunopharmacol. 2007; 7: 1765–1775. doi: 10.1016/j.intimp.2007.09.017 PMID: 17996687

40. Siracusa MC, Overstreet MG, Housseau F, Scott AL, Klein SL. 17beta-estradiol alters the activity of conventional and IFN-producing killer dendritic cells. J Immunol Baltim Md 1950. 2008; 180: 1423–1431.
41. Arruvito L, Sanz M, Banham AH, Fainboim L. Expansion of CD4+CD25+and FOXP3+ regulatory T cells during the follicular phase of the menstrual cycle: implications for human reproduction. J Immunol Baltim Md 1950. 2007; 178: 2572–2578.
42. Robert C, Thomas L, Bondarenko I, O'Day S, Weber J, Garbe C, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. N Engl J Med. 2011; 364: 2517–2526. doi: 10.1056/NEJMoa1104621 PMID: 21639810
43. Brahmer J, Reckamp KL, Baas P, Crinò L, Eberhardt WEE, Poddubskaya E, et al. Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. N Engl J Med. 2015; 373: 123–135. doi: 10.1056/NEJMoa1504627 PMID: 26028407
44. Giuliano M, Trivedi MV, Schiff R. Bidirectional Crosstalk between the Estrogen Receptor and Human Epidermal Growth Factor Receptor 2 Signaling Pathways in Breast Cancer: Molecular Basis and Clinical Implications. Breast Care Basel Switz. 2013; 8: 256–262.
45. Sabatier R, Finetti P, Mamessier E, Raynaud S, Cervera N, Lambaudie E, et al. Kinome expression profiling and prognosis of basal breast cancers. Mol Cancer. 2011; 10: 86. doi: 10.1186/1476-4598-10-86 PMID: 21777462
46. Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol. 2007; 8: R157. doi: 10.1186/gb-2007-8-8-r157 PMID: 17683518
47. Karn T, Pusztai L, Holtrich U, Iwamoto T, Shiang CY, Schmidt M, et al. Homogeneous datasets of triple negative breast cancers enable the identification of novel prognostic and predictive signatures. PloS One. 2011; 6: e28403. doi: 10.1371/journal.pone.0028403 PMID: 22220191
48. West NR, Milne K, Truong PT, Macpherson N, Nelson BH, Watson PH. Tumor-infiltrating lymphocytes predict response to anthracycline-based chemotherapy in estrogen receptor-negative breast cancer. Breast Cancer Res BCR. BioMed Central Ltd; 2011; 13: R126.
49. Rody A, Holtrich U, Pusztai L, Liedtke C, Gaetje R, Ruckhaeberle E, et al. T-cell metagene predicts a favorable prognosis in estrogen receptor-negative and HER2-positive breast cancers. Breast Cancer Res BCR. 2009; 11: R15. doi: 10.1186/bcr2234 PMID: 19272155
50. Loi S, Michiels S, Salgado R, Sirtaine N, Jose V, Fumagalli D, et al. Tumor infiltrating lymphocytes are prognostic in triple negative breast cancer and predictive for trastuzumab benefit in early breast cancer: results from the FinHER trial. Ann Oncol Off J Eur Soc Med Oncol ESMO. 2014; 25: 1544–1550.
51. Mahmoud SMA, Paish EC, Powe DG, Macmillan RD, Grainge MJ, Lee AHS, et al. Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2011; 29: 1949–1955.
52. Salgado R, Denkert C, Campbell C, Savas P, Nucifero P, Aura C, et al. Tumor-Infiltrating Lymphocytes and Associations With Pathological Complete Response and Event-Free Survival in HER2-Positive Early-Stage Breast Cancer Treated With Lapatinib and Trastuzumab: A Secondary Analysis of the NeoALTTO Trial. JAMA Oncol. 2015; 1: 448–454. doi: 10.1001/jamaoncol.2015.0830 PMID: 26181252
53. Perez E. Stromal tumor-infiltrating lymphocytes(S-TILs): In the alliance N9831 trial S-TILs are associated with chemotherapy benefit but not associated with trastuzumab benefit [Internet]. 2014. http://www.abstracts2view.com/sabcs14/view.php?nu=SABCS13L_1455
54. Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. Ann Oncol Off J Eur Soc Med Oncol ESMO. 2015; 26: 259–271.
55. Stagg J, Loi S, Divisekera U, Ngiow SF, Duret H, Yagita H, et al. Anti-ErbB-2 mAb therapy requires type I and II interferons and synergizes with anti-PD-1 or anti-CD137 mAb therapy. Proc Natl Acad Sci U S A. 2011; 108: 7142–7147. doi: 10.1073/pnas.1016569108 PMID: 21482773
56. Wang Q, Li S-H, Wang H, Xiao Y, Sahin O, Brady SW, et al. Concomitant targeting of tumor cells and induction of T-cell response synergizes to effectively inhibit trastuzumab-resistant breast cancer. Cancer Res. 2012; 72: 4417–4428. doi: 10.1158/0008-5472.CAN-12-1339-T PMID: 22773664

6

Collaborations outside of RT2 Lab

6.1 No evidence for TSLP pathway activity in human breast cancer

In this collaboration with the team of Vassili Soumelis, I conducted the analysis of gene expression data in cell lines from the Cancer Cell Line Encyclopedia (CCLE) and human tumors from the Cancer Genome Atlas, TCGA database. I also contributed to the interpretation of the results.

Thymic stromal lymphopoietin (TSLP) is a cytokine derived from epithelial cells involved in initiating the differentiation of T-helper type 2 (Th2) cells. Th2 cells are essential in the activation and growth of cytotoxic T cells by releasing T cell cytokine. The presence of Th2 type responses in some tumors lead previous studies to investigate the role of TSLP and reported that it is a key cytokine produced by tumor epithelial cells in breast cancer (131, 135). The team of Vassili Soumelis has been studying for several years the expression of TLP in various tumor types by by immunohistochemistry and especially in human breast tumors. Their results do not agree with previous studies because they do not detect TSLP staining in breast adenocarcinomas. They decided to systematically assess the presence of TSLP at mRNA and protein levels, in several human breast cancer cell lines, large-scale public transcriptomics data sets and human primary breast tumors.

Expression of TSLP in breast cancer cell lines (BCCLs) was first conduct by quan-

titative PCR in 11 BCCLs of different molecular subtypes. Lung fibroblast sarcoma cell line MRC5 and peripheral blood mononuclear cell (PBMC) were used respectively as positive and negative control. All breast cancer cell lines were negative for TSLP expression. We extended the analysis by looking at TSLP expression in a set of 58 BCCLs from the CCLE (16). Expression of IL-8 and SDF1 were used as controls. In this dataset, TSLP was absent or expressed at a very low level in the BCCLs. These results suggest that BCCLs do not express TSLP mRNA.
We used the 591 samples of primary breast cancer and normal breast tissue from TCGA (183) to investigate the level of expression of TSLP in tumors. TSLP was significantly more expressed only in normal breast tissue than in any breast cancer molecular subtypes. We can not conclude that TSLP expression is a specific feature of breast tumor since it was observed at higher levels in normal breast tissue.
Breast tumors from Institut Curie Pathology Department were available to perform quantitative PCR analysis for TSLP mRNA on 19 independent tumors coupled to their juxta-tumor non-involved counterpart. TSLP expression was negative or very low (< 20 %) in all tumor samples, and was systematically lower in the tumor as compared to the juxta-tumor non-involved counterpart. These results are concordant with those derived from normal TCGA samples.
Finally, we investigated the amount of TSLP proteins in breast cancer tissue. 19 primary tumor samples were analyzed by ELISA. Only three tumor samples released TSLP at very low levels. In addition, only two of the 16 breast tumors appeared to be slightly positive for TSLP when evaluated by immunohistochemistry.

Our results suggest that TSLP is not expressed by breast tumors, unlike previous studies reporting TSLP expression in breast cancer (131, 135). The presence of TSLP in normal breast suggests that TSLP may contribute to physiological mechanisms and is not a feature developed by breast tumors. Our negative results should encourage further work to assess TSLP pathway activity in different types of cancer, as well as the relative contribution of Th2-promoting factors in individual tumor samples.

ONCOIMMUNOLOGY
2016, VOL. 5, NO. 8, e1178438 (9 pages)
http://dx.doi.org/10.1080/2162402X.2016.1178438

Taylor & Francis
Taylor & Francis Group

ORIGINAL RESEARCH

No evidence for TSLP pathway activity in human breast cancer

Cristina Ghirelli[a,b,c], Benjamin Sadacca[a,c,d,e], Fabien Reyal[a,c,d,f], Raphaël Zollinger[a,b,c], Paula Michea[a,b,c], Philémon Sirven[a,b,c], Lucia Pattarini[a,b,c], Carolina Martínez-Cingolani[a,b,c], Maude Guillot-Delost[a,b,c], André Nicolas[g], Alix Scholer-Dahirel[a,b,c,*], and Vassili Soumelis[a,b,c,*]

[a]U932 Immunity and Cancer, INSERM, Institut Curie, Paris, France; [b]Department of Immunology, Institut Curie, Paris, France; [c]Inserm Center of Clinical Investigations, CIC IGR Curie, Paris, France; [d]Residual Tumor & Response to Treatment Laboratory, RT2Lab, Translational Research Department, Institut Curie, Paris, France; [e]Statistics and Genome team of the Laboratoire de Mathématiques et Modélisation d'Évry, University of Évry val d'Essonne/UMR CNRS 8071/USC INRA, Evry, France; [f]Department of Surgery, Institut Curie, Paris, France; [g]Platform of Investigative Pathology of Biopathology Department, Curie Institute, Paris, France

ABSTRACT

Thymic stromal lymphopoietin (TSLP) is an epithelial cell-derived cytokine that primes dendritic cells for Th2 induction. It has been implicated in different types of allergic diseases. Recent work suggested that TSLP could play an important role in the tumor microenvironment and influence tumor progression, in particular in breast cancer. In this study we systematically assessed the production of TSLP at the mRNA and protein levels in several human breast cancer cell lines, large-scale public transcriptomics data sets, and primary human breast tumors. We found that TSLP production was marginal, and concerned less than 10% of the tumors, with very low mRNA and protein levels. In most cases TSLP was undetectable and found to be expressed at lower levels in breast cancer as compared to normal breast tissue. Last, we could not detect any functional TSLP receptor (TSLPR) expression neither on hematopoietic cells nor on stromal cells within the primary tumor microenvironment. We conclude that TSLP-TSLPR pathway activity is not significantly detected within human breast cancer. Taken together, these observations do not support TSLP targeting in breast cancer.

ARTICLE HISTORY
Received 22 January 2016
Revised 8 April 2016
Accepted 8 April 2016

KEYWORDS
Breast cancer; cytokines; dendritic cells; thymic stromal lymphopoietin; tumor microenvironment

Introduction

Within the tumor microenvironment, a diversity of immune-modulating factors can shape antitumor immunity, either by inducing and strengthening it, or by shifting a protective cytotoxic response toward an inappropriate regulatory response.[1] In particular, cytokines mediate complex cross talks between tumor cells and immune cells. Immune cell-derived cytokines may affect tumor cell differentiation, invasion and metastasis, hence participating in the oncogenic process. For example inflammatory cell-derived TNF can promote tumor epithelial cell survival through the induction of genes encoding NF-κF—dependent antiapoptotic molecules.[2] Conversely tumor cell-derived cytokines are critical to shape the state and effector functions of tumor-infiltrating immune cells. IL6 produced by renal cell carcinoma inhibits dendritic cell function and differentiation.[3] Tumor-derived TNF may contribute to different functions related to the type of tumor and the timing.[4] We have shown that tumor-derived GM-CSF promotes plasmacytoid pre-dendritic cell (pDC) survival and activation, with subsequent priming of a regulatory Th2 response.[5]

Thymic stromal lymphopoietin (TSLP) is an epithelial cell-derived cytokine that promotes Th2 polarization through dendritic cell activation.[6,7] TSLP is central to the physiopathology of allergic inflammation, such as atopic dermatitis and asthma.[6,8,9] Recently, it was shown to contribute to autoimmune inflammation, in particular in psoriasis.[10] Because of its role in linking epithelial cells to immune cells, TSLP has been explored in the past few years as potentially contributing to the tumor-immune cell crosstalk.[11-14] TSLP investigation in cancer was also motivated by the presence of Th2 type responses in some tumors, raising the hypothesis that Th2 promoting factors may be produced by tumor cells. Based on this rationale, TSLP was suggested to play a role in human breast cancer as a key cytokine produced by tumor epithelial cells and promoting T helper cells to produce IL-13, a prototypical Th2 cytokine.[12,13] TSLP was also implicated in pancreatic cancer where it was shown to be secreted by tumor-infiltrating fibroblasts.[11] In these two clinical settings and in a mouse model of breast and pancreatic cancer, TSLP was associated with tumor progression and metastasis.[13,15] In contrast, a recent study reported systemic TSLP impairs breast cancer and pancreatic development in mice through direct stimulation of Th2 cells.[16] In these studies, association between TSLP and prognosis in clinical cohorts of breast cancer patients was not assessed. In addition, results obtained in various skin cancer models suggested that TSLP may be of good prognosis and favor tumor regression.[17,18] Hence, the role of TSLP in cancer and the underlying mechanisms of action remain controversial.

In an effort to map TSLP expression in various tumor types, we had initiated several years ago a screening by immunohistochemistry in human primary tumors. Breast cancer was used as a main model for adenocarcinomas and head and neck cancer as a model for epidermoid tumor. While we observed high TSLP expression in head and neck tumors (see Guillot-Delost

CONTACT Vassili Soumelis vassili.soumelis@curie.net U932 Immunity and Cancer, INSERM, Institut Curie, Paris, France.
*These authors contributed equally to this work.

et al., companion paper), we did not detect TSLP staining in breast adenocarcinomas. Because of the conflicting data published in breast cancer, we sought to analyze in depth, in a systematic manner, the TSLP-TSLP receptor axis in human breast cancer at different levels and using a variety of methods. Although low TSLP was found in few tumor samples, most results indicated a lack of TSLP expression, both at the mRNA and protein levels, *ex vivo* and *in situ*. We also observed a lack of TSLP-receptor expression in the breast cancer microenvironment. We therefore, conclude that there is no evidence for TSLP pathway activity in human breast cancer.

Results

Breast cancer cell lines do not express TSLP mRNA

First, we screened TSLP mRNA expression in breast cancer cell lines using quantitative PCR (Fig. 1A). The lung fibroblast sarcoma cell line MRC5, which was initially used to clone human TSLP,[19] was used as a positive control for TSLP mRNA, with PBMC being our negative control (Fig. 1A). In comparison, we analyzed 11 breast cancer cell lines of different molecular subtypes. All breast cancer cell lines were negative for TSLP expression (Fig. 1A). In order to further increase the diversity of cell lines in an unbiased manner, we mined the Cancer Cell Line Encyclopedia, CCLE public database, which includes RNAseq expression data of about 1036 human cell lines from different anatomical sites.[20] We analyzed TSLP expression, and also assessed IL-8 and SDF1 expression as controls (Fig. 1 B–D). TSLP was absent or expressed at very low level in breast cancer cell lines (Fig. 1B, right panel). Higher levels were observed in triple negative (TN) cell lines, as compared to luminal A (ER^+Her2^-) and $Her2^+$ cell lines, although expression remained close to detection limit for most cell lines (Fig. 1B, right panel). In addition, these gene expression levels were marginal as compared to cell lines expressing significant levels (>2) of TSLP (Fig. 1B, left panel). IL-8 expression was detected at low levels in the different cell lines, with again a slightly higher expression in TN breast cancer cell lines (Fig. 1C). In comparison, SDF1 was most significantly expressed in luminal A and TN breast cancer cell lines (Fig. 1D). In summary, although our own assessments of TSLP expression by quantitative PCR was negative on all 11 breast cancer cell lines tested, data mining of transcriptomic profiles in 58 breast cancer cell lines raised the possibility of a very low TSLP expression in TN breast cancer subtype.

Transcriptomic analysis reveals that TSLP mRNA level is higher in normal breast than in primary breast cancer tissue

In order to get closer to primary tumors, we went on data mining The Cancer Genome Atlas, TCGA database, comprising transcriptional profiles of about 591 primary breast tumors.[21] Importantly, this large-scale data set also included normal breast tissue, which is key to interpret any results related to a neoplastic tissue. Surprisingly, we found significant TSLP expression only in normal breast tissue, which was statistically higher than in any of the three breast cancer molecular subtypes (Fig. 2A). In this primary tumor expression dataset, higher TSLP in TN tumors was not observed, since all three tumor types had expression levels close to zero (Fig. 2A). By comparison IL-8 expression was significantly higher in luminal A and TN tumors, as compared to $Her2^+$ tumors and normal breast tissue (Fig. 2B). SDF1 was higher in normal breast tissue, parallel to TSLP (Fig. 1A and C). These results do not support significant TSLP expression in primary breast cancer, and excluded TSLP expression as a specific feature of the breast tumor inflammation since it was observed at higher levels in normal breast tissue.

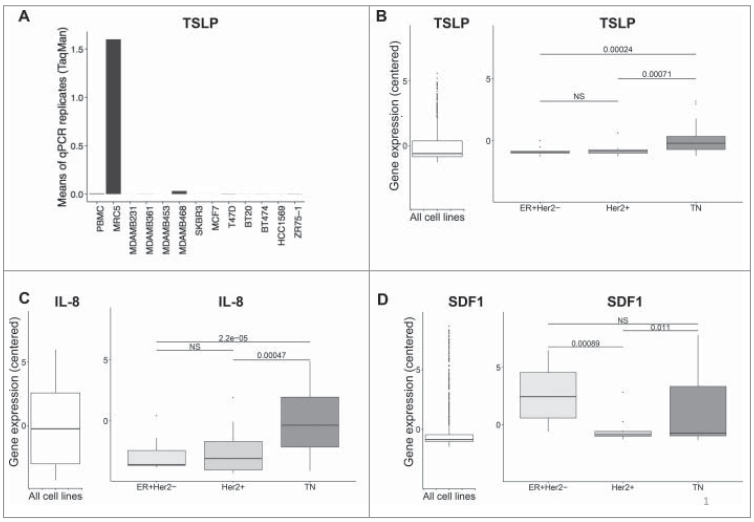


Figure 1. Breast cancer cell lines do not express TSLP mRNA. (A) TSLP mRNA expression quantified by quantitative PCR (TaqMan) in 11 breast cancer cell lines. PBMC and MRC5 correspond to negative and positive control respectively. N = 4. (B, C, D) Boxplots in the left panels represent mRNA expression in 1036 cancer cell lines from CCLE of TSLP, IL-8 and SDF1 respectively. (B, C, D) Boxplots in the right panels show the gene expression of TSLP, IL-8 and SDF1 respectively for breast cancer cell lines, which were grouped according to their corresponding molecular subtype. *p* values were calculated with a t test comparing different cancer cell subtypes.

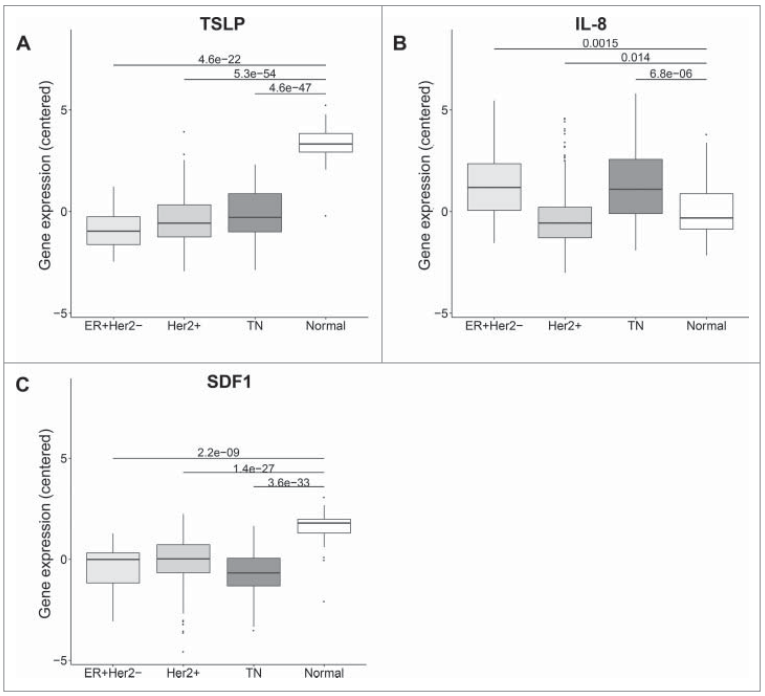


Figure 2. TSLP mRNA level is higher in normal breast than breast cancer tissue. (A, B, C) mRNA expression in 58 breast cancer patients from TCGA of TSLP, IL-8 and SDF1 respectively. Boxplots represent data of tumors classified in three different subtypes, namely Luminal ($ER^{+}Her2^{-}$), $Her2^{+}$ and Triple negative (TN) and normal breast tissue as indicated. *p* values were calculated with a t test comparing different clinical groups.

qPCR analysis reveals that TSLP mRNA expression is higher in juxta-tumor tissue than primary breast cancer tissue

In order to get a more reliable and controlled assessment of TSLP expression in primary breast cancer, we prospectively collected and analyzed primary breast tumors obtained from the Institut Curie Pathology Department. We performed quantitative PCR analysis for TSLP mRNA on 19 independent tumors coupled to their juxta-tumor non-involved counterpart (Fig. 3A). As for the cell lines, PBMC were used as negative control and MRC5 as a positive control for TSLP (Fig. 3A). TSLP expression was negative or very low (<20 %) in all tumor samples, and was systematically lower in the tumor as compared to the juxta-tumor non-involved counterpart (Fig. 3A). This result was in accordance with the higher levels observed in normal breast tissue from the TCGA database (Fig. 2A). Considering all tumor samples, higher TSLP levels were observed in the non-involved tumor counterpart in a statistically significant manner (Fig. 3B).

Breast cancer tissue do not express detectable amount of TSLP protein

In order to get a first assessment of TSLP protein expression levels, we cultured for 24 h each of the 19 primary tumor samples, as well as their juxta-tumor counterpart, and analyzed the tissue-conditioned supernatants for TSLP secretion by ELISA (Fig. 3C). Three tumor samples out of 19 (15%) released TSLP, although at very low levels, slightly above detection limit, which was set by the ELISA manufacturer at 32.5 pg/mL (Fig. 3C). Surprisingly, juxta-tumor samples did not release any detectable TSLP protein while they expressed TSLP at the mRNA level. This could be due to TSLP retention within the cytoplasm in this tissue type (Fig. 3C), as was noted in skin TSLP studies.[22]

As tissue culture performed in serum-containing medium could artificially alter TSLP expression, we decided to next assess TSLP expression *in situ*. TSLP expression was assessed by immunohistochemistry on 16 primary breast tumors that were frozen within 15 min following resection. We used a previously validated monoclonal antibody (by us and others).[10,23] On human tonsil sections, we could verify that the majority (>80 %) of epithelial cells were TSLP positive, as previously published[6] (Fig. 4A and C). On the contrary only two out of 16 breast tumors (12.5%) showed a slight positivity for TSLP, with less than 10% of the tumor cells harboring a specific staining pattern (Fig. 4B and C). For each tumor, pathological examination and hematoxylin-eosin-safran (HES) staining in consecutive tissue sections confirmed the presence of epithelial tumor cells (Fig. 4B). Thus, using two complementary approaches, we showed that only a minority of tumors was positive for TSLP protein and the TSLP levels we could detect remained marginal.

TSLP receptor is not expressed in the breast cancer microenvironment

Considering the possibility that low levels of TSLP protein may be secreted in some tumors, downstream function would

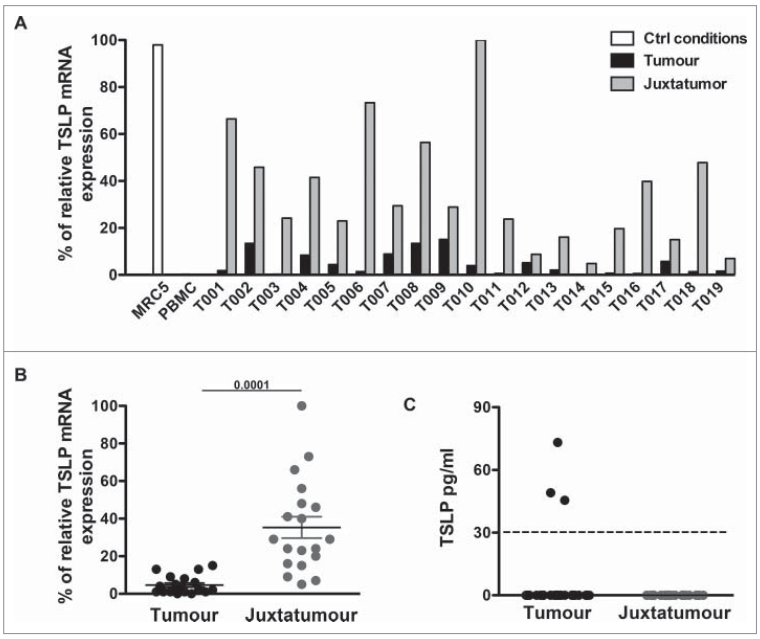


Figure 3. TSLP mRNA expression is higher in juxta-tumor tissue than breast cancer tissue. (A) TSLP transcripts were measured by quantitative PCR (TaqMan) in 19 breast cancer tissues (black bars) and 19 corresponding juxta-tumor tissues (gray bars). White bars represent TSLP levels detected in MRC5 and PBMC used as positive and negative control respectively. (B) Quantification of TSLP mRNA transcripts shown as percentage of housekeeping gene expression. Four housekeeping genes were used for these experiments: Actin Beta (ACTB), Hypoxanthine Phosphoribosyltransferase 1 (HPRT1), Ribosomal Protein L31 (RPL31) and Beta-2-Microglobulin (B2M). Lines represent mean +/− the Standard Error of the Mean (SEM). Wilcoxon matched pairs test was used to calculate p value. N = 19. (C) Quantification of soluble TSLP measured by ELISA in the supernatants generated from primary breast tumor tissues and corresponding juxta-tumor samples as described in the Material and Methods section. ELISA sensitivity detection limit, which is represented by the dashed line, was 31 pg/mL as recommended by the manufacturer instructions. N = 40.

depend on the expression of the TSLP receptor within the tumor microenvironment. TSLP signals through its receptor only when the TSLP receptor specific chain (TSLP-R) dimerizes with the IL-7 receptor α chain (IL-7-Rα). We thus analyzed by flow cytometry the co-expression of TSLP-R and IL-7-Rα on different cellular compartments. A single cell suspension was obtained from freshly resected primary breast tumors after tissue digestion (Fig. 5). Virtually no IL-7-Rα and TSLP receptor co-expressing cells were detected within the $CD45^+$ compartment, similar to the $CD45^+Lineage^-$ compartment (Fig. 5A). To exclude the possibility that a very low percentage of rare dendritic cells could express the receptor, we also gated on $CD45^+Lineage^-CD11c^+HLA\text{-}DR^+$ dendritic cells (Fig. 5B quadrant 6) and quantified the expression of the TSLP receptor heterodimer in six independent tumors (Fig. 5C). TSLP-R and IL-7-Rα double positive dendritic cells ranged from 0 to 0.03%, which we can consider not significant, and close to background (Fig. 5B). Similar levels were found on dendritic cells from juxta-tumor non-involved samples (Fig. 5C). Comparing digested and non-digested DC showed that the tumor sample digestion protocol did not affect the detection of the two TSLP-R chains (Fig. 5D). We conclude that even if few tumors may show low TSLP positivity, the absence of TSLP receptor-expressing cells in the breast tumor microenvironment excludes the possibility of downstream TSLP functionality.

Discussion

In the present study, we systematically assessed the presence of TSLP at the mRNA and protein levels, in several human breast cancer cell lines, large-scale public transcriptomics data sets and human primary breast tumors. We found that TSLP production was marginal, and concerned less than 10% of the tumors, with very low mRNA and protein levels. In most cases, TSLP was undetectable and found to be expressed at lower levels in breast cancer as compared to normal breast tissue. Last, we could not detect any functional TSLPR expression neither on haematopoietic cells nor on stromal cells within the primary tumor microenvironment. We conclude that TSLP-TSLPR pathway activity is not significantly detected within human breast cancer. Those results are in contrast with previous studies reporting TSLP expression in breast cancer.[12,13]

Expression of immune modulating cytokines in the tumor microenvironment is most of the time interpreted as being associated to the tumoral process, and being part of pro- or antitumor immune mechanisms. This view only stands if the normal tissue counterpart is devoid of expression of that specific cytokine, or harbors a much lower expression, implying that the cytokine expression is a specific feature of the tumor. In our results, we analyzed normal breast TSLP mRNA expression from public databases, and found that TSLP levels were higher than in breast tumors. This is in accordance with recent

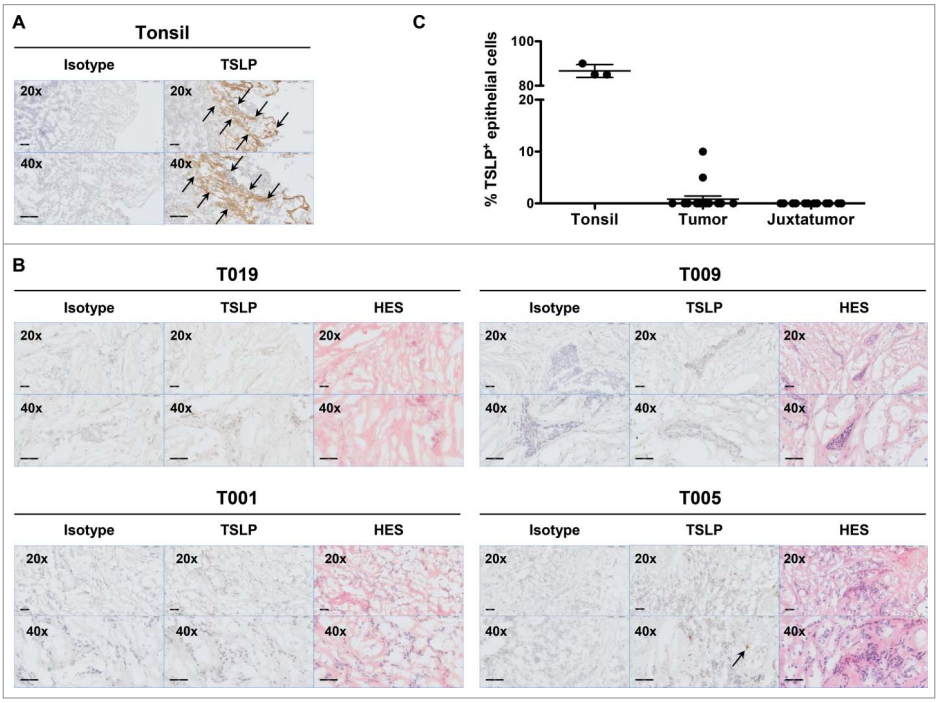


Figure 4. Breast cancer tissues do not express TSLP. (A) Tonsil sections were used as positive control tissue to validate TSLP staining by immunohistochemistry. TSLP and matched isotype staining was performed in two consecutive tonsil sections. Arrows indicate positive staining. (B) Isotype, TSLP and H&S staining in three consecutive slides of four representative breast cancer specimens. Arrows indicate positive staining. All tissue sections are shown at 20x and 40x magnification. (C) Percentages of TSLP staining in epithelial cells. Each symbol represents a different sample.

reports detecting TSLP protein in breast milk[24] with a potential role in the intestinal immunity of the neonates.[25] The presence of TSLP in normal breast suggests that TSLP may contribute to physiological mechanisms in this context, and is not a feature developed by breast tumors or related to breast cancer inflammation. This is an important aspect to take into consideration in the interpretation of TSLP role in cancer.

TSLP assessment *in situ* in inflamed tissue has provided a very efficient and unbiased manner to detect TSLP in various diseases.[8,9] Using immunohistology, we and others have shown strong TSLP staining in the keratinocytes of atopic dermatitis and psoriasis[6,10,26] as well as in the tonsillar epithelial cells.[6] Conversely, normal skin has repeatedly been found as negative for any TSLP staining.[6,10,27] Such TSLP positive and negative tissue types constitute valuable controls for any study of TSLP expression in disease. In the present study, we have systematically compared TSLP analysis in breast cancer to human tonsils, and the TSLP levels we could detect remained marginal in terms of percentage of TSLP positive epithelial cells, as well as in the intensity of TSLP staining. In a previous study about TSLP role in breast cancer, TSLP expression was analyzed by immunofluorescence and was found positive on most tumor epithelial cells,[13] in discrepancy with our own results. However, that study lacked a negative control, and used normal skin as a positive control, when other studies could not detect any TSLP expression in this context.[6,10,27] This raises questions on the interpretation of the results. In contrast, in our companion paper, we report that TSLP is highly expressed in head and neck squamous cell carcinoma, at levels similar to atopic dermatitis. This indicates that although TSLP is absent from the breast cancer microenvironment, it can be expressed in other cancer types.

Assessing the protein expression of cytokines in human primary tumors has many intrinsic difficulties, and multiple complementary strategies should be considered. Analysis of tumor-derived supernatants has been used extensively to analyze the soluble tumor microenvironment. It has the potential drawback that tissue culture manipulation may induce the secretion of factors that are not being spontaneously secreted. Culture conditions may also influence the amounts of cytokine production. In this study, we have used basic serum-containing medium without any activator in order to avoid as much as possible artificial induction of cytokine synthesis and secretion. Other studies have used PMA/ionomycin in order to stimulate immune cells, which may generate direct and indirect effects promoting cytokine production within the tumor microenvironment.[13] Although such immune activators may be helpful to analyzed T cell-derived cytokines,[28] they may also have effects on non-immune cells, either directly or through paracrine activating loops. For example PMA/ionomycin may induce TNF production by T cells,[29] which can subsequently activate TSLP production by neighboring epithelial cells.[22] Hence, tumor-conditioned media obtained in the presence of any type of activating signal should be interpreted with caution.

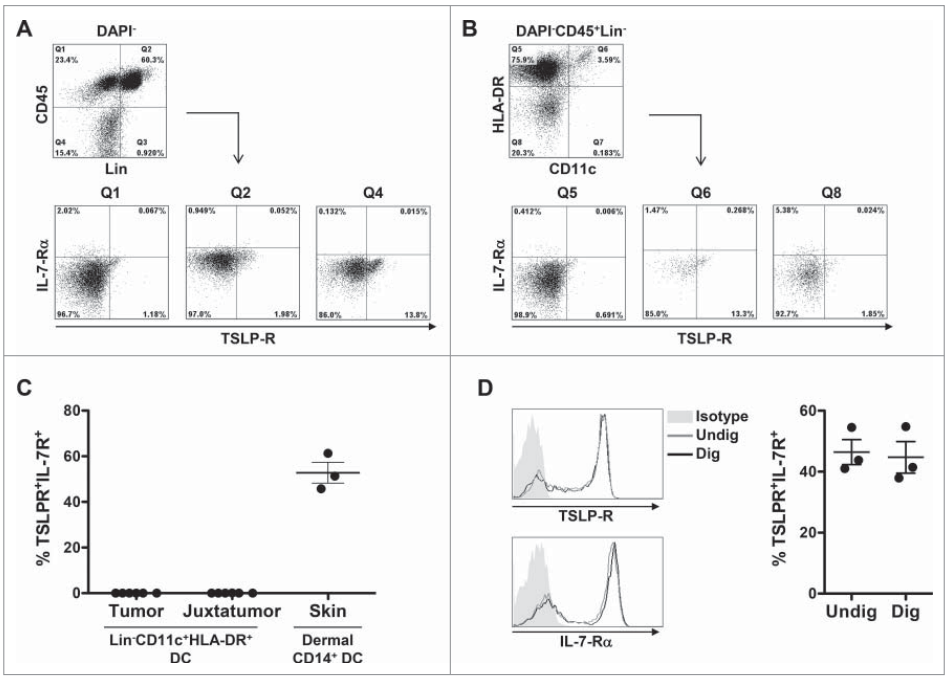


Figure 5. TSLP receptor is not detected in primary breast cancer by flow cytometry. (A–B) Flow cytometry dot plots from one representative human primary breast tumor. (A) Gating strategy used to investigate the expression of TSLP-R and IL-7-Rα in viable $CD45^+Lin^-$ (Q1), $CD45^+Lin^+$ (Q2) and $CD45^-Lin^-$ (Q4). (B) Gating strategy designed to detect DC in primary breast cancer. DC were defined as $DAPI^-CD45^+Lin^-HLA\text{-}DR^{high}CD11c^{high}$. The expression of TSLP-R and IL-7-Rα was assessed in DC (Q6) as well as $HLA\text{-}DR^{high}CD11c^-$ (Q5) and $HLA\text{-}DR^-CD11c^-$ (Q8). (C) Percentages of $TSLP\text{-}R^+IL\text{-}7\text{-}R\alpha^+$ DC in tumors, corresponding juxta-tumor tissues and in the dermis of normal skin. Lines represent mean +/− the Standard Error of the Mean (SEM). N = 6 (D) TSLP-R and IL-7-R expression in DC enriched from PBMC without digestion step (gray histogram, Undig) and after mechanical and enzymatic digestion (dark gray histogram, Dig), left panels. One representative background histogram staining is shown in light gray. Quantification of $TSLP\text{-}R^+IL\text{-}7\text{-}R\alpha^+$DC in undigested and digested enriched DC is shown in the left panel. N = 3.

Signaling pathway activity requires expression of all its components: ligand, receptor and downstream signaling molecules. In the case of TSLP, a functional pathway would require the presence not only of TSLP but also of TSLP receptor heterodimer (TSLP-R/IL7Rα). However, in previous studies suggesting a TSLP role in human cancer, TSLP receptor heterodimer expression in the tumor microenvironment was not assessed[11,13] raising the possibility that this pathway is not functional despite the presence of TSLP itself. Along this line, we show in our companion paper that although head and neck tumors express high levels of TSLP, TSLP-R-expressing cells are absent from the tumor microenvironment. These observations highlight that such a dissociated ligand-receptor expression can result in an inactive TSLP pathway. Here, we report for the first time that TSLP-R is not present in the microenvironment of human breast primary tumors. This suggests that even if low levels of TSLP are present (below the detection threshold, or at low levels in rare cases), TSLP pathway is unlikely to be active as its signaling is disrupted by the absence of TSLP receptor.

Th2 cytokines were reported to contribute to tumorigenesis in several mouse models.[30-32] Human breast cancer has also been associated to a Th2 response, which was shown to be promoted by TSLP.[13] However, in that study, a large proportion of tumors (61%) were negative for TSLP expression. In our study, TSLP could not be detected in the vast majority of breast tumors and its receptor was also absent. Those observations raise the question of parallel or alternative pathways to induce Th2 response. We have recently shown that GM-CSF produced by tumor epithelial cells activates pDC to promote a regulatory Th2 response in human primary breast tumors.[5] Concomitant increase of GM-CSF and pDC was found in 11.8% of breast tumors studied (14 out of 118), and associated to more aggressive breast cancer subtypes. In addition, CCL5 was also reported to promote Th2 polarization in breast cancer. Indeed CCL5 depletion in MMTV-PyMT mouse model leads to a deficit in Th2 cell associated with reduced tumor burden and metastasis.[33] Furthermore, CCL5 and IL-4 expression were correlated and associated with aggressiveness of human luminal breast cancer.[33] Further studies will be needed to evaluate the relative role of these pathways within the breast cancer microenvironment and their corresponding roles in supporting Th2 cells.

Our work has potential implications for therapy and drug development. Indeed, it does not support TSLP as a relevant target in breast cancer as we do not detect any evidence of TSLP-TSLPR pathway activity in this clinical setting. Although it is almost impossible to completely rule out a biological pathway implication in disease, our negative results should encourage further work to assess TSLP pathway activity in different types of cancer, as well as the relative contribution of Th2-promoting factors in individual tumor samples.

Table 1. Clinical information of patients included in the study.

	N	%
Demography		
Female	44	100
Age		
<40	3	6.8
41–55	17	38.6
>56	24	54.5
Extension		
Size		
< 20	18	40.1
21–40	24	54.5
> 41	2	4.5
Lymph nodes involvement		
LN^+	14	31.8
LN^-	28	63.6
Unknown	2	4.5
Histological subtype		
Invasive ductal	28	63.6
Invasive lobular	12	27.3
Mixed ductal/lobular	4	9.1
Elston Ellis (Ee) grade		
I	6	13.6
II	18	40.9
III	20	45.5
Molecular subgroup		
Triple negative (TN)	8	18.2
$HER2^+$	0	0
Luminal B (LB)	3	6.8
Luminal A (LA)	33	75

Material and methods

Human samples and patients' characteristics

Tumor and juxta-tumor (adjacent to the tumor and exempt of malignant tumor cells) tissues were collected during standard surgical procedures as surgical residues from untreated breast cancer patients, from the department of Pathology (Institut Curie, Paris). Patients signed an informed consent after approval of the study. This study was approved by the Internal Review Board and Clinical Research Committee of the Institut Curie. Patient characteristics are summarized in Table 1.

Tonsils sections were obtained after surgical resection from children undergoing tonsillar resection (Necker Hospital, Paris) after informed consent of the parents. Tissues were transported in CO2-independant medium (Gibco) and processed within the next 3 h after resection.

Healthy donor human blood buffy coats were obtained from "Etablissement Français du Sang," Paris, Saint-Antoine Crozatier blood bank through an approved convention with the Institut Curie.

Normal skin samples considered as surgical wastes were obtained from healthy donors undertaking esthetic or reconstructive surgery and processed within 6 h of resection. This discarded human surgical material was obtained anonymously according to the institutional regulations, in compliance with French legislation.

Cell line culture

All cell lines were cultured without stimulation at the density of 0.5×10^6 cells/mL in complete RPMI GlutaMAX (Gibco) containing 10% FBS (HyClone) for 48 h. Cells were then washed with PBS, detached with trypsin (Gibco), pelleted and lysed in RLT buffer (Qiagen) to allow RNA extraction. All cell lines were mycoplasma-free.

Large-scale public database mining

The gene level expression of TSLP, IL8 and SDF1 have been analyzed using the Breast Cancer Cell Lines Encyclopedia. Data were downloaded from the CCLE website (http://www.broadinstitute.org/ccle)[20] and normalized with RMA. This data set was composed of 1036 cell lines from 24 tissues. Patient data were retrieve from a sample of 591 breast cancer patients from TCGA (Level 3).[34] Breast cancer subtypes were defined using a bimodal mixture of 2 gaussian distributions for ER, PR and HER2 gene expression. TN breast cancer samples were defined by the absence of estrogen and progesterone receptor expression and a lack of HER2 overexpression/amplification. In CCLE dataset, breast cancer cell lines were composed of 31 TN, 12 ER^+Her2^- and 15 $Her2^+$. Breast tumors from TCGA data set were composed of 26 ER^+Her2^-, 410 $Her2^+$, 94 TN and 61 Normal.

Primary tumor processing for RNA extraction

Tumor and juxta-tumor tissues were cryopreserved in Tissue-Tek (Sakura Finetek USA, Inc., Torrance, Calif) at −80°C. Tissues were cryosectioned with a Cryostat. Ten sections of 20 μm thickness were collected for every tissue sample. Tissues were lysed in RLT buffer (Qiagen) supplemented with β-mercaptoethanol (Sigma) immediately after cutting. RNA was extracted using RNAeasy mini kit (Qiagen) following manufacturer instructions and processed as described above.

Quantitative PCR

RNA was extracted from cell line and tumor section lysates using RNeasy micro and RNeasy mini kit (Qiagen) respectively following manufacturer instructions including on-column DNase digestion. RNA concentration and absence of protein contamination were determined using the NanoDrop instrument. All RNA samples had 260 nm/280 nm absorbance ratios between 1.9 and 2.1, indicating high purity. RNA quality was assessed using RNA 6000 Nano chips on the Agilent 2100 Bioanalyzer. Only samples with RIN>7 were further processed for reverse transcription. cDNA was synthesized with a mix containing random hexamers (Promega), oligo(dT)15 (Promega) and SuperScript II reverse transcriptase (Invitrogen). TSLP mRNA transcripts, as well as ACTB (Actin β), B2M (Beta-2 Microglobulin), HPRT (hypoxanthine phosphoribosyltransferase 1) and RPL34 (ribosomal protein L34), which were used as housekeeping genes were quantified by real-time quantitative reverse transcription PCR on Light Cycler 480 (Roche) with Applied Biosystems predesigned TaqMan Gene Expression Assays and Absolute qPCR ROXmix (Thermo Fisher Scientific).

Primary breast tumor and juxta-tumor digestion for flow cytometry analysis and generation of tissue-conditioned supernatants

Breast tumor and juxta-tumor tissues were cut in three pieces. One piece was frozen in Tissue-Tek (Sakura Finetek USA, Inc.,

Torrance, Calif) for further histological analysis. The second piece was carefully minced into smaller pieces in CO_2 independent medium (Gibco) containing 5% FBS (HyClone). Those pieces were digested with collagenase (1 mg/mL; Roche) and DNAse (25 μg/mL; Roche) in a total volume of 3 mL CO_2 independent medium (Gibco) for 1 h at 37°C under agitation at 180 rpm. Cell suspension was then filtered through a 40 μm nylon cell strainer (Falcon BD) and washed twice in cold PBS containing 5% of human serum (Biowest) and EDTA 2 mM (Gibco). Skin digestion was performed as described in ref.[10]. Cells were stained with the following mouse anti-human antibodies and corresponding matched isotype controls: CD3-FITC, CD16-FITC, CD45-APC-Cy7, CD11c-PE-Cy5 (BD Biosciences), CD14-FITC, CD20-FITC (Miltenyi Biotec), HLADR-Alexa700, TSLPR-APC (Biolegend), IL-7Rα-PE, (eBiosciences). Dead cells were excluded based on side and forward scatter characteristics and positivity for 4,6-diamidino-2-phenylindole (DAPI) (Invitrogen). Sample acquisition was performed on a LSRII flow cytometer (Becton Dickinson) and data analysis was performed using FlowJo software version 9.4.7. The third piece of tissue was cut in smaller pieces of 40 mg. Each piece of tissue was put in one well of a 48 well plate in 250 μl of complete RPMI GlutaMAX (Gibco) containing 10% FBS (HyClone) without any stimulation. Supernatants were harvested after 24 h of culture and tissues were discarded. Supernatants were spun for 5 min at maximum speed to remove dead cells and debris and stored at −80°C for further ELISA measurements.

Soluble TSLP quantification

Soluble TSLP was quantified using the DuoSet Kit from R&D following manufacturer instructions. Detection limit was set at 31.25 pg/mL as recommended.

Immunohistochemistry

Tissues were embedded in Tissu-Tek (Sakura Finetek USA, Inc., Torrance, Calif) and cryopreserved at −80°C. Acetone-fixed cryosections of 4 μm thickness were stained with monoclonal rat anti-human TSLP 5 μg/mL (kind gift from Pr. Yong-Jun Liu), and corresponding matched isotype control antibody (BD Pharmigen), followed by a biotinylated goat anti-rat secondary antibody (Vector Laboratories). The staining was revealed using a Vectastain ABC peroxidase system (Vector Laboratories) and it was detected using 3-3Œ-diamino-benzidine-tetrahydrochloride (DAB) revelation (Vector Laboratories). The sections were counterstained with haematoxylin and mounted with Perthex mounting media (Histolab). Tonsil sections were treated using the same procedure and they served as positive control for TSLP staining. The staining was performed using the Autostainer 480 (Labvision). Tissue images were taken on the Philips Digital Pathology Ultra-Fast Scanner.

DC enrichment from human blood

Peripheral blood mononuclear cells (PBMC) were isolated using Ficoll-gradient (GE Healthcare). DC were enriched using the EasySep ^{TM} human Pan-DC Pre-Enrichment kit (Stem Cell Technologies) following manufacturer instructions.

Statistical analysis

Unpaired *t* tests were used to determine statistical significance. Statistical significance was retained for *p* values lower than 0.05. Symbols used: NS, not significant.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Funding

This work was supported by funding from Institut National de la Santé et de la Recherche Médicale (BIO2012-02 and BIO2014-08), Fondation pour la Recherche Médicale, Association de la Recherche Contre le Cancer (PJA 20131200436), INCA (2011-1-PL BIO-12-IC-1 and 2012-1-GYN-04-IC-1), ANR-13-BSV1-0024-02, ANR-10-IDEX-0001-02 PSL* and ANR-11-LABX-0043, CIC IGR-Curie 1428.

References

1. Chen DS, Mellman I. Oncology meets immunology: the cancer-immunity cycle. Immunity 2013; 39:1-10; PMID:23890059; http://dx.doi.org/10.1016/j.immuni.2013.07.012
2. Luo JL, Maeda S, Hsu LC, Yagita H, Karin M. Inhibition of NF-kappaB in cancer cells converts inflammation- induced tumor growth mediated by TNFalpha to TRAIL-mediated tumor regression. Cancer Cell 2004; 6:297-305; PMID:15380520; http://dx.doi.org/10.1016/j.ccr.2004.08.012
3. Song EY, Shurin MR, Tourkova IL, Chatta G, Shurin GV. Human renal cell carcinoma inhibits dendritic cell maturation and functions. Urologe A 2004; 43 Suppl 3:S128-130; PMID:15150693; http://dx.doi.org/10.1007/s00120-004-0599-1
4. Balkwill F. Tumour necrosis factor and cancer. Nat Rev Cancer 2009; 9:361-71; PMID:19343034; http://dx.doi.org/10.1038/nrc2628
5. Ghirelli C, Reyal F, Jeanmougin M, Zollinger R, Sirven P, Michea P, Caux C, Bendriss-Vermare N, Donnadieu MH, Caly M et al. Breast cancer cell-derived GM-CSF licenses regulatory Th2 induction by plasmacytoid predendritic cells in aggressive disease subtypes. Cancer Res 2015; 75:2775-87; PMID:25977333; http://dx.doi.org/10.1158/0008-5472.CAN-14-2386
6. Soumelis V, Reche PA, Kanzler H, Yuan W, Edward G, Homey B, Gilliet M, Ho S, Antonenko S, Lauerma A et al. Human epithelial cells trigger dendritic cell mediated allergic inflammation by producing TSLP. Nat Immunol 2002; 3:673-80; PMID:12055625; http://dx.doi.org/10.1038/nrm910
7. Liu YJ, Soumelis V, Watanabe N, Ito T, Wang YH, Malefyt Rde W, Omori M, Zhou B, Ziegler SF. TSLP: an epithelial cell cytokine that regulates T cell differentiation by conditioning dendritic cell maturation. Annu Rev Immunol 2007; 25:193-219; PMID:17129180; http://dx.doi.org/10.1146/annurev.immunol.25.022106.141718
8. Shikotra A, Choy DF, Ohri CM, Doran E, Butler C, Hargadon B, Shelley M, Abbas AR, Austin CD, Jackman J et al. Increased expression of immunoreactive thymic stromal lymphopoietin in patients with severe asthma. J Allergy Clin Immunol 2012; 129:104-11 e101-109; PMID:21975173; http://dx.doi.org/10.1016/j.jaci.2011.08.031
9. Ying S, O'Connor B, Ratoff J, Meng Q, Mallett K, Cousins D, Robinson D, Zhang G, Zhao J, Lee TH et al. Thymic stromal lymphopoietin expression is increased in asthmatic airways and correlates with expression of Th2-attracting chemokines and disease severity. J Immunol 2005; 174:8183-90; PMID:15944327; http://dx.doi.org/10.4049/jimmunol.174.12.8183
10. Volpe E, Pattarini L, Martinez-Cingolani C, Meller S, Donnadieu MH, Bogiatzi SI, Fernandez MI, Touzot M, Bichet JC, Reyal F et al. Thymic stromal lymphopoietin links keratinocytes and dendritic cell-derived IL-

23 in patients with psoriasis. J Allergy Clin Immunol 2014; 134:373-81; PMID:24910175; http://dx.doi.org/10.1016/j.jaci.2014.04.022
11. De Monte L, Reni M, Tassi E, Clavenna D, Papa I, Recalde H, Braga M, Di Carlo V, Doglioni C, Pia Protti M. Intratumor T helper type 2 cell infiltrate correlates with cancer-associated fibroblast thymic stromal lymphopoietin production and reduced survival in pancreatic cancer. J Exp Med 2011; 208:469-78; PMID:21339327; http://dx.doi.org/10.1084/jem.20101876
12. Olkhanud PB, Rochman Y, Bodogai M, Malchinkhuu E, Wejksza K, Xu M, Gress RE, Hesdorffer C, Leonard WJ, Biragyn A. Thymic stromal lymphopoietin is a key mediator of breast cancer progression. J Immunol 2011; 186:5656-62; PMID:21490155; http://dx.doi.org/10.4049/jimmunol.1100463
13. Pedroza-Gonzalez A, Xu K, Wu TC, Aspord C, Tindle S, Marches F, Gallegos M, Burton EC, Savino D, Hori T et al. Thymic stromal lymphopoietin fosters human breast tumor growth by promoting type 2 inflammation. J Exp Med 2011; 208:479-90; PMID:21339324; http://dx.doi.org/10.1084/jem.20102131
14. Wu TC, Xu K, Banchereau R, Marches F, Yu CI, Martinek J, Anguiano E, Pedroza-Gonzalez A, Snipes GJ, O'Shaughnessy J et al. Reprogramming tumor-infiltrating dendritic cells for CD103+ CD8+ mucosal T-cell differentiation and breast cancer rejection. Cancer Immunol Res 2014; 2:487-500; PMID:24795361; http://dx.doi.org/10.1158/2326-6066.CIR-13-0217
15. Lo Kuan E, Ziegler SF. Thymic stromal lymphopoietin and cancer. J Immunol 2014; 193:4283-8; PMID:25326546; http://dx.doi.org/10.4049/jimmunol.1400864
16. Demehri S, Cunningham TJ, Manivasagam S, Ngo KH, Moradi Tuchayi S, Reddy R, Meyers MA, DeNardo DG, Yokoyama WM. Thymic stromal lymphopoietin blocks early stages of breast carcinogenesis. J Clin Invest 2016; 126:1458-70; PMID:26927668; http://dx.doi.org/10.1172/JCI83724
17. Di Piazza M, Nowell CS, Koch U, Durham AD, Radtke F. Loss of cutaneous TSLP-dependent immune responses skews the balance of inflammation from tumor protective to tumor promoting. Cancer Cell 2012; 22:479-93; PMID:23079658; http://dx.doi.org/10.1016/j.ccr.2012.08.016
18. Demehri S, Cunningham TJ, Manivasagam S, Ngo KH, Moradi Tuchayi S, Reddy R, Meyers MA, DeNardo DG, Yokoyama WM. Elevated epidermal thymic stromal lymphopoietin levels establish an antitumor environment in the skin. Cancer Cell 2012; 22:494-505; PMID:23079659; http://dx.doi.org/10.1016/j.ccr.2012.08.017
19. Reche PA, Soumelis V, Gorman DM, Clifford T, Liu Mr, Travis M, Zurawski SM, Johnston J, Liu YJ, Spits H et al. Human thymic stromal lymphopoietin preferentially stimulates myeloid cells. J Immunol 2001; 167:336-43; PMID:11418668; http://dx.doi.org/10.4049/jimmunol.167.1.336
20. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012; 483:603-7; PMID:22460905; http://dx.doi.org/10.1038/nature11003
21. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature 2012; 490:61-70; PMID:23000897; http://dx.doi.org/10.1038/nature11412
22. Bogiatzi SI, Fernandez I, Bichet JC, Marloie-Provost MA, Volpe E, Sastre X, Soumelis V. Cutting Edge: Proinflammatory and Th2 cytokines synergize to induce thymic stromal lymphopoietin production by human skin keratinocytes. J Immunol 2007; 178:3373-7; PMID:17339431; http://dx.doi.org/10.4049/jimmunol.178.6.3373
23. Fontenot D, He H, Hanabuchi S, Nehete PN, Zhang M, Chang M, Nehete B, Wang YH, Wang YH, Ma ZM et al. TSLP production by epithelial cells exposed to immunodeficiency virus triggers DC-mediated mucosal infection of CD4+ T cells. Proc Natl Acad Sci U S A 2009; 106:16776-81; PMID:19805372; http://dx.doi.org/10.1073/pnas.0907347106
24. Macfarlane TV, Seager AL, Moller M, Morgan G, Thornton CA. Thymic stromal lymphopoietin is present in human breast milk. Pediatr Allergy Immunol 2010; 21:e454-456; PMID:20444169; http://dx.doi.org/10.1111/j.1399-3038.2009.00916.x
25. Thornton CA, Morgan G. Innate and adaptive immune pathways to tolerance. Nestle Nutr Workshop Ser Pediatr Program 2009; 64:45-57; discussion 57-61, 251-257; PMID:19710514; http://dx.doi.org/10.1159/000235782
26. Luo Y, Zhou B, Zhao M, Tang J, Lu Q. Promoter demethylation contributes to TSLP overexpression in skin lesions of patients with atopic dermatitis. Clin Exp Dermatol 2014; 39:48-53; PMID:24341479; http://dx.doi.org/10.1111/ced.12206
27. Briot A, Deraison C, Lacroix M, Bonnart C, Robin A, Besson C, Dubus P, Hovnanian A. Kallikrein 5 induces atopic dermatitis-like lesions through PAR2-mediated thymic stromal lymphopoietin expression in Netherton syndrome. J Exp Med 2009; 206:1135-47; PMID:19414552; http://dx.doi.org/10.1084/jem.20082242
28. Truneh A, Albert F, Golstein P, Schmitt-Verhulst AM. Early steps of lymphocyte activation bypassed by synergy between calcium ionophores and phorbol ester. Nature 1985; 313:318-20; PMID:3918270; http://dx.doi.org/10.1038/313318a0
29. Kim TK, St John LS, Wieder ED, Khalili J, Ma Q, Komanduri KV. Human late memory CD8+ T cells have a distinct cytokine signature characterized by CC chemokine production without IL-2 production. J Immunol 2009; 183:6167-74; PMID:19841187; http://dx.doi.org/10.4049/jimmunol.0902068
30. Kobayashi M, Kobayashi H, Pollard RB, Suzuki F. A pathogenic role of Th2 cells and their cytokine products on the pulmonary metastasis of murine B16 melanoma. J Immunol 1998; 160:5869-73; PMID:9637498
31. Berzofsky JA, Terabe M. A novel immunoregulatory axis of NKT cell subsets regulating tumor immunity. Cancer Immunol Immunother 2008; 57:1679-83; PMID:18369622; http://dx.doi.org/10.1007/s00262-008-0495-4
32. DeNardo DG, Barreto JB, Andreu P, Vasquez L, Tawfik D, Kolhatkar N, Coussens LM. CD4(+) T cells regulate pulmonary metastasis of mammary carcinomas by enhancing protumor properties of macrophages. Cancer Cell 2009; 16:91-102; PMID:19647220; http://dx.doi.org/10.1016/j.ccr.2009.06.018
33. Zhang Q, Qin J, Zhong L, Gong L, Zhang B, Zhang Y, Gao WQ. CCL5-mediated Th2 immune polarization promotes metastasis in luminal breast cancer. Cancer Res 2015; 75:4312-21; PMID:26249173; http://dx.doi.org/10.1158/0008-5472.CAN-14-3590
34. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, Fulton LL, Dooling DJ, Ding L, Mardis ER, et al. Comprehensive molecular portraits of human breast tumours. Nature 2012; 490:61-70; PMID:23000897; http://dx.doi.org/10.1038/nature11412

6.2 Impaired PRC2 activity promotes transcriptional instability and favors breast tumorigenesis

In this study, I contributed to the analysis of METABRIC copy number data. We compared the prognosis of the patient according to EZH2 gene copy number status and tried to define the region surrounding EZH2 showing a correlation between genetic loss and outcome.

Epigenetic changes are by now a well-recognized cancer hallmark. One of the main epigenetic mechanisms is DNA methylation which modifies the ability of DNA to be transcribed. The gene EZH2 is in charge of the methylation activity of Polycomb Repressive Complex 2 (PRC2) which plays a key role in maintaining transcriptional programs during development. PRC2 was assumed to be related to cancer when its function is deregulated (29). In addition, EZH2 is involved in the methylation of Lys27 of histone H3 (H3K27me3), a mark linked to transcriptional silencing. Several studies have suggested that a high level of EZH2 in prostate and breast tumors is associated with a poor prognosis (95, 192). However, recent studies have shown that the levels of H3K27m3 are low in these cancers and although high expression of EZH2 is correlated with a poor prognosis, high levels of H3K27me3 correlate with a good prognosis (12, 80). This has led several groups to propose that EZH2 could play PRC2-independent roles in carcinomas and that abnormally high levels of this enzyme contribute to malignant transformation. Thus, the role of EZH2 in solid tumors remains unclear.

This work investigates the role of EZH2 in solid tumors in mouse and human samples. We have shown that EZH2 plays a key role in the development of carcinomas and tumorigenesis increases when the enzyme is absent. We have studied precisely the prognostic value of EZH2 expression by correcting for proliferation and demonstrating that in this case, low EZH2 expression is associated wit poor prognosis in breast cancer. We have further showed that mutations in PRC2 genes can be found in breast cancer metastases and associated with poor prognosis. Finally, we demonstrated that altered PRC2 activity promotes transcriptomic instability with irreversible consequences on the gene expression program. Our study suggests that high expression of EZH2 is a consequence rather than a cause of cancer. Moreover, our results report that the alteration of the PRC2 machinery is likely to favor the development of the tumor.

Impaired PRC2 activity promotes transcriptional instability and favors breast tumorigenesis

Michel Wassef,[1,2,3] Veronica Rodilla,[1,2,3] Aurélie Teissandier,[1,4,5] Bruno Zeitouni,[1,4,5] Nadege Gruel,[1] Benjamin Sadacca,[1] Marie Irondelle,[1] Margaux Charruel,[1,2,3] Bertrand Ducos,[6,7,8] Audrey Michaud,[1,2,3] Matthieu Caron,[1,2,3] Elisabetta Marangoni,[1] Philippe Chavrier,[1] Christophe Le Tourneau,[9,10] Maud Kamal,[9] Eric Pasmant,[11,12,13] Michel Vidaud,[11,12,13] Nicolas Servant,[1,4,5] Fabien Reyal,[1] Dider Meseure,[1,14] Anne Vincent-Salomon,[1] Silvia Fre,[1,2,3] and Raphaël Margueron[1,2,3]

[1]Institut Curie, Paris Sciences et Lettres Research University, 75005 Paris, France; [2]U934, Institut National de la Santé et de la Recherche Médicale, 75005 Paris, France; [3]UMR3215, Centre National de la Recherche Scientifique, 75005 Paris, France; [4]U900, Institut National de la Santé et de la Recherche Médicale, 75005 Paris, France; [5]Mines ParisTech, 77300 Fontainebleau, France; [6]Laboratoire de Physique Statistique-Ecole Normale Supérieure de Paris, Centre National de la Recherche Scientifique, 75005 Paris, France; [7]UMR 8550, Centre National de la Recherche Scientifique, 75005 Paris, France; [8]Plateforme de PCR Quantitative à Haut Débit Genomic Paris Centre, Institut de Biologie de l'École Normale Supérieure, 75005 Paris, France; [9]Department of Medical Oncology, Institut Curie, 75005 Paris, France; [10]EA7285, Université de Versailles, Saint-Quentin-en-Yvelines, 78000 Versailles, France; [11]UMR_S745, EA7331, Institut National de la Santé et de la Recherche Médicale, 75006 Paris, France; [12]Facultée des Sciences Pharmaceutiques et Biologiques, Université Paris Descartes, Sorbonne Paris Cité, 75006 Paris, France; [13]Service de Biochimie et Génétique Moléculaire, Assistance Publique-Hôpitaux de Paris, Hôpital Cochin, 75014 Paris, France; [14]Platform of Investigative Pathology, 75005 Paris, France

Alterations of chromatin modifiers are frequent in cancer, but their functional consequences often remain unclear. Focusing on the Polycomb protein EZH2 that deposits the H3K27me3 (trimethylation of Lys27 of histone H3) mark, we showed that its high expression in solid tumors is a consequence, not a cause, of tumorigenesis. In mouse and human models, EZH2 is dispensable for prostate cancer development and restrains breast tumorigenesis. High EZH2 expression in tumors results from a tight coupling to proliferation to ensure H3K27me3 homeostasis. However, this process malfunctions in breast cancer. Low EZH2 expression relative to proliferation and mutations in Polycomb genes actually indicate poor prognosis and occur in metastases. We show that while altered EZH2 activity consistently modulates a subset of its target genes, it promotes a wider transcriptional instability. Importantly, transcriptional changes that are consequences of EZH2 loss are predominantly irreversible. Our study provides an unexpected understanding of EZH2's contribution to solid tumors with important therapeutic implications.

[*Keywords*: cancer; chromatin; Polycomb; EZH2]

Supplemental material is available for this article.

Received July 31, 2015; revised version accepted November 13, 2015.

Eukaryotic cells have developed sophisticated mechanisms to prevent or correct genetic mutations that could result in cell transformation. These mechanisms are often altered during tumor progression, leading to increased genome instability. In addition to genetic lesions, the chromatin undergoes dramatic changes that are routinely used by pathologists to characterize tumor aggressiveness. Consistently, key determinants of chromatin structure and gene regulation are mutated or misregulated in numerous cancer types (You and Jones 2012). Hence, both genetic and epigenetic alterations seem to contribute to deregulation of gene expression programs, favoring the malignant evolution of transformed cells.

The Polycomb group of proteins plays a key role in maintaining transcriptional programs during development (Simon and Kingston 2013), and deregulations of its function has been hypothesized to be involved in cancer (Bracken and Helin 2009). Two multiprotein complexes, Polycomb-repressive complex 1 (PRC1) and PRC2, catalyze a specific modification on the histone tails. The PRC2 complex, through its enzymatic subunits EZH1 and EZH2, is in charge of di- and trimethylation of

Corresponding author: raphael.margueron@curie.fr
Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.269522.115. Freely available online through the *Genes & Development* Open Access option.

Lys27 of histone H3 (H3K27me3), a mark linked to transcriptional silencing. Several types of alteration of PRC2 have been reported in tumors. Heterozygous gain-of-function mutations in EZH2 are found in follicular lymphoma and diffuse large cell B-cell lymphoma (Morin et al. 2010), in which the mutant enzyme is proposed to cooperate with its wild-type counterpart to increase the levels of H3K27me3 (Sneeringer et al. 2010). Conversely, loss-of-function mutations in PRC2 genes occur in malignant peripheral nerve sheath tumors (MPNSTs), myelodysplasia, and T-cell acute lymphoblastic leukemia (T-ALL) (Nikoloski et al. 2010; Ntziachristos et al. 2012; De Raedt et al. 2014).

More relevant to the present work, previous studies reported high levels of EZH2 in carcinomas such as prostate and breast cancer (Varambally et al. 2002; Kleer et al. 2003). In these tumor types, high levels of EZH2 are associated with advanced stages of cancer and poor prognosis. Subsequent studies extended these observations to many other tumor types (for review, see Chase and Cross 2011). Overexpression of EZH2 in cancer was proposed to result from gene amplification (Bracken et al. 2003), down-regulation of microRNA 101 (miRNA-101) (Varambally et al. 2008), and stimulation of its expression by the pRB–E2F (Bracken et al. 2003) and MEK–ERK pathways. In addition, the MYC oncogene can also stimulate EZH2 expression (Koh et al. 2011) and has been suggested to interact with the Polycomb machinery at multiple levels in cancer (for review, see Benetatos et al. 2014). Overexpressed EZH2 was proposed to participate in aberrant silencing of tumor suppressor genes such as *DAB2IP* (Min et al. 2010), *ADRB2*, and *SLIT2*.

Paradoxically, recent studies have reported that the levels of H3K27me3 are decreased in several solid tumor types, including breast and prostate (Wei et al. 2008; Holm et al. 2012; Xu et al. 2012; Healey et al. 2014; Bae et al. 2015). Even more surprising, the levels of the enzyme and the mark were found to be anti-correlated between the different breast cancer subtypes (Holm et al. 2012), and, while high expression of EZH2 correlates with poor prognosis, high levels of H3K27me3 correlate with good prognosis (Holm et al. 2012; Bae et al. 2015). This has led several groups to propose that EZH2 might play PRC2-independent roles in carcinomas (Lee et al. 2011; Xu et al. 2012). However, no clear picture has emerged from these studies on the involvement of EZH2 in solid tumors. Thus, whether elevated expression of EZH2 in carcinomas actively contributes to tumor progression or is simply a consequence of malignant evolution remains an open question.

Here, we set out to investigate the role of EZH2 in carcinomas using genetic tools in mouse and human model systems. We discovered that Ezh2 is largely dispensable for development of solid tumors and that the absence of the enzyme can actually enhance tumorigenesis. Consistently, when corrected for proliferation, the prognostic value of EZH2 expression is inverted; low EZH2 expression relative to proliferation is associated with poor prognosis in breast cancer. In addition, we found that mutations in PRC2 genes are linked to poor prognosis and are found in breast cancer metastases. Importantly, we showed that impaired PRC2 activity promotes transcriptomic instability with irreversible consequences on the gene expression program. Altogether, our study sheds a new light on the interplay between the Polycomb machinery and cancer and calls for caution concerning disruption of PRC2 as a therapeutic strategy.

Results

Ezh2 is dispensable in genetically engineered mouse models of prostate and breast cancers

Given the prior links made between Ezh2 overexpression and the more aggressive forms of prostate cancer (Varambally et al. 2002, 2008), we used genetically engineered mouse models of prostate cancer to investigate the role of the enzyme in carcinogenesis. Both amplification of the *c-MYC* oncogene and loss of the *PTEN* tumor suppressor are common features of human prostate cancer, and corresponding alterations in the mouse prostate result in adenocarcinomas.

We first examined 9- to 12-mo-old Hi-Myc mice, driving c-Myc expression in the prostate, which developed invasive prostate adenocarcinomas with 100% penetrance ($n = 6$). These mice exhibited high levels of Ezh2 and proliferation marker PCNA relative to normal prostates, as shown by both Western blot and immunohistochemistry (IHC) (Fig. 1A,B). Unlike Ezh2, the expression of Ezh1 did not significantly change, while Eed and Suz12, two core PRC2 components, were modestly up-regulated (Fig. 1A; Supplemental Fig. S1A). This Hi-Myc mouse line was then crossed to an *Ezh2* conditional knockout mouse (Su et al. 2003), and genetic deletion of *Ezh2* was induced in prostate epithelium with a *Probasin*-driven Cre recombinase (PB4-Cre) (Wu et al. 2001). Of note, presumably due to the postnatal expression of the Cre, prostate-specific deletion of *Ezh2* in *PB4-Cre;Ezh2*$^{fl/fl}$ males had no noticeable consequences on normal prostate tissue (data not shown). Ezh2 was efficiently depleted in *Hi-Myc; PB4-Cre;Ezh2*$^{fl/fl}$, as assessed by IHC (Fig. 1B) and Western blot (Supplemental Fig. S1B). Importantly, although H3K27me3 was heavily reduced in tumors lacking Ezh2 in 9- to 12-mo-old mice (Fig. 1B), invasive adenocarcinomas still formed with full penetrance ($n = 6$). The invasiveness is evidenced by the disruption of the fibromuscular layer stained by smooth muscle actin (SMA) (Fig. 1B). In addition, *Ezh2* knockout tumors, like *Ezh2* wild-type tumors, retained high levels of PCNA (Fig. 1B), androgen receptor (AR) (Supplemental Fig. S1C), and the epithelial marker E-cadherin (Supplemental Fig. S1C) and were negative for the expression of the tumor suppressor Nkx3.1 (Supplemental Fig. S1C), as previously shown for Hi-Myc tumors (Ellwood-Yen et al. 2003). To determine whether tumors progressively adapt to lack of Ezh2 or whether Ezh2 is overall dispensable in this model, we knocked down Ezh2 through shRNA interference in a cell line derived from advanced Hi-Myc tumors (MycCaP) (Watson et al. 2005). Despite strong down-regulation of H3K27me3, proliferation was unimpaired. Prior studies

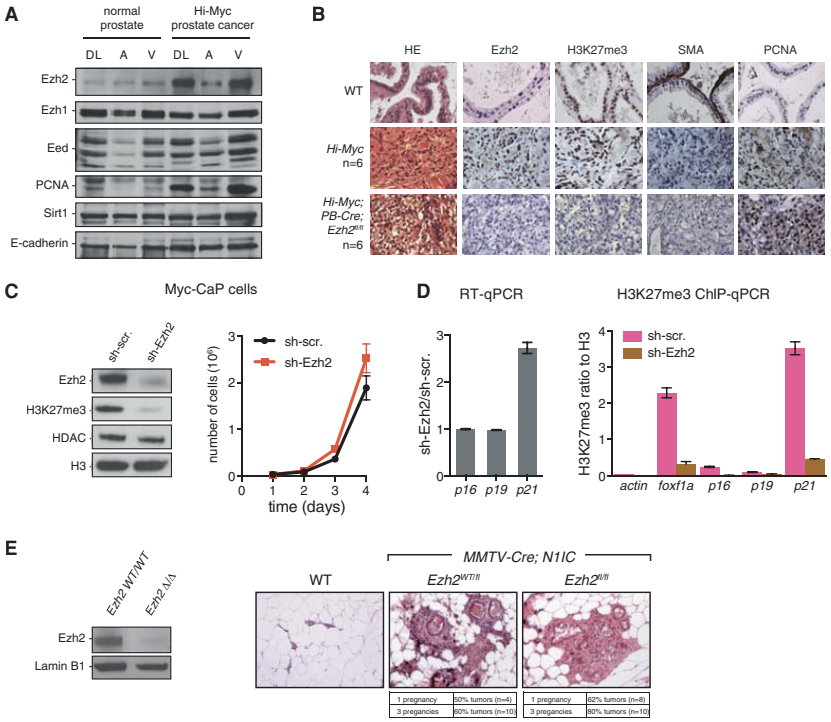


Figure 1. Ezh2 is dispensable in genetically engineered mouse models of prostate and breast cancers. (*A*) Western blot on whole-tissue lysates from dorso–lateral (DL), anterior (A), and ventral (V) lobes of a normal prostate and a Hi-Myc invasive prostate cancer. The specific antibodies used are indicated at the *left*. (*B*) Hematoxylin and eosin (HE) and IHC staining of different proteins, as indicated, in a normal mouse prostate (wild type) and *Hi-Myc* and *Hi-Myc;PB-Cre;Ezh2*$^{fl/fl}$ invasive prostate cancer. (*C*) Impact of Ezh2 knockdown in Myc-CaP cells (derived from Hi-Myc mouse prostates) on Ezh2, H3K27me3, HDAC, and H3 (Western blot; *left* panel) as well as cell proliferation (*right* panel). Mean ± SD. $n = 3$. (*D*, *left* panel) RT-qPCR analysis of *p16*, *p19*, and *p21* expression in Myc-CaP cells. RT-qPCR values indicate relative expression in sh-Ezh2 compared with sh-scramble cells after normalization to TBP. (*Right* panel) Enrichment of H3K27me3 by chromatin immunoprecipitation (ChIP) and qPCR (ChIP-qPCR) at the corresponding loci. *actin* and *foxf1a* were used as negative and positive controls, respectively. ChIP-qPCR values indicate relative enrichment compared with histone H3. Mean ± SD. $n = 3$. (*E*, *left* panel) Western blot showing loss of Ezh2 protein in *MMTV-Cre;N1IC;Ezh2*$^{fl/fl}$ FACS-sorted luminal cells. *MMTV-Cre;N1IC;Ezh2*$^{wt/wt}$ cells were used as a control. (*Right* panel) Representative HE staining on mammary glands of wild-type, *MMTV-Cre;N1IC;Ezh2*$^{wt/fl}$, and *MMTV-Cre;N1IC;Ezh2*$^{fl/fl}$ mice showing the presence of tumors in the presence or absence of Ezh2.

suggest that Ezh2 can control cell proliferation in part through silencing of the *Ink4a/Arf* and *p21* tumor suppressor loci (Bracken et al. 2007; Seward et al. 2013). However, the levels of *p16/p19* transcripts, already detectable in sh-scramble Myc-CaP cells, were not affected upon Ezh2 knockdown in this model (Fig. 1C,D; data not shown). The *p21* transcript was nevertheless significantly up-regulated (Fig. 1D). H3K27me3 was present at low, close to background, levels at the *Ink4a/Arf* locus in comparison with *foxf1a* (an established PRC2 target) and *p21* loci (Fig. 1D, right panel). Thus, in the context of c-Myc-induced prostate cancer, cell proliferation and malignant evolution appear unaffected by the absence of Ezh2.

We analyzed a second model of prostate cancer, generated by deletion of the *Pten* tumor suppressor. Conditional deletion of this gene in mouse prostates leads to prostate adenocarcinomas with varying degrees of severity (Wang et al. 2003; Ma et al. 2005), presumably due to differences in the genetic background and/or mutant allele used. In our mixed strain, PB4-Cre-induced deletion of *Pten* led to intraepithelial neoplasia at 6–9 mo of age showing no sign of invasion ($n = 7$) (Supplemental Fig. S1D). Relative to normal prostates, Ezh2 expression was nonetheless up-regulated in these tumors (Supplemental Fig. S1D, left panel). However, similar to the Hi-Myc model, deletion of *Ezh2* did not prevent tumor development ($n = 7$) (Supplemental Fig. S1D, right panel).

Since high EZH2 expression has also been reported in breast cancer (Kleer et al. 2003), we next turned to a mouse model eliciting mammary tumors upon mammary-specific expression of the activated form of Notch1 (N1ICD). Aberrant Notch signal activation is a common feature of human breast cancers (Stylianou et al. 2006) and has been shown to induce mammary tumors in mice (Bolos

et al. 2013). We induced ectopic Notch activation by targeting an inducible *Rosa26*^{flox}*N1ICD*^{flox} allele (Murtaugh et al. 2003) to the mammary epithelium with MMTV-Cre mice, as previously described (Bolos et al. 2013). *Rosa26-N1ICD;MMTV-Cre* compound female mice developed hormone-dependent mammary tumors. They were subjected to one or three rounds of pregnancy and analyzed for the presence of tumors. In this model, penetrance of tumor development was incomplete even after three rounds of pregnancy. Mammary-specific deletion of *Ezh2* did not impair tumor development but in fact resulted in an increased penetrance of tumor formation (Fig. 1E).

Altogether, our findings based on three different mouse models indicate that solid tumors can develop in the absence of Ezh2.

H3K27me3 homeostasis is compromised in breast cancer

Since Ezh2 is dispensable for mouse prostate and mammary cancer development, we wondered why the enzyme is nonetheless highly up-regulated in tumors. High EZH2 expression has been repeatedly found to be associated with proliferating tissues (e.g., Margueron et al. 2008), and its expression was shown to be under the influence of key cell proliferation pathways (Bracken et al. 2003). In addition, EZH2 expression in several solid tumor types was shown to be correlated with proliferation (Bachmann et al. 2006). Thus, elevated expression of EZH2 in cancer may simply result from abnormally high cell proliferation rates in tumors rather than deregulated expression.

To obtain further insight into the expression of EZH2 in cancer, we analyzed transcriptome data from a publicly available study on 131 primary prostate tumors and 19 metastases (Taylor et al. 2010). As expected, hierarchical clustering of the transcriptome data revealed that the *EZH2* transcript is part of a cluster of genes highly expressed in metastatic prostate cancer (Fig. 2A). Importantly, cell cycle and proliferation genes (e.g., *Ki67* and *PCNA*) are overly represented in this cluster. It is noteworthy that several transcripts (e.g., *Ki67*) (Supplemental Fig. S2A) display stronger differential expression between primary and metastatic cancer than *EZH2*. This further suggests that association of high *EZH2* with cancer aggressiveness might reflect the increased cell proliferation occurring in advanced stages of prostate cancer (e.g., see Tomlins et al. 2007). A similar analysis on a breast cancer cohort comprising 146 samples from the four main molecular subtypes (Maire et al. 2013) confirmed that EZH2 expression correlates with proliferation markers (Supplemental Fig. S2B). In addition, analysis of copy number data from the same cohort revealed that amplification of *EZH2* is a rare event, since no instances were found in this data set. Gains occur in proportions similar to losses (Supplemental Fig. S2C; data not shown), arguing against a major role for copy number gains or amplifications in driving high EZH2 levels. Thus, high EZH2 expression seems to be predominantly linked to proliferation.

To assess why EZH2 expression is associated with cell proliferation, we turned to a cell-based system allowing modulation of proliferation rate through increased serum concentration and addition of growth factors. Modulation of Myc-CaP proliferation in vitro revealed that, while Ezh2 expression shows a near-perfect correlation to the rate of cell division, H3K27me3 remains constant (Fig. 2B, left). This result is consistent with a previous study monitoring EZH2 and H3K27me3 upon serum stimulation of quiescent cells (Hansen et al. 2008). It further suggests that proliferation-induced Ezh2 levels may serve to oppose cell division-mediated dilution of H3K27me3. To test this hypothesis, we altered Ezh2 expression using shRNA-mediated knockdown. Ezh2 expression was reduced as expected, but, more importantly, the rate of increase of Ezh2 with proliferation was also diminished (Supplemental Fig. S2D). This resulted in a gradual drop of H3K27me3 (Fig. 2B, right), suggesting that the increase of Ezh2 was no longer sufficient to counteract cell division-mediated dilution of the mark. Thus, while Ezh2 is not an obligate modulator of cell proliferation, the tight coupling of Ezh2 expression levels to the rate of cell division is required to ensure homeostatic maintenance of H3K27me3.

This result prompted us to hypothesize that the anti-correlated levels of EZH2 and H3K27me3 observed in several solid tumor types might stem from a failure to properly counteract cell division-mediated dilution of the histone mark. We thus sought to assess the impact of modulating proliferation on the levels of EZH2 and H3K27me3 in the context of human breast cancers. We analyzed EZH2 and H3K27me3 levels by IHC in two previously characterized patient-derived xenografts (PDXs) of estrogen-positive breast cancer (Cottu et al. 2014). The engrafted mice were treated with various combinations of endocrine therapies and the mTOR inhibitor everolimus, the impact of which on tumor proliferation was evaluated by Ki67 staining (Cottu et al. 2014). As previously reported (Supplemental Table S1; Cottu et al. 2014), some drug combinations led to a near complete inhibition of cell proliferation (e.g., everolimus + fulvestrant), while other treatments only reduced proliferation (e.g., everolimus alone or everolimus + tamoxifen) or failed to impair proliferation (e.g., ovariectomy). Quantification of EZH2 signal revealed that it was highly correlated to Ki67 (Fig. 2C, left; Supplemental Table S1), confirming that, in the context of tumors, EZH2 expression is under the control of proliferation cues. Importantly, though, H3K27me3 signal was significantly anti-correlated to both Ki67 (Fig. 2C, right) and EZH2 (Supplemental Fig. S2E).

Although the drugs used are likely to impact processes other than proliferation, which might lead to confounding effects on H3K27me3 homeostasis, these data suggest that, in spite of higher EZH2 levels, the PRC2 complex might not be able to match the abnormally high proliferation of breast cancer cells, leading to down-regulation of H3K27me3.

Genetic loss of EZH2 *is linked to poor prognosis in breast cancer*

Our results question the contribution of proliferation to EZH2's prognostic value. Indeed, such an association

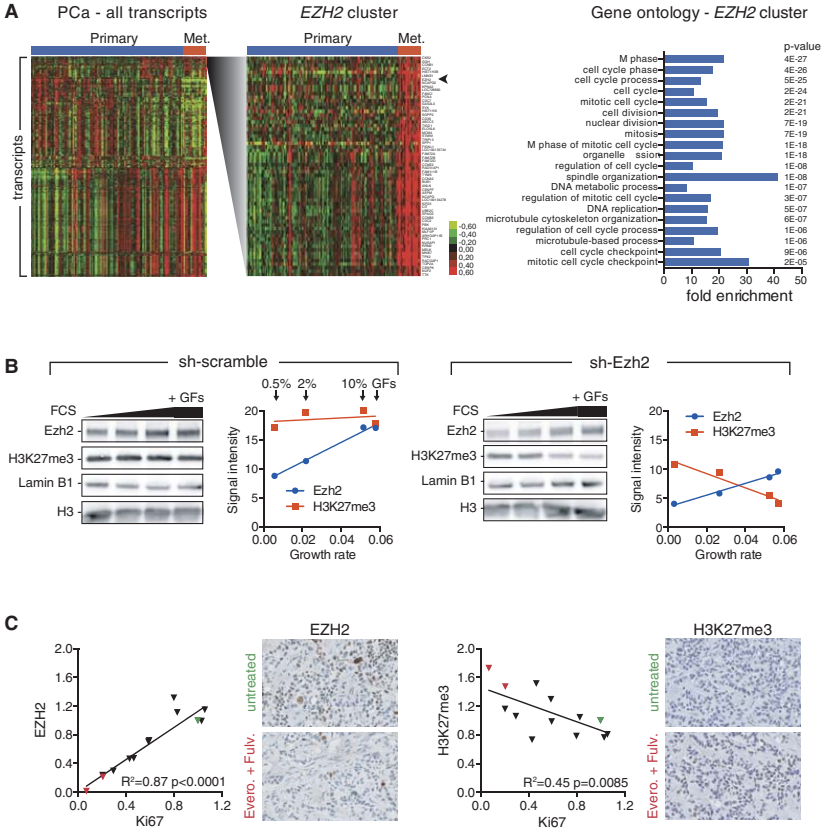


Figure 2. Coupling of EZH2 expression to proliferation is required for H3K27me3 homeostasis but perturbed in breast cancers. (*A*, *left* panel) Heat map of hierarchical clustering of the most significantly varying transcripts in primary and metastatic (Met.) prostate cancers (PCa). Data are from Taylor et al. (2010). Samples are arranged horizontally, and transcripts are arranged vertically. The cluster containing the *EZH2* transcript is shown in greater detail at the *right*. (*Right* panel) Gene ontology (DAVID, http://david.abcc.ncifcrf.gov) of the *EZH2* cluster showing the 20 most significantly enriched categories, their fold enrichment, and corresponding *P*-values. (*B*) Western blot probed with antibodies recognizing Ezh2, H3K27me3, Lamin B1, or histone H3 in sh-scramble (*left* panel) and sh-Ezh2 (*right* panel) Myc-CaP cells. In order to modulate proliferation in vitro, cells were cultured in the presence of 0.5%, 2%, or 10% fetal calf serum (FCS) or 10% FCS medium plus a cocktail of growth factors (bovine pituitary extract, insulin, and epidermal growth factor, indicated as +GFs in the last lane). Corresponding dot plots show signal quantification of Ezh2 (blue) and H3K27me3 (red) abundance (arbitrary units) as a function of the growth rate (number of divisions per cell per hour), as assessed by proliferation assays carried out in parallel for each culture condition. (*C*) EZH2 and H3K27me3 IHC staining quantifications across two patient-derived xenografts (PDXs) treated with various combinations of drugs. Correlation plots of EZH2 versus Ki67 and H3K27me3 versus Ki67 signal intensities are shown. Intensity values were normalized to the control (untreated) condition (green triangles). The everolimus + fulvestrant-treated PDXs show strongly reduced proliferation (red triangles). Each dot corresponds to the mean of six measurements (two stainings on three biological replicates). The corresponding coefficient of determination (R^2) and *P*-value of the linear regression are shown. Representative IHC staining for EZH2 and H3K27me3 in untreated and everolimus + fulvestrant treated PDXs are shown. Nuclei are counterstained in blue/purple.

was found in many gene expression-based signatures associated with clinical outcome (Venet et al. 2011).

To address this issue, we used transcriptome/CNV data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al. 2012), which collected information on transcript levels and copy number as well as long-term clinical follow-ups from 2000 breast cancers (Curtis et al. 2012). We first investigated the prognostic value of the *EZH2* transcript in comparison with that of the *ORC6* transcript (a proliferation-associated transcript used as a control) and a proliferation metagene consisting of the median expression of 54 proliferation-associated transcripts used as a molecular readout for proliferation (Nagalla et al. 2013). As expected, all three variables displayed a significant association with outcome as assessed by Receiver Operating

Characteristic (ROC) analysis (Supplemental Fig. S3A) or Kaplan-Meier analysis (Fig. 3A, top panels), although the prognostic value of *EZH2* expression was the least powerful. We then sought to evaluate the prognostic value of the *EZH2* and *ORC6* transcripts independently of proliferation. For this purpose, we calculated residual ("adjusted") values of *EZH2* and *ORC6* to the proliferation metagene (shown for *EZH2* in Fig. 3A, left panels). Strikingly, adjusted *EZH2* was now negatively associated with outcome (Fig. 3A, bottom middle panel; Supplemental Fig. S3B), suggesting that low *EZH2* expression relative to proliferation is linked to a poor prognosis. By comparison, the adjusted *ORC6* transcript no longer bore any association with outcome (Fig. 3A, bottom right panel; Supplemental Fig. S3C), confirming that its prognostic value is mainly proliferation-dependent. Thus, the association of *EZH2* expression with prognosis comprises both a positive component linked to proliferation and a negative component independent of proliferation.

Copy number variations seemed to largely account for variations of *EZH2* levels independently of proliferation (Fig. 3A). We therefore stratified tumors according to the copy number status of the gene. Hemizygous loss of *EZH2* was indeed linked to a significantly worse prognosis in comparison with normal *EZH2* copy number (Fig. 3B). Conversely, gain of *EZH2* was associated with better prognosis than normal *EZH2* copy number. Association of *EZH2* CNV with outcome was independent of estrogen status (Supplemental Fig. S3D,E), although more pronounced in estrogen-positive tumors. In order to determine the size of the region surrounding *EZH2* showing a correlation between genetic loss and outcome, we analyzed the prognostic association of all annotated genes on chromosome 7. Loss of the long arm of chromosome 7 was significantly linked to poor prognosis, with the end of the arm (encompassing the *EZH2* locus) having the strongest association (Fig. 3C).

We next assessed alterations of *EZH2* or other genes encoding core PRC2 components by targeted sequencing in breast cancer metastases previously analyzed by Affymetrix CytoScan arrays (Le Tourneau et al. 2014). Interestingly, in addition to metastases having missense mutations in *SUZ12* (two samples) and *EZH2* (one sample), one metastasis harbored a critical splice site mutation in *EZH2* in the −1 position relative to exon 11 (Fig. 3D; Supplemental Table S2). This mutation is predicted to abolish splicing at this intron/exon junction and to result in a truncated protein lacking the catalytic SET domain. Since it is found in a metastasis having a hemizygous loss of *EZH2*, this mutation could drive a complete PRC2 loss of function.

Finally, we investigated the presence of homozygous loss or mutation of PRC2 core component genes in The Cancer Genome Atlas (TCGA) breast cancer cohort. Strikingly, the group of tumors harboring such mutations (3% of all tumors) (Fig. 3E) displayed a significantly worse prognosis than the non-PRC2 mutant tumors (Fig. 3F).

In summary, our data show that the association of EZH2 expression to prognosis results from its correlation to proliferation. Strikingly, low levels of EZH2 relative to proliferation, resulting from genetic loss of the gene and mutations in PRC2 genes, are in fact associated with poor prognosis.

Genetic disruption of EZH2 *in a breast cancer cell line promotes tumorigenesis*

Since our analyses indicated that impaired PRC2 function is associated with poor prognosis, we used CRISPR/CAS9-based genome engineering tools to delete *EZH2* in a cellular model of human breast cancer. We chose the MDA-MB-231 cell line, a widely studied near-triploid cell line derived from a metastatic triple-negative (estrogen-, progesterone-, and HER2-negative) breast cancer. We sequentially targeted all three alleles of *EZH2* and confirmed that the resulting cell line no longer expressed EZH2 when compared with the parental clone carrying only one mutant allele, leading to a near-complete erasure of H3K27me3 (Fig. 4A). In contrast to a previous report using RNAi (Gonzalez et al. 2009), loss of EZH2 did not have an impact on cell proliferation (Fig. 4B). However, we observed an increased three-dimensional (3D) cell migration through type I collagen, indicative of metastatic potential ($n = 52$ for control cells, and $n = 45$ for *EZH2*-null cells) (Fig. 4C). Importantly, we confirmed these results on proliferation and 3D cell migration using an inhibitor targeting both EZH1 and EZH2 (Supplemental Fig. S4; Konze et al. 2013), indicating that genetic deletion of EZH2 recapitulates pharmacological inhibition of the enzyme and that EZH1 does not compensate for loss of EZH2 in this model. Finally, we analyzed the consequences of *EZH2* deletion on orthotropic tumor growth in mammary fat pads of immune-deficient host mice. Strikingly, tumors originating from *EZH2*-null xenografts were significantly bigger than the control tumors ($n = 12$ for control xenografts, and $n = 13$ for *EZH2*-null xenografts) (Fig. 4D).

These results suggest that PRC2-mediated gene silencing might have a protective function in breast tumorigenesis.

Impaired PRC2 function selectively affects H3K27me3-low genes

Our analysis suggests that partial impairment of PRC2 might be sufficient to promote tumorigenesis. We therefore analyzed how incomplete disruption of PRC2 affects transcription of Polycomb target genes. For this purpose, we used a c-Myc transformed, Ezh2 conditional mouse embryonic fibroblast (iMEF) clonal cell line. This model allows OHT-dependent deletion of *Ezh2* and results in a drastic reduction of H3K27me3 and subsequent up-regulation of a small cohort of H3K27me3-positive genes, which we refer to as direct responsive targets (Fig. 5A). To assess the impact of a milder down-regulation of H3K27me3, we analyzed gene expression at an early time point after OHT treatment such that the mark was only partially depleted (day 5 after OHT treatment) (Fig. 5B). Only a subset (9%) of responsive genes was up-regulated at this time point (Fig. 5C; Supplemental Fig. S5A). Strikingly, early responsive genes were characterized by

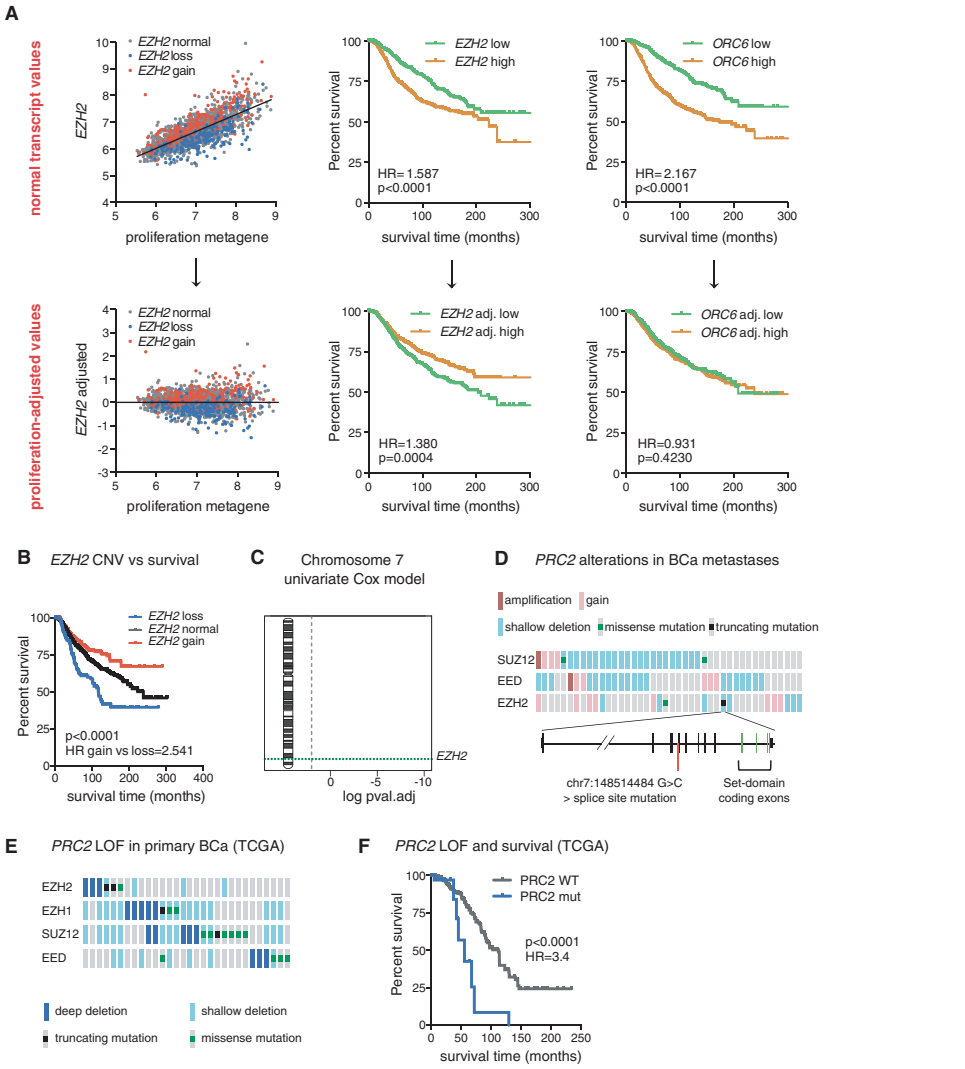


Figure 3. Genetic loss of *EZH2* is linked to poor prognosis in breast cancer. (*A*, *top left* panel) Correlation plot of *EZH2* transcript and a proliferation metagene. Residual (adjusted) values of *EZH2* transcripts to the proliferation metagene are shown in the *bottom left* panel. *EZH2* copy number variations are color-coded, with normal copy number in gray, hemizygous loss in blue, and gain in red. The same procedure was applied to adjust *ORC6* transcript values (data not shown). Kaplan-Meier plots of breast cancer-specific survival for patients with primary tumors with high (above median) or low (below median) *EZH2* and *ORC6* transcript levels are shown in the *middle* and *right top* panels. Kaplan-Meier plots of breast cancer-specific survival for patients with primary tumors with high versus low proliferation-adjusted levels of *EZH2* and *ORC6* transcripts are shown in the *middle* and *right bottom* panels. The hazard ratio (HR) between the highest and lowest survival groups and *P*-values are displayed on Kaplan-Meier plots. (*B*) Kaplan-Meier plot of breast cancer-specific survival for patients with primary tumors with normal *EZH2* or hemizygous loss or gain of *EZH2*. (*C*) Univariate analysis showing the association between genetic loss and death from breast cancer on all genes of chromosome 7. False discovery rate (FDR)-corrected *P*-values ($\log_{10}$ scale) are plotted for all chromosome 7 genes, and significant values are highlighted in red (threshold of 0.15). A dashed green line indicates the position of the *EZH2* locus. The analyses shown in *A*–*C* were performed on data from 2000 primary breast cancers of the METABRIC cohort. (*D*, *top*) Oncoprint generated on the cBioPortal OncoPrinter showing genomic alterations and mutations in genes encoding PRC2 core components in 58 breast cancer (BCa) metastases. Only altered cases are shown. (*Bottom*) Schematic representation of the *EZH2* locus showing the position of a splice site mutation in position −1 of exon 11. (*E*) Oncoprint (cBioPortal) showing loss-of-function (LOF) mutations of core PRC2 genes in The Cancer Genome Atlas (TCGA) breast cancer data set. (*F*) Kaplan-Meier plot of overall survival associated with the corresponding tumors compared with the remaining (PRC2 wild-type) tumors.

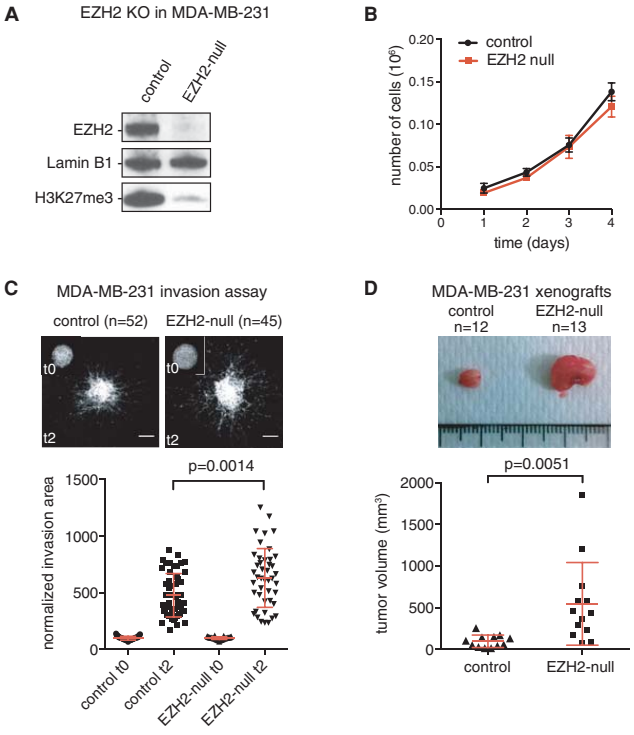


Figure 4. Genetic disruption of *EZH2* in a breast cancer cell line promotes tumorigenesis. (*A*) Western blot showing loss of the EZH2 protein and H3K27me3 mark in the EZH2 full knockout MDA-MB-231 cell line (indicated as EZH2-null) compared with the parental clone mutant for one allele out of three. (*B*) Proliferation curve of control and EZH2-null cells. (*C*) Multicellular spheroids of control or EZH2-null MDA-MB-231 cells were embedded in 3D acid-extracted type I collagen (T0) and further incubated for 2 d (T2). Images show representative phalloidin-labeled spheroids collected at T0 (*inset*) or T2. Bars, 200 µm. Data represent mean invasion area in type I collagen at T2 normalized to the mean invasion area at T0 ±SEM. $n = 3$; 15–20 spheroids were analyzed for each cell line, with a total of 52 and 45 measurements for control and EZH2-null cells, respectively. Red bars indicate mean ± SD. The *P*-value of the two-tailed unpaired *t*-test is indicated. (*D*) Representative pictures of orthotopic tumor xenografts developed from the control MDA-MB-231 clone and EZH2-null clone (*top* panel) and a plot showing corresponding tumor volumes (*bottom* panel). Red bars indicate mean ± SD. The *P*-value of the two-tailed unpaired *t*-test is indicated.

a low level of H3K27me3 specifically in the promoter region as compared with late responsive genes (Fig. 5D). This result suggests that accumulation of the mark in the promoter region controls the robustness of transcriptional repression.

To confirm that partial loss of H3K27me3 indeed releases the silencing of a subset of PRC2 target genes, we performed a complementary experiment in which we rescued the loss of Ezh2 by Ezh1, an enzyme that was previously reported to have a reduced enzymatic activity relative to Ezh2 (Margueron et al. 2008). In this experiment, Ezh1 or Ezh2 was stably expressed in cells before OHT-induced deletion of endogenous *Ezh2*. As expected, the global level of H3K27me3 was significantly lower in the Ezh1 rescue condition than in the control Ezh2 rescue condition (Fig. 5E, cf. lanes 5 and 6), and the genomic distribution of H3K27me3 was uniformly weaker in the Ezh1 rescue condition (Supplemental Fig. S5B). Responsive genes that could not be rescued by Ezh1 (26%) (Fig. 5F) had an initial lower enrichment of the mark in their promoter region compared with genes for which expression was rescued (Fig. 5G), thus corroborating our time-course analysis (Fig. 5D). Altogether, these results indicate that H3K27me3 accumulation in the promoter region is linked to robustness toward depletion of the mark; a mild decrease of H3K27me3 selectively impairs silencing of genes that have a low level of the mark in their promoter.

Impaired PRC2 function leads to transcriptional instability

Alterations of PRC2 have been observed in cancers of different origin, indicating a fundamental, tissue-independent role in tumor suppression. However, disruption of PRC2 only results in the detectable up-regulation of a minority of tissue-specific genomic targets (Fig. 5A; Ezhkova et al. 2009; Woodhouse et al. 2013). We reasoned that, in addition, low-frequency (e.g., stochastic) responses might occur at the level of nonresponsive targets, leading to increased transcriptional instability. We thus asked how responsive and nonresponsive targets would be expressed in the presence or absence of Ezh2 at the level of individual cells. RNA FISH analysis indicated that responsive targets are expressed at low frequency in the presence of the enzyme (Fig. 6A). Single-cell RT-qPCR confirmed this observation and revealed that responsive target genes are detected in a subset of Ezh2 wild-type cells expressing a low level of the enzyme (Fig. 6B). Interestingly, while some nonresponsive genes were insensitive to Ezh2 status (Fig. 6B, bottom genes), a number of genes became activated in *Ezh2*-low cells in a sparse fashion (Fig. 6B, middle genes). Deletion of *Ezh2* resulted in a full derepression of responsive genes, while the frequency of expression was increased for nonresponsive genes. Remarkably, while responsive targets were expressed in a concerted—i.e.,

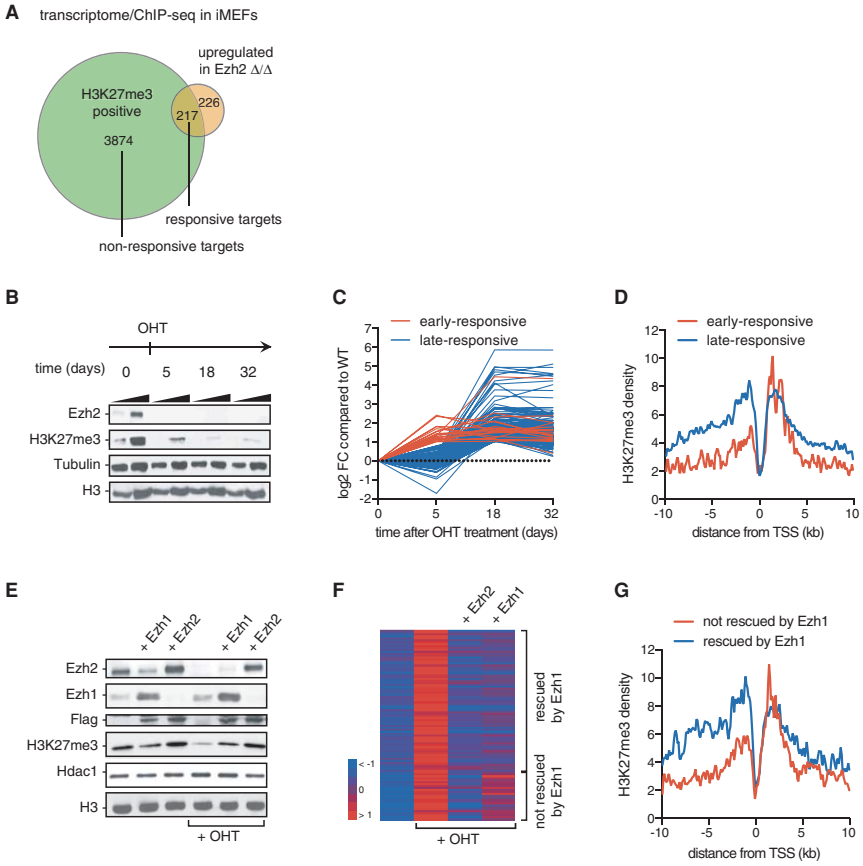


Figure 5. Impaired PRC2 function selectively affects H3K27me3-low genes. (*A*) Venn diagram showing the overlap between up-regulated transcripts upon *Ezh2* deletion and Ezh2 targets defined by the presence of H3K27me3 in the promoter of wild-type versus Ezh2$^{\Delta/\Delta}$ iMEFs. Up-regulated transcripts were identified with a minimum adjusted *P*-value of 15% and a minimum fold change of two. Targets that are H3K27me3-enriched and up-regulated upon loss of Ezh2 ($n = 217$) are defined as "responsive targets" as opposed to H3K27me3-enriched "nonresponsive targets" ($n = 3974$). Genes that are up-regulated but are H3K27me3-negative ($n = 226$) correspond to indirect targets. (*B*) Time-course analysis by Western blot of Ezh2 and H3K27me3 before (time 0) and at different times after OHT-induced *Ezh2* deletion. A two-point titration is provided for each condition. Tubulin and H3 were used as loading controls. (*C*) Time-course expression analysis of Ezh2-responsive direct targets. Red represents early responsive targets, and blue indicates late responsive genes. All values were normalized to initial (time 0) values. (*D*) H3K27me3 density plot around the transcription start site (TSS) of early responsive (in red) and late responsive targets (in blue). (*E*) Western blot of control, Ezh1-overexpressing, and Ezh2-overexpressing iMEFs untreated or treated with OHT to remove endogenous Ezh2 expression. Specific antibodies are indicated at the *left*. (*F*) Heat map representing the mean-centered expression of Ezh2-responsive targets in wild-type, Ezh2$^{\Delta/\Delta}$, and Ezh2$^{\Delta/\Delta}$ iMEFs overexpressing either Ezh2 or Ezh1. (*G*) H3K27me3 density plot around the TSSs of genes for which the absence of endogenous Ezh2 is rescued by Ezh1 (blue line) or not (red line). Analyses presented in *C*, *D*, *F*, and *G* were performed using measurements on two biological replicates.

deterministic—fashion (e.g., cells expressed either all or none of the responsive targets), the expression of nonresponsive targets seemed probabilistic, with each cell expressing a different combination of genes. Since Polycomb target genes represent on the order of 3000–4000 genes, the observed effects of PRC2 disruption are expected to translate into widespread transcriptional instability. Thus, gene expression analysis at the single-cell level reveals that changes occurring at the level of Polycomb target genes are much more profound than previously appreciated.

We next asked whether disruption of PRC2 would translate into a long-lasting impact on gene expression. We therefore inquired whether reintroduction of Ezh2

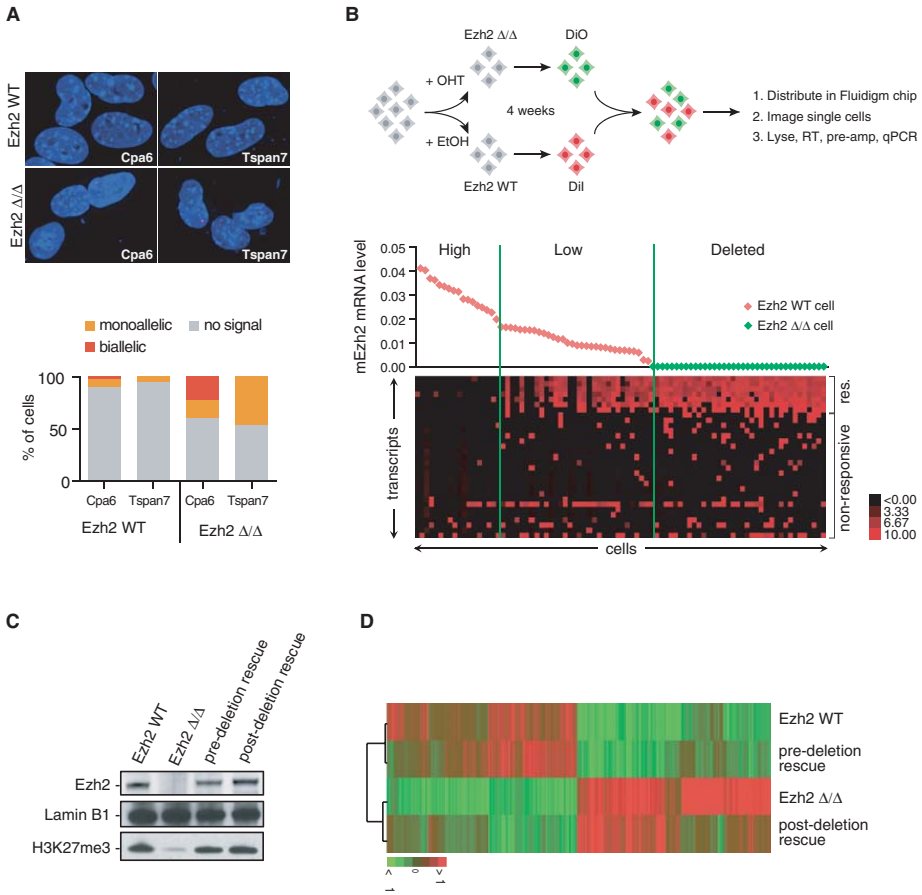


Figure 6. Impaired PRC2 function leads to transcriptional instability. (*A*) Nascent RNA FISH analysis of two responsive Ezh2 targets in Ezh2 wild-type and Ezh2 mutant iMEFs. The *Cpa6* gene is autosomal, while the *Tspan7* gene is localized on the X chromosome and thus only presents in one copy in this male cell line. The *top* panel shows representative examples of RNA FISH signals, and the *bottom* graph shows relative proportions of nuclei with no signal, one pinpoint (monoallelic), and two pinpoints (biallelic) over a minimum of 50 nuclei. (*B*, *top*) Experimental scheme for the single-cell analysis of PRC2 target genes. (*Bottom*) Single-cell analysis of the *Ezh2* transcript and selected responsive (res.) and nonresponsive genes. Forty-nine Ezh2 wild-type and 37 Ezh2 mutant cells were analyzed by RT-qPCR on a Biomark-HD system. *Ezh2* mRNA level in individual cells is plotted at the *top*; red diamonds represent Ezh2 wild-type cells (DiI-positive), and green diamonds indicate Ezh2$^{\Delta/\Delta}$ cells (DiO-positive). A heat map representing the mean-centered, $\log_2$ transformed expression of selected target genes is displayed at the *bottom*. (*C*) Western blot of Ezh2, H3K27me3, and Lamin B1 as a loading control in different conditions as indicated at the *top* of each lane. (*D*) Heat map showing hierarchical clustering of transcripts in Ezh2 wild-type, Ezh2$^{\Delta/\Delta}$, and pre- and post-deletion rescue conditions.

in Ezh2 knockout iMEFs could revert loss of gene silencing. We compared Ezh2 wild-type, Ezh2 knockout, and Ezh2 rescue before and after deletion of the endogenous gene (hereafter called predeletion and post-deletion rescues). In both pre- and post-deletion rescue conditions, the global levels of Ezh2 and associated H3K27me3 were similar to that of wild-type cells (Fig. 6C). Strikingly, although the predeletion rescue prevented most transcriptional changes resulting from the absence of endogenous Ezh2, the post-deletion rescue failed to revert the transcriptional status of the majority of transcripts and clustered with the Ezh2 knockout condition (Fig. 6D). This indicates that transient disruption of PRC2 results in a permanent epigenetic switch in gene expression.

Thus, PRC2 safeguards genome-wide silencing through fine-tuned H3K27me3. Perturbation of this equilibrium results in both predictable, deterministic responses and stochastic loss of gene silencing with irreversible consequences on gene expression programs.

Discussion

The current paradigm concerning the role of EZH2 in solid tumors postulates that abnormally high levels of this enzyme contribute to malignant transformation. Our study challenges this hypothesis. We present evidence indicating that high expression of EZH2 is a consequence rather than a cause of cancer and that, in breast cancer, disruption of the PRC2 machinery is likely to promote tumor development.

Using two mouse models of prostate cancer and one model of mammary tumorigenesis, we found that EZH2, although highly up-regulated in cancerous tissue, is dispensable for tumor progression. The strong correlation of EZH2 levels with proliferation markers in transcriptome analyses and in our tumorgraft series suggest that high EZH2 expression in cancers is predominantly a consequence of increased cell proliferation rate. Through fine-tuning of Ezh2 expression in vitro, we demonstrate that the tight coupling of Ezh2 expression to proliferation is required to oppose cell division-mediated dilution of H3K27me3, as previously hypothesized (Hansen et al. 2008). This is evidenced by the failure to maintain H3K27me3 with rising proliferation rates when the increase of EZH2 is not sufficient. These data imply that relative levels compared with the proliferation rate rather than absolute levels of EZH2 are a key factor in determining H3K27me3 levels. It also suggests that anti-correlated levels of the enzyme and the mark can be caused by a relative reduction of PRC2 activity compared with proliferation. Our analysis of EZH2 and H3K27me3 levels in our PDX series indeed suggests that this is likely to be the case, since proliferation, although positively influencing EZH2 levels, negatively modulates H3K27me3 abundance. Thus, PRC2 activity might not be sufficient to maintain the mark in rapidly dividing breast cancer cells. We propose that these observations reconcile contradictory data reporting inverse variations of EZH2 and H3K27me3 in several tumor types (Wei et al. 2008; Holm et al. 2012; Xu et al. 2012; Healey et al. 2014; Bae et al. 2015).

In addition, we found that while high *EZH2* expression is overall correlated to a poor prognosis in breast cancer, this association can be subdivided into two opposite components. The first component, originating from the coupling of *EZH2* expression to proliferation, associates high *EZH2* with an adverse outcome. However, we found that copy number-driven, proliferation-independent expression of *EZH2* displays an inverse association with tumor outcome, with low expression of *EZH2* being linked to a poor prognosis. This finding emphasizes the need to carefully account for the effect of proliferation when assessing the prognostic value of a given marker in cancers. In addition, this result suggests that decreased levels of EZH2 relative to proliferation might accelerate tumor development. In support of a protective role for PRC2 in breast cancer, we found that mutations in PRC2 core components are associated with a poor prognosis and documented several mutations in *EZH2* and PRC2 core component *SUZ12* in breast cancer metastases. One mutation in *EZH2* is predicted to profoundly affect its function. Finally, inactivation of *EZH2* in a prototypical human breast cancer cell line promotes in vitro invasion and in vivo tumor growth. Together, our findings indicate that EZH2 is likely to constrain breast tumorigenesis.

Although several studies have assessed the role of EZH2 in prostate and breast tumorigenesis, our study is, to our knowledge, the first to use genetic tools in both mouse and humans models. Of note, a recent study investigating the role of Ezh2 in a Brca1 deficiency-based model of mammary tumorigenesis found that deletion of *Ezh2* shortens the latency of tumor formation (Bae et al. 2015), further reinforcing the view that the enzyme might inhibit breast tumorigenesis.

In addition to leukemia and MPNST, a tumor-suppressive role for PRC2 has been suggested in a mouse model of pancreatic cancer (Mallen-St Clair et al. 2012) and in renal cancer (Vanharanta et al. 2013). Moreover, in pediatric glioblastomas, point mutations resulting in a change from lysine to methionine at position 27 of histone H3 (H3K27M) have been shown to inhibit PRC2 activity (Lewis et al. 2013), suggesting that disruption of the Polycomb machinery might be a recurring theme in cancers. However, how PRC2 impairment is linked to tumor progression is currently unclear. Our transcriptomic analysis revealed that alterations of PRC2 activity result in both a deterministic activation of a subset of PRC2 target genes and a broader stochastic activation of gene expression. While the former is expected to control the immediate biological response to Ezh2 inhibition, the latter, by increasing the plasticity of gene expression programs, might lead to long-term responses. Such a distinction between early and late response to Ezh2 inhibition was recently reported in a model of glioblastoma in which prolonged knockdown of Ezh2 results in the emergence of "escaper" tumors characterized by an aggressive phenotype (de Vries et al. 2015). It is tempting to speculate that the transcriptional instability of Polycomb targets as a consequence of Ezh2 knockdown might have fueled the emergence of escaper tumors. Given the high mutation rates reported for other chromatin regulators in cancer, it will be interesting to determine whether they also results in increased transcriptomic instability.

Several EZH2 inhibitors are entering clinical trials. It is expected that tumor types in which EZH2 gain-of-function mutations occur (e.g., DLBCL and FL) (Campbell et al. 2015) as well as tumors harboring mutations in SWI/SNF components (Wilson et al. 2010; Knutson et al. 2013; Bitler et al. 2015) might benefit from these molecules. However, the long-term impact of such inhibition should be carefully examined in light of transcriptional instability and irreversible changes resulting from PRC2 disruption.

Finally, our analysis prompts a careful examination of the contribution to tumorigenesis of genes whose expression is linked to proliferation. We propose that applying a similar analysis to other proliferation-associated genes, including key players of epigenetic modifications, could help clarify their contribution to cancer.

Materials and methods

Plasmids

The MSCVhygro-Flag-Ezh2 retroviral vector was obtained from Addgene (24926). MSCVhygro-Flag-Ezh1 was generated by subcloning. Following retroviral infection, cells were selected with 400 µg/mL Hygromycin B (Life Technologies). pLKO.1-shEzh2 was purchased from Dharmacon (clone ID TRCN0000039042; antisense sequence TTTCTTTCAGTTCTTCTGCGG). Oligonucleotides corresponding to a scramble shRNA (antisense sequence CGAGGGCGACTTAACCTTAGG) were cloned into the pLKO.1 vector. Following lentiviral infection, cells were selected with 2 µg/µL puromycin (Life Technologies).

Cell lines

iMEF cells were grown in DMEM supplemented with 10% FCS, 100 mM nonessential amino acids, 1 mM L-glutamine. *Ezh2*$^{flox/flox}$;*ROSA26-CreERT2* or *Ezh2*$^{flox/\Delta}$;*ROSA26-CreERT2* MEF cells were isolated from 13.5-d-old embryos and subsequently infected with the following retroviral constructs: pMXs-hc-MYC (Addgene, 17220) to generate c-Myc iMEFs, pBABE-hygro p53 DD (Addgene, 9058) to generate p53-DN iMEFs, and Ndy1-MigR1 (kindly provided by Philip N. Tsichlis) to generate Ndy1 iMEFs. A clone was obtained by limiting dilution of a pool of c-Myc *Ezh2*$^{flox/\Delta}$;*ROSA26-CreERT2* iMEFs. For conditional deletion of *Ezh2*, cells were treated with 4-hydroxytamoxifen (Sigma) at a final concentration ranging from 1 nM to 1 µM. For Ezh1 and Ezh2 rescue experiments, cells were infected with MSCVhygro-Flag-Ezh1 or MSCVhygro-Flag-Ezh2 ecotropic retroviruses.

The Myc-CaP mouse prostate cancer cell line was generously provided by Charles L. Sawyers, and cells were grown in DMEM supplemented with 10% FCS, 100 mM nonessential amino acids, and 1 mM L-glutamine. To obtain optimal growth conditions, the growth medium was supplemented with a growth factor cocktail composed of 25 µg/mL bovine pituitary extract (Life Technologies), 5 µg/mL bovine insulin (Sigma), and 6 ng/mL recombinant human epidermal growth factor (Sigma).

Cell growth assay

Twenty-thousand cells were plated in six-well dishes in triplicates and counted every 24 h over 4 d using a Vi cells counter (Beckman-coulter).

Mice

All animals used in the studies were handled with care, and experiments were done according to the guidelines from French legislation (Pten and N1ic models) or U.S. legislation (Hi-Myc model) and institutional policies. For the prostate cancer study, *PB4-Cre* (kindly provided by M. Chen), *Pten*$^{flox/+}$ (kindly provided by O. Lantz), *Hi-Myc* (Jackson laboratory), and *Ezh2*$^{flox/+}$ (generously provided by A. Tarakhovsky) mice were used. For the breast cancer study, *MMTV-Cre*, *Rosa-N1ic*, and *Ezh2*$^{flox/+}$ mice were used.

For xenografting experiments, 6-wk-old female CB17-SCID mice were purchased from Janvier Laboratories. MDA-MB-231 cells (1×10^6) were orthotopically injected into the mammary fat pads of 20 mice. Control cells (one of three alleles of *EZH2* mutated) were grafted on the left side of each mouse, and EZH2$^-$ cells were grafted on the right side of each mouse. Mice were sacrificed 10 wk after injection, and the resulting tumors were measured and collected.

Constitutive knockout of EZH2 *in the MDA-MB-231 cell line*

Mutation of all three alleles of *EZH2* in the MDA-MB-231 cell line was performed using CRISPR/CAS9 technology. A donor template encoding a puromycin selection cassette was nucleofected at a 1:1:1 ratio with CAS9 (Addgene, 41815) and *EZH2*-specific guide RNA (gRNA; build from gRNA cloning vector; Addgene, 41824). Homology-directed repair of the double-strand break resulted in the insertion of an FRT-flanked puromycin resistance gene in-frame with the first exon of EZH2. The first allele of *EZH2* was targeted using the following guide sequence: GTATACCTAATTCCTGTAAT. Sequencing of the remaining alleles revealed a single nucleotide insertion at the target site. Consequently, after flippase-mediated excision of the puromycin resistance cassette from the first allele, the remaining alleles were targeted using the following guide sequence: GTATACC TAATTCCTGTTAA. The resulting clone was thus used as a constitutive *EZH2* knockout, using the parental clone (one allele targeted) as a control in all experiments.

Inhibition of EZH1 and EZH2

Inhibition of EZH1 and EZH2 was achieved by a treatment with 1 µM UNC1999 (Sigma).

Formation of the spheroids and type I collagen invasion assay

Multicellular spheroids of MDA-MB-231 cells were prepared using the hanging droplet method (Kelm et al. 2003), with 3×10^3 cells in 20-µL droplets in complete L15 medium + 1% volume of collagen I for 3 d. Next, spheroids were embedded in 2.2 mg/mL type I collagen gel (T0) prepared from acid extract of rat tail tendon (from BD Biosciences) and incubated for 2 d (T2).

Imaging and quantification of the area of invasion

Samples were fixed at T0 and T2 and costained with fluorescent phalloidin to label F-actin and DAPI. Images were taken with a confocal LSM 510 (Zeiss) microscope with a 5× dry objective, collecting stacks of images along the *Z*-axis with 10-µm intervals between optical sections.

Quantification of invasion was done with ImageJ software (http:// rsb.info.nih.gov/ij) by estimating the diameter of spheroids at T0 and T2 as described (Rey et al. 2011). These values were averaged and used to calculate the mean invasion area (πr^2). The mean invasion area at T2 was normalized to the mean invasion area at T0.

Antibodies

Antibodies against Ezh1, Ezh2, Eed, Suz12, and H3K27me2/3 (Western blot and ChIP [chromatin immunoprecipitation]/ ChIP-seq [ChIP combined with deep sequencing]) were previously described (Margueron et al. 2008); total H3 (39163) and H3K27me3 (39155) for ChIP-seq were purchased from Active Motif; Lamin B1 (ab16048) was purchased from Abcam; Ezh2 (NCL-L-EZH2) for IHC on PDXs was purchased from Novocastra; H3K27me3 (C36B11) for IHC on PDXs was purchased from Cell Signaling; Flag M2 was purchased from Sigma (F1804); Nkx3.1 antibody was a generous gift from Dr C. Abate-Shen; SMA antibody was purchased from Dako; PCNA, AR, and Sirt1 antibodies were purchased from Santa Cruz Biotechnology; and tubulin antibody was purchased from Sigma.

Nuclear extracts

For nuclear extract preparation, cells were incubated with buffer A (10 mM Hepes at pH 7.9, 2.5 mM $MgCl_2$, 0.25 M sucrose, 0.1% NP40, 0.5 mM DTT, 1 mM PSMF) for 10 min on ice, centrifuged at 8000 rpm for 10 min, resuspended in buffer B (25 mM Hepes at pH 7.9, 1.5 mM $MgCl_2$, 700 mM NaCl, 0.5 mM DTT, 0.1 mM EDTA, 20% glycerol), sonicated, and centrifuged at 14,000 rpm for 15 min.

Western blot quantification

Image acquisition of Western blots was performed on a LAS 4000 imager (Leica), and signal intensity was measured using ImageJ software.

Tissue extracts

Protein extracts from tissues were prepared as previously described (Margueron et al. 2008).

RT-qPCR

Total RNA was isolated using the RNeasy minikit (Qiagen). cDNA was synthetized using the High-Capacity cDNA RT kit (Applied Biosystems, 4368814), and qPCR was performed with technical triplicate using SYBR Green reagent (Roche) on ViiA7 equipment (Applied Biosystems). At least three independent biological experiments were performed for each assay, and negative control RTs were always included. Primers sequences are in Supplemental Table S3.

IHC

IHC on mouse tissue was performed as previously described (Margueron et al. 2008).

IHC analysis on PDXs was performed on tissue microarrays obtained from treated xenografts as described in Cottu et al. (2014). Samples were dewaxed, and antigen retrieval was performed for 20 min in pH 6 citrate buffer. Developing was performed with the "Bond refine detection" kit (Leica biosystems); samples were incubated with diaminobenzidine for 7 min followed by counterstaining with hematoxylin for 4 min.

IHC quantification

IHC images were first processed with the ImageJ Colour Deconvolution plug-in (http://www.mecourse.com/landinig/software/cdeconv/cdeconv.html) in order to separate HE and DAB signals. For each tissue microarray (TMA), signal intensity of DAB signal over HE signal was then calculated. Alternatively, when the level of background was too high (i.e., for EZH2, showing areas of nonspecific staining), staining intensity was scored on a scale of 0 to 3 in a blind fashion. Importantly, both software-assisted and visual scoring yielded highly correlated results.

ChIP

ChIPs were performed as described previously (Margueron et al. 2008). Cell confluence and amount of starting material were kept constant by plating a defined number of cells the day before cross-linking. Quantification was done as described for the RT-qPCR. Primers sequences are in Supplemental Table S3.

ChIP-seq

ChIP was performed as described above, starting from 25 µg of chromatin; magnetic Dynabeads coupled to Protein A were used for the immunoprecipitation (Invitrogen). Incubation with micrococcal nuclease was performed to obtain mostly mononucleosome-sized fragments (150 base pairs [bp]). Libraries were prepared according to the manufacturer's instructions (TruSeq ChIP sample preparation kit, Illumina). High-throughput sequencing was performed on a SOLiD 5500 and an Illumina Hi-Seq 2500. Single-end 75-bp (SOLiD) or 100-bp (Illumina) reads were mapped on the mouse reference genome (mm9) using Bowtie 2 (version 2.1.0) (Langmead and Salzberg 2012), allowing one mismatch in the seed (22 bp) and reporting one location in case of multiple mapping hits. PCR duplicates were then removed using PicardTools (version 1.65; http://picard.sourceforge.net).

ChIP-seq analysis

Chip-seq data for H3K27me3 were generated in duplicates by the next-generation sequencing (NGS) platform at the Institut Curie (T. Rio Frio). From ~50 million SOLiD 5500 75-bp reads per sample, 50% were uniquely mapped (minimum mapQ = 10) onto the mouse reference genome (mm9) using Bowtie (parameters: -S -C -p 8 -q -y –col-keepends -l 28 -n 2 -e 70 -a –best –strata -m 1). From 50 million Illumina HiSeq 2500 100-bp reads per sample, ~98% were uniquely mapped onto the mouse reference genome (mm9) using Bowtie 2, allowing one mismatch in the seed (22 bp long). Given the high proportion of mapped read duplicates observed (until 22%), the duplicates were removed using the rmdup function of SAMTools. The mapped read data were then normalized to the total number of reads (25 million) by performing a down-sampling with Picard, and a coefficient factor was applied in some conditions ($Ezh2^{-/-}$: 0.05; Ezh1 rescue: 0.25) according to the amount of immunoprecipitated DNA obtained during the ChIP preparation. Peak calling for each sample was done with MACS 2 (2.0.10) with default parameters for histone analysis by specifying a fragment size of 150 bp and the input DNA as control. From the significant peak regions, the differential binding analysis was performed with the R package DiffBind. Only reproducible peaks between replicates were kept for read counting. Significant differential peaks were identified with a maximum false discovery rate (FDR) of 1% and a minimum value of fold change equal to 2 for each comparison. Peaks were quantified using HOMER.

Transcriptome data analysis

Microarray data were generated in duplicates by the microarray core facility of the Institut Curie (D. Gentien) using Affymetrix Mouse Gene 1.1 ST arrays (targeting 21041 genes). Raw data were normalized with the Robust Multiarray Average (RMA) method available in the Bioconductor R package oligo and the "pd.mogene.1.1.st.v1" annotation package. For rescue experiments, given the observed batch bias between the first and second replicates, the data were then batch-corrected using a linear model (Limma R package). Differential gene expression analysis was done using the RankProduct R package, and significantly underexpressed or overexpressed genes were identified with a minimum adjusted *P*-value of 15% and a minimum value of fold change equal to 2. Hierarchical clustering analysis was performed using Cluster 3.0, and heat maps were generated with TreeView.

Gene expression, copy number, and survival analysis of breast primary tumors

This study makes use of data generated by METABRIC (first described in Curtis et al. (2012). Funding for the project was provided by Cancer Research UK and the British Columbia Cancer Agency Branch. Upon access request, single-nucleotide

polymorphism (SNP) 6.0 copy number and Illumina HT-12 expression data for nearly 2000 primary breast tumors were available through the European Genome–Phenome Archive (http://www.ebi.ac.uk/ega) under accession number EGAS00000000083.

For survival analyses, disease-specific survival was used as the end point. Follow-up time was defined as time from diagnosis until death from breast cancer or time of last follow-up if the patient was not known to have died. Kaplan-Meier and ROC analyses were performed using Prism 6. Statistical significance was evaluated with the log-rank test. The hazard ratio of the highest to the lowest survival rate was calculated using the log-rank method. For survival analysis on all genes located on chromosome 7, the univariate Cox model was applied to each gene by comparing the outcome linked to tumors with genetic loss with that of the remaining samples. A FDR was used to correct for multiple testing.

Targeted sequencing of PRC2 genes in breast cancer metastases

The coding sequences of the EED, EZH2, and SUZ12 genes were analyzed using a targeted NGS approach. Experiments were performed on the NGS platform of the Cochin Hospital, Paris (Assistance Publique, Hopitaux de Paris, France). Briefly, the custom primer panel targeting the three genes (coding exons and IVS boundaries) was designed using AmpliSeq Designer (Life Technologies). For NGS library preparation, the Ion AmpliSeq 2.0 library kit was used according to the manufacturer's instructions. Amplified libraries were purified using Agencourt AMPure XP beads (Beckman Coulter). Prior to library pooling and sequencing sample preparation, amplified libraries were quantified using the Qubit fluorometer system (Agilent Technologies). Emulsion PCR and enrichment were performed on the Ion OneTouch and Ion OneTouch ES instruments with the Ion PGM template OT2 400 and Ion PGM sequencing 400 kits (Life Technologies). The template-positive ion sphere particles were loaded on Ion 318 chips and sequenced with an Ion PGM system (Life Technologies). Sequence alignment and extraction of SNPs and short insertions/deletions were performed using the Variant Caller plug-in on Ion Torrent suite version 4.4 and Ion Reporter version 4.4 (Life Technologies). DNA sequences were visualized using the Integrated Genomics Viewer (version 2.3.3) from the Broad Institute.

RNA FISH

RNA FISH was performed as described elsewhere (Chaumeil et al. 2008). The following BAC probes (CHORI) were used: RP23-333D4 (*Cpa6*) and RP23-40H14 (*Tspan7*).

Single-cell RT-qPCR analysis

To discriminate OHT and vehicle-treated Ezh2-conditional iMEFs, DiI (vehicle-treated) or DiO (OHT-treated) was added in growth medium for 4 h. Cells were then trypsinized and counted on a Vi cells counter (Beckman-Coulter). The average diameter of iMEFs was 13 µm. After mixing OHT- and vehicle-treated cells in equal proportions, 250,000 cells per milliliter were mixed at a 3:2 ratio in C1 cell suspension reagent (Fluidigm) before being loaded on a primed C1 Single-Cell Auto Prep Integrated Fluidic Circuit (Fluidigm). Cells were then visualized under an inverted fluorescent microscope (Leica) to assess viability and assignment of red (OHT-treated) and green (vehicle-treated) cells. Lysis, RT, and preamplifications were performed according to the manufacturer's protocol using Ambion Single Cell-to-CT kit (LifeTechnologies). Preamplified cDNA was analyzed by high-throughput qPCR on a Biomark-HD system (Fluidigm). The complete list of primers is in Supplemental Table S4. A qPCR primer pair designed on the region of the set domain that is deleted upon OHT treatment served as an independent genotype assignment and was found to closely match the color assignment. Only single cells with matching genotype/color assignments were considered for analysis.

Data access

All ChIP-seq data sets have been deposited in the Gene Expression Omnibus repository (GSE59427).

Acknowledgments

We thank all members of the Margueron laboratory for helpful discussions, and Marc-Henry Stern, Pascale Gilardi, Elphège Nora, and Edith Heard for critical reading of the manuscript. The ATIP-Avenir program, the Fondation pour la Recherche Médicale (FRM), the European Research Council (ERC-Stg, REPODDID), the Labex DEEP, and the Institut Curie supported work in R.M.'s laboratory. M.W. was a recipient of a post-doctoral fellowship from the Association pour la Recherche contre le Cancer (ARC). High-throughput sequencing was performed by the NGS platform of the Institut Curie, supported by grants ANR-10-EQPX-03 and ANR10-INBS-09-08 from the Agence Nationale de le Recherche (Investissements d'Avenir) and by the Cancéropôle Ile-de-France. High-throughput qPCR was carried out on the qPCR-HD-Genomic Paris Centre platform, supported by grants from the Region Ile-de-France. We acknowledge the Institut Curie imagery facility for microscopy, the Institut Curie animal facility for mouse care, and the Institut Curie genomic facility for gene expression profiles.

References

Bachmann IM, Halvorsen OJ, Collett K, Stefansson IM, Straume O, Haukaas SA, Salvesen HB, Otte AP, Akslen LA. 2006. EZH2 expression is associated with high proliferation rate and aggressive tumor subgroups in cutaneous melanoma and cancers of the endometrium, prostate, and breast. *J Clin Oncol* **24:** 268–273.

Bae WK, Yoo KH, Lee JS, Kim Y, Chung IJ, Park MH, Yoon JH, Furth PA, Hennighausen L. 2015. The methyltransferase EZH2 is not required for mammary cancer development, although high EZH2 and low H3K27me3 correlate with poor prognosis of ER-positive breast cancers. *Mol Carcinog* **54:** 1172–1180.

Benetatos L, Vartholomatos G, Hatzimichael E. 2014. Polycomb group proteins and MYC: the cancer connection. *Cell Mol Life Sci* **71:** 257–269.

Bitler BG, Aird KM, Garipov A, Li H, Amatangelo M, Kossenkov AV, Schultz DC, Liu Q, Shih IeM, Conejo-Garcia JR, et al. 2015. Synthetic lethality by targeting EZH2 methyltransferase activity in ARID1A-mutated cancers. *Nat Med* **21:** 231–238.

Bolos V, Mira E, Martinez-Poveda B, Luxan G, Canamero M, Martinez AC, Manes S, de la Pompa JL. 2013. Notch activation stimulates migration of breast cancer cells and promotes tumor growth. *Breast Cancer Res* **15:** R54.

Bracken AP, Helin K. 2009. Polycomb group proteins: navigators of lineage pathways led astray in cancer. *Nat Rev Cancer* **9:** 773–784.

Bracken AP, Pasini D, Capra M, Prosperini E, Colli E, Helin K. 2003. EZH2 is downstream of the pRB–E2F pathway, essential for proliferation and amplified in cancer. *EMBO J* **22:** 5323–5335.

Bracken AP, Kleine-Kohlbrecher D, Dietrich N, Pasini D, Gargiulo G, Beekman C, Theilgaard-Monch K, Minucci S, Porse BT, Marine JC, et al. 2007. The Polycomb group proteins bind throughout the INK4A-ARF locus and are disassociated in senescent cells. *Genes Dev* **21:** 525–530.

Campbell JE, Kuntz KW, Knutson SK, Warholic NM, Keilhack H, Wigle TJ, Raimondi A, Klaus CR, Rioux N, Yokoi A, et al. 2015. EPZ011989, a potent, orally-available EZH2 inhibitor with robust in vivo activity. *ACS Med Chem Lett* **6:** 491–495.

Chase A, Cross NC. 2011. Aberrations of EZH2 in cancer. *Clin Cancer Res* **17:** 2613–2618.

Chaumeil J, Augui S, Chow JC, Heard E. 2008. Combined immunofluorescence, RNA fluorescent in situ hybridization, and DNA fluorescent in situ hybridization to study chromatin changes, transcriptional activity, nuclear organization, and X-chromosome inactivation. *Methods Mol Biol* **463:** 297–308.

Cottu P, Bieche I, Assayag F, El Botty R, Chateau-Joubert S, Thuleau A, Bagarre T, Albaud B, Rapinat A, Gentien D, et al. 2014. Acquired resistance to endocrine treatments is associated with tumor-specific molecular changes in patient-derived luminal breast cancer xenografts. *Clin Cancer Res* **20:** 4314–4325.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486:** 346–352.

De Raedt T, Beert E, Pasmant E, Luscan A, Brems H, Ortonne N, Helin K, Hornick JL, Mautner V, Kehrer-Sawatzki H, et al. 2014. PRC2 loss amplifies Ras-driven transcription and confers sensitivity to BRD4-based therapies. *Nature* **514:** 247–251.

de Vries NA, Hulsman D, Akhtar W, de Jong J, Miles DC, Blom M, van Tellingen O, Jonkers J, van Lohuizen M. 2015. Prolonged Ezh2 depletion in glioblastoma causes a robust switch in cell fate resulting in tumor progression. *Cell Rep* **10:** 383–397.

Ellwood-Yen K, Graeber TG, Wongvipat J, Iruela-Arispe ML, Zhang J, Matusik R, Thomas GV, Sawyers CL. 2003. Myc-driven murine prostate cancer shares molecular features with human prostate tumors. *Cancer Cell* **4:** 223–238.

Ezhkova E, Pasolli HA, Parker JS, Stokes N, Su IH, Hannon G, Tarakhovsky A, Fuchs E. 2009. Ezh2 orchestrates gene expression for the stepwise differentiation of tissue-specific stem cells. *Cell* **136:** 1122–1135.

Gonzalez ME, Li X, Toy K, DuPrie M, Ventura AC, Banerjee M, Ljungman M, Merajver SD, Kleer CG. 2009. Downregulation of EZH2 decreases growth of estrogen receptor-negative invasive breast carcinoma and requires BRCA1. *Oncogene* **28:** 843–853.

Hansen KH, Bracken AP, Pasini D, Dietrich N, Gehani SS, Monrad A, Rappsilber J, Lerdrup M, Helin K. 2008. A model for transmission of the H3K27me3 epigenetic mark. *Nat Cell Biol* **10:** 1291–1300.

Healey MA, Hu R, Beck AH, Collins LC, Schnitt SJ, Tamimi RM, Hazra A. 2014. Association of H3K9me3 and H3K27me3 repressive histone marks with breast cancer subtypes in the Nurses' Health Study. *Breast Cancer Res Treat* **147:** 639–651.

Holm K, Grabau D, Lovgren K, Aradottir S, Gruvberger-Saal S, Howlin J, Saal LH, Ethier SP, Bendahl PO, Stal O, et al. 2012. Global H3K27 trimethylation and EZH2 abundance in breast tumor subtypes. *Mol Oncol* **6:** 494–506.

Kelm JM, Timmins NE, Brown CJ, Fussenegger M, Nielsen LK. 2003. Method for generation of homogeneous multicellular tumor spheroids applicable to a wide variety of cell types. *Biotechnol Bioeng* **83:** 173–180.

Kleer CG, Cao Q, Varambally S, Shen R, Ota I, Tomlins SA, Ghosh D, Sewalt RG, Otte AP, Hayes DF, et al. 2003. EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc Natl Acad Sci* **100:** 11606–11611.

Knutson SK, Warholic NM, Wigle TJ, Klaus CR, Allain CJ, Raimondi A, Porter Scott M, Chesworth R, Moyer MP, Copeland RA, et al. 2013. Durable tumor regression in genetically altered malignant rhabdoid tumors by inhibition of methyltransferase EZH2. *Proc Natl Acad Sci* **110:** 7922–7927.

Koh CM, Iwata T, Zheng Q, Bethel C, Yegnasubramanian S, De Marzo AM. 2011. Myc enforces overexpression of EZH2 in early prostatic neoplasia via transcriptional and post-transcriptional mechanisms. *Oncotarget* **2:** 669–683.

Konze KD, Ma A, Li F, Barsyte-Lovejoy D, Parton T, Macnevin CJ, Liu F, Gao C, Huang XP, Kuznetsova E, et al. 2013. An orally bioavailable chemical probe of the lysine methyltransferases EZH2 and EZH1. *ACS Chem Biol* **8:** 1324–1334.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Lee ST, Li Z, Wu Z, Aau M, Guan P, Karuturi RK, Liou YC, Yu Q. 2011. Context-specific regulation of NF-κB target gene expression by EZH2 in breast cancers. *Mol cell* **43:** 798–810.

Le Tourneau C, Paoletti X, Servant N, Bieche I, Gentien D, Rio Frio T, Vincent-Salomon A, Servois V, Romejon J, Mariani O, et al. 2014. Randomised proof-of-concept phase II trial comparing targeted therapy based on tumour molecular profiling vs conventional therapy in patients with refractory cancer: results of the feasibility part of the SHIVA trial. *Br J Cancer* **111:** 17–24.

Lewis PW, Muller MM, Koletsky MS, Cordero F, Lin S, Banaszynski LA, Garcia BA, Muir TW, Becher OJ, Allis CD. 2013. Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma. *Science* **340:** 857–861.

Ma X, Ziel-van der Made AC, Autar B, van der Korput HA, Vermeij M, van Duijn P, Cleutjens KB, de Krijger R, Krimpenfort P, Berns A, et al. 2005. Targeted biallelic inactivation of Pten in the mouse prostate leads to prostate cancer accompanied by increased epithelial cell proliferation but not by reduced apoptosis. *Cancer Res* **65:** 5730–5739.

Maire V, Nemati F, Richardson M, Vincent-Salomon A, Tesson B, Rigaill G, Gravier E, Marty-Prouvost B, De Koning L, Lang G, et al. 2013. Polo-like kinase 1: a potential therapeutic option in combination with conventional chemotherapy for the management of patients with triple-negative breast cancer. *Cancer Res* **73:** 813–823.

Mallen-St Clair J, Soydaner-Azeloglu R, Lee KE, Taylor L, Livanos A, Pylayeva-Gupta Y, Miller G, Margueron R, Reinberg D, Bar-Sagi D. 2012. EZH2 couples pancreatic regeneration to neoplastic progression. *Genes Dev* **26:** 439–444.

Margueron R, Li G, Sarma K, Blais A, Zavadil J, Woodcock CL, Dynlacht BD, Reinberg D. 2008. Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. *Mol Cell* **32:** 503–518.

Min J, Zaslavsky A, Fedele G, McLaughlin SK, Reczek EE, De Raedt T, Guney I, Strochlic DE, Macconaill LE, Beroukhim R, et al. 2010. An oncogene–tumor suppressor cascade drives metastatic prostate cancer by coordinately activating Ras and nuclear factor-κB. *Nat Med* **16:** 286–294.

Morin RD, Johnson NA, Severson TM, Mungall AJ, An J, Goya R, Paul JE, Boyle M, Woolcock BW, Kuchenbauer F, et al. 2010. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat Genet* **42:** 181–185.

Murtaugh LC, Stanger BZ, Kwan KM, Melton DA. 2003. Notch signaling controls multiple steps of pancreatic differentiation. *Proc Natl Acad Sci* **100:** 14920–14925.

Nagalla S, Chou JW, Willingham MC, Ruiz J, Vaughn JP, Dubey P, Lash TL, Hamilton-Dutoit SJ, Bergh J, Sotiriou C, et al. 2013. Interactions between immunity, proliferation and molecular subtype in breast cancer prognosis. *Genome Biol* **14:** R34.

Nikoloski G, Langemeijer SM, Kuiper RP, Knops R, Massop M, Tonnissen ER, van der Heijden A, Scheele TN, Vandenberghe P, de Witte T, et al. 2010. Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. *Nat Genet* **42:** 665–667.

Ntziachristos P, Tsirigos A, Van Vlierberghe P, Nedjic J, Trimarchi T, Flaherty MS, Ferres-Marco D, da Ros V, Tang Z, Siegle J, et al. 2012. Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. *Nat Med* **18:** 298–301.

Rey M, Irondelle M, Waharte F, Lizarraga F, Chavrier P. 2011. HDAC6 is required for invadopodia activity and invasion by breast tumor cells. *Eur J Cell Biol* **90:** 128–135.

Seward S, Semaan A, Qazi AM, Gruzdyn OV, Chamala S, Bryant CC, Kumar S, Cameron D, Sethi S, Ali-Fehmi R, et al. 2013. EZH2 blockade by RNA interference inhibits growth of ovarian cancer by facilitating re-expression of p21(waf1/cip1) and by inhibiting mutant p53. *Cancer Lett* **336:** 53–60.

Simon JA, Kingston RE. 2013. Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. *Mol Cell* **49:** 808–824.

Sneeringer CJ, Scott MP, Kuntz KW, Knutson SK, Pollock RM, Richon VM, Copeland RA. 2010. Coordinated activities of wild-type plus mutant EZH2 drive tumor-associated hypertrimethylation of lysine 27 on histone H3 (H3K27) in human B-cell lymphomas. *Proc Natl Acad Sci* **107:** 20980–20985.

Stylianou S, Clarke RB, Brennan K. 2006. Aberrant activation of notch signaling in human breast cancer. *Cancer Res* **66:** 1517–1525.

Su IH, Basavaraj A, Krutchinsky AN, Hobert O, Ullrich A, Chait BT, Tarakhovsky A. 2003. Ezh2 controls B cell development through histone H3 methylation and Igh rearrangement. *Nat Immunol* **4:** 124–131.

Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, et al. 2010. Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18:** 11–22.

Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, et al. 2007. Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* **39:** 41–51.

Vanharanta S, Shu W, Brenet F, Hakimi AA, Heguy A, Viale A, Reuter VE, Hsieh JJ, Scandura JM, Massague J. 2013. Epigenetic expansion of VHL–HIF signal output drives multiorgan metastasis in renal cancer. *Nat Med* **19:** 50–56.

Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG, Ghosh D, Pienta KJ, Sewalt RG, Otte AP, et al. 2002. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419:** 624–629.

Varambally S, Cao Q, Mani RS, Shankar S, Wang X, Ateeq B, Laxman B, Cao X, Jing X, Ramnarayanan K, et al. 2008. Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. *Science* **322:** 1695–1699.

Venet D, Dumont JE, Detours V. 2011. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* **7:** e1002240.

Wang S, Gao J, Lei Q, Rozengurt N, Pritchard C, Jiao J, Thomas GV, Li G, Roy-Burman P, Nelson PS, et al. 2003. Prostate-specific deletion of the murine Pten tumor suppressor gene leads to metastatic prostate cancer. *Cancer Cell* **4:** 209–221.

Watson PA, Ellwood-Yen K, King JC, Wongvipat J, Lebeau MM, Sawyers CL. 2005. Context-dependent hormone-refractory progression revealed through characterization of a novel murine prostate cancer cell line. *Cancer Res* **65:** 11565–11571.

Wei Y, Xia W, Zhang Z, Liu J, Wang H, Adsay NV, Albarracin C, Yu D, Abbruzzese JL, Mills GB, et al. 2008. Loss of trimethylation at lysine 27 of histone H3 is a predictor of poor outcome in breast, ovarian, and pancreatic cancers. *Mol Carcinog* **47:** 701–706.

Wilson BG, Wang X, Shen X, McKenna ES, Lemieux ME, Cho YJ, Koellhoffer EC, Pomeroy SL, Orkin SH, Roberts CW. 2010. Epigenetic antagonism between polycomb and SWI/SNF complexes during oncogenic transformation. *Cancer Cell* **18:** 316–328.

Woodhouse S, Pugazhendhi D, Brien P, Pell JM. 2013. Ezh2 maintains a key phase of muscle satellite cell expansion but does not regulate terminal differentiation. *J Cell Sci* **126:** 565–579.

Wu X, Wu J, Huang J, Powell WC, Zhang J, Matusik RJ, Sangiorgi FO, Maxson RE, Sucov HM, Roy-Burman P. 2001. Generation of a prostate epithelial cell-specific Cre transgenic mouse model for tissue-specific gene ablation. *Mech Dev* **101:** 61–69.

Xu K, Wu ZJ, Groner AC, He HH, Cai C, Lis RT, Wu X, Stack EC, Loda M, Liu T, et al. 2012. EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* **338:** 1465–1469.

You JS, Jones PA. 2012. Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* **22:** 9–20.

7

Discussion and Conclusion

Different contribution to understand triple-negative breast cancer drug response and computational biology were presented here. We will discuss in the following chapter our three main contributions.

7.1 Large scale pharmacogenomic studies

Our first contribution was to develop analytical approaches to exploit large pharmacogenenomic dataset. We introduced a new cell line classification based on transcriptomic profiles of cell lines, according to a biological network-driven gene selection process. The identified clusters appeared robust in two large-scale cell line panels and we also demonstrated that our clustering is more relevant than tissue of origin when the drug sensitivity is considered. We further robustly associated cell line cluster with pharmacological data in four different dataset, independently of the drug sensitivity measure used.

The findings of Haibe-Kains et al. study on variation in drug sensitivity data should not be minimized. The lack of standardization in the biological assays is a major challenge for two main reasons. First, from the point of view of bioinformatics and biostatistics. The main goal of these studies is to develop predictive models that could lead to therapeutic strategies according to the genomic profile of the patient's tumor. The poor correlation of these studies is a major obstacle to the identification or validation of predictive markers of the response to treatments. Important works are urgently needed to standardize the biological assays, otherwise it would be unwise to reliably identify genomic drug response predictors. While our study is encouraging, these are limitations that the bioinformatic and biostatistics can not overcome alone. From the biological

point of view, one can ask how reliable the results of these studies are. Preclinical models are essential for finding and developing new drugs, but it is important to keep in mind the strengths and weaknesses of each model. One might consider that cell-line pharmacological data should only be used for hypothesis generation and for elaborating on existing hypotheses. If we wish to continue this kind of predictive pharmacogenomic study, it is essential to carry out important standardization work of pharmacological assays among researchers. The identification of an effective drug in cell lines, which is validated in mice for use in humans, remains very challenging. Many studies have proposed to explain this high attrition rate by the difference in complexity and representativeness of each of the model. However, if major differences are found in a model as simple and robust as cell line, it is not surprising to have discrepancies with more complex models.

The published results of Haibe-Kains et al. have caused numerous reactions in the biomedical community. The authors of CCLE and GDSC proposed a novel analysis of their data and reported a "reasonable" agreement between their respective studies and the predictors that can be derived (172). However, Safikhani et al. (from Haibe-Kains' lab) (154) have demonstrated fundamental differences with their initial study (72) . One of their main criticisms concerns the analytical design to identify the predictors of drug sensitivity found with each dataset. Stransky at al. have used an ElasticNet using only the CCLE genomic data across the two datasets that were trained with different measures of drug sensitivity (IC_{50} and AUC, see Figure 1 in (154)). These analytical choices, which have not been explained, may contribute to an artificial concordance of biomarkers between the two studies. In our work, we used the molecular data and drug sensitivity data specific of each dataset. We demonstrated that our cell line clustering is able to find significant and meaningful associations with drugs efficacy in four different datasets, using three different drug sensitivity measures. The goal of our study was not to identify molecular predictors of drug sensitivity. Nevertheless, the identification of sensitive or resistant populations constitutes a meaningful insights for pharmacogenomics studies.

Pharmacogenomics studies are likely to remain a crucial preclinical source for generating hypotheses about the mechanism of drugs and their potential molecular targets. We have demonstrated that groups of cell lines from different tissue of origin may have a common drug response if they share similar transcriptomic profiles. This classification seems particularly interesting to identify the mechanism of drugs. Therapeutic agents frequently act by unknown mechanisms or have hidden phenotypes that result from un-

expected activities. Such unexpected activities can be harmful, leading to toxicity or beneficial effects, suggesting new therapeutic indications. Off-targets are challenging and often linked to toxicity, which is responsible for the high failure rate of the drug development. Methods to identify these unexpected effects are needed. By comparing cell line profiles within and between clusters, pathways contributing to drug activity can be obtained and hidden phenotypes can be studied. These strategies can help to better understand the mechanism of drug action and improve their development.

Finally, one might wonder if it was worth the effort to devote so much time and money to these two studies ? I do not have the answer but it raises several issues. First, we do not know which pharmacological assay represents the 'truth' ? The probable answer is either both or neither. Each biological assay measures a different reaction and it is likely that some are more sensitive to the reaction induced by a specific drug. However, these data have provided a huge amount of information to the scientific community that should not be minimized. The complementary study of Haibe-Kains et al. raised the important issue of the publication of negative results. To quote Matosin et al. (116) : 'negative findings are a valuable component of the scientific literature because they force us to critically evaluate and validate our current thinking, and fundamentally move us towards unabridged science". There has been a lot of enthusiasm at the time of publication of the CCLE and GDSC that has lagged one year later with the publication of Haibe-Kains et al. (207, 208). Although it is necessary to encourage this kind of initiative, which are the CCLE and the GDSC, we must not lose our critical thinking about them. We can see in this story that it is important to have several datasets emerging from different sources. Diversity is necessary to validate our results, which forces us to question the degree of confidence in the results obtained by other large-scale omic databases.

7.2 Challenges in our understanding of neoadjuvant-resistant triple-negative breast cancer

These last years have been marked by an unprecedented increase in our understanding of the molecular landscape of TNBCs. Similarly to the breast cancer which has long been considered a unique disease, TNBCs today appear to be composed of many different pathological entities. The studies of Lehmann et al. (100) and Burstein et al. (31) dedicated to the study of TNBCs have identified distinct molecular subtypes based on several hundred tumors. Although the greater molecular heterogeneity of TNBCs limits the reproductibility between different classification methods, four main subtypes were

recurrently found: (i) androgen receptor (AR), (ii) mesenchymal (MES), (iii) basal-like (BLS) and (iv) immunomodulatory (IM). The presence of the AR subtype suggests that the agents targeting AR could benefit a subgroup of TNBC patients. Likewise, immune-based therapies (e.g., PD1 antibodies) may be useful treatments for IM tumors. However, no reliable predictive marker has yet been identified. Lehmann et al. have developed a website (http://cbc.mc.vanderbilt.edu/tnbc/) for the classification of TNBC samples on the basis of their gene expression profiles. The authors used a very large number of genes (2,188 genes) to establish their molecular signature, which can be a source of instability, because of the noise potentially introduced. In addition, their tool begins by normalizing the gene expression of the studied cohort. Many discrepancies can then be observed for the same tumors depending on the set of samples used as input. Several potentially therapeutically important TNBC subgroups are emerging but further analysis to defined robust markers and the evaluation of their clinical relevance is still needed.

Most of the biological changes occurring under NAC are largely unknown. So far, we do not know whether the resistance of TNBCs to the NAC is innate or acquired: for this, we will need large cohorts of core biopsies with their residual disease to follow the molecular evolution during treatment. If resistance is innate, studies comparing sensitive and resistant primary tumors are well suited. If resistance is acquired, the study of primary tumors and their residual diseases should be favored. In both cases, larger cohorts of tumors are essential. First, evaluate the proportion of tumors with innate and acquired resistance, and secondly, clearly define their heterogeneity. All studies agree that resistant tumors appear to have a unique molecular profile. If several hundred patients were needed to evaluate the heterogeneity of all TNBCs, it is likely that at least as many patients will be needed to define the different TNBC subtypes resistant to NAC.

Whole exome sequencing may not be well suited for this type of study. By studying only about 1% of the genome, it is likely that we underestimate the subclonal diversity present in tumors. In addition, breast cancer and TNBC are known to have more chromosomal rearrangements than mutational events. Detection of subclonal copy number events could reveal more meaningful information. Previous work has sought to investigate the resistance of TNBC to NAC with gene expression and genomic data. The study of noncoding element, the epigenome or the proteome could make important contributions on the evolution and the heterogeneity of the tumors, unidentified until then.

7.3 Post hoc approaches to large-scale multiple testing

Our last main contribution provides a tool to integrate RNA-seq results of the differential alternative exon analysis with information of RBP motif binding sites. This work proposes to use an innovative multiple testing procedure to control the proportion of false discoveries. Post-hoc inference calculates the error rates that are not invalidated by multiple looks at the data, and the user is free to study the data in every possible way before finally deciding what rejections to make.

This kind of problem is frequently encountered in genomics. Indeed, the emergence of high-throughput technologies generates huge amount of data easily and quickly, leading scientist to a different way of thinking. We have moved from hypothesis-driven research to data-driven research. Hypothesis-driven (or confirmatory) research consists of formulating a hypothesis and then testing it to refute or not the hypothesis. It consists of a structured and rigorous data analysis. Data-driven research is explorative rather than confirmatory and therefore more open-minded. The findings are not meant to be reported as end results but help scientists formulate hypotheses to be followed up by independent validation experiments.
In addition, it is common to refine the results of a differential analysis using *prior knowledge*, once the list of candidates is established. Most of us are also interested by looking for differential expressed genes in a set of genes corresponding to a specific biological process before investigating other related pathways. Classical methods based on FWER or FDR have been developed for confirmatory analysis and their use is limited in exploratory research. These methen dépit ods assume that all data analysis decisions must be made before seeing the data, otherwise no formal risk assessment can generally be performed on the resulting set of markers.

In a context where the problem of poor reproducibility of the scientific research has been recognized by the statistical community (91, 158), it is important to promote dedicated methods that avoid reporting spurious findings as “significant”.

7.4 Conclusion

The studies presented in this thesis explored different aspects of cancer research and computational biology.

As a first step, we analyzed data from the Genomics of Drug Sensitivity in Cancer and Cancer Cell Line Encyclopedia studies. We propose a novel classification based on transcriptomic profiles of cell lines, according to a biological network-driven gene selection process. Our robust molecular classification displays greater homogeneity of drug sensitivity than cancer cell line grouped based on tissue of origin. We identified significant associations between cell line cluster and drug response robustly found between both datasets. We further demonstrate the relevance of our method using two additional external datasets and distinct sensitivity metrics. Some associations were still found robust, despite cell lines and drug responses' variations.

Our second contribution studies the resistance of triple negative breast cancer to neoadjuvant chemotherapy. Molecular analysis of 16 paired samples before and after treatment revealed large-scale heterogeneity in transcription, mutation, and copy number, with no frequently recurring mutations other than TP53. Consistent with this molecular heterogeneity, no systematic pathway was associated with the tumor before or after treatment. However, we observed numerous changes specific to each patient suggesting a clonal evolution and showed that several patients were eligible for targeted therapy.

The third main contribution presented here concerns motif enrichment analysis, through improvement of the rMAPS method. Our method identifies RBPs whose pattern is overrepresented in a given set of regulated sequences as well as with their precise binding site to guide biological validations. We have demonstrated that post hoc approaches are adapted to this type of problem, in order to control the proportion of false discoveries. We hope that this study will contribute to the diffusion of post hoc approaches in order to avoid erroneous results.

Appendices

Appendix A

New insight for pharmacogenetics studies from the transcriptional analysis of two large-scale cancer cell line panels

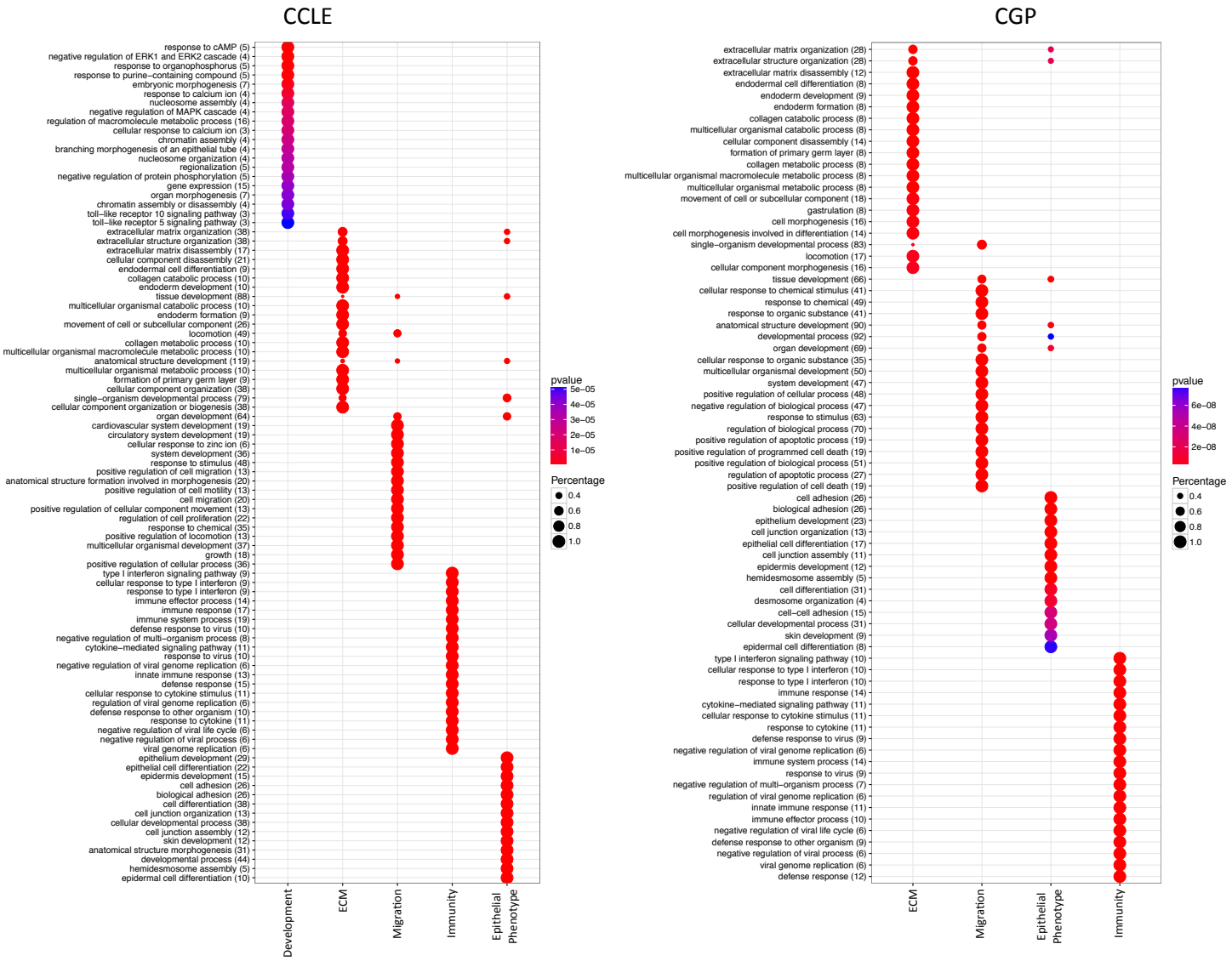
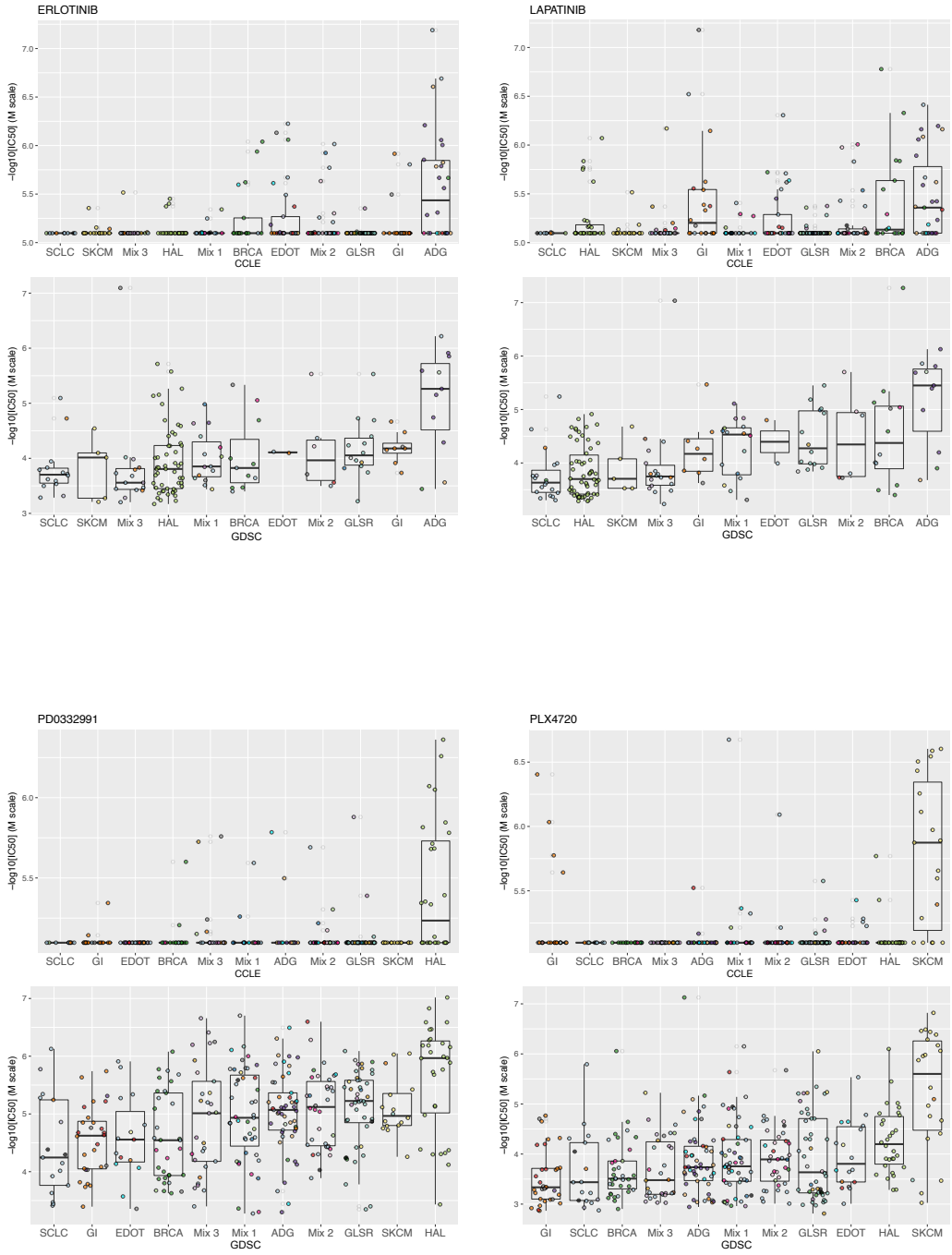


Figure A.1: Gene ontology analysis for gene clusters identified in CCLE and GDSC.

Figure A.2: Distribution of IC50 values for each cluster in CCLE and GDSC. Ordered according to mean IC50 for the cluster. From resistant (left) to sensitive (right).



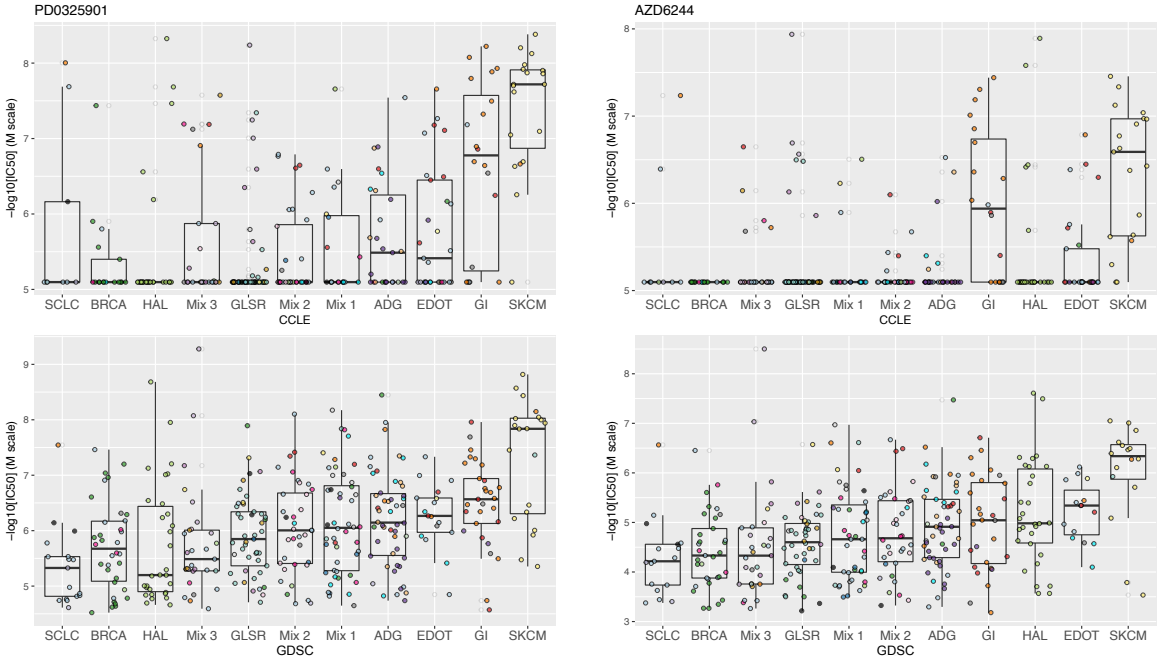

PD0325901
AZD6244
−log10[IC50] (M scale)
SCLC BRCA HAL Mix 3 GLSR Mix 2 Mix 1 ADG EDOT GI SKCM
CCLE
GDSC

Figure A.3: Barplot of the number of mutations in cell lines The eight genes in bold were used for cluster characterization (A) Proportion of cell lines mutated for each gene. Red in CCLE and blue in GDSC. (B) Proportion of same cell lines mutated in CCLE and GDSC. (C) Among the mutated cell lines in CCLE and GDSC, proportion of same cell lines mutated. (D) Among the mutated cell lines in CCLE and GDSC, proportion of cell lines with exactly the same mutation.

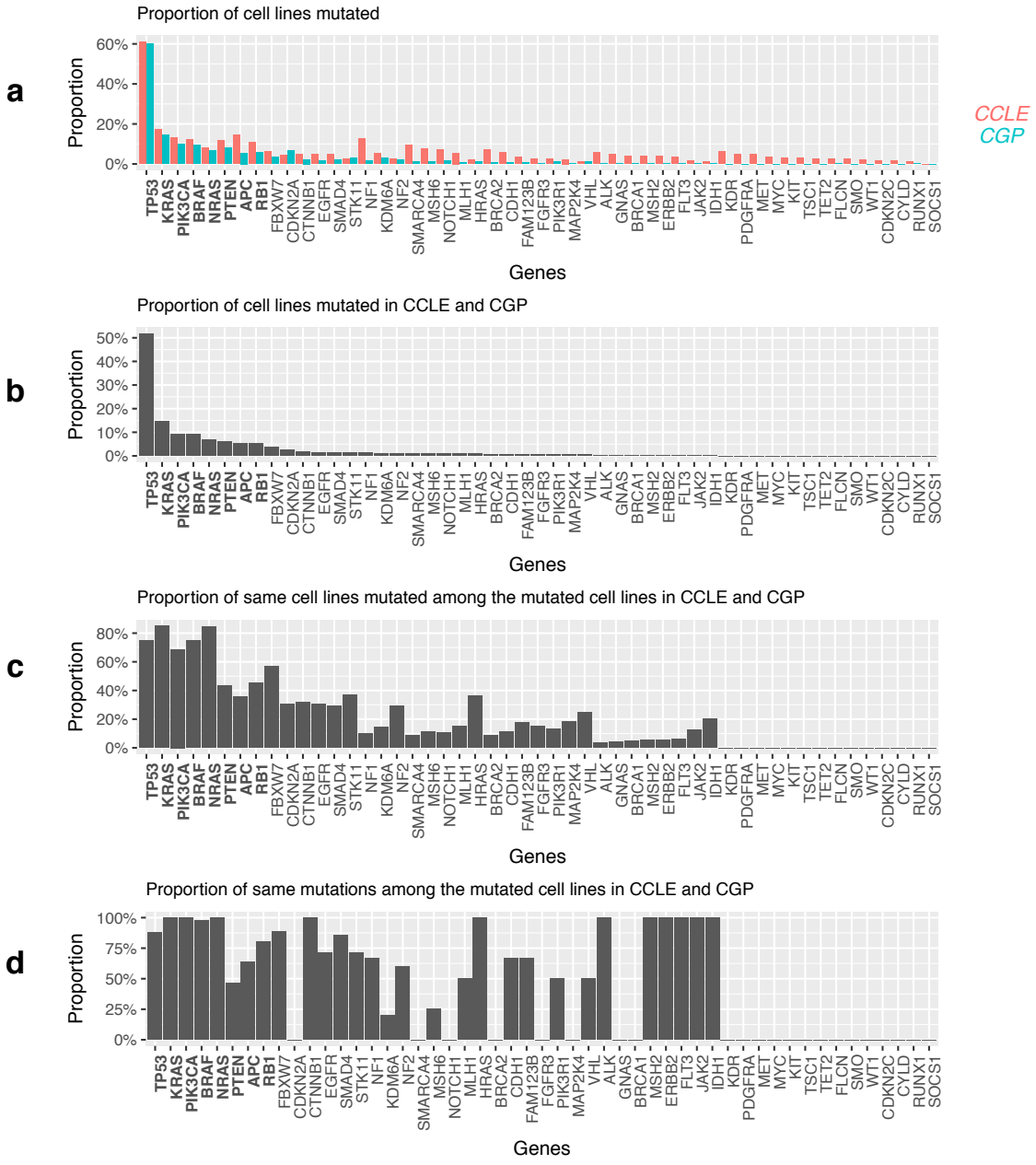
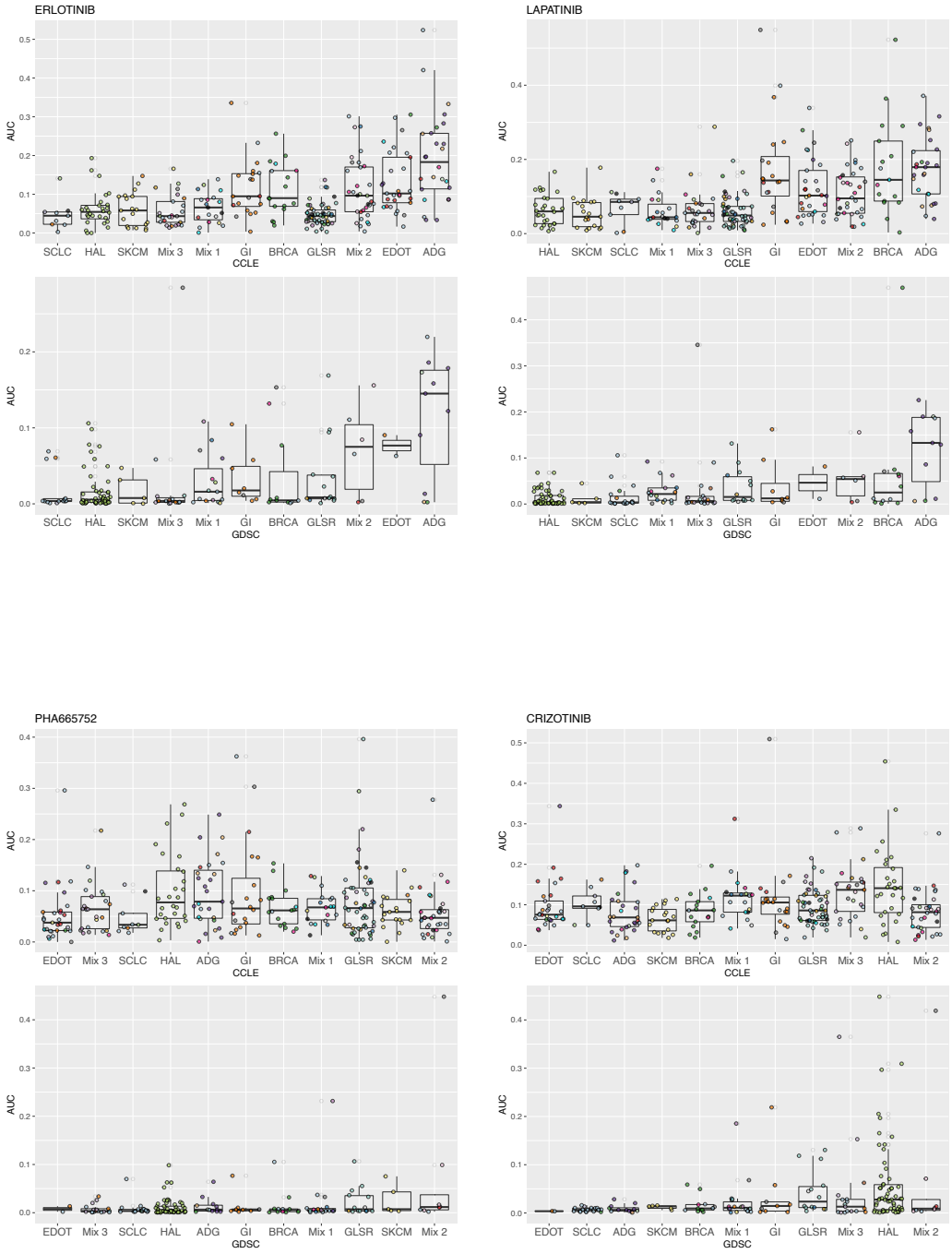
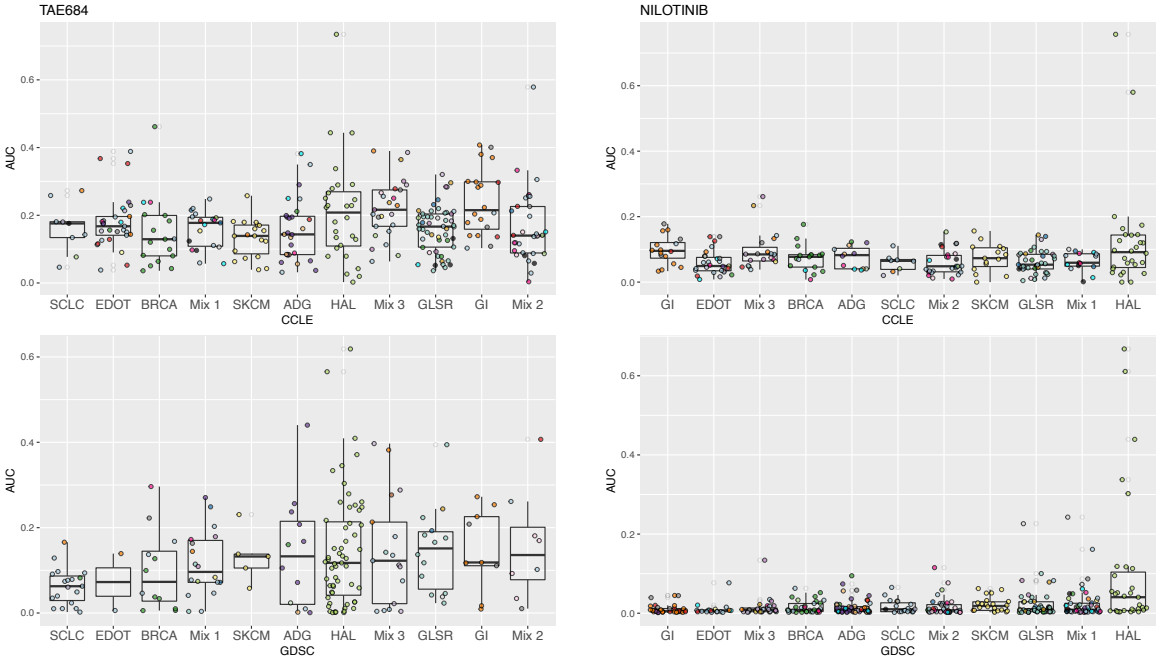
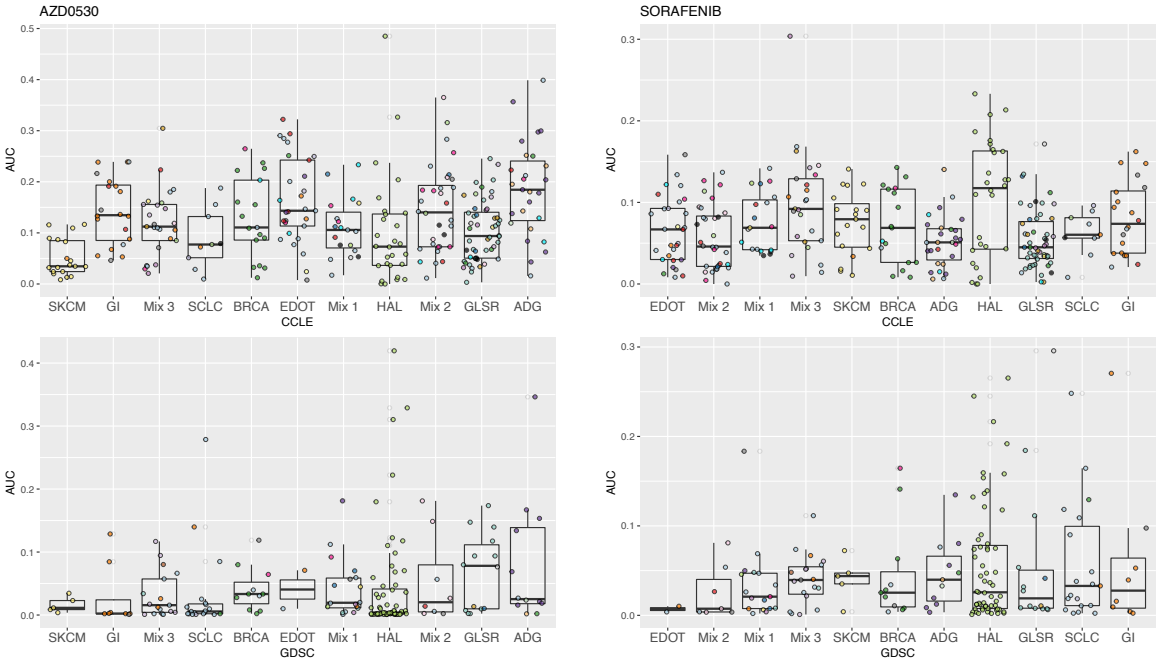
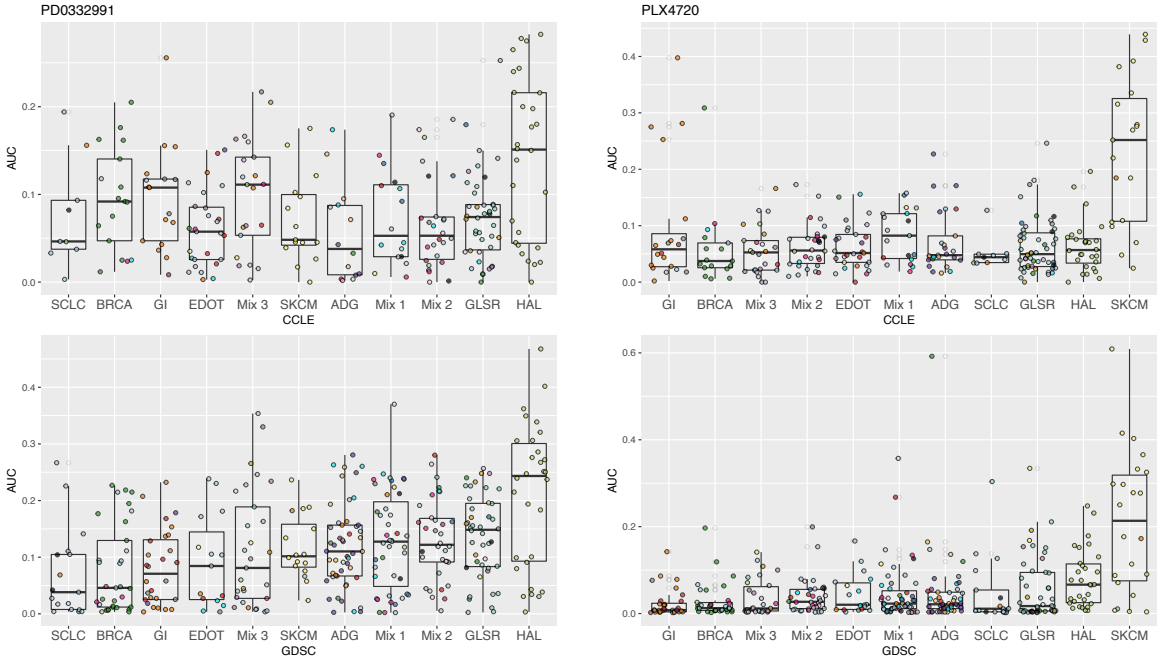


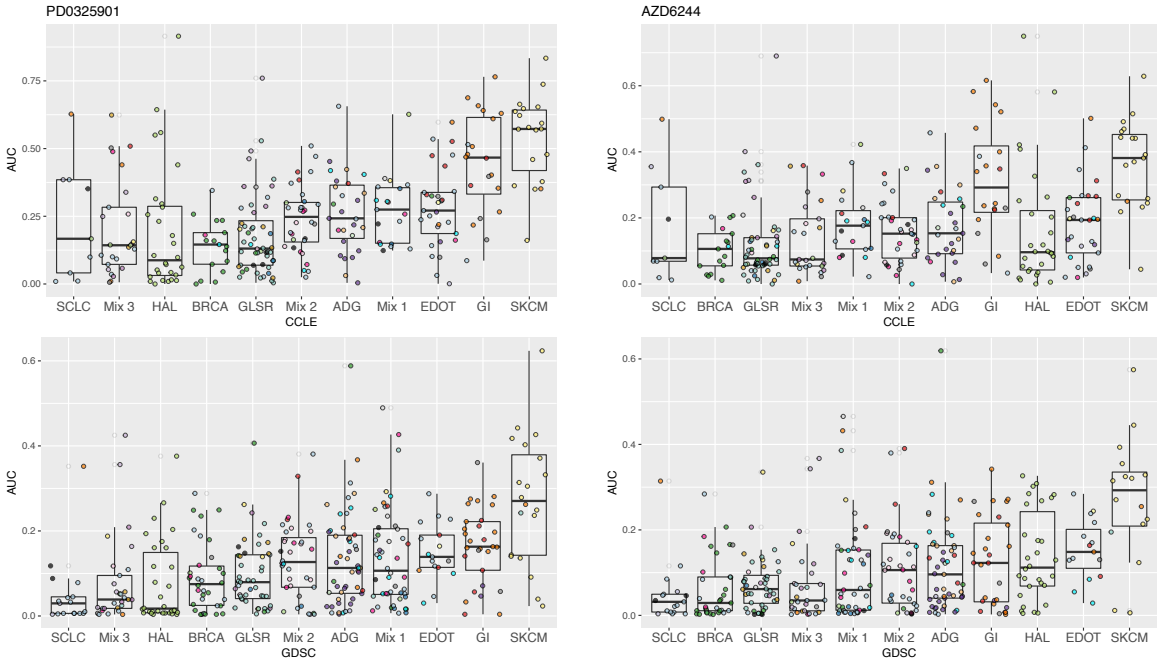
Figure A.4: AUC distribution for each cluster in CCLE and GDSC. Ordered according to the mean AUC. From resistant (left) to sensitive (right).

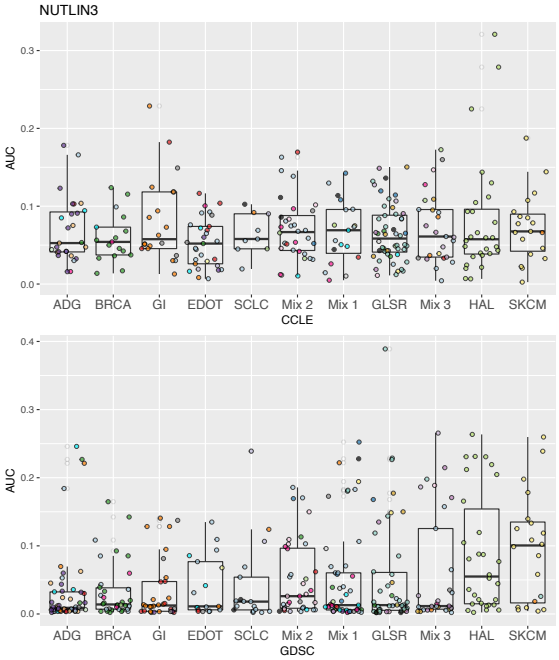


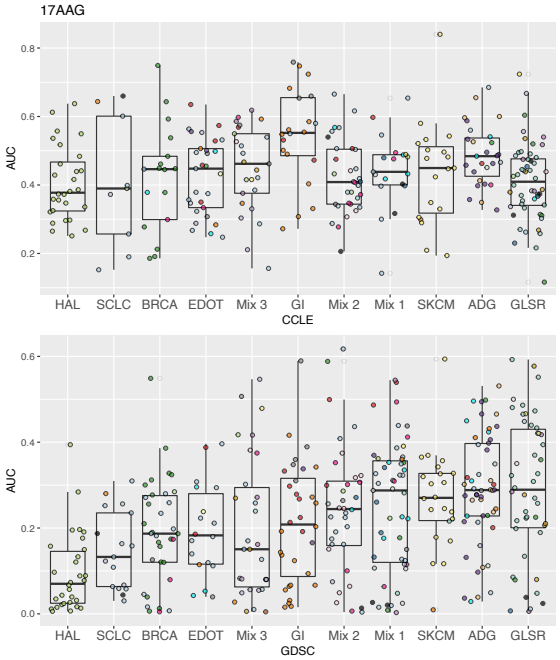

TAE684
AUC
SCLC EDOT BRCA Mix 1 SKCM ADG HAL Mix 3 GLSR GI Mix 2
CCLE
SCLC EDOT BRCA Mix 1 SKCM ADG HAL Mix 3 GLSR GI Mix 2
GDSC
NILOTINIB
AUC
GI EDOT Mix 3 BRCA ADG SCLC Mix 2 SKCM GLSR Mix 1 HAL
CCLE
GI EDOT Mix 3 BRCA ADG SCLC Mix 2 SKCM GLSR Mix 1 HAL
GDSC


AZD0530
AUC
SKCM GI Mix 3 SCLC BRCA EDOT Mix 1 HAL Mix 2 GLSR ADG
CCLE
SKCM GI Mix 3 SCLC BRCA EDOT Mix 1 HAL Mix 2 GLSR ADG
GDSC
SORAFENIB
AUC
EDOT Mix 2 Mix 1 Mix 3 SKCM BRCA ADG HAL GLSR SCLC GI
CCLE
EDOT Mix 2 Mix 1 Mix 3 SKCM BRCA ADG HAL GLSR SCLC GI
GDSC


PD0332991
PLX4720
AUC
SCLC BRCA GI EDOT Mix 3 SKCM ADG Mix 1 Mix 2 GLSR HAL
GI BRCA Mix 3 Mix 2 EDOT Mix 1 ADG SCLC GLSR HAL SKCM
CCLE
GDSC


PD0325901
AZD6244
AUC
SCLC Mix 3 HAL BRCA GLSR Mix 2 ADG Mix 1 EDOT GI SKCM
SCLC BRCA GLSR Mix 3 Mix 1 Mix 2 ADG GI HAL EDOT SKCM
CCLE
GDSC


NUTLIN3
AUC
ADG BRCA GI EDOT SCLC Mix 2 Mix 1 GLSR Mix 3 HAL SKCM
CCLE
ADG BRCA GI EDOT SCLC Mix 2 Mix 1 GLSR Mix 3 HAL SKCM
GDSC


17AAG
AUC
HAL SCLC BRCA EDOT Mix 3 GI Mix 2 Mix 1 SKCM ADG GLSR
CCLE
HAL SCLC BRCA EDOT Mix 3 GI Mix 2 Mix 1 SKCM ADG GLSR
GDSC

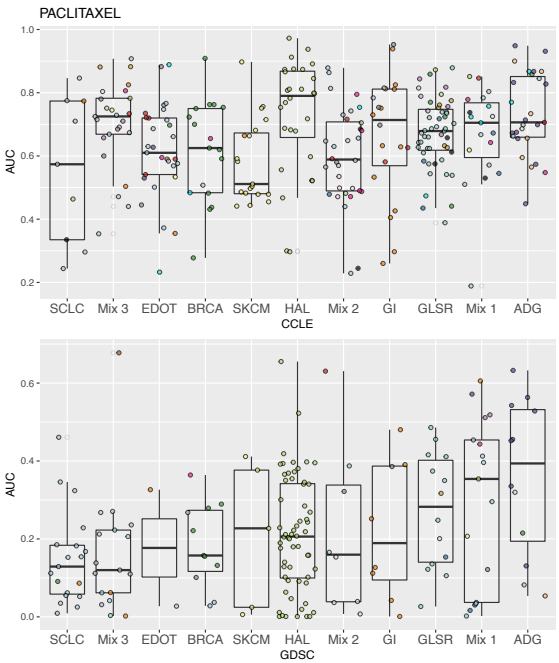

PACLITAXEL
AUC
SCLC Mix 3 EDOT BRCA SKCM HAL Mix 2 GI GLSR Mix 1 ADG
CCLE
SCLC Mix 3 EDOT BRCA SKCM HAL Mix 2 GI GLSR Mix 1 ADG
GDSC

Figure A.5: IC50 distribution for each cluster in CCLE and GSK. Ordered according to the mean IC50. From resistant (left) to sensitive (right).

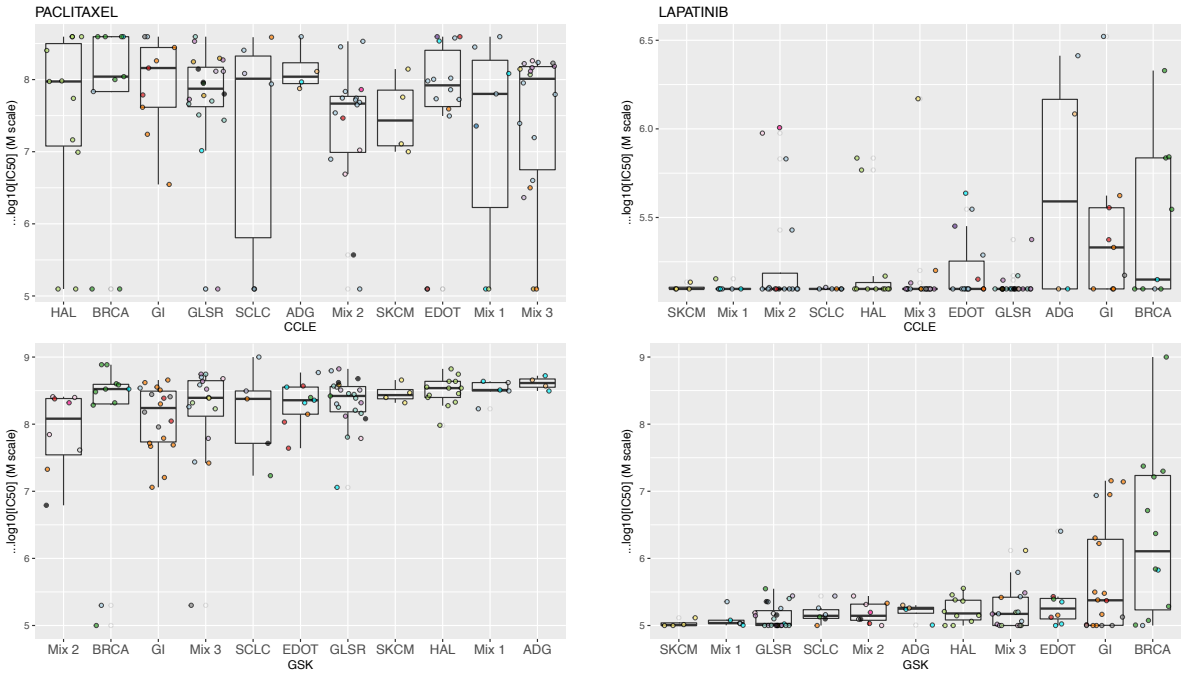


Figure A.6: IC50 distribution for each cluster in GDSC and GSK. Ordered according to the mean IC50. From resistant (left) to sensitive (right).

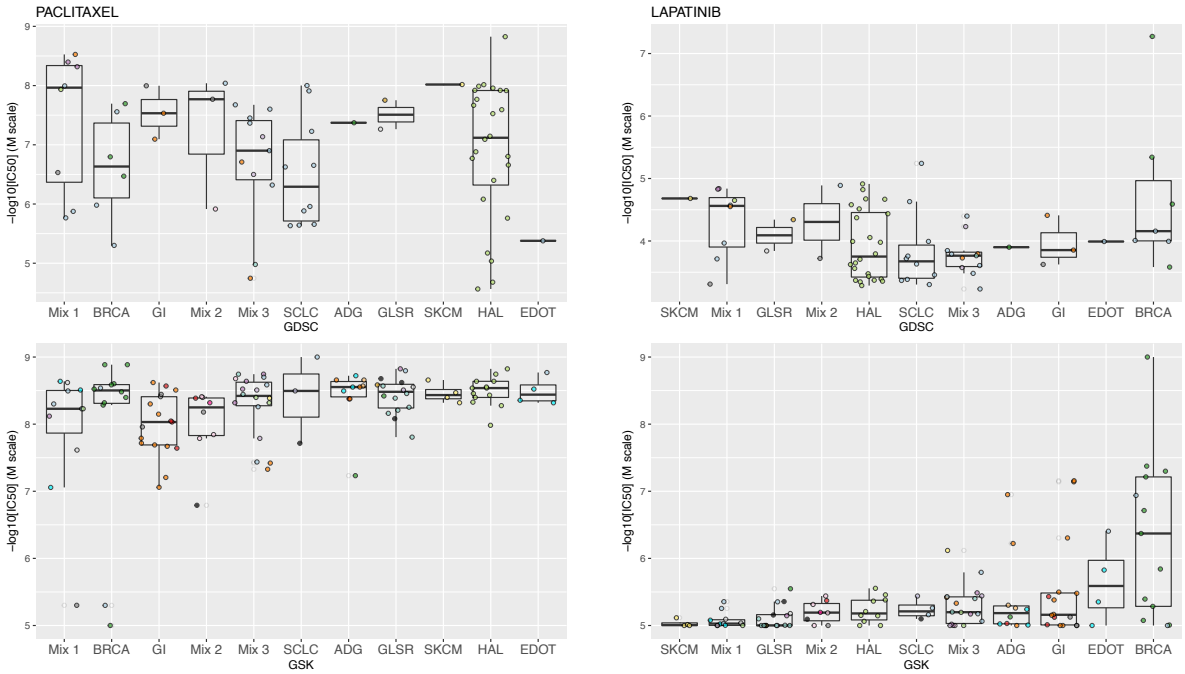
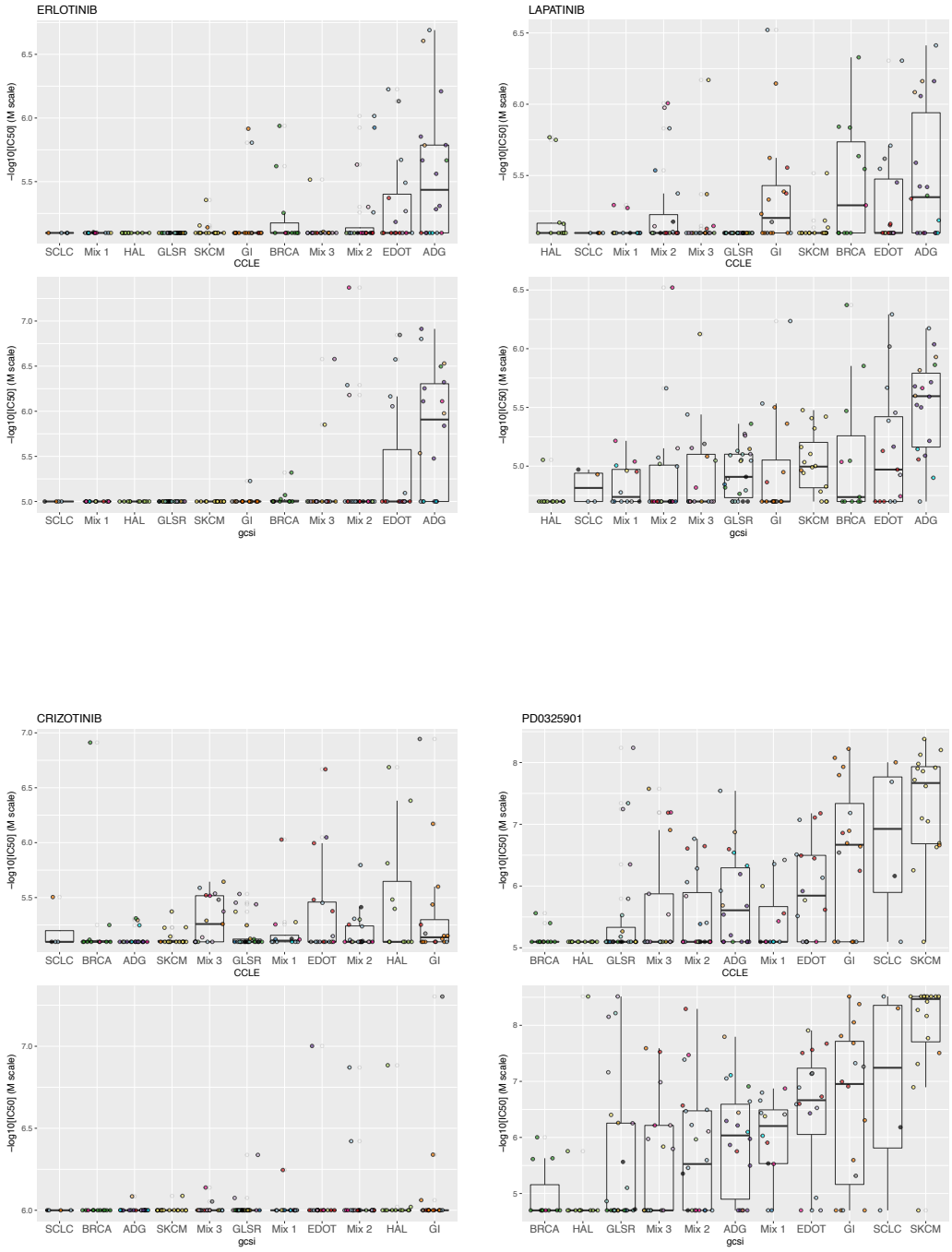


Figure A.7: IC50 distribution for each cluster in CCLE and gCSI. Ordered according to the mean IC50. From resistant (left) to sensitive (right).



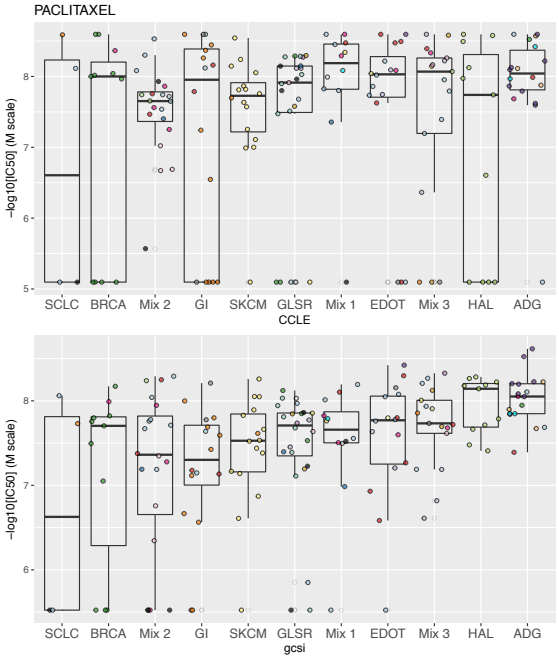
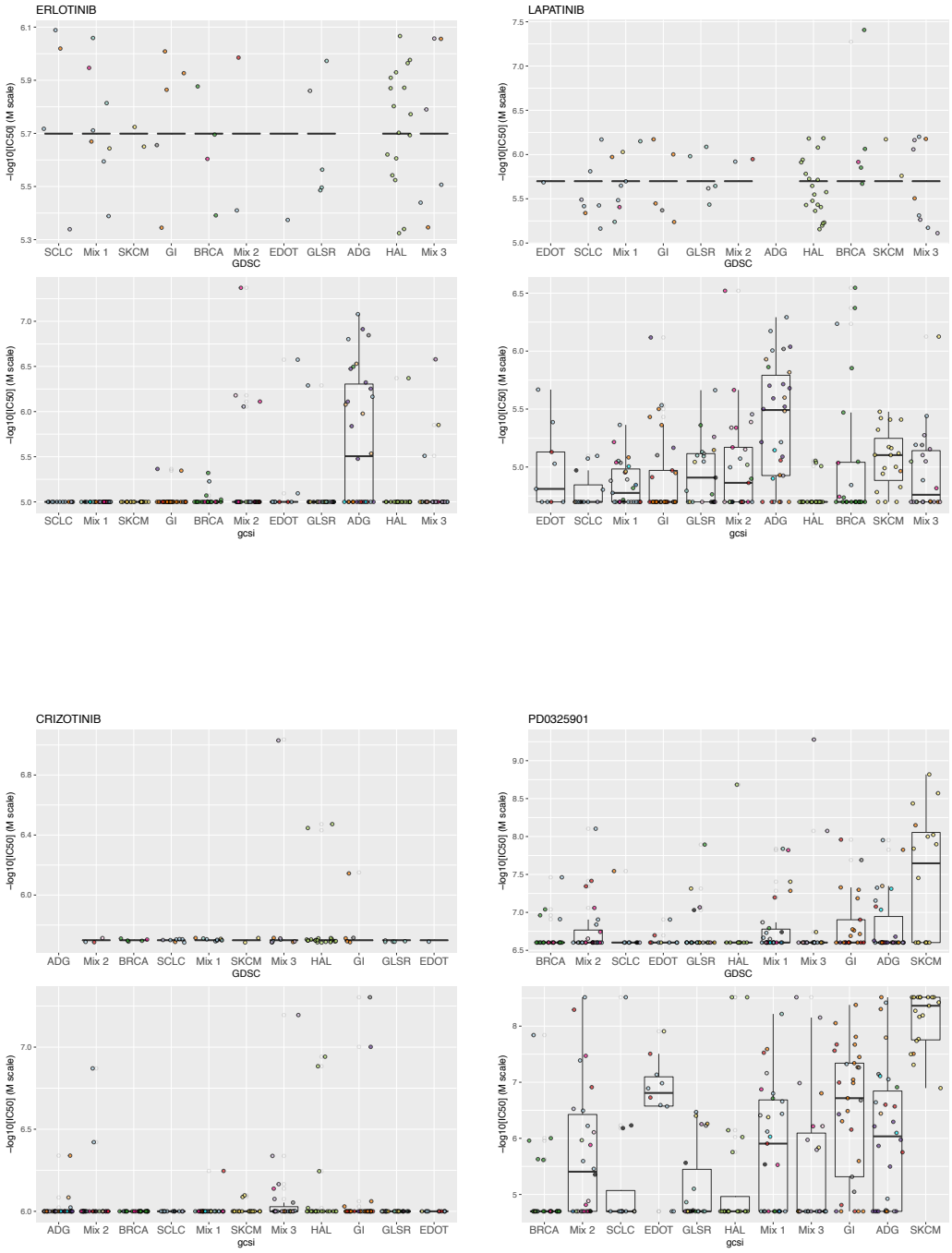

PACLITAXEL
−log10[IC50] (M scale)
8
7
6
5
SCLC
BRCA
Mix 2
GI
SKCM
GLSR
Mix 1
EDOT
Mix 3
HAL
ADG
CCLE
−log10[IC50] (M scale)
8
7
6
SCLC
BRCA
Mix 2
GI
SKCM
GLSR
Mix 1
EDOT
Mix 3
HAL
ADG
gcsi

Figure A.8: IC50 distribution for each cluster in GDSC and gCSI. Ordered according to the mean IC50. From resistant (left) to sensitive (right).



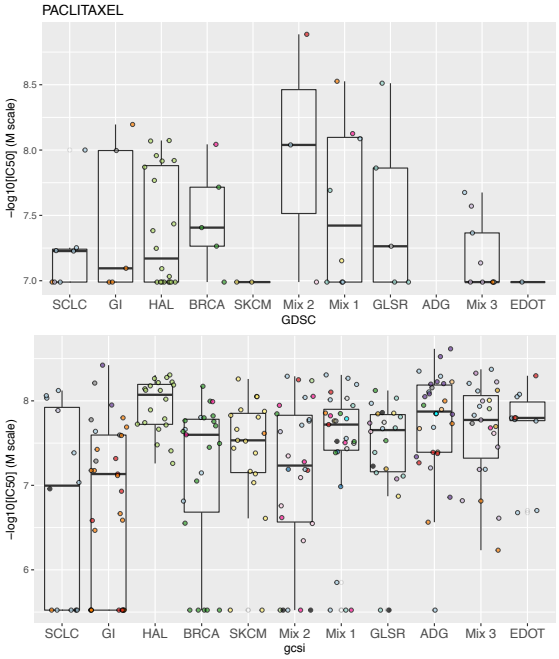
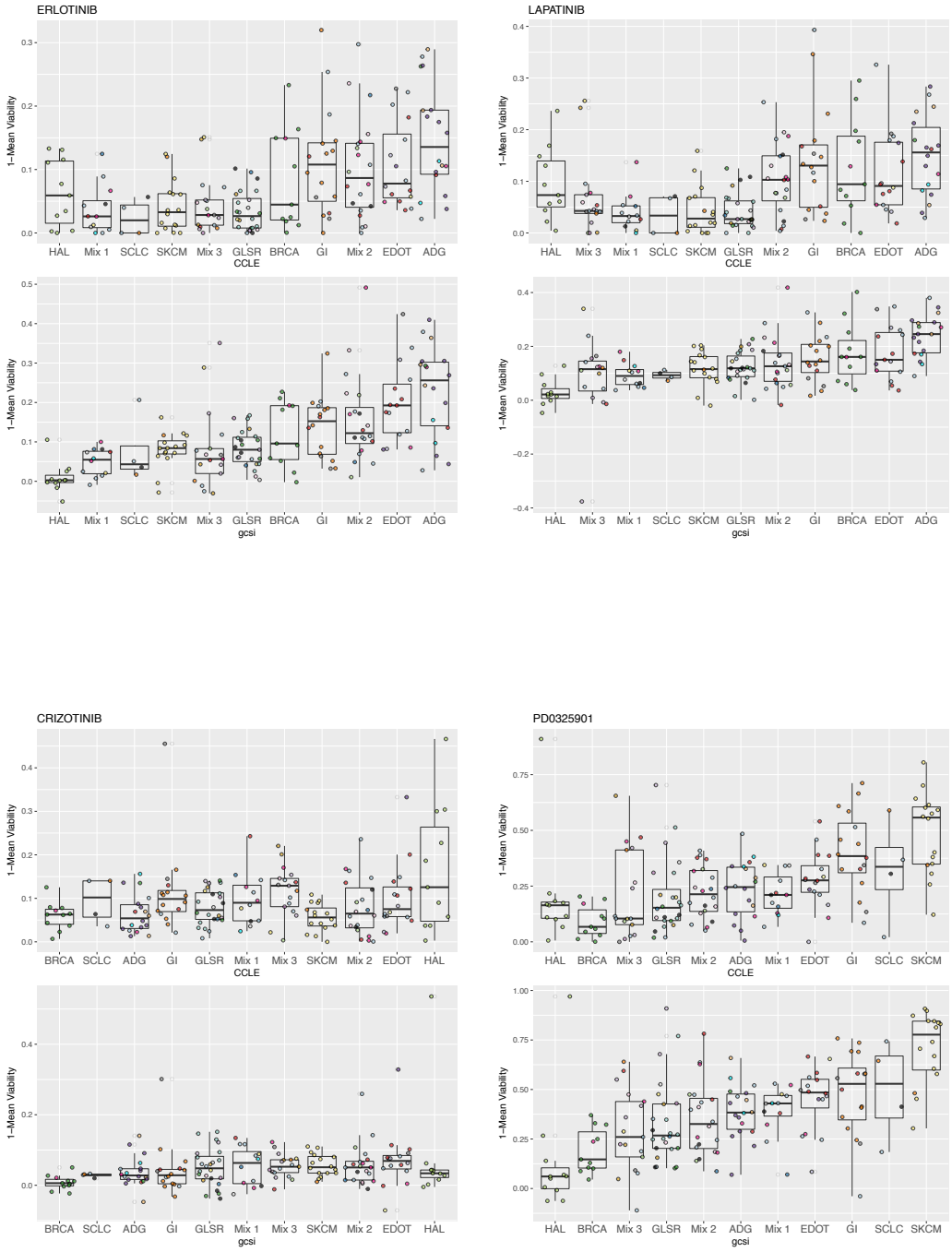

PACLITAXEL
−log10[IC50] (M scale)
8.5
8.0
7.5
7.0
SCLC
GI
HAL
BRCA
SKCM
Mix 2
Mix 1
GLSR
ADG
Mix 3
EDOT
GDSC
−log10[IC50] (M scale)
8
7
6
SCLC
GI
HAL
BRCA
SKCM
Mix 2
Mix 1
GLSR
ADG
Mix 3
EDOT
gcsi

Figure A.9: Mean viability distribution for each cluster in CCLE and gCSI. Ordered according to the average mean viability. From resistant (left) to sensitive (right).



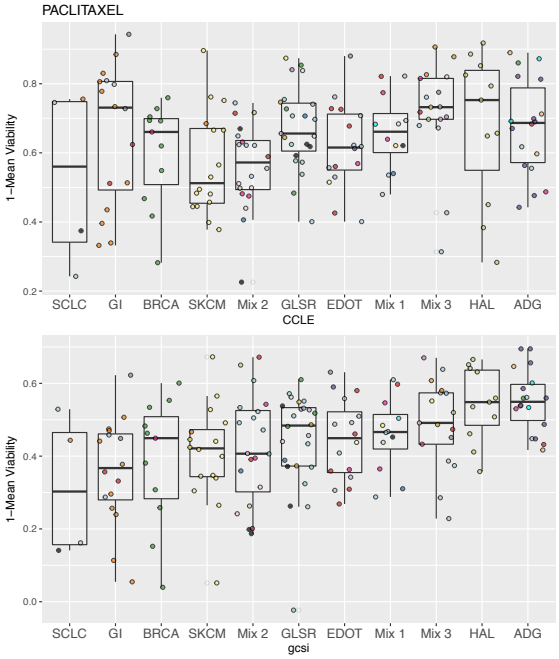
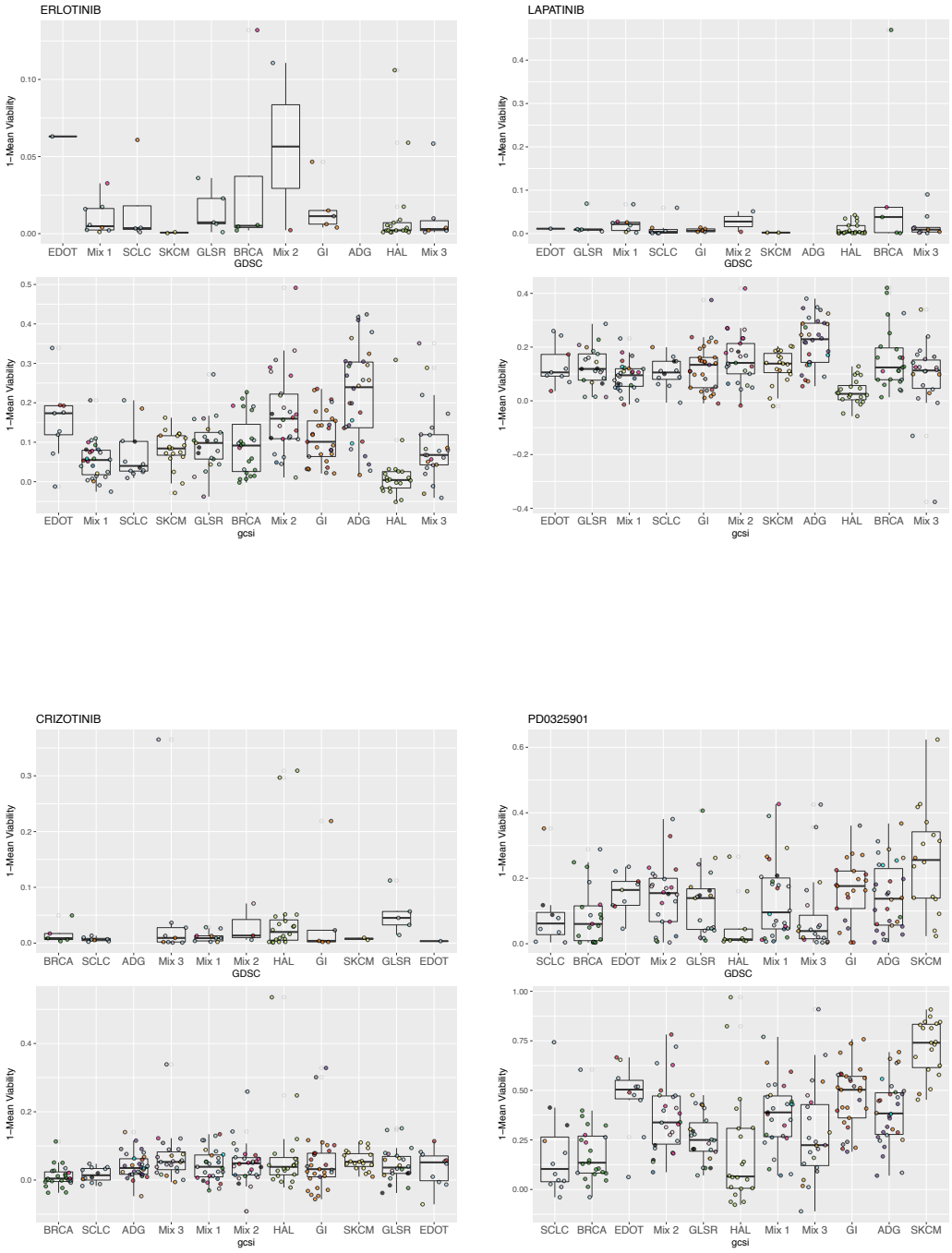

PACLITAXEL
1−Mean Viability
0.2
0.4
0.6
0.8
SCLC
GI
BRCA
SKCM
Mix 2
GLSR
EDOT
Mix 1
Mix 3
HAL
ADG
CCLE
1−Mean Viability
0.0
0.2
0.4
0.6
SCLC
GI
BRCA
SKCM
Mix 2
GLSR
EDOT
Mix 1
Mix 3
HAL
ADG
gcsi

Figure A.10: Mean viability distribution for each cluster in GDSC and gCSI. Ordered according to the average mean viability. From resistant (left) to sensitive (right).



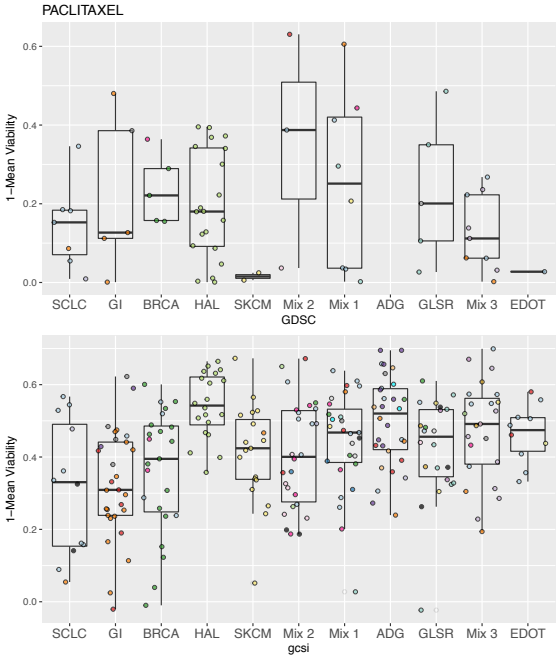

PACLITAXEL
1−Mean Viability
0.0
0.2
0.4
0.6
SCLC
GI
BRCA
HAL
SKCM
Mix 2
Mix 1
ADG
GLSR
Mix 3
EDOT
GDSC
gcsi

Figure A.11: Mutations for each cluster in CCLE dataset. The proportion of cell lines mutated, with raw and adjusted Fisher p-values.

SCLC	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	84	8	14	5	8	5	24	41
Fisher p-value	0,00443	0,957	0,662	0,978	0,492	0,864	0,174	0,000307
FDR-adjusted p-value	0,0297	1	1	1	0,983	1	0,959	0,00338

GI	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	60	57	31	23	0	9	63	0
Fisher p-value	0,736	1,72E-07	0,00289	0,0521	1	0,684	5,62E-12	1
FDR-adjusted p-value	1	1,89E-06	0,0318	0,287	1	1	6,18E-11	1

EDOT	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	82	44	9	4	7	0	7	16
Fisher p-value	0,0054	0,0000125	0,811	1	0,861	1	0,956	0,538
FDR-adjusted p-value	0,0297	0,0000688	1	1	1	1	0,997	1

Mixed 1	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	48	48	10	19	19	5	14	5
Fisher p-value	0,96	0,00164	0,786	0,593	0,0762	1	0,609	1
FDR-adjusted p-value	1	0,00601	1	1	0,419	1	0,997	1

HAL	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	64	8	3	3	23	11	8	5
Fisher p-value	0,584	0,983	1	0,999	0,0000602	0,738	0,997	0,941
FDR-adjusted p-value	1	1	1	1	0,000662	1	0,997	1

Mixed 2	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	51	29	18	10	0	12	10	4
Fisher p-value	0,951	0,00862	0,157	0,744	1	0,703	0,862	0,949
FDR-adjusted p-value	1	0,0237	0,432	1	1	1	0,997	1

GLSR	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	58	5	11	9	9	24	12	18
Fisher p-value	0,847	0,998	0,902	0,935	0,29	0,0236	0,865	0,208
FDR-adjusted p-value	1	1	1	1	0,797	0,259	0,997	0,763

Mixed 3	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	59	11	14	14	0	16	11	19
Fisher p-value	0,801	0,818	0,563	0,766	1	0,505	0,829	0,0432
FDR-adjusted p-value	1	1	1	1	1	1	0,997	0,238

ADG	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	77	3	23	3	8	8	13	0
Fisher p-value	0,0911	0,994	0,147	0,978	0,536	0,864	0,77	1
FDR-adjusted p-value	0,334	1	0,432	1	0,983	1	0,997	1

SKCM	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	31	0	0	81	15	15	15	4
Fisher p-value	1	1	1	1,42E-10	0,163	0,375	0,578	1
FDR-adjusted p-value	1	1	1	1,56E-09	0,599	1	0,997	1

BRCA	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	73	3	24	6	0	21	3	12
Fisher p-value	0,168	0,988	0,0713	0,838	1	0,152	0,992	0,534
FDR-adjusted p-value	0,462	1	0,392	1	1	0,836	0,997	1

Figure A.12: Mutations for each cluster in GDSC dataset. The proportion of cell lines mutated, with raw and adjusted Fisher p-values.

SCLC	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	72	6	6	0	3	9	6	31
Fisher p-value	0,0389	1	0,805	1	0,864	0,482	0,403	7,72E-06
FDR-adjusted p-value	0,139	1	1	1	1	1	1	0,000232

GI	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	65	62	16	11	0	3	49	0
Fisher p-value	0,622	1,87E-08	0,124	0,783	1	0,916	5,68E-14	1
FDR-adjusted p-value	0,629	1,09E-06	0,319	1	1	1	6,63E-11	1

EDOT	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	72	56	6	11	17	0	6	11
Fisher p-value	0,466	0,00104	0,85	0,769	0,0983	1	1	0,548
FDR-adjusted p-value	0,629	0,00014	1	1	1	1	1	1

Mixed 1	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	58	28	21	15	8	8	2	6
Fisher p-value	0,515	0,0107	0,0141	0,649	0,509	0,71	0,947	0,438
FDR-adjusted p-value	0,782	0,131	0,105	1	1	1	1	1

HAL	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	69	3	3	3	21	8	0	2
Fisher p-value	0,309	0,995	0,988	0,965	0,000334	0,498	1	0,952
FDR-adjusted p-value	0,629	1	1	1	0,00627	1	1	1

Mixed 2	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	54	23	15	5	0	8	0	8
Fisher p-value	0,98	0,0776	0,359	0,783	1	0,668	1	0,22
FDR-adjusted p-value	0,999	0,966	0,424	1	1	1	1	1

GLSR	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	60	0	4	10	4	21	0	6
Fisher p-value	0,83	1	0,94	0,414	0,781	0,185	1	0,878
FDR-adjusted p-value	0,999	1	1	1	1	0,57	1	1

Mixed 3	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	52	5	2	2	8	8	5	8
Fisher p-value	0,895	0,958	0,99	1	0,526	0,931	0,612	0,48
FDR-adjusted p-value	0,999	1	1	1	1	1	1	1

ADG	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	76	9	11	4	4	4	4	4
Fisher p-value	0,00281	0,73	0,436	0,936	0,882	0,977	0,743	0,926
FDR-adjusted p-value	0,139	1	1	1	1	1	1	1

SKCM	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	38	0	0	75	17	12	0	0
Fisher p-value	0,998	1	1	5,01E-09	0,0725	0,674	1	1
FDR-adjusted p-value	0,999	1	1	1,48E-06	0,465	1	1	1

BRCA	TP53	KRAS	PIK3CA	BRAF	NRAS	PTEN	APC	RB1
Proportion in %	69	2	21	2	0	10	0	2
Fisher p-value	0,133	0,997	0,0175	0,981	1	0,573	1	1
FDR-adjusted p-value	0,602	1	0,105	1	1	1	1	1

Figure A.13: Significant associations between CCLE and GDSC found in both dataset with IC50.

Drug	Cluster	Effect (-Log10(IC50(M)))	95% CI (-Log10(IC50(M)))	pvalue	adjusted pvalue	Drug response relative to other cell lines
Erlotinib	ADG	1.07	(0.44 to 1.71)	3.5E-03	3.9E-02	Sensitive
Lapatinib	SCLC	-0.41	(-0.67 to -0.14)	3.6E-03	3.2E-02	Resistant
	ADG	1.12	(0.53 to 1.69)	1.5E-03	1.5E-02	Sensitive
PD0332991	GI	-0.46	(-0.71 to -0.21)	7.2E-04	7.2E-03	Resistant
	HAL	0.74	(0.38 to 1.1)	2.0E-04	2.2E-03	Sensitive
PLX4720	SKCM	1.39	(0.86 to 1.93)	1.9E-05	2.1E-04	Sensitive
PD0325901	GI	0.46	(0.15 to 0.78)	5.4E-03	4.9E-02	Sensitive
	SKCM	1.31	(0.79 to 1.82)	3.3E-05	3.6E-04	Sensitive
AZD6244	SKCM	1.25	(0.69 to 1.81)	2.3E-04	2.6E-03	Sensitive
	BRCA	-0.45	(-0.74 to -0.16)	3.4E-03	3.4E-02	Resistant

Figure A.14: Number of cell lines with sensitivity measure in each dataset. The numbers were computed based on the cell lines common to each pair of dataset.

	CCLE vs GDSC		GSK				gCSI			
	IC50/AUC		IC50				IC50/Mean Viability			
			CCLE vs GSK		GDSC vs GSK		CCLE vs gCSI		GDSC vs gCSI	
	CCLE	GDSC	CCLE	GSK	GDSC	GSK	CCLE	gCSI	GDSC	gCSI
PACLITAXEL	276	173	118	116	70	116	164	165	65	241
LAPATINIB	277	169	118	114	70	114	165	161	64	237
ERLOTINIB	276	154					165	163	54	239
CRIZOTINIB	277	175					165	165	66	241
PD0325901	277	355					165	165	200	241
PHA665752	276	175								
TAE684	277	175								
NILOTINIB	220	369								
AZD0530	277	174								
SORAFENIB	276	172								
PD0332991	232	338								
PLX4720	273	366								
AZD6244	276	338								
NUTLIN3	277	358								
17AAG	277	366								

Figure A.15: List of drugs significantly associated with a sensitivity phenotype for each cluster in CCLE.

GI	Drug	Target
	drugid_17AAG	Other
	drugid_AZD6244	S/T Kinase
	drugid_PD0325901	S/T Kinase

EDOT	Drug	Target
	drugid_DOXORUBICIN	Chemotherapy

Mixed 1	Drug	Target
	drugid_GEMCITABINE	Chemotherapy
	drugid_BLEOMYCIN	Chemotherapy
	drugid_VINBLASTINE	Chemotherapy
	drugid_NVP-BEZ235	S/T Kinase, Other

HAL	Drug	Target
	drugid_IRINOTECAN	Chemotherapy
	drugid_L685458	Other
	drugid_PANOBINOSTAT	Other
	drugid_TOPOTECAN	Chemotherapy
	drugid_SUNITINIB	RTK
	drugid_CRIZOTINIB	RTK
	drugid_XMD8-85	S/T Kinase
	drugid_DMOG	Other
	drugid_PAC-1	Other
	drugid_IPA-3	S/T Kinase
	drugid_BAY-61-3606	CTK
	drugid_AICAR	S/T Kinase
	drugid_CAMPTOTHECIN-WTSI	Chemotherapy
	drugid_METHOTREXATE	Chemotherapy
	drugid_ATRA	Other
	drugid_GEFITINIB	RTK
	drugid_ABT-263	Other
	drugid_VORINOSTAT	Other
	drugid_NILOTINIB	CTK
	drugid_TEMSIROLIMUS	S/T Kinase
	drugid_AZD-2281	Other
	drugid_ABT-888	Other
	drugid_BOSUTINIB	CTK
	drugid_LENALIDOMIDE	Other
	drugid_AZD7762	S/T Kinase
	drugid_CEP-701	RTK, CTK
	drugid_VX-702	S/T Kinase
	drugid_SL-0101-1	S/T Kinase
	drugid_681640	S/T Kinase, CTK
	drugid_PD-173074	RTK
	drugid_ZM-447439	S/T Kinase
	drugid_PD0332991	S/T Kinase
	drugid_SB590885	S/T Kinase

Mixed 2	Drug	Target
	drugid_MIDOSTAURIN	RTK

GLSR	Drug	Target
	drugid_DASATINIB	CTK, RTK
	drugid_BORTEZOMIB	Other
	drugid_MIDOSTAURIN	RTK
	drugid_CHIR-99021	S/T Kinase
	drugid_AZD6482	Other
	drugid_BEXAROTENE	Other
	drugid_LFM-A13	CTK
	drugid_PAZOPANIB	RTK
	drugid_VINBLASTINE	Chemotherapy
	drugid_DOCETAXEL	Chemotherapy
	drugid_TEMSIROLIMUS	S/T Kinase
	drugid_ELESCLOMOL	Other
	drugid_BX-795	S/T Kinase
	drugid_NVP-BEZ235	S/T Kinase, Other
	drugid_GDC0941	Other

Mixed 3	Drug	Target
	drugid_VORINOSTAT	Other

ADG	Drug	Target
	drugid_AZD0530	RTK
	drugid_VANDETANIB	RTK
	drugid_ERLOTINIB	RTK
	drugid_LAPATINIB	RTK
	drugid_VINORELBINE	Chemotherapy
	drugid_BICALUTAMIDE	Other
	drugid_PF-562271	S/T Kinase
	drugid_OBATOCLAX-MESYLATE	Other
	drugid_EPOTHILONE-B	Other
	drugid_AICAR	S/T Kinase
	drugid_VINBLASTINE	Chemotherapy
	drugid_DOCETAXEL	Chemotherapy
	drugid_GEFITINIB	RTK
	drugid_BOSUTINIB	CTK
	drugid_17AAG	Other
	drugid_BIBW2992	RTK

SKCM	Drug	Target
	drugid_AZD6244	S/T Kinase
	drugid_EMBELIN	Other
	drugid_FH535	Other
	drugid_OBATOCLAX-MESYLATE	Other
	drugid_RDEA119	S/T Kinase
	drugid_CI-1040	S/T Kinase
	drugid_CEP-701	RTK, CTK
	drugid_PLX4720	S/T Kinase
	drugid_NUTLIN3	Other
	drugid_PD0325901	S/T Kinase
	drugid_SB590885	S/T Kinase
	drugid_AZD6244	S/T Kinase

BRCA	Drug	Target
	drugid_AKT-INHIBITOR-VIII	S/T Kinase
	drugid_MK-2206	S/T Kinase

Figure A.16: List of drugs significantly associated with a resistant phenotype for each cluster in GDSC.

SCLC	Drug	Target
	drugid_ERLOTINIB	RTK
	drugid_LBW242	#N/A
	drugid_PD0332991	S/T Kinase
	drugid_PHA665752	RTK
	drugid_PLX4720	S/T Kinase
	drugid_SORAFENIB	RTK
	drugid_SUNITINIB	RTK
	drugid_MG-132	Other
	drugid_CYCLOPAMINE	GPCR
	drugid_AZ628	S/T Kinase
	drugid_CRIZOTINIB	RTK
	drugid_Z-LLNLE-CHO-	Other
	drugid_DASATINIB	CTK, RTK
	drugid_WH-4-023	CTK
	drugid_BMS-536924	RTK
	drugid_JW-7-52-1	S/T Kinase
	drugid_BORTEZOMIB	Other
	drugid_XMD8-85	S/T Kinase
	drugid_LAPATINIB	RTK
	drugid_GEMCITABINE	Chemotherapy
	drugid_MIDOSTAURIN	RTK
	drugid_CHIR-99021	S/T Kinase
	drugid_AP-24534	CTK
	drugid_AZD6482	Other
	drugid_PF-562271	S/T Kinase
	drugid_FTI-277	Other
	drugid_SHIKONIN	Other
	drugid_EMBELIN	Other
	drugid_OBATOCLAX-MESYLATI	Other
	drugid_BEXAROTENE	Other
	drugid_TIPIFARNIB	Other
	drugid_RDEA119	S/T Kinase
	drugid_CI-1040	S/T Kinase
	drugid_PD0325901	S/T Kinase

GI	Drug	Target
	drugid_AP-24534	CTK
	drugid_AZD6482	Other
	drugid_DMOG	Other
	drugid_IPA-3	S/T Kinase
	drugid_PAZOPANIB	RTK
	drugid_CAMPTOTHECIN-WTSI	Chemotherapy
	drugid_VINBLASTINE	Chemotherapy
	drugid_CISPLATIN	Chemotherapy
	drugid_DOCETAXEL	Chemotherapy
	drugid_ATRA	Other
	drugid_NILOTINIB	CTK
	drugid_TEMSIROLIMUS	S/T Kinase
	drugid_AZD-2281	Other
	drugid_BOSUTINIB	CTK
	drugid_AXITINIB	RTK
	drugid_CEP-701	RTK, CTK
	drugid_AMG-706	RTK
	drugid_ELESCLOMOL	Other
	drugid_PLX4720	S/T Kinase
	drugid_BX-795	S/T Kinase
	drugid_SL-0101-1	S/T Kinase
	drugid_BI-D1870	S/T Kinase
	drugid_JNK-INHIBITOR-VIII	S/T Kinase
	drugid_681640	S/T Kinase, CTK
	drugid_PD-173074	RTK
	drugid_ZM-447439	S/T Kinase
	drugid_RO-3306	S/T Kinase
	drugid_PD0332991	S/T Kinase
	drugid_NVP-BEZ235	S/T Kinase, Other
	drugid_GDC0941	Other
	drugid_AZD8055	S/T Kinase

EDOT	Drug	Target
	drugid_L685458	#N/A
	drugid_PD0332991	S/T Kinase
	drugid_PHA665752	RTK
	drugid_PLX4720	S/T Kinase
	drugid_TKI258	RTK
	drugid_CYCLOPAMINE	GPCR
	drugid_CRIZOTINIB	RTK
	drugid_PYRIMETHAMINE	Other
	drugid_ATRA	Other
	drugid_TEMSIROLIMUS	S/T Kinase
	drugid_VX-702	S/T Kinase

Mixed 1	Drug	Target
	drugid_ERLOTINIB	RTK
	drugid_L685458	#N/A
	drugid_LBW242	#N/A
	drugid_NILOTINIB	CTK
	drugid_VANDETANIB	#N/A
	drugid_GEFITINIB	RTK
	drugid_BIBW2992	RTK

HAL	Drug	Target
	drugid_ERLOTINIB	RTK
	drugid_KIN001-135	S/T Kinase
	drugid_LAPATINIB	RTK
	drugid_NSC-87877	Other
	drugid_BICALUTAMIDE	Other
	drugid_FTI-277	Other
	drugid_FH535	Other
	drugid_BMS-754807	RTK
	drugid_BRYOSTATIN-1	S/T Kinase
	drugid_DOCETAXEL	Chemotherapy
	drugid_AZD-2281	Other
	drugid_AXITINIB	RTK
	drugid_GW-441756	RTK
	drugid_VX-702	S/T Kinase
	drugid_MK-2206	S/T Kinase

Mixed 2	Drug	Target
	drugid_AZD6244	S/T Kinase
	drugid_IRINOTECAN	Chemotherapy
	drugid_L685458	#N/A
	drugid_NILOTINIB	CTK
	drugid_PANOBINOSTAT	#N/A
	drugid_AICAR	S/T Kinase
	drugid_17AAG	Other

GLSR	Drug	Target
	drugid_AZD0530	CTK
	drugid_ERLOTINIB	RTK
	drugid_LAPATINIB	RTK
	drugid_PANOBINOSTAT	#N/A
	drugid_PLX4720	S/T Kinase
	drugid_SORAFENIB	RTK
	drugid_VANDETANIB	#N/A
	drugid_PAC-1	Other
	drugid_METHOTREXATE	Chemotherapy
	drugid_GEFITINIB	RTK

Mixed 3	Drug	Target
	drugid_ERLOTINIB	RTK
	drugid_LBW242	#N/A
	drugid_PHA665752	RTK
	drugid_PLX4720	S/T Kinase
	drugid_Z-LLNLE-CHO-	Other
	drugid_DASATINIB	CTK, RTK
	drugid_BICALUTAMIDE	Other
	drugid_BIBW2992	RTK

ADG	Drug	Target
	drugid_CRIZOTINIB	RTK
	drugid_PLX4720	S/T Kinase
	drugid_SORAFENIB	RTK
	drugid_CYCLOPAMINE	GPCR
	drugid_GSK269962A	S/T Kinase
	drugid_PAZOPANIB	RTK
	drugid_SB-216763	S/T Kinase

SKCM	Drug	Target
	drugid_AZD0530	CTK
	drugid_ERLOTINIB	RTK
	drugid_L685458	#N/A
	drugid_LAPATINIB	RTK
	drugid_PD0332991	S/T Kinase
	drugid_SORAFENIB	RTK
	drugid_TAE684	RTK
	drugid_VANDETANIB	#N/A
	drugid_WZ-1-84	CTK
	drugid_AKT-INHIBITOR-VIII	S/T Kinase
	drugid_AICAR	S/T Kinase
	drugid_GEFITINIB	RTK
	drugid_BIBW2992	RTK

BRCA	Drug	Target
	drugid_AZD6244	S/T Kinase
	drugid_PLX4720	S/T Kinase
	drugid_TOPOTECAN	Chemotherapy
	drugid_GEMCITABINE	Chemotherapy
	drugid_AP-24534	CTK
	drugid_DMOG	Other
	drugid_BAY-61-3606	CTK
	drugid_OBATOCLAX-MESYLAT	Other
	drugid_CAMPTOTHECIN-WTSI	Chemotherapy
	drugid_CYTARABINE	Chemotherapy
	drugid_METHOTREXATE	Chemotherapy
	drugid_RDEA119	S/T Kinase
	drugid_CI-1040	S/T Kinase
	drugid_AZD-2281	Other
	drugid_CEP-701	RTK, CTK
	drugid_SB-216763	S/T Kinase
	drugid_BX-795	S/T Kinase
	drugid_BI-D1870	S/T Kinase
	drugid_681640	S/T Kinase, CTK
	drugid_AZD6244	S/T Kinase

SKCM	Drug	Target
	drugid_AZD0530	CTK
	drugid_ERLOTINIB	RTK
	drugid_L685458	#N/A
	drugid_LAPATINIB	RTK
	drugid_PD0332991	S/T Kinase
	drugid_SORAFENIB	RTK
	drugid_TAE684	RTK
	drugid_VANDETANIB	#N/A
	drugid_WZ-1-84	CTK
	drugid_AKT-INHIBITOR-VIII	S/T Kinase
	drugid_AICAR	S/T Kinase
	drugid_GEFITINIB	RTK
	drugid_BIBW2992	RTK

BRCA	Drug	Target
	drugid_AZD6244	S/T Kinase
	drugid_PLX4720	S/T Kinase
	drugid_TOPOTECAN	Chemotherapy
	drugid_GEMCITABINE	Chemotherapy
	drugid_AP-24534	CTK
	drugid_DMOG	Other
	drugid_BAY-61-3606	CTK
	drugid_OBATOCLAX-MESYLAT	Other
	drugid_CAMPTOTHECIN-WTSI	Chemotherapy
	drugid_CYTARABINE	Chemotherapy
	drugid_METHOTREXATE	Chemotherapy
	drugid_RDEA119	S/T Kinase
	drugid_CI-1040	S/T Kinase
	drugid_AZD-2281	Other
	drugid_CEP-701	RTK, CTK
	drugid_SB-216763	S/T Kinase
	drugid_BX-795	S/T Kinase
	drugid_BI-D1870	S/T Kinase
	drugid_681640	S/T Kinase, CTK
	drugid_AZD6244	S/T Kinase

Figure A.17: Significant associations between CCLE and GDSC found in both dataset with AUC.

CCLE

Drug	Cluster	Effect (AUC)	95% CI (AUC)	pvalue	adjusted pvalue	Drug response relative to other cell lines
Erlotinib	ADG	0,05	(0,015 to 0,08)	0.0052	0.0360	Sensitive
Lapatinib	HAL	-0,04	(0,06 to -0,02)	1,00E-04	9,00E-04	Resistant
AZD6244	SKCM	0,2	(0,13 to 0,27)	5,00E-06	5,00E-05	Sensitive
	BRCA	-0,079	(-0,12 to -0,04)	2,00E-04	2,00E-03	Resistant
PLX4720	SKCM	0,16	(0,10 to 0,23)	3.1e-05	3.4e-04	Sensitive
Crizotinib	SKCM	-0,04	(-0,05 to -0,02)	3.8e-05	4.2e-04	Resistant
AZD0530	SKCM	-0,08	(-0,09 to -0,06)	2.4e-09	2.7e-08	Resistant

GDSC

Drug	Cluster	Effect (AUC)	95% CI (AUC)	pvalue	adjusted pvalue	Drug response relative to other cell lines
Erlotinib	ADG	0,09	(0,04 to 0,14)	0.003	0.0360	Sensitive
Lapatinib	HAL	-0,04	(-0,04 to -0,02)	2,39E-05	3,00E-04	Resistant
AZD6244	SKCM	0,17	(0,09 to 0,25)	3,00E-03	5,00E-05	Sensitive
	BRCA	-0,05	(-0,08 to -0,021)	0.001	0.01	Resistant
PLX4720	SKCM	0,16	(0,10 to 0,23)	1,00E-04	1,00E-03	Sensitive
Crizotinib	SKCM	-0,03	(-0,04 to -0,01)	7.9e-06	6.3e-05	Resistant
AZD0530	SKCM	-0,03	(-0,05 to -0,01)	0.002	0.020	Resistant

Figure A.18: Significant associations between CCLE and GSK found in both dataset with IC50.

CCLE

Drug	Cluster	Effect (-Log10(IC50(M)))	95% CI (-Log10(IC50(M)))	pvalue	adjusted pvalue	Drug response relative to other cell lines
Lapatinib	Mix 1	-0,14	(-0.2 to -0.075)	0,000024	0,00024	Resistant
	SKCM	-0,13	(-0.19 to -0.071)	0,000044	0,00039	Resistant

GSK

Drug	Cluster	Effect (-Log10(IC50(M)))	95% CI (-Log10(IC50(M)))	pvalue	adjusted pvalue	Drug response relative to other cell lines
Lapatinib	Mix 1	-0,31	(-0.5 to -0.12)	0,004	0,036	Resistant
	SKCM	-0,38	(-0.52 to -0.24)	8,5E-07	0,0000093	Resistant

Figure A.19: Significant associations between CCLE and gCSI found in both dataset with IC50.

CCLE

Drug	Cluster	Effect (-Log10(IC50(M)))	95% CI (-Log10(IC50(M)))	pvalue	adjusted pvalue	Drug response relative to other cell lines
Erlotinib	Mix 1	-0,12	(-0.17 to -0.073)	0,0000016	0,000015	Resistant
	GLSR	-0,13	(-0.18 to -0.08)	0,0000014	0,000015	Resistant
	ADG	0,4	(0.13 to 0.66)	0,0053	0,032	Sensitive
	SKCM	-0,099	(-0.16 to -0.04)	0,0011	0,0079	Resistant
Lapatinib	Mix 1	-0,14	(-0.2 to -0.069)	0,00015	0,0014	Resistant
	ADG	0,27	(0.033 to 0.5)	0,027	0,19	Sensitive
PD0325901	SKCM	1,6	(1.1 to 2.1)	0,0000013	0,000012	Sensitive
	BRCA	-0,73	(-0.92 to -0.54)	9,2E-12	9,2E-11	Resistant

gCSI

Drug	Cluster	Effect (-Log10(IC50(M)))	95% CI (-Log10(IC50(M)))	pvalue	adjusted pvalue	Drug response relative to other cell
Erlotinib	Mix 1	-0,19	(-0.26 to -0.11)	0,0000049	0,000051	Resistant
	GLSR	-0,2	(-0.28 to -0.12)	0,0000046	0,000051	Resistant
	ADG	0,7	(0.36 to 1)	0,00044	0,0022	Sensitive
	SKCM	-0,19	(-0.27 to -0.11)	0,0000048	0,000051	Resistant
Lapatinib	Mix 1	-0,19	(-0.32 to -0.06)	0,0061	0,055	Resistant
	ADG	0,55	(0.34 to 0.76)	0,000023	0,00023	Sensitive
PD0325901	SKCM	2,1	(1.6 to 2.7)	1,7E-07	0,0000019	Sensitive
	BRCA	-1,1	(-1.5 to -0.74)	0,0000038	0,000038	Resistant

Figure A.20: Significant associations between CCLE and gCSI found in both dataset with mean viability.

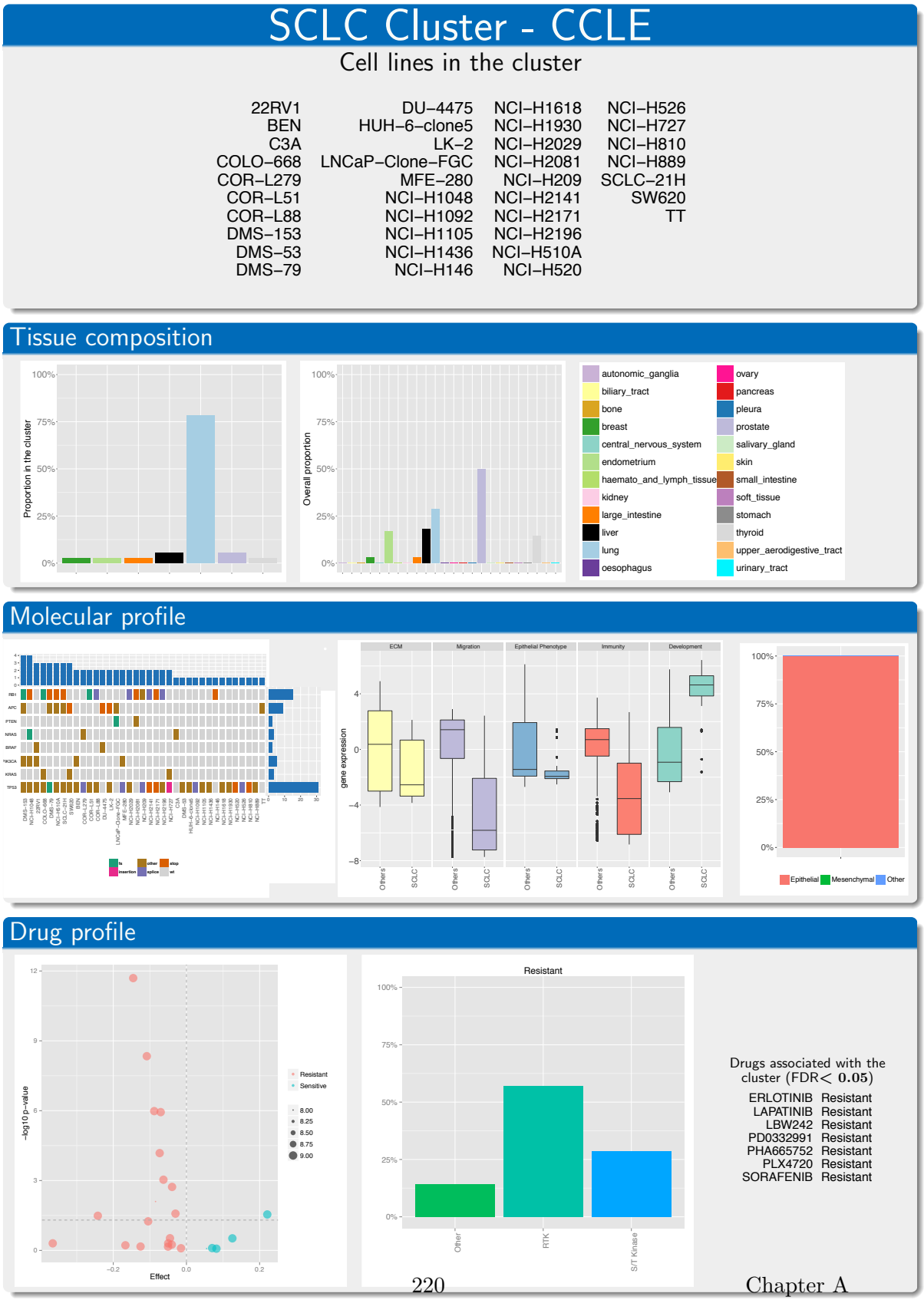
Drug	Cluster	Effect (1-Mean viability)	95% CI (1-Mean viability)	pvalue	adjusted pvalue	Drug response relative to other cell lines
Erlotinib	SCLC	-0,054	(-0.096 to -0.011)	0,025	0,15	Resistant
	Mix 1	-0,042	(-0.069 to -0.016)	0,0038	0,031	Resistant
	HAL	-0,014	(-0.051 to 0.022)	0,41	0,83	Resistant
	GLSR	-0,049	(-0.067 to -0.032)	0,0000005	0,0000055	Resistant
	ADG	0,081	(0.037 to 0.12)	0,0011	0,011	Sensitive
	SKCM	-0,039	(-0.063 to -0.014)	0,0031	0,028	Resistant
PD0325901	SKCM	0,25	(0.15 to 0.35)	0,000051	0,00051	Sensitive
	BRCA	-0,18	(-0.23 to -0.13)	8,5E-07	0,0000094	Resistant

Drug	Cluster	Effect (1-Mean viability)	95% CI 1-Mean viability)	pvalue	adjusted pvalue	Drug response relative to other cell lines
Erlotinib	SCLC	-0,042	(-0.18 to 0.094)	0,41	0,89	Resistant
	Mix 1	-0,076	(-0.1 to -0.049)	0,0000028	0,000028	Resistant
	HAL	-0,12	(-0.15 to -0.088)	3,2E-08	3,6E-07	Resistant
	GLSR	-0,041	(-0.068 to -0.015)	0,0028	0,022	Resistant
	ADG	0,12	(0.064 to 0.18)	0,00035	0,0032	Sensitive
	SKCM	-0,044	(-0.073 to -0.014)	0,0046	0,029	Resistant
PD0325901	SKCM	0,36	(0.26 to 0.46)	3,7E-07	0,000004	Sensitive
	BRCA	-0,2	(-0.29 to -0.12)	0,000064	0,00064	Resistant

A.1 Summary of cell line clusters

Summary of each cell line cluster with information regarding tissue composition, molecular profile and drug profile in respectively CCLE and GDSC. The first block shows the cell lines belonging to the given cluster. The second block shows tissue composition. First window: proportion of lines for the cluster originating from a given tissue. Second window: proportion of the lines from a given tissue belonging to the cluster. Third window: Molecular profile of the cluster. Mutation type of each cell line. Genes and cell lines are sorted according to the number of events. The expression profile of the cluster relative to all the other cell lines for the defined gene clusters is shown. The proportion of epithelial or mesenchymal cells, according to the signature defined by Taub et al. The last window illustrates the drug profile associated with the cluster. A volcano plot representation of t-test results showing the magnitude (effect; x-axis) and significance ($-log10(p - value)$; y-axis) of all drug-cluster associations in the dataset. Each circle represents a single drug-cluster interaction and its size is proportional to the number of cell lines screened. The horizontal dashed line indicates the threshold of statistical significance ($-log10(0.05)$). Barplot representation of the t-test results showing the proportion of drug types significantly associated with the cluster. List of drugs associated with the cluster (FDR < 0.05) and their phenotype.

Figure A.21

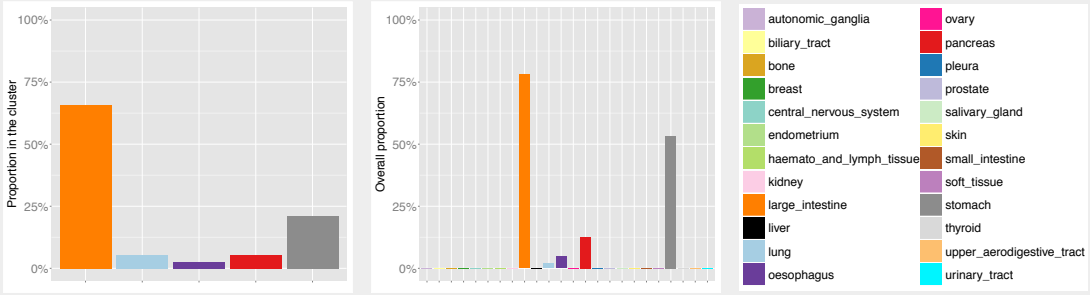

SCLC Cluster - CCLE
Cell lines in the cluster
22RV1
BEN
C3A
COLO−668
COR−L279
COR−L51
COR−L88
DMS−153
DMS−53
DMS−79
DU−4475
HUH−6−clone5
LK−2
LNCaP−Clone−FGC
MFE−280
NCI−H1048
NCI−H1092
NCI−H1105
NCI−H1436
NCI−H146
NCI−H1618
NCI−H1930
NCI−H2029
NCI−H2081
NCI−H209
NCI−H2141
NCI−H2171
NCI−H2196
NCI−H510A
NCI−H520
NCI−H526
NCI−H727
NCI−H810
NCI−H889
SCLC−21H
SW620
TT
Tissue composition
Proportion in the cluster
Overall proportion
autonomic_ganglia
biliary_tract
bone
breast
central_nervous_system
endometrium
haemato_and_lymph_tissue
kidney
large_intestine
liver
lung
oesophagus
ovary
pancreas
pleura
prostate
salivary_gland
skin
small_intestine
soft_tissue
stomach
thyroid
upper_aerodigestive_tract
urinary_tract
Molecular profile
ECM
Migration
Epithelial Phenotype
Immunity
Development
gene expression
Others
SCLC
Epithelial
Mesenchymal
Other
Drug profile
−log10 p−value
Effect
Resistant
Sensitive
Resistant
Drugs associated with the cluster (FDR< 0.05)
ERLOTINIB Resistant
LAPATINIB Resistant
LBW242 Resistant
PD0332991 Resistant
PHA665752 Resistant
PLX4720 Resistant
SORAFENIB Resistant
Other
RTK
S/T Kinase

Gastrointestinal tract Cluster - CCLE

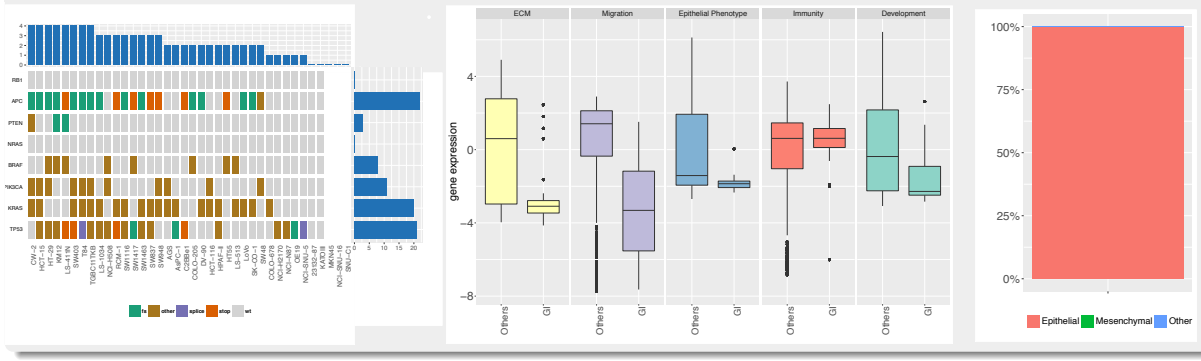
Cell lines in the cluster

23132–87	HPAF–II	NCI–H2170	SW1417
AGS	HT–29	NCI–H508	SW1463
AsPC–1	HT55	NCI–N87	SW403
C2BBe1	KATOIII	OE19	SW48
COLO–205	KM12	RCM–1	SW837
COLO–678	LS–1034	SK–CO–1	SW948
CW–2	LS–411N	NCI–SNU–16	T84
DV–90	LS–513	NCI–SNU–5	TGBC11TKB
HCT–116	LoVo	SNU–C1	
HCT–15	MKN45	SW1116	

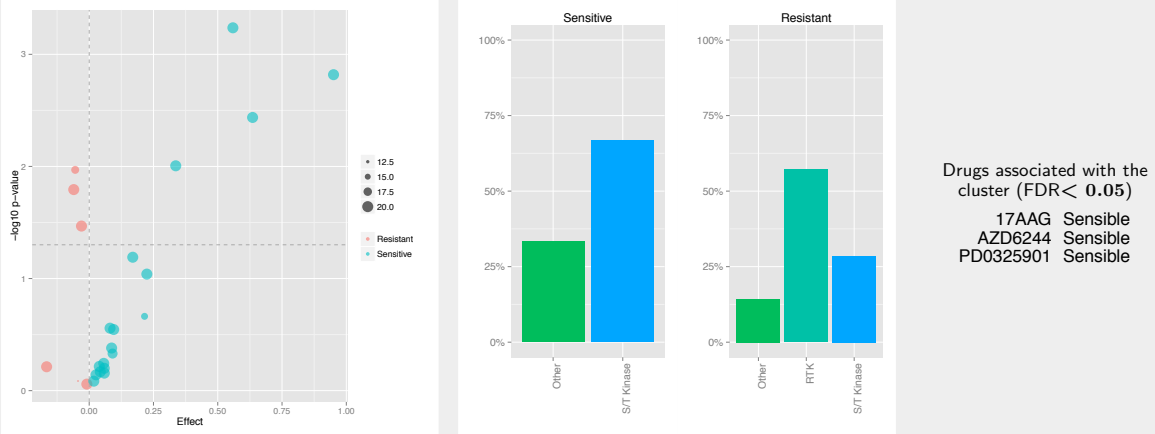
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR< **0.05**)

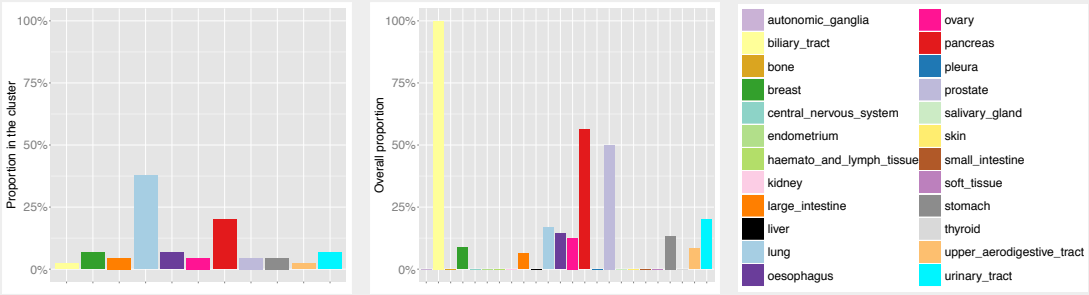
17AAG	Sensible
AZD6244	Sensible
PD0325901	Sensible

Endodermal origin Cluster - CCLE

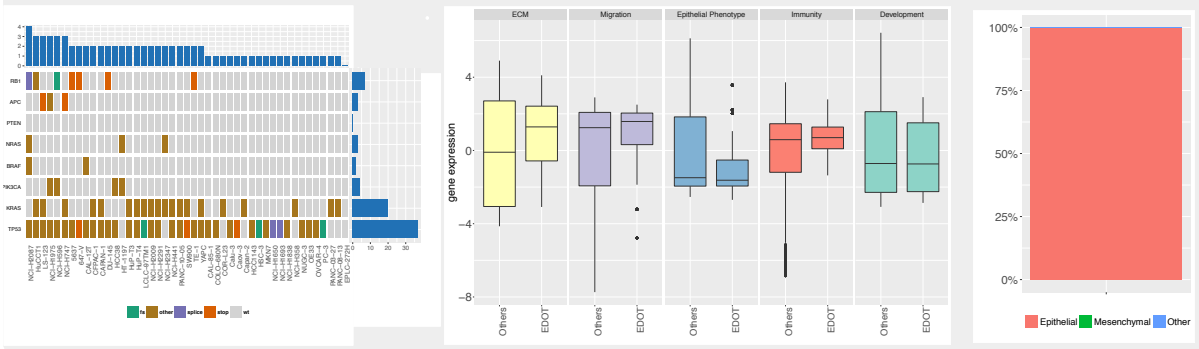
Cell lines in the cluster

5637	Capan−2	LCLC−97TM1	NCI−H2347	PANC−08−13
647−V	DU−145	LS−123	NCI−H358	PANC−10−05
CAL−12T	EPLC−272H	MKN7	NCI−H441	SW900
CAL−85−1	HCC1143	NCI−H1650	NCI−H596	TE−1
CFPAC−1	HCC38	NCI−H1693	NCI−H747	YAPC
COLO−680N	HSC−3	NCI−H1838	NUGC−3	
COR−L23	HT−1197	NCI−H1975	OE33	
Calu−3	HuP−T3	NCI−H2009	OVCAR−4	
Caov−3	HuP−T4	NCI−H2087	PC−3	
CAPAN−1	HuCCT1	NCI−H2291	PANC−03−27	

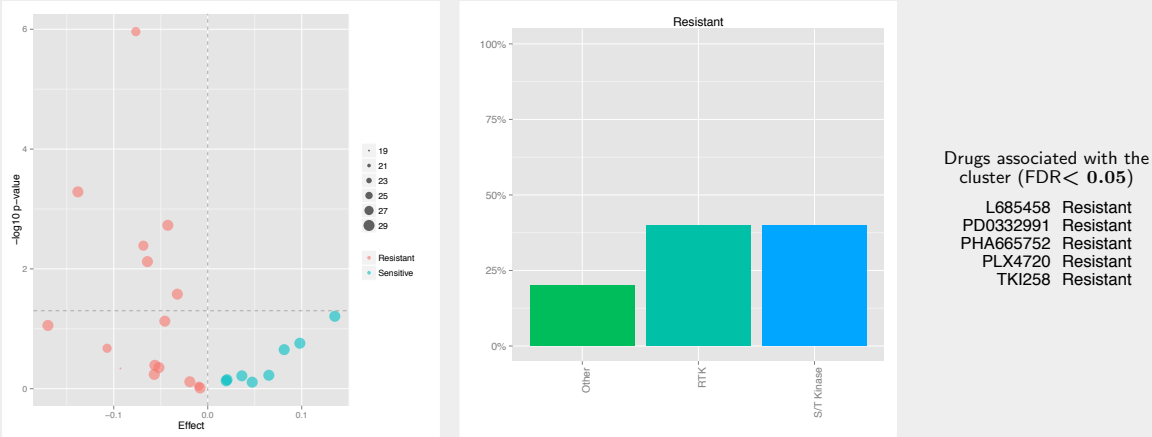
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR$<$ **0.05**)

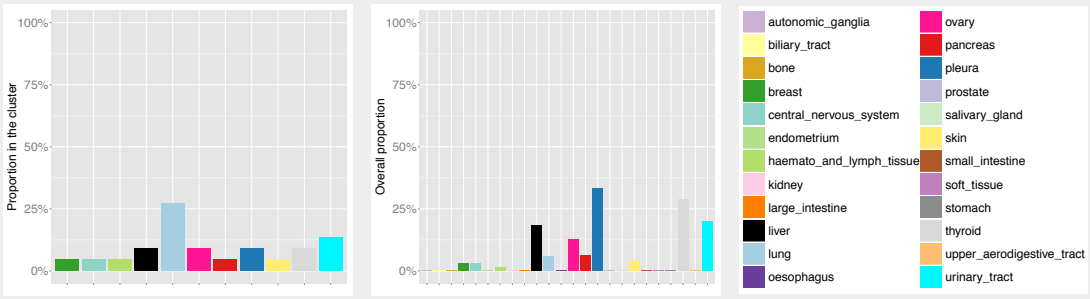
L685458	Resistant
PD0332991	Resistant
PHA665752	Resistant
PLX4720	Resistant
TKI258	Resistant

Mixed 1 Cluster - CCLE

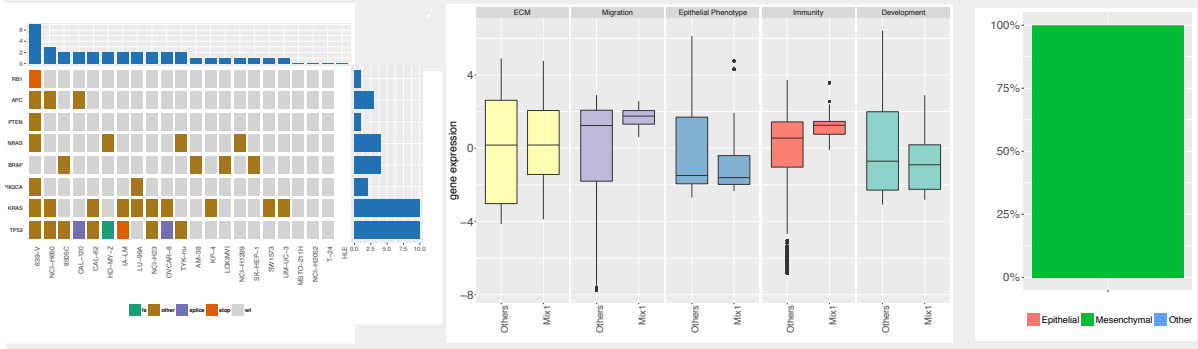
Cell lines in the cluster

639–V	LU–99A	TYK–nu
8505C	MSTO–211H	UM–UC–3
AM–38	NCI–H1299	
CAL–120	NCI–H2052	
CAL–62	NCI–H23	
HD–MY–Z	NCI–H650	
HLE	OVCAR–8	
IA–LM	SK–HEP–1	
KP–4	SW1573	
LOXIMVI	T–24	

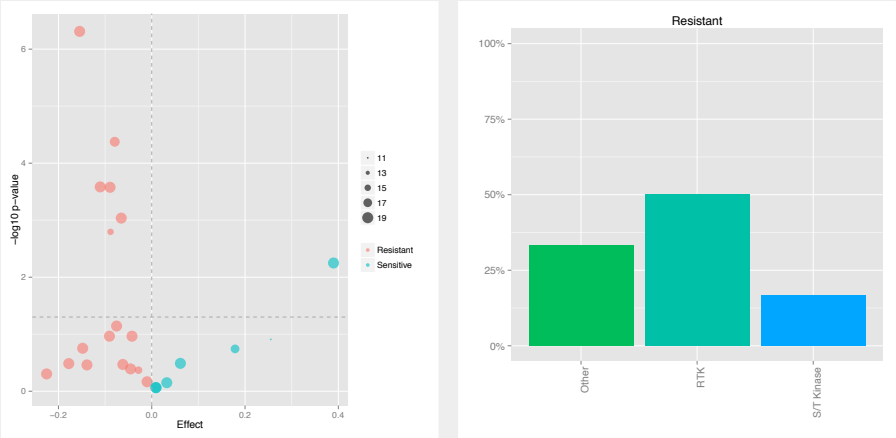
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR< **0.05**)

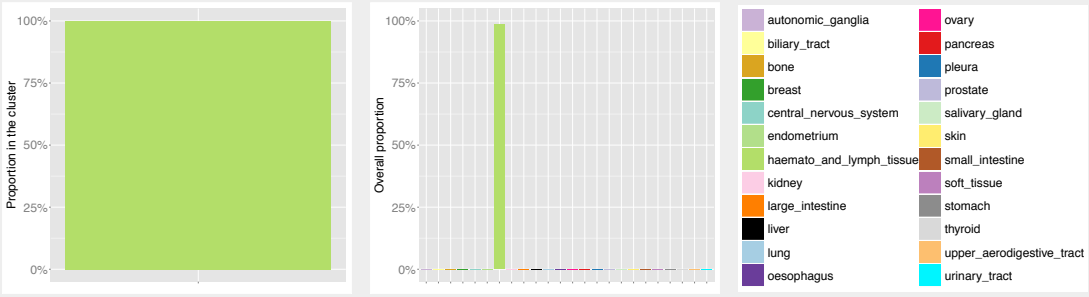
ERLOTINIB	Resistant
L685458	Resistant
LAPATINIB	Resistant
LBW242	Resistant
NILOTINIB	Resistant
VANDETANIB	Resistant

Haematopoietic and Lymphoid tissue Cluster - CCLE

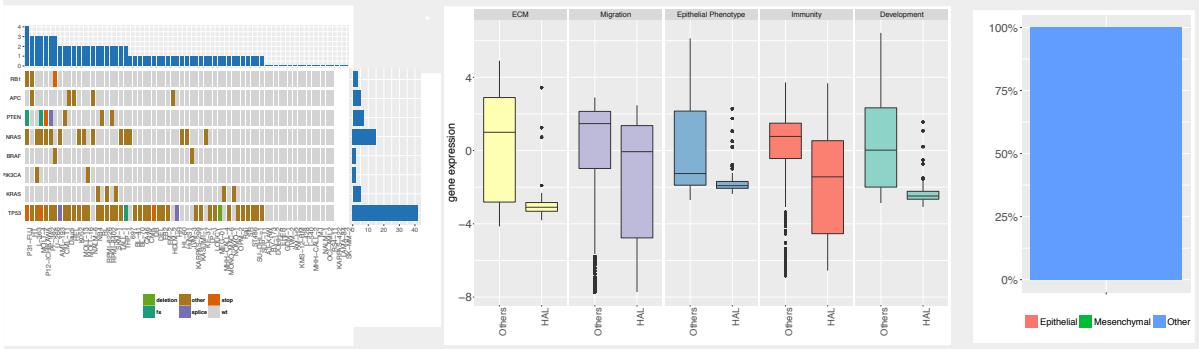
Cell lines in the cluster

697	DEL	HL–60	KMS–12–BM	MHH–CALL–4	OCI–AML2	SK–MM–2
A3–KAW	DOHH–2	HT	K052	MJ	OPM–2	SKM–1
AML–193	Daudi	HuNS1	L–363	MOLT–13	P12–ICHIKAWA	SR
BL–41	EB2	JVM–2	L–428	MOLT–16	P31–FUJ	ST486
BL–70	EHEB	JVM–3	L–540	MOLT–4	PF–382	SU–DHL–1
BV–173	EM–2	KARPAS–299	LAMA–84	MONO–MAC–6	RL	SUP–T1
CA46	GDM–1	KARPAS–422	LP–1	NALM–1	RPMI–8226	TALL–1
CMK	HDLM–2	KASUMI–1	LOUCY	NALM–6	RPMI–8402	THP–1
CML–T1	HEL	KE–37	MEG–01	NB4	RS4–11	U–266
DB	HH	KM–H2	MHH–CALL–2	NOMO–1	Raji	

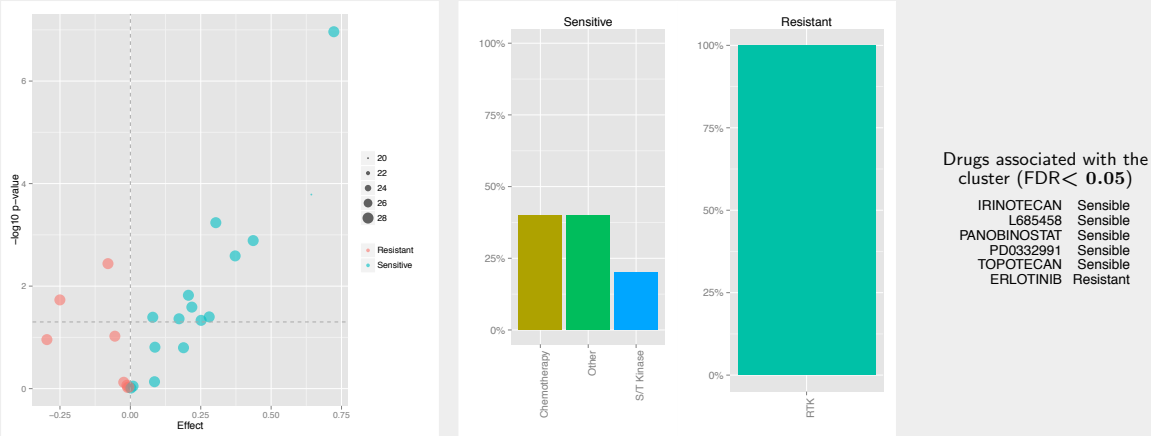
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR< **0.05**)

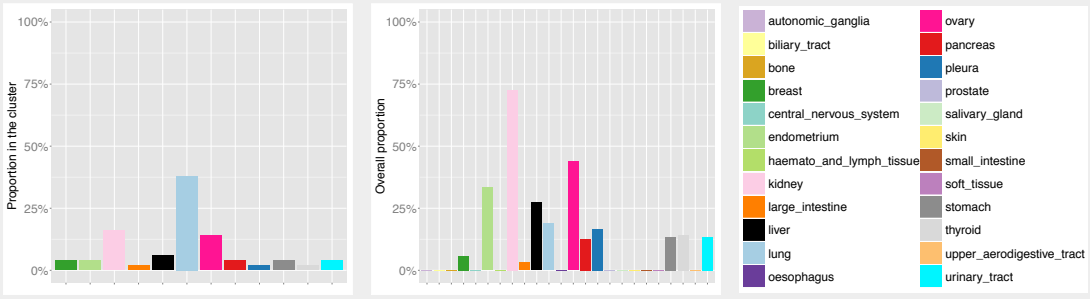
IRINOTECAN	Sensible
L685458	Sensible
PANOBINOSTAT	Sensible
PD0332991	Sensible
TOPOTECAN	Sensible
ERLOTINIB	Resistant

Mixed 2 Cluster - CCLE

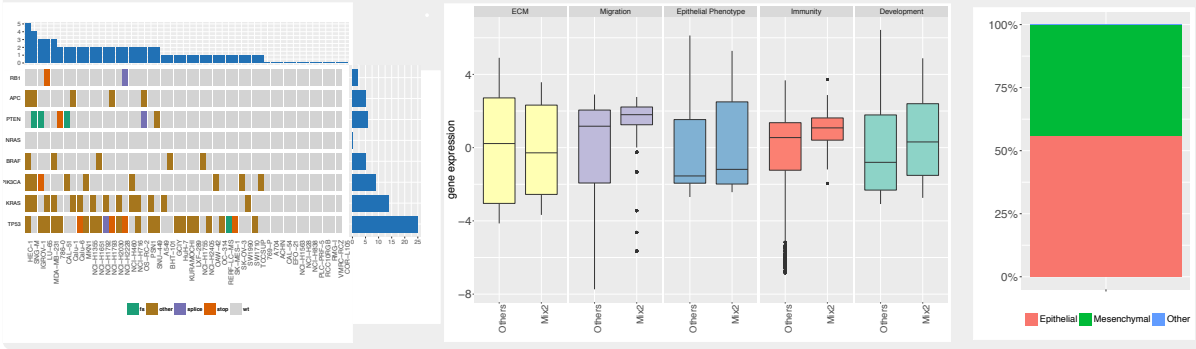
Cell lines in the cluster

769–P	Calu–6	MKN1	NCI–H28	RERF–LC–MS
786–0	EFO–21	NCI–H1355	NCI–H460	RMG–I
A704	GCIY	NCI–H1563	NCI–H716	SK–MES–1
A549	HEC–1	NCI–H1651	NCI–H838	SK–OV–3
ACHN	HuH–7	NCI–H1755	OAW–42	SNG–M
BHT–101	IGROV–1	NCI–H1792	OC–314	SNU–449
CAL–51	KURAMOCHI	NCI–H1793	OS–RC–2	SW1990
CAL–54	LU–65	NCI–H2030	PLC–PRF–5	SW1710
COR–L105	LXF–289	NCI–H2228	PSN1	TCCSUP
Calu–1	MDA–MB–231	NCI–H2405	RCC10RGB	VMRC–RCZ

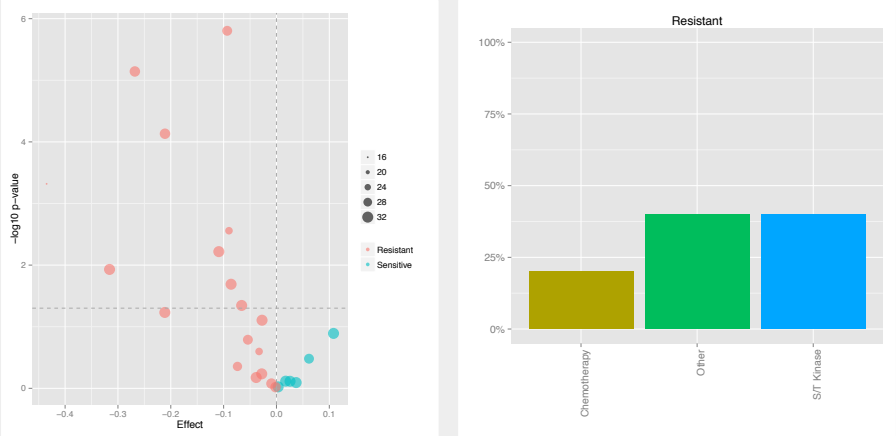
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR< **0.05**)

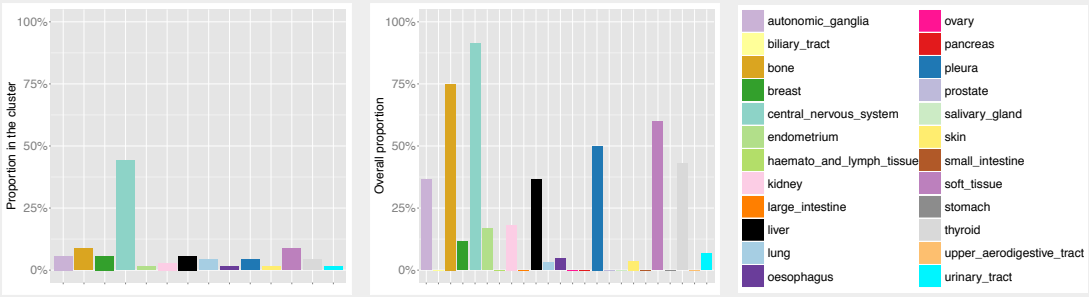
AZD6244	Resistant
IRINOTECAN	Resistant
L685458	Resistant
NILOTINIB	Resistant
PANOBINOSTAT	Resistant

Glioma and Sarcoma Cluster - CCLE

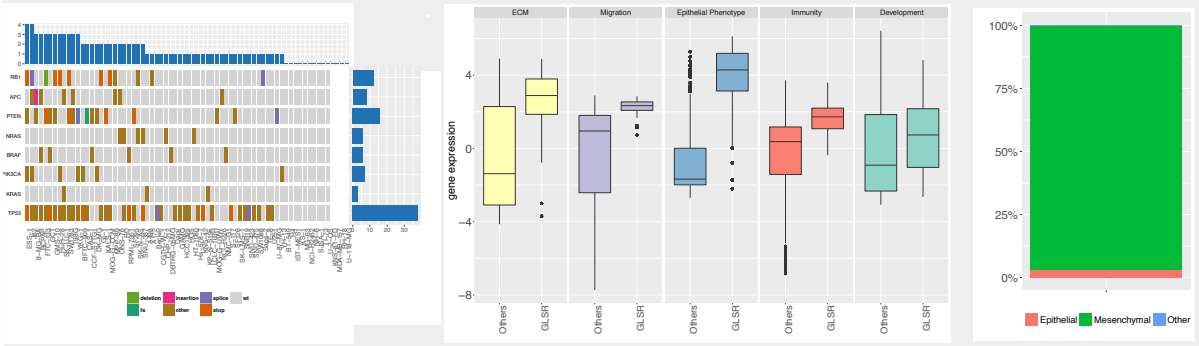
Cell lines in the cluster

8–MG–BA	CHP–212	H4	KNS–81–FD	NCI–H2452	SJSA–1	Saos–2
A498	DBTRG–05MG	HCC1395	KP–N–S19s	NH–6	SK–LMS–1	TI–73
A172	DK–MG	HOS	KS–1	NMC–G1	SK–LU–1	T98G
BCPAP	Daoy	HT–1080	LCLC–103H	ONS–76	SK–N–AS	TE–8
BFTC–909	ESS–1	Hs–578–T	MDA–MB–157	RD	SNB19	U–118–MG
BT–549	FTC–133	HuH–28	MG–63	RH–18	SNU–387	U–2–OS
Becker	GAMG	IST–MES1	MOG–G–CCM	RPMI–7951	SNU–423	U251
CAS–1	GCT	J82	MOG–G–UVW	S–117	SNU–475	U–87–MG
CCF–STTG1	GI–1	KALS–1	MPP–89	SF295	SW1088	YH–13
CGTH–W–1	GMS–10	KNS–42	NCI–H226	SF126	SW1783	YKG–1

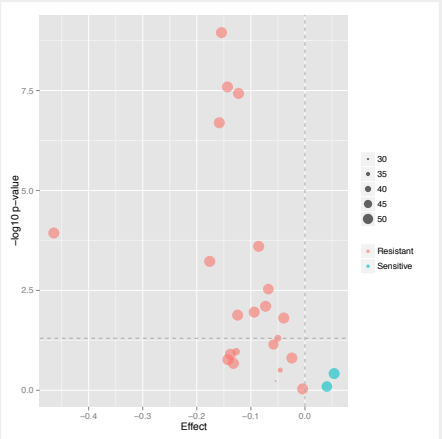
Tissue composition

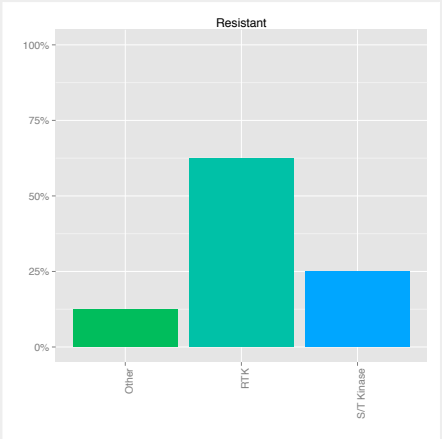


Molecular profile



Drug profile





Drugs associated with the cluster (FDR< **0.05**)

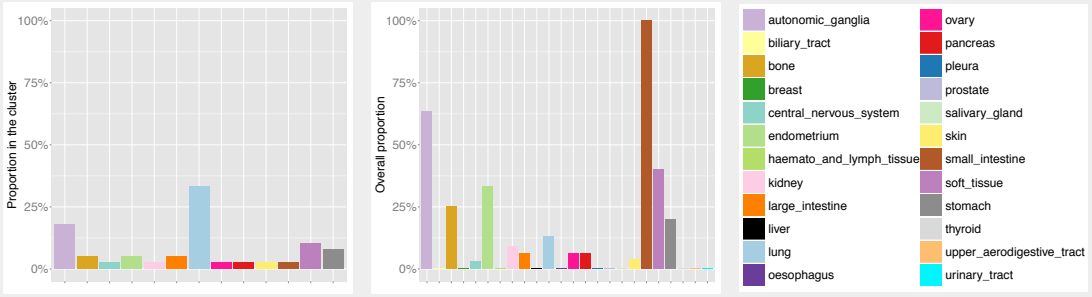
AZD0530	Resistant
ERLOTINIB	Resistant
LAPATINIB	Resistant
PANOBINOSTAT	Resistant
PD0325901	Resistant
PLX4720	Resistant
SORAFENIB	Resistant
VANDETANIB	Resistant

Mixed 3 Cluster - CCLE

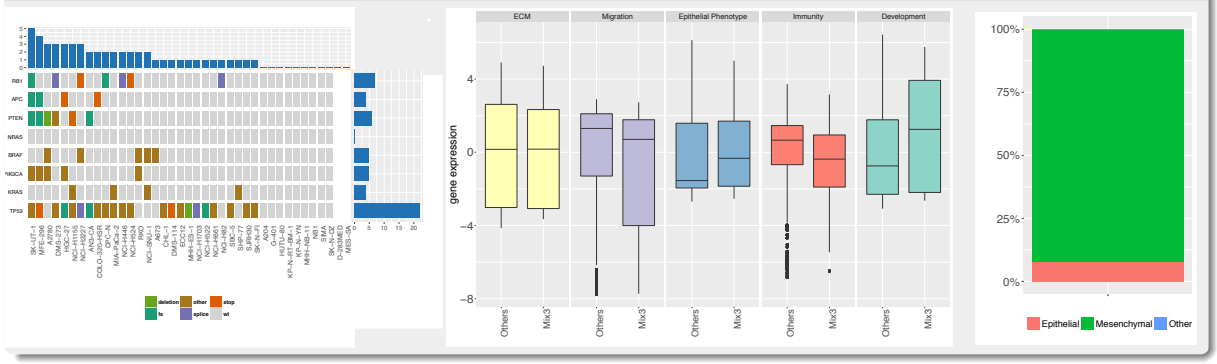
Cell lines in the cluster

A204	ECC12	MIA–PaCa–2	RKO
A673	G–401	NB1	SBC–5
A2780	HGC–27	NCI–H1155	SHP–77
AN3–CA	HUTU–80	NCI–H1703	SIMA
CHL–1	KP–N–RT–BM–1	NCI–H2227	SJRH30
COLO–320–HSR	KP–N–YN	NCI–H446	SK–N–DZ
CPC–N	MES–SA	NCI–H522	SK–N–FI
D–283MED	MFE–296	NCI–H524	SK–UT–1
DMS–114	MHH–ES–1	NCI–H661	NCI–SNU–1
DMS–273	MHH–NB–11	NCI–H82	

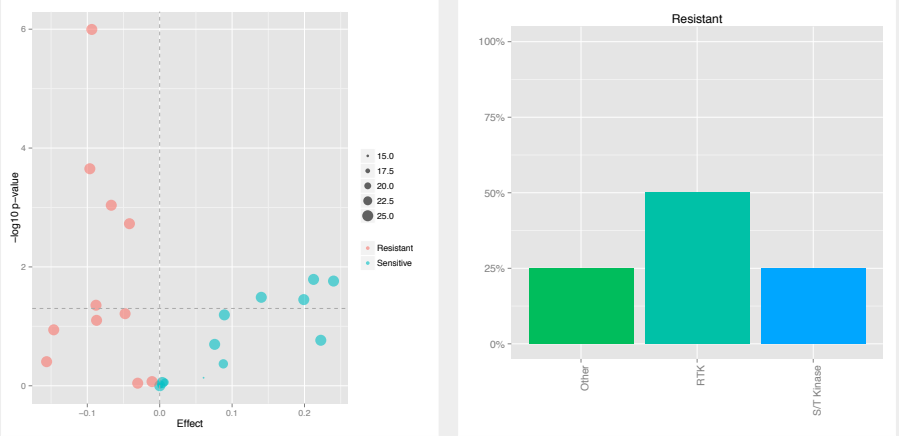
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR< **0.05**)

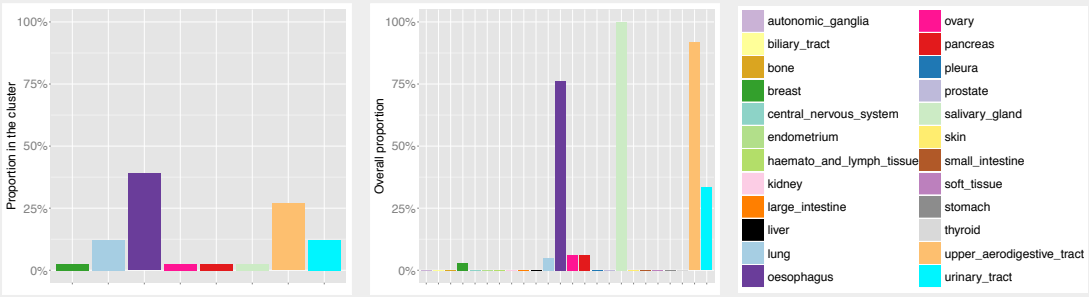
ERLOTINIB	Resistant
LBW242	Resistant
PHA665752	Resistant
PLX4720	Resistant

Aerodigestive tract Cluster - CCLE

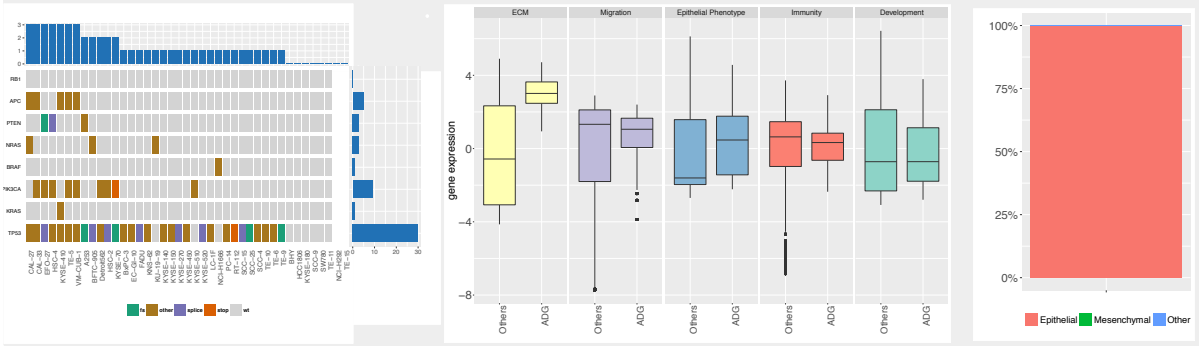
Cell lines in the cluster

A253	HCC1806	KYSE−450	SCC−25	VM−CUB−1
BFTC−905	HSC−2	KYSE−510	SCC−4	
BHY	HSC−4	KYSE−520	SCC−9	
BxPC−3	KNS−62	KYSE−70	SW780	
CAL−27	KU−19−19	LC−1F	TE−10	
CAL−33	KYSE−140	NCI−H1666	TE−11	
Detroit562	KYSE−150	NCI−H292	TE−15	
EC−GI−10	KYSE−180	PC−14	TE−5	
EFO−27	KYSE−270	RT−112	TE−6	
FADU	KYSE−410	SCC−15	TE−9	

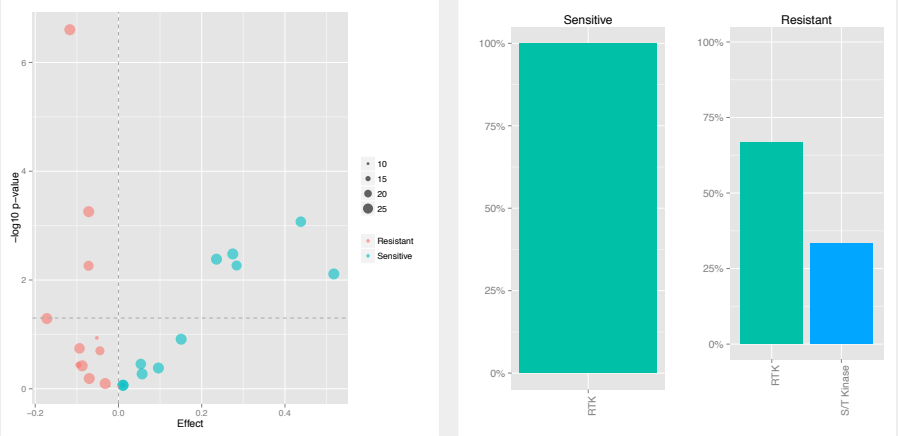
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR< **0.05**)

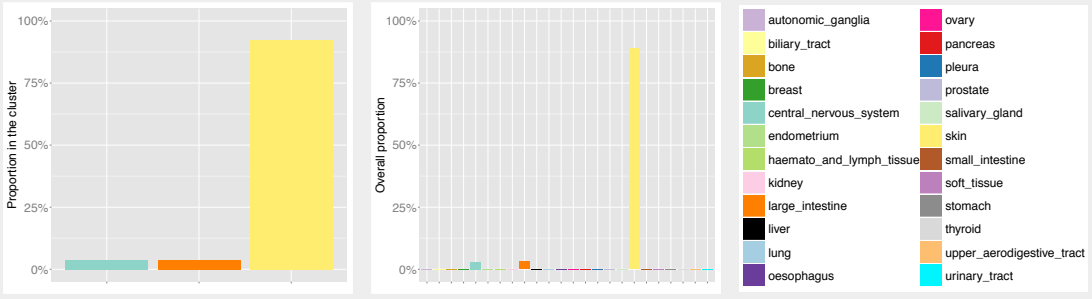
AZD0530	Sensible
ERLOTINIB	Sensible
LAPATINIB	Sensible
VANDETANIB	Sensible
CRIZOTINIB	Resistant
PLX4720	Resistant
SORAFENIB	Resistant

Skin Cluster - CCLE

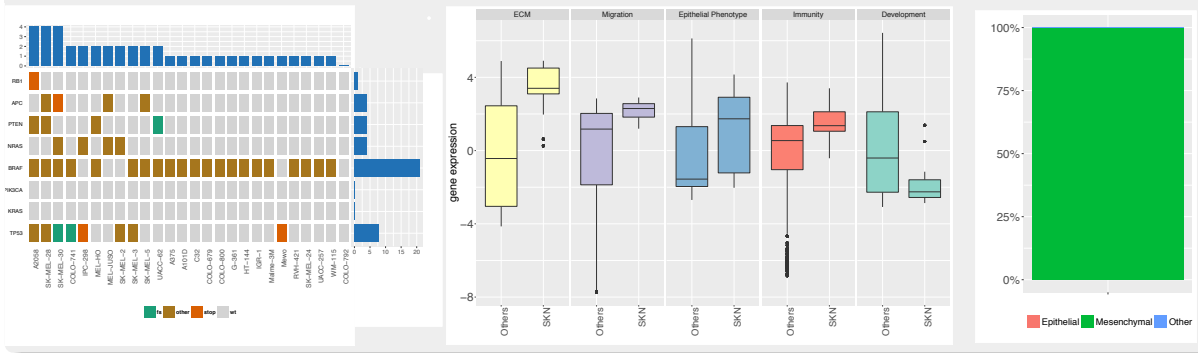
Cell lines in the cluster

A375	IGR−1	SK−MEL−3
A101D	IPC−298	SK−MEL−30
A2058	MEL−HO	SK−MEL−5
C32	MEL−JUSO	UACC−257
COLO−741	Malme−3M	UACC−62
COLO−792	Mewo	WM−115
COLO−679	RVH−421	
COLO−800	SK−MEL−2	
G−361	SK−MEL−24	
HT−144	SK−MEL−28	

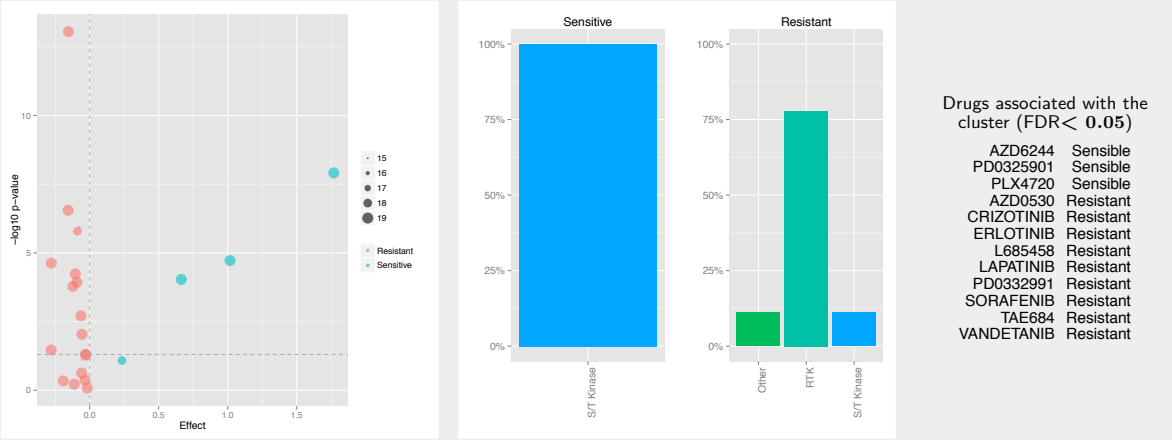
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR< **0.05**)

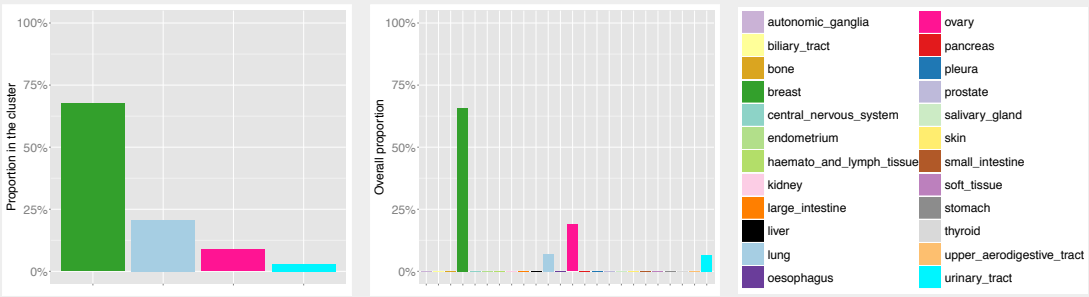
AZD6244	Sensible
PD0325901	Sensible
PLX4720	Sensible
AZD0530	Resistant
CRIZOTINIB	Resistant
ERLOTINIB	Resistant
L685458	Resistant
LAPATINIB	Resistant
PD0332991	Resistant
SORAFENIB	Resistant
TAE684	Resistant
VANDETANIB	Resistant

Breast Cluster - CCLE

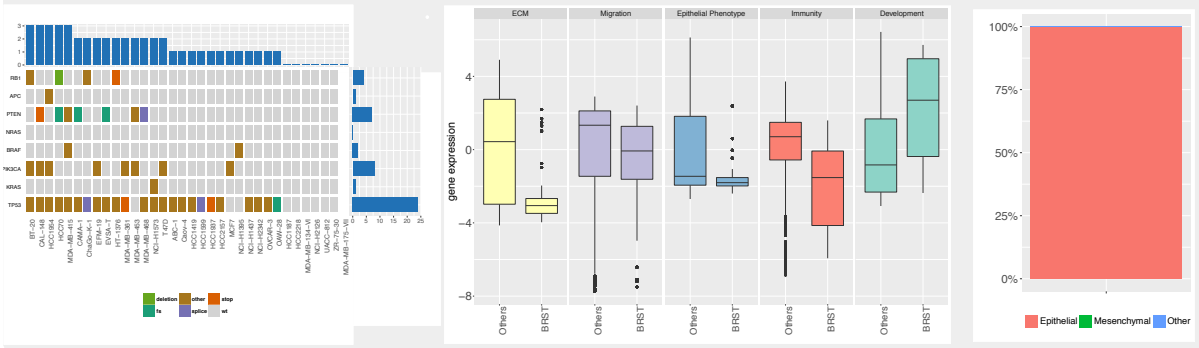
Cell lines in the cluster

ABC–1	HCC1599	MDA–MB–361	OAW–28
BT–20	HCC1937	MDA–MB–415	T47D
CAL–148	HCC1954	MDA–MB–453	UACC–812
CAMA–1	HCC2157	MDA–MB–468	ZR–75–30
Caov–4	HCC2218	NCI–H1395	
ChaGo–K–1	HCC70	NCI–H1437	
EFM–19	HT–1376	NCI–H1573	
EVSA–T	MCF7	NCI–H2126	
HCC1187	MDA–MB–134–VI	NCI–H2342	
HCC1419	MDA–MB–175–VII	OVCAR–3	

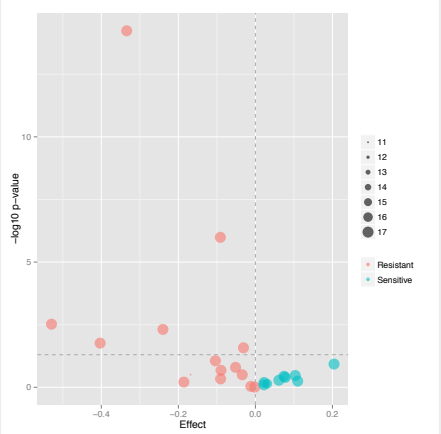
Tissue composition

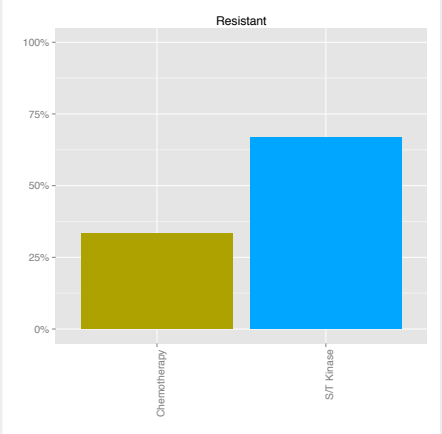


Molecular profile



Drug profile





Drugs associated with the cluster (FDR< **0.05**)

AZD6244	Resistant
PLX4720	Resistant
TOPOTECAN	Resistant

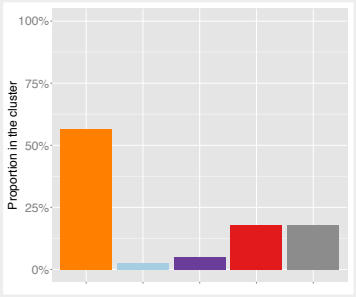
Figure A.22: Summary of each cell line cluster with GDSC data.

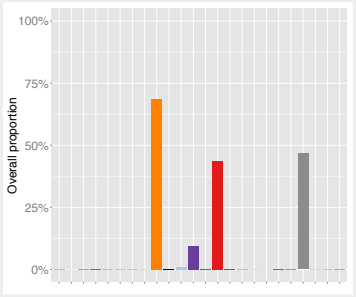
Gastrointestinal tract Cluster - GDSC

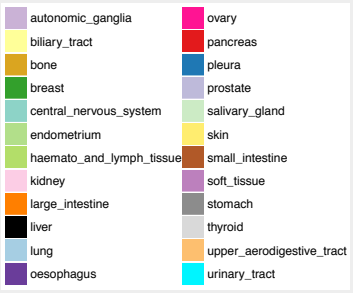
Cell lines in the cluster

23132–87	HPAF–II	MKN45	SNU–C1
AGS	HT–29	NCI–H508	SW1116
CFPAC–1	HT55	NCI–H747	SW1417
COLO–205	HuP–T4	NCI–N87	SW1463
COLO–678	KATOIII	OE19	SW837
CW–2	KM12	OE33	SW948
CAPAN–1	LS–1034	PANC–08–13	T84
Capan–2	LS–411N	RCM–1	TGBC11TKB
DV–90	LS–513	SK–CO–1	YAPC
HCT–15	LoVo	NCI–SNU–16	

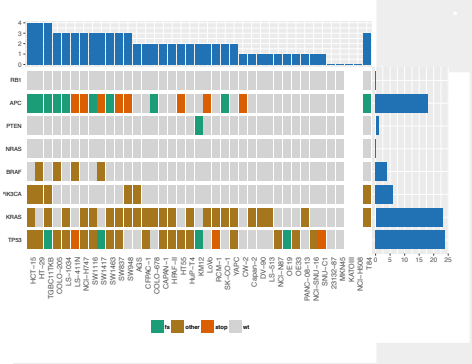
Tissue composition

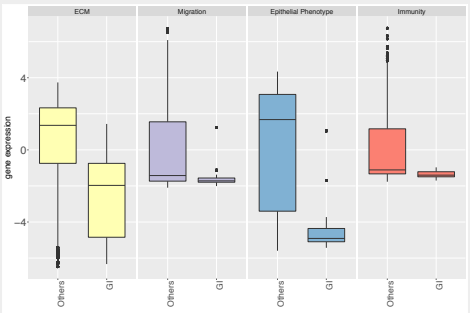


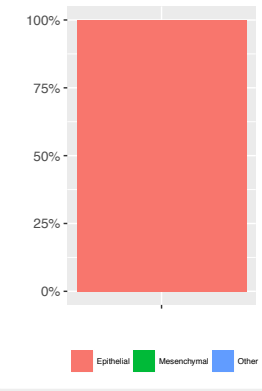




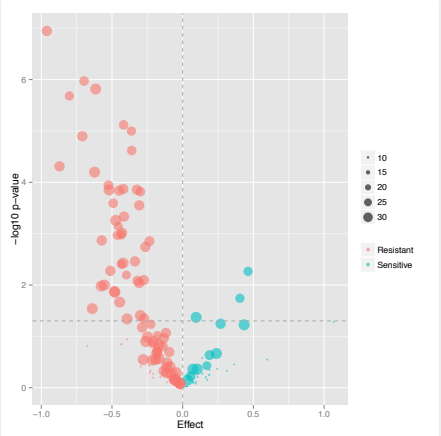
Molecular profile

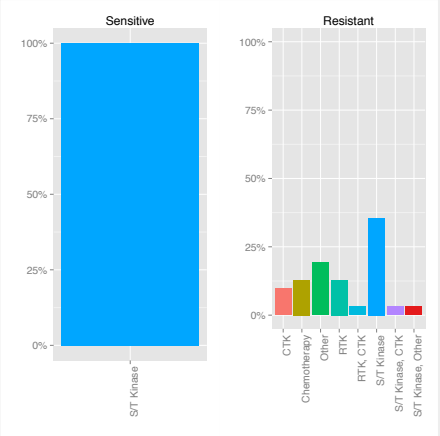






Drug profile





Drugs associated with the cluster (FDR$<$ **0.05**)

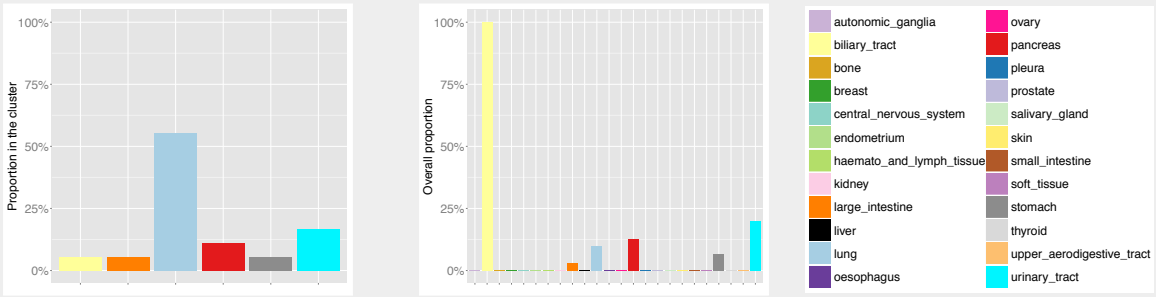
PD0325901	Sensible
AP–24534	Resistant
AZD6482	Resistant
DMOG	Resistant
IPA–3	Resistant
PAZOPANIB	Resistant
CAMPTOTHECIN–WTSI	Resistant
VINBLASTINE	Resistant
CISPLATIN	Resistant
DOCETAXEL	Resistant
ATRA	Resistant
NILOTINIB	Resistant
TEMSIROLIMUS	Resistant
AZD–2281	Resistant
BOSUTINIB	Resistant
AXITINIB	Resistant
CEP–701	Resistant
AMG–706	Resistant
ELESCLOMOL	Resistant
PLX4720	Resistant
BX–795	Resistant
SL–0101–1	Resistant
BI–D1870	Resistant
JNK–INHIBITOR–VIII	Resistant
681640	Resistant
PD–173074	Resistant
ZM–447439	Resistant
RO–3306	Resistant
PD0332991	Resistant
NVP–BEZ235	Resistant
GDC0941	Resistant
AZD8055	Resistant

Endodermal origin Cluster - GDSC

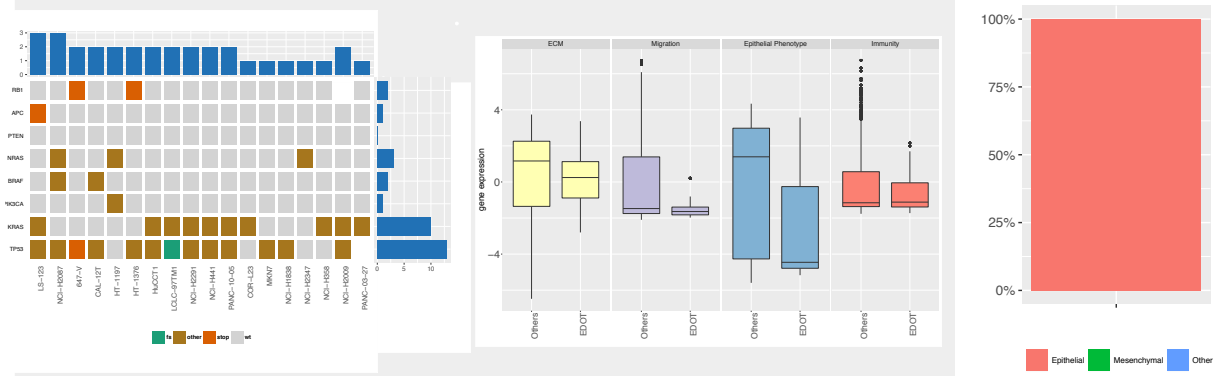
Cell lines in the cluster

647–V	NCI–H2009
CAL–12T	NCI–H2087
COR–L23	NCI–H2291
HT–1197	NCI–H2347
HT–1376	NCI–H358
HuCCT1	NCI–H441
LCLC–97TM1	PANC–03–27
LS–123	PANC–10–05
MKN7	
NCI–H1838	

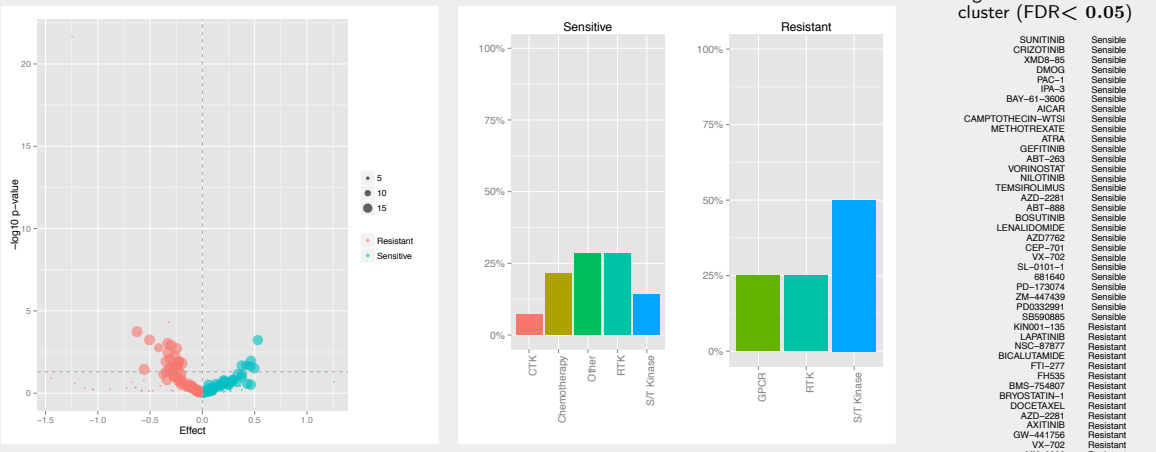
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR< **0.05**)

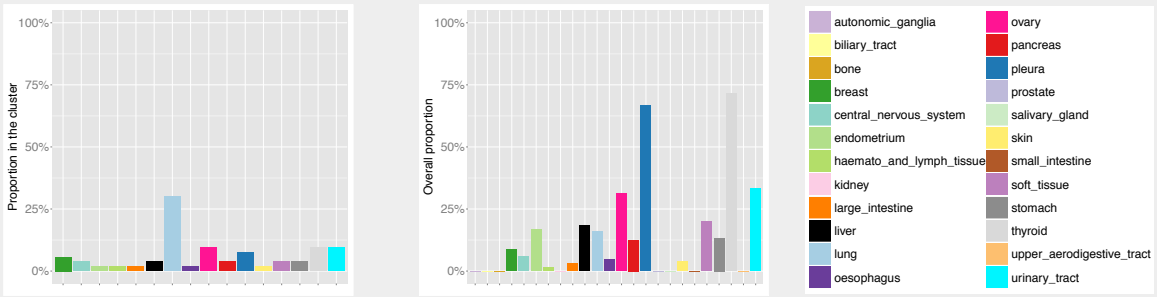
SUNITINIB	Sensible
CRIZOTINIB	Sensible
XMD8–85	Sensible
DMOG	Sensible
PAC–1	Sensible
IPA–3	Sensible
BAY–61–3606	Sensible
AICAR	Sensible
CAMPTOTHECIN–WTSI	Sensible
METHOTREXATE	Sensible
ATRA	Sensible
GEFITINIB	Sensible
ABT–263	Sensible
VORINOSTAT	Sensible
NILOTINIB	Sensible
TEMSIROLIMUS	Sensible
AZD–2281	Sensible
ABT–888	Sensible
BOSUTINIB	Sensible
LENALIDOMIDE	Sensible
AZD7762	Sensible
CEP–701	Sensible
VX–702	Sensible
SL–0101–1	Sensible
681640	Sensible
PD–173074	Sensible
ZM–447439	Sensible
PD0332991	Sensible
SB590885	Sensible
KIN001–135	Resistant
LAPATINIB	Resistant
NSC–87877	Resistant
BICALUTAMIDE	Resistant
FTI–277	Resistant
FH535	Resistant
BMS–754807	Resistant
BRYOSTATIN–1	Resistant
DOCETAXEL	Resistant
AZD–2281	Resistant
AXITINIB	Resistant
GW–441756	Resistant
VX–702	Resistant
MK–2206	Resistant

Mixed 1 Cluster - GDSC

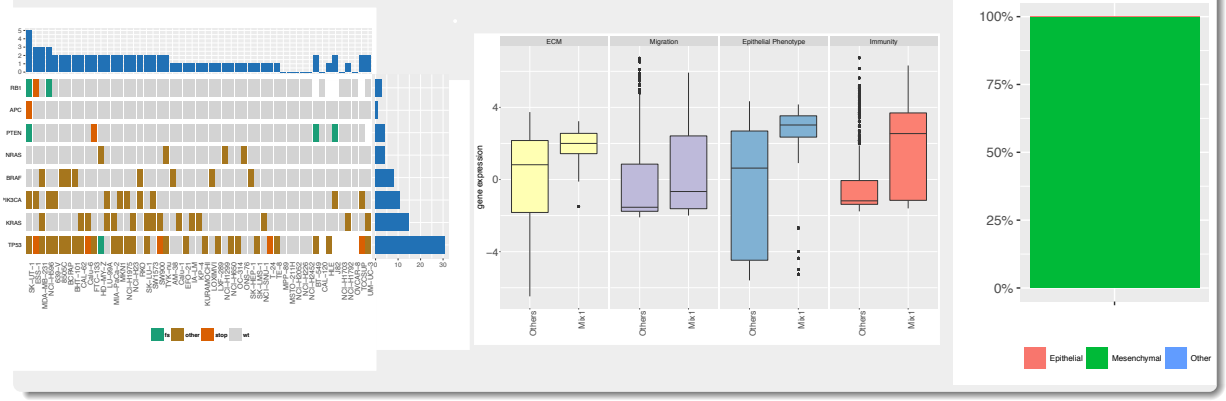
Cell lines in the cluster

639–V	EFO–21	LU–99A	NCI–H1975	RKO	TE–8
8505C	ESS–1	LXF–289	NCI–H2052	SK–HEP–1	TYK–nu
AM–38	FTC–133	MDA–MB–231	NCI–H226	SK–LMS–1	UM–UC–3
BCPAP	HD–MY–Z	MIA–PaCa–2	NCI–H23	SK–LU–1	
BHT–101	HLE	MKN1	NCI–H2452	SK–UT–1	
BT–549	IA–LM	MPP–89	NCI–H596	NCI–SNU–1	
CAL–120	J82	MSTO–211H	NCI–H650	SW1573	
CAL–62	KP–4	NCI–H1299	OC–314	SW900	
Calu–1	KURAMOCHI	NCI–H1703	ONS–76	T–24	
Calu–6	LOXIMVI	NCI–H1792	OVCAR–8	TCCSUP	

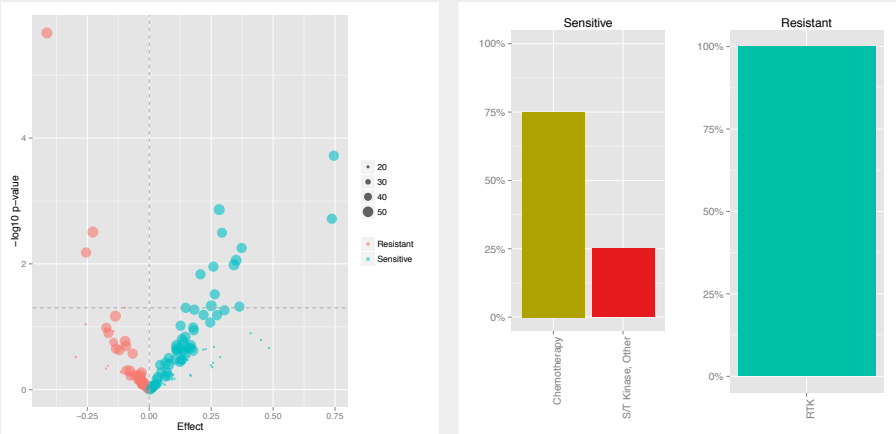
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR$<$ **0.05**)

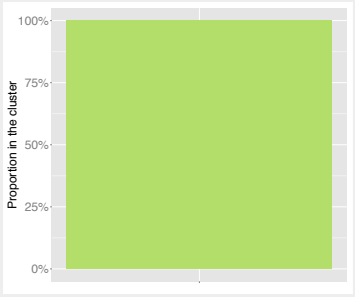
GEMCITABINE	Sensible
BLEOMYCIN	Sensible
VINBLASTINE	Sensible
NVP–BEZ235	Sensible
GEFITINIB	Resistant
BIBW2992	Resistant

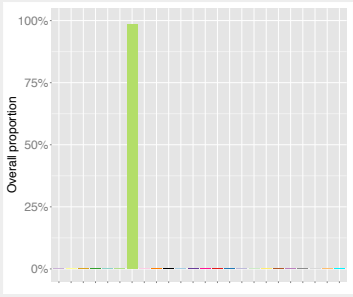
Haematopoietic and Lymphoid tissue Cluster - GDSC

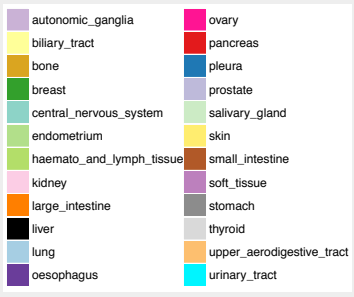
Cell lines in the cluster

697	DEL	HL–60	KMS–12–BM	MHH–CALL–4	OCI–AML2	SK–MM–2
A3–KAW	DOHH–2	HT	K052	MJ	OPM–2	SKM–1
AML–193	Daudi	HuNS1	L–363	MOLT–13	P12–ICHIKAWA	SR
BL–41	EB2	JVM–2	L–428	MOLT–16	P31–FUJ	ST486
BL–70	EHEB	JVM–3	L–540	MOLT–4	PF–382	SU–DHL–1
BV–173	EM–2	KARPAS–299	LAMA–84	MONO–MAC–6	RL	SUP–T1
CA46	GDM–1	KARPAS–422	LP–1	NALM–1	RPMI–8226	TALL–1
CMK	HDLM–2	KASUMI–1	LOUCY	NALM–6	RPMI–8402	THP–1
CML–T1	HEL	KE–37	MEG–01	NB4	RS4–11	U–266
DB	HH	KM–H2	MHH–CALL–2	NOMO–1	Raji	

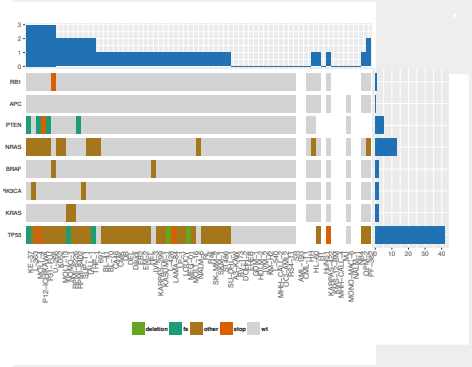
Tissue composition

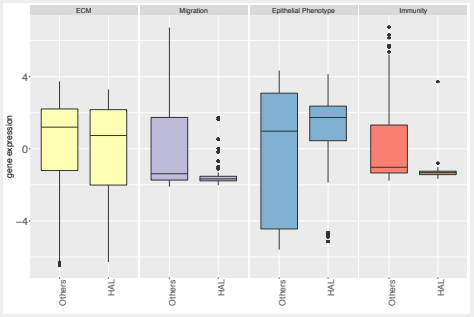


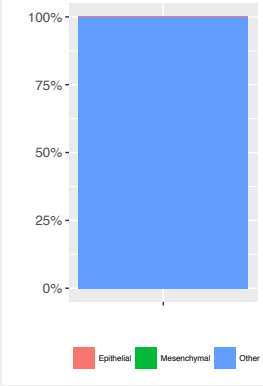




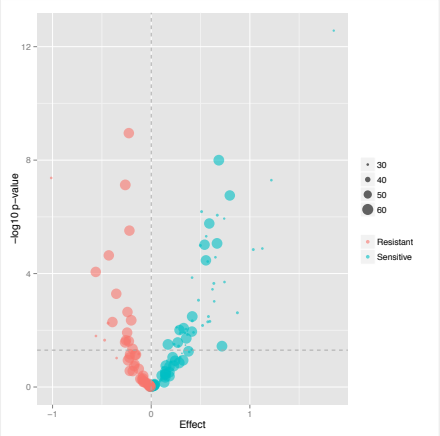
Molecular profile

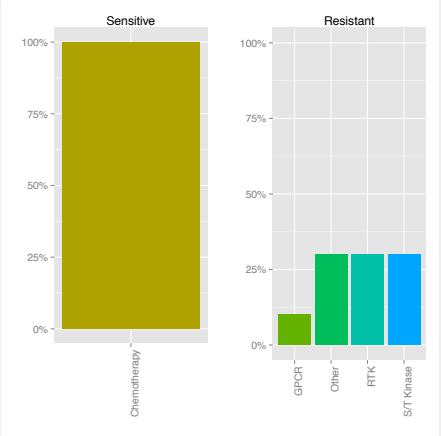






Drug profile





Drugs associated with the cluster (FDR< **0.05**)

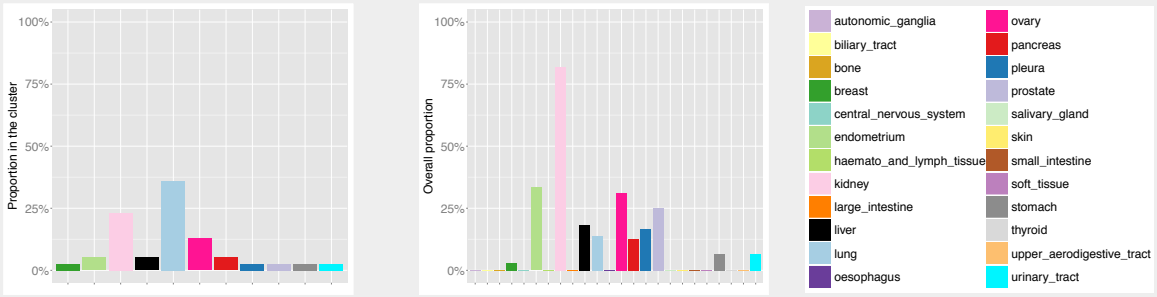
MIDOSTAURIN	Sensible
AICAR	Resistant
17AAG	Resistant

Mixed 2 Cluster - GDSC

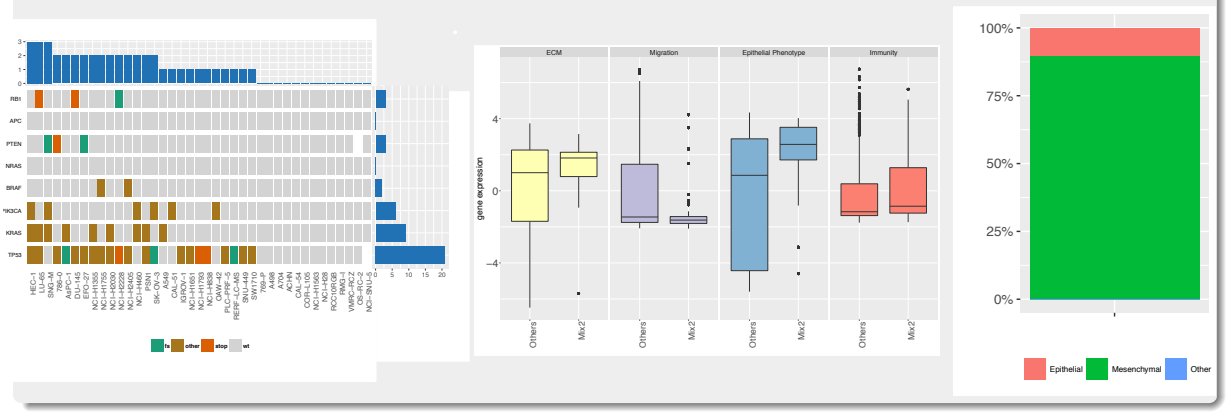
Cell lines in the cluster

769–P	DU–145	NCI–H2030	RCC10RGB
786–0	EFO–27	NCI–H2228	RERF–LC–MS
A498	HEC–1	NCI–H2405	RMG–I
A704	IGROV–1	NCI–H28	SK–OV–3
A549	LU–65	NCI–H460	SNG–M
ACHN	NCI–H1355	NCI–H838	SNU–449
AsPC–1	NCI–H1563	OAW–42	NCI–SNU–5
CAL–51	NCI–H1651	OS–RC–2	SW1710
CAL–54	NCI–H1755	PLC–PRF–5	VMRC–RCZ
COR–L105	NCI–H1793	PSN1	

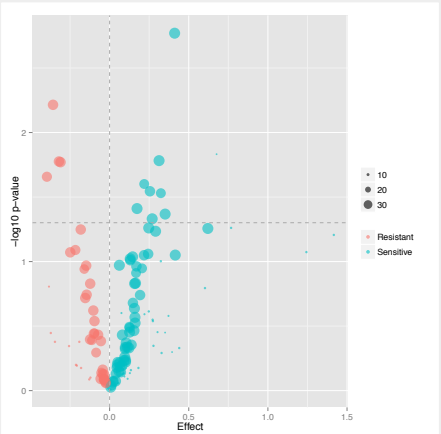
Tissue composition

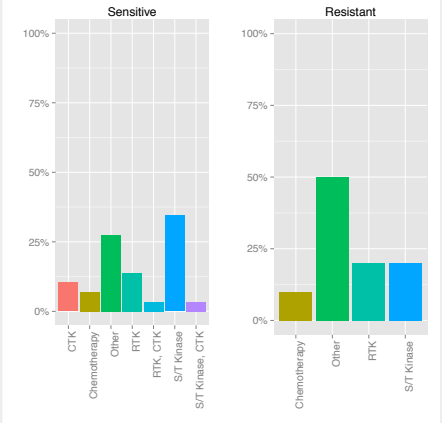


Molecular profile



Drug profile





Drugs associated with the cluster (FDR< **0.05**)

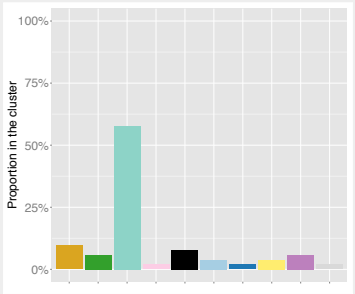
DASATINIB	Sensible
BORTEZOMIB	Sensible
MIDOSTAURIN	Sensible
CHIR–99021	Sensible
AZD6482	Sensible
BEXAROTENE	Sensible
LFM–A13	Sensible
PAZOPANIB	Sensible
VINBLASTINE	Sensible
DOCETAXEL	Sensible
TEMSIROLIMUS	Sensible
ELESCLOMOL	Sensible
BX–795	Sensible
NVP–BEZ235	Sensible
GDC0941	Sensible
PAC–1	Resistant
METHOTREXATE	Resistant
GEFITINIB	Resistant

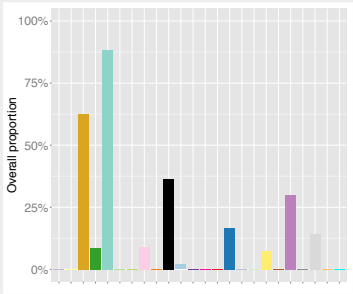
Glioma and Sarcoma Cluster - GDSC

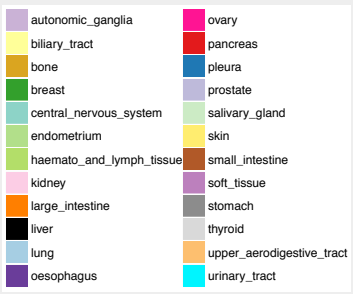
Cell lines in the cluster

8–MG–BA	GAMG	IST–MES1	NMC–G1	SNU–475	YH–13
A172	GCT	KALS–1	RPMI–7951	SW1088	YKG–1
BFTC–909	GI–1	KNS–42	S–117	SW1783	
Becker	GMS–10	KNS–81–FD	SF295	TI–73	
CAS–1	H4	KS–1	SF126	T98G	
CCF–STTG1	HCC1395	LCLC–103H	SJSA–1	U–118–MG	
CGTH–W–1	HOS	MDA–MB–157	SK–MES–1	U–2–OS	
DBTRG–05MG	HT–1080	MG–63	SNB19	U251	
DK–MG	Hs–578–T	MOG–G–CCM	SNU–387	U–87–MG	
Daoy	HuH–28	MOG–G–UVW	SNU–423	WM–115	

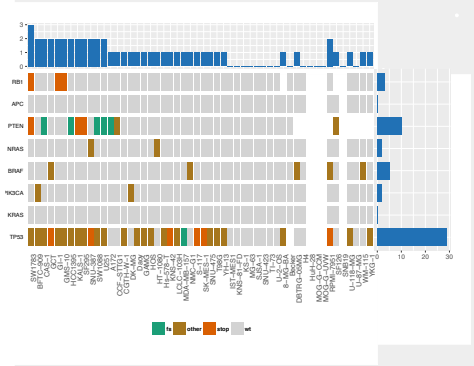
Tissue composition

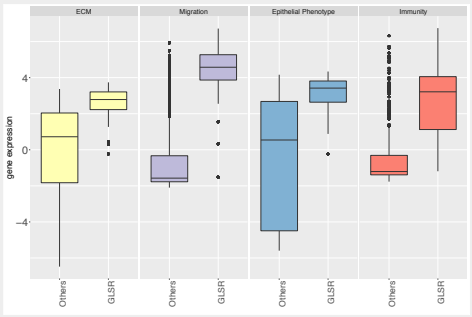


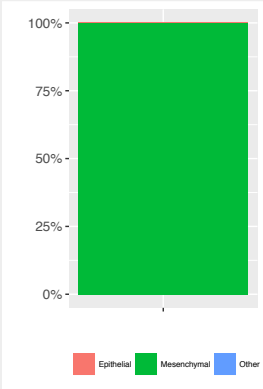




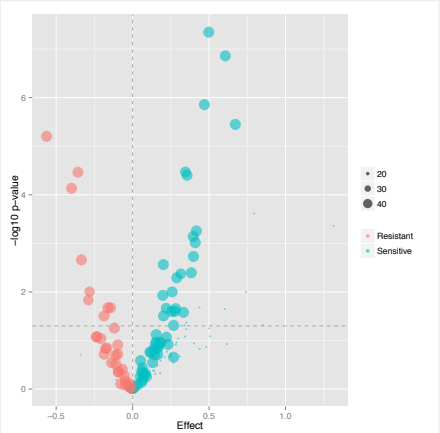
Molecular profile

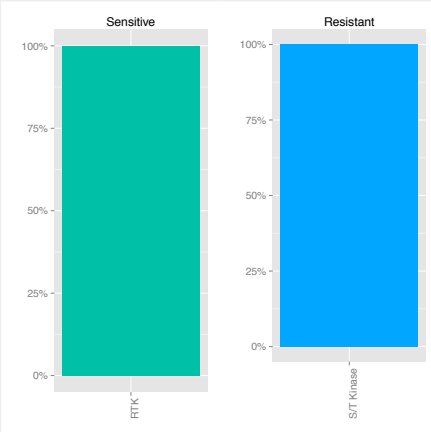






Drug profile





Drugs associated with the cluster (FDR< **0.05**)

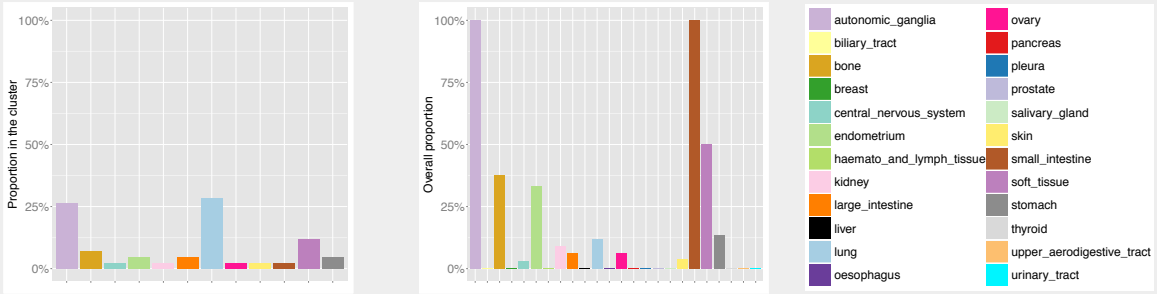
VORINOSTAT	Sensible
Z–LLNLE–CHO–	Resistant
DASATINIB	Resistant
BICALUTAMIDE	Resistant
BIBW2992	Resistant

Mixed 3 Cluster - GDSC

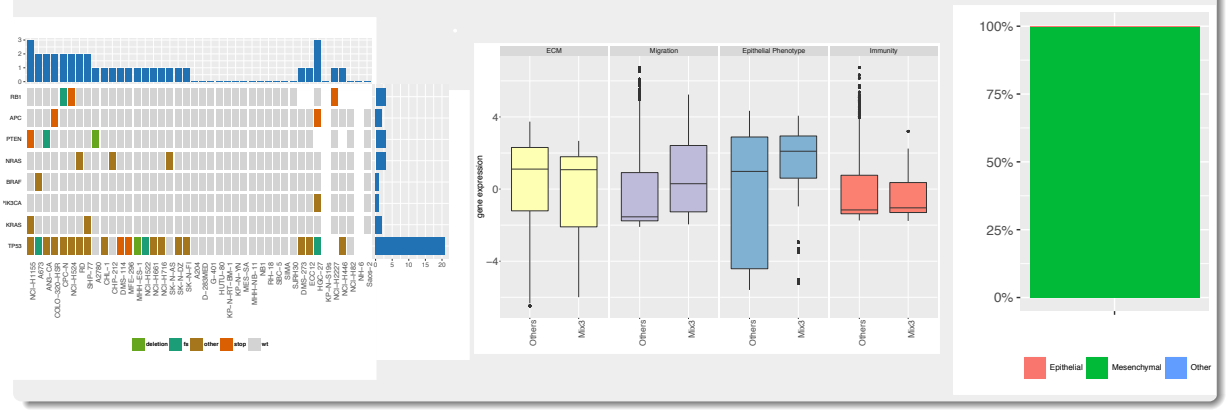
Cell lines in the cluster

A204	DMS–273	MHH–ES–1	NCI–H82	SK–N–FI
A673	ECC12	MHH–NB–11	NH–6	Saos–2
A2780	G–401	NB1	RD	
AN3–CA	HGC–27	NCI–H1155	RH–18	
CHL–1	HUTU–80	NCI–H2227	SBC–5	
CHP–212	KP–N–RT–BM–1	NCI–H446	SHP–77	
COLO–320–HSR	KP–N–S19s	NCI–H522	SIMA	
CPC–N	KP–N–YN	NCI–H524	SJRH30	
D–283MED	MES–SA	NCI–H661	SK–N–AS	
DMS–114	MFE–296	NCI–H716	SK–N–DZ	

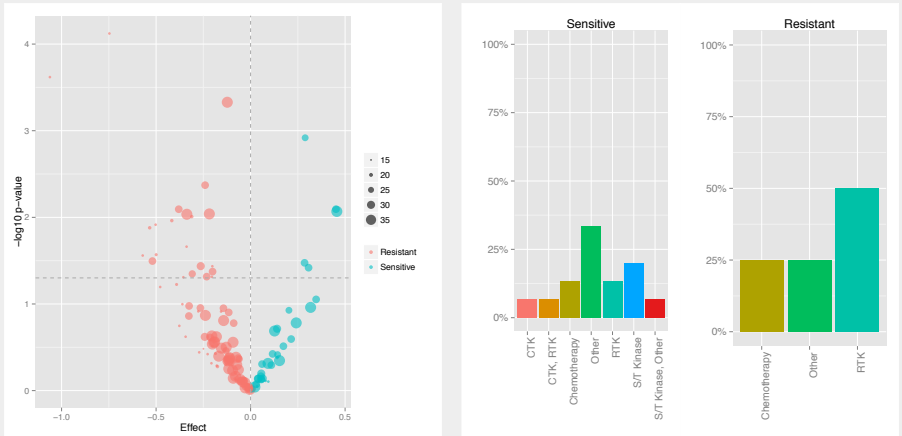
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR< 0.05)

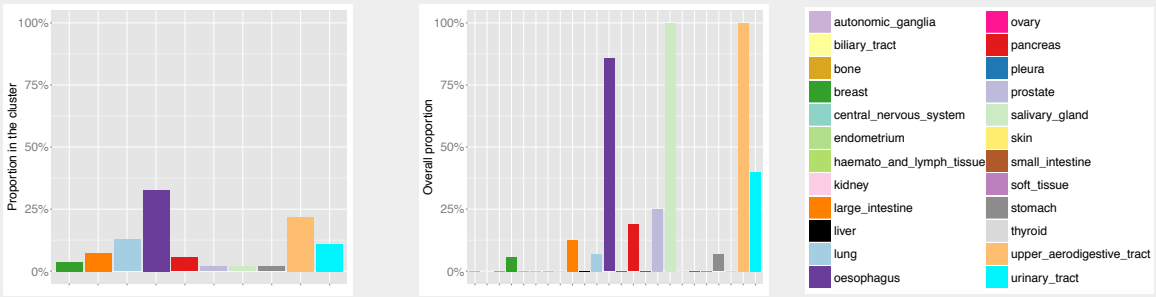
ERLOTINIB	Sensible
LAPATINIB	Sensible
VINORELBINE	Sensible
BICALUTAMIDE	Sensible
PF–562271	Sensible
OBATOCLAX–MESYLATE	Sensible
EPOTHILONE–B	Sensible
AICAR	Sensible
VINBLASTINE	Sensible
DOCETAXEL	Sensible
GEFITINIB	Sensible
BOSUTINIB	Sensible
17AAG	Sensible
BIBW2992	Sensible
CYCLOPAMINE	Resistant
GSK269962A	Resistant
PAZOPANIB	Resistant
SB–216763	Resistant

Aerodigestive tract Cluster - GDSC

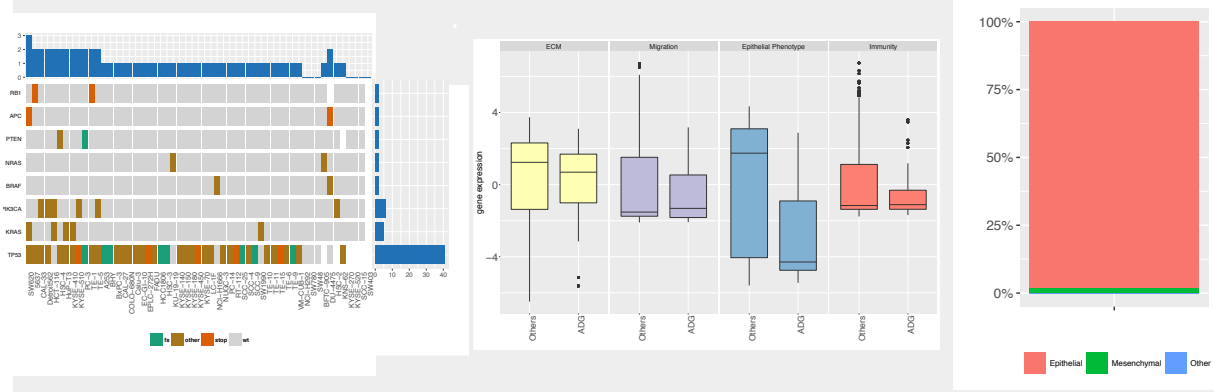
Cell lines in the cluster

5637	Detroit562	KNS–62	KYSE–70	SCC–4	TE–15
A253	EC–GI–10	KU–19–19	LC–1F	SCC–9	TE–5
BFTC–905	EPLC–272H	KYSE–140	NCI–H1666	SW1990	TE–6
BHY	FADU	KYSE–150	NCI–H292	SW780	TE–9
BxPC–3	HCC1806	KYSE–180	NUGC–3	SW403	VM–CUB–1
CAL–27	HCT–116	KYSE–270	PC–14	SW48	
CAL–33	HSC–2	KYSE–410	PC–3	SW620	
COLO–680N	HSC–3	KYSE–450	RT–112	TE–1	
Calu–3	HSC–4	KYSE–510	SCC–15	TE–10	
DU–4475	HuP–T3	KYSE–520	SCC–25	TE–11	

Tissue composition



Molecular profile



Drug profile

Drugs associated with the cluster (FDR< **0.05**)

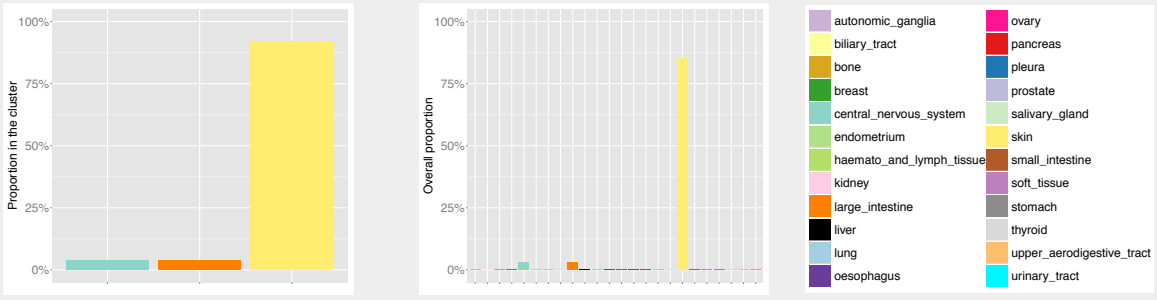
DOXORUBICIN	Sensible
CYCLOPAMINE	Resistant
CRIZOTINIB	Resistant
PYRIMETHAMINE	Resistant
ATRA	Resistant
TEMSIROLIMUS	Resistant
VX–702	Resistant

Skin Cluster - GDSC

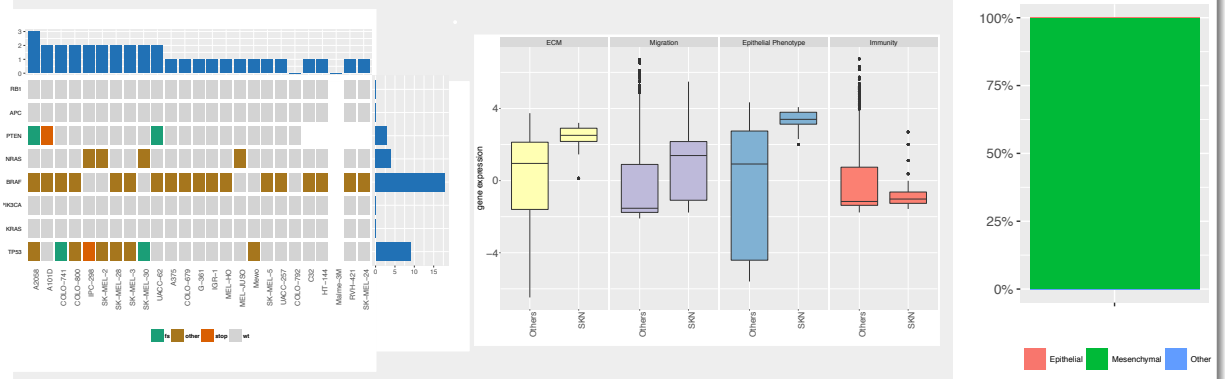
Cell lines in the cluster

A375	IGR−1	SK−MEL−3
A101D	IPC−298	SK−MEL−30
A2058	MEL−HO	SK−MEL−5
C32	MEL−JUSO	UACC−257
COLO−741	Malme−3M	UACC−62
COLO−792	Mewo	
COLO−679	RVH−421	
COLO−800	SK−MEL−2	
G−361	SK−MEL−24	
HT−144	SK−MEL−28	

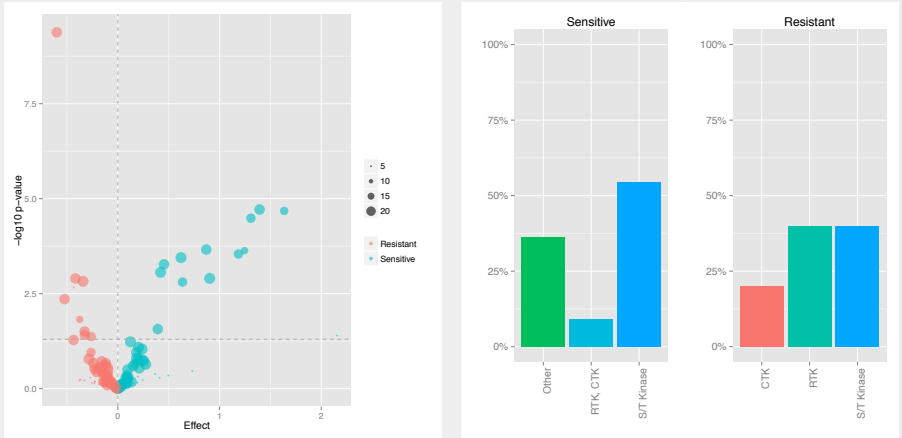
Tissue composition



Molecular profile



Drug profile



Drugs associated with the cluster (FDR< **0.05**)

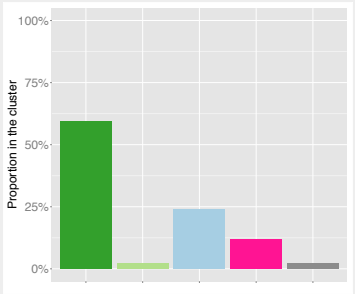
EMBELIN	Sensible
FH535	Sensible
OBATOCLAX−MESYLATE	Sensible
RDEA119	Sensible
CI−1040	Sensible
CEP−701	Sensible
PLX4720	Sensible
NUTLIN3	Sensible
PD0325901	Sensible
SB590885	Sensible
AZD6244	Sensible
WZ−1−84	Resistant
AKT−INHIBITOR−VIII	Resistant
AICAR	Resistant
GEFITINIB	Resistant
BIBW2992	Resistant

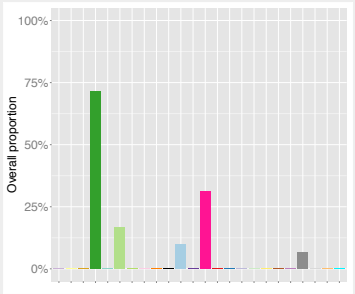
Breast Cluster - GDSC

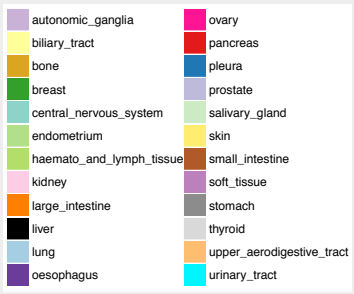
Cell lines in the cluster

ABC–1	GCIY	HCC70	NCI–H1573	UACC–812
BT–20	HCC1143	MCF7	NCI–H1650	ZR–75–30
CAL–148	HCC1187	MDA–MB–175–VII	NCI–H1693	
CAL–85–1	HCC1419	MDA–MB–361	NCI–H2126	
CAMA–1	HCC1599	MDA–MB–415	NCI–H2170	
Caov–3	HCC1937	MDA–MB–453	NCI–H2342	
Caov–4	HCC1954	MDA–MB–468	OVCAR–3	
ChaGo–K–1	HCC2157	MFE–280	OAW–28	
EFM–19	HCC2218	NCI–H1395	OVCAR–4	
EVSA–T	HCC38	NCI–H1437	T47D	

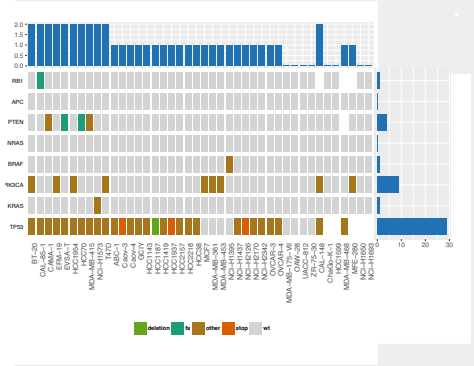
Tissue composition

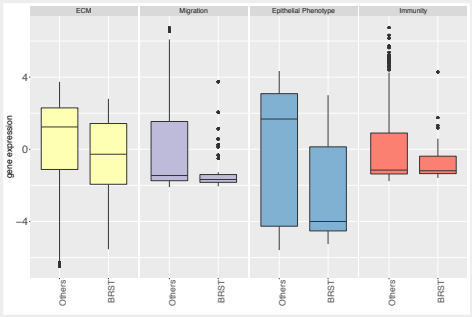


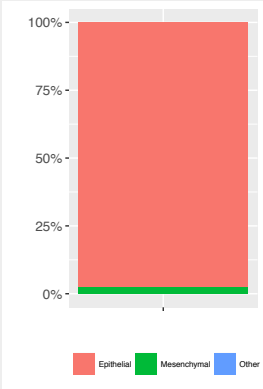




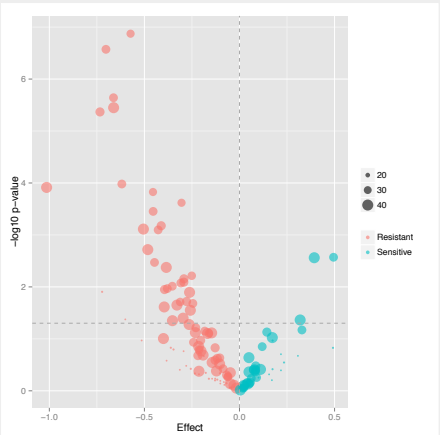
Molecular profile

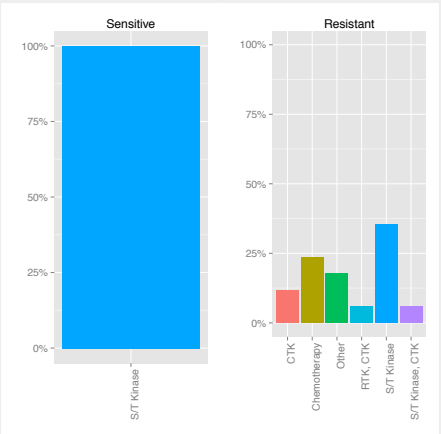






Drug profile





Drugs associated with the cluster (FDR< **0.05**)

AKT–INHIBITOR–VIII	Sensible
MK–2206	Sensible
GEMCITABINE	Resistant
AP–24534	Resistant
DMOG	Resistant
BAY–61–3606	Resistant
OBATOCLAX–MESYLATE	Resistant
CAMPTOTHECIN–WTSI	Resistant
CYTARABINE	Resistant
METHOTREXATE	Resistant
RDEA119	Resistant
CI–1040	Resistant
AZD–2281	Resistant
CEP–701	Resistant
SB–216763	Resistant
BX–795	Resistant
BI–D1870	Resistant
681640	Resistant
AZD6244	Resistant

A. NEW INSIGHT FOR PHARMACOGENETICS STUDIES FROM THE TRANSCRIPTIONAL ANALYSIS OF TWO LARGE-SCALE CANCER CELL LINE PANELS

Appendix B

The genomic and transcriptomic landscape of neoadjuvant-resistant triple-negative breast cancer

Figure B.1: List of genes differentially expressed. A) Between PT and RD, B) between PT and LN, C) between RD and LN.

A

Genes	logFC	p-value	adjusted p-value
AIM2	-1.67	0.00093	0.0476
ARC	1.78	0.000752	0.0443
BANK1	-2.12	0.000185	0.0336
CD19	-2.49	0.00062	0.0431
CLEC17A	-2.44	0.00033	0.0393
CLEC3A	3.08	0.000513	0.0423
COL6A5	-1.94	0.000148	0.031
CPB1	3.97	0.000878	0.047
CPS1	-1.54	1.93e-05	0.0141
CYR61	1.59	1.46e-06	0.0034
DNM3OS	-1.91	2.75e-05	0.0141
DTHD1	-2.06	0.000905	0.0476
DUSP1	2.61	1.04e-09	1.7e-05
EDN2	1.52	0.000298	0.0387
EGR1	2.29	4.29e-06	0.00749
FCRL1	-3.72	5.67e-05	0.0207
FCRL2	-2.96	2.66e-05	0.0141
FCRL3	-2.58	0.000559	0.0431
FOS	3.06	1.86e-07	0.00101
FOSB	3.83	6.32e-08	0.000514
GAD1	1.6	6.74e-05	0.0219
GDF15	1.51	0.000218	0.0356
GPR174	-2.24	0.000489	0.0412
IGLL5	-2.51	0.000111	0.0277
INSM1	2.39	0.000887	0.0472
KIAA0125	-2.85	1.06e-05	0.0139
LRRTM2	-1.59	0.000475	0.0412
MMRN1	-2.26	0.000854	0.0465
MS4A1	-4.14	4.67e-05	0.0185
NKX2-5	2.02	0.000606	0.0431
NR4A1	2.2	8.57e-07	0.00233
OSM	1.81	5.06e-06	0.00749
PARP15	-1.78	0.000395	0.0409
PAX5	-3.47	1.19e-05	0.0139
PTGS2	1.57	0.000145	0.031
RGS1	1.96	2.99e-06	0.00608
SLC30A8	2.82	0.000281	0.0385
SULT1E1	1.96	0.000583	0.0431
TERT	1.68	0.000727	0.0443
TTN	-1.7	0.000337	0.0393
ZNF208	-1.91	1.41e-05	0.0141
ZNF676	-1.88	7.6e-05	0.0233

B

Genes	logFC	p-value	adjusted p-value
AMPH	-2.38	0.000124	0.0394
ATP1A3	1.51	0.000166	0.0499
B4GALNT2	-2.61	3.35e-07	0.000317
C11orf96	1.54	6.02e-06	0.00366
C7	3.16	1.06e-05	0.00605
CPS1	-2.36	5.15e-07	0.000448
CYR61	1.92	1.95e-08	3.12e-05
DUSP1	2.88	2.22e-16	1.79e-12
EGR1	2.79	5.16e-08	6.81e-05
EGR2	1.57	7.5e-05	0.0286
F2RL2	-5.24	1.15e-06	0.000908
FAM3D	2.28	1.45e-06	0.00111
FOS	4.11	9.61e-14	5.15e-10
FOSB	4.81	2.22e-16	1.79e-12
FREM1	-2.69	2.79e-05	0.0139
GDF15	1.64	0.000113	0.0371
HMCN2	2.75	9.33e-11	3,00E-07
KCNT1	2.07	6.66e-05	0.0265
KRT20	1.99	7.03e-05	0.0274
KRT4	3.27	3,00E-07	0.00029
LINC00473	-3.2	3.38e-06	0.00231
MAGEA4	6.23	1.86e-08	3.05e-05
MMP13	-4.11	4.21e-05	0.0192
NCAM2	-1.68	1.65e-05	0.00887
NKX2-1	-6.96	1.78e-11	7.17e-08
NR4A1	1.75	4.84e-05	0.0212
OSM	2.28	9.98e-09	2.12e-05
PLD5	3.13	8.54e-05	0.031
PTGS2	2.29	2.32e-08	3.39e-05
RGS1	2.41	8.61e-09	1.92e-05
RNF157-AS1	1.54	0.000105	0.0355
SALL1	-4.31	1.44e-08	2.65e-05
SCIN	-1.56	6.42e-06	0.00383
SERHL	-1.84	5.77e-07	0.000489
SERHL2	-1.61	4.27e-05	0.0194
SLC24A2	-3.45	3.76e-05	0.0176
TMEM132E	2.17	1.58e-05	0.00852
TNN	-3.29	9.86e-05	0.034
ZFP36	1.79	1.28e-07	0.000138

C

Genes	logFC	p-value	adjusted p-value
HMCN2	-2.79	3.88e-06	0.0211
NKX2-1	7.41	4.45e-07	0.00725
PAX5	-4.44	2.26e-06	0.0184

Figure B.2: RBP detected as regulator of splicing events between PTs and RDs. Motif enrichment detected in upstream or downstream intronic sequences. The color indicates the role of the RBP on the regulated exons when binding to the given location. Blue for exclusion, red for inclusion and purple for both.

Figure B.3: RBP detected as regulator of splicing events between PTs and LNs. Motif enrichment detected in upstream or downstream intronic sequences. The color indicates the role of the RBP on the regulated exons when binding to the given location. Blue for exclusion, red for inclusion and purple for both.

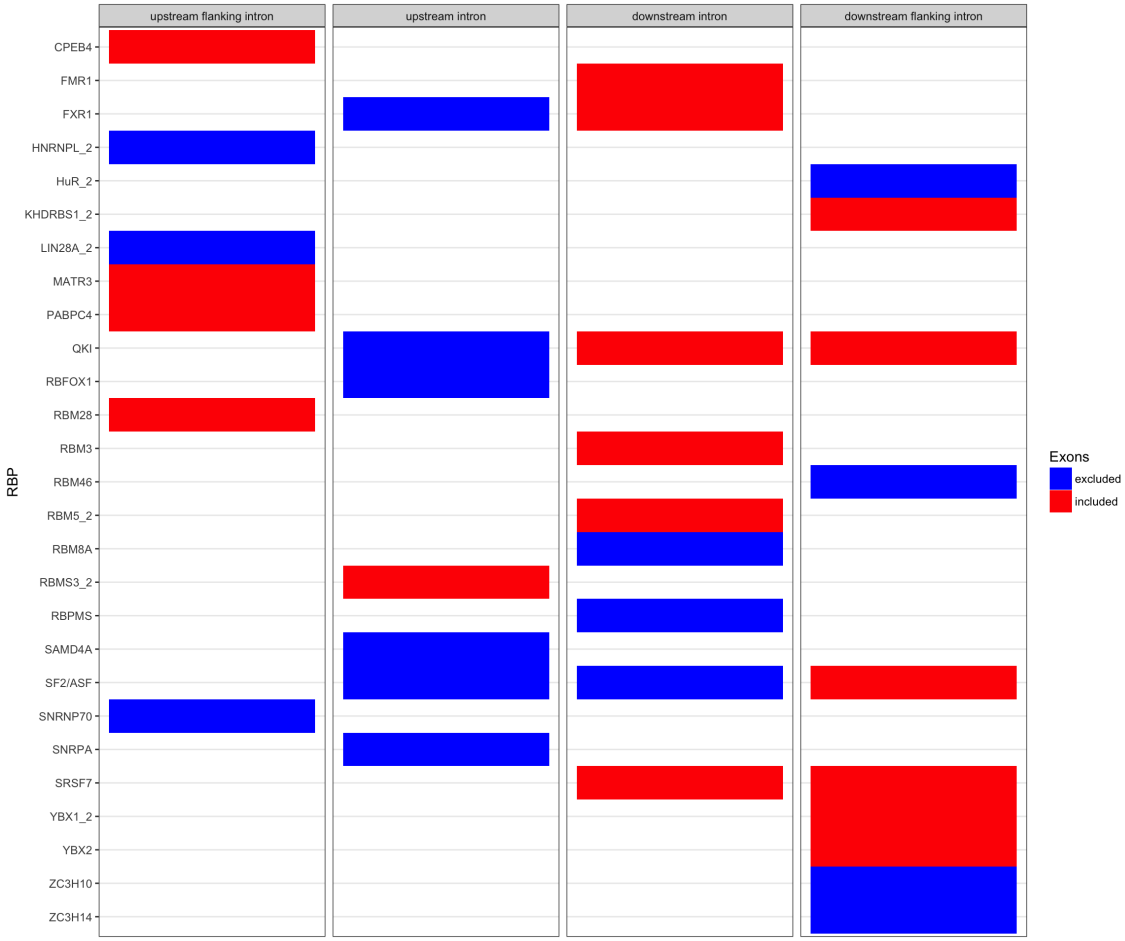


Figure B.4: RBP detected as regulator of splicing events between RDs and LNs. Motif enrichment detected in upstream or downstream intronic sequences. The color indicates the role of the RBP on the regulated exons when binding to the given location. Blue for exclusion, red for inclusion and purple for both.

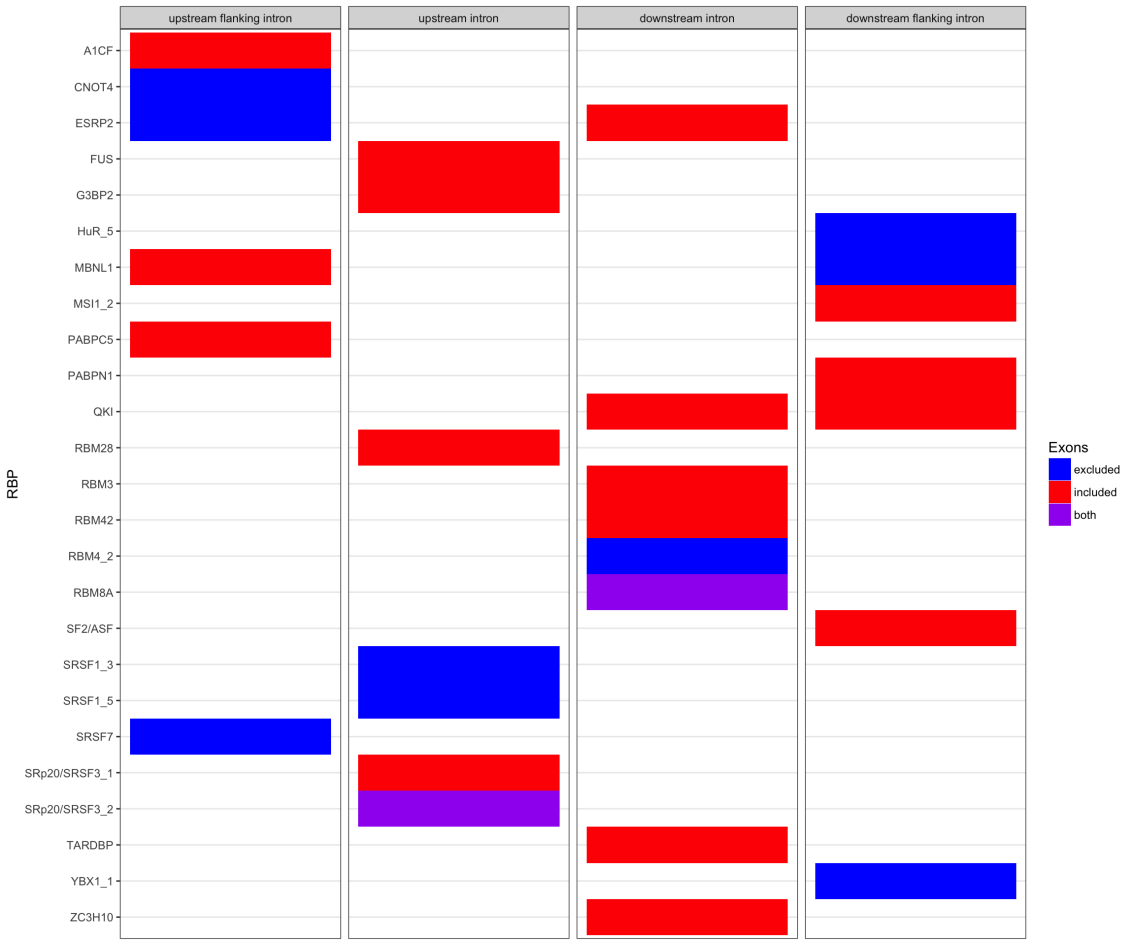


Figure B.5: Estimated ploidy by FACETS.

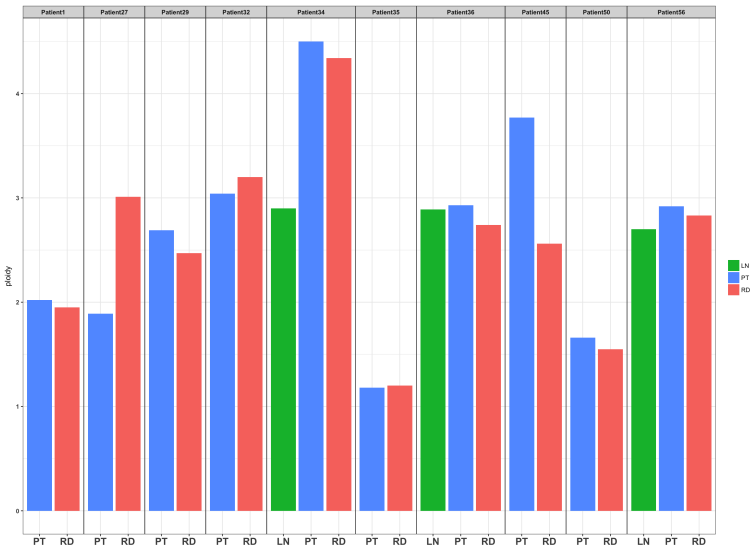
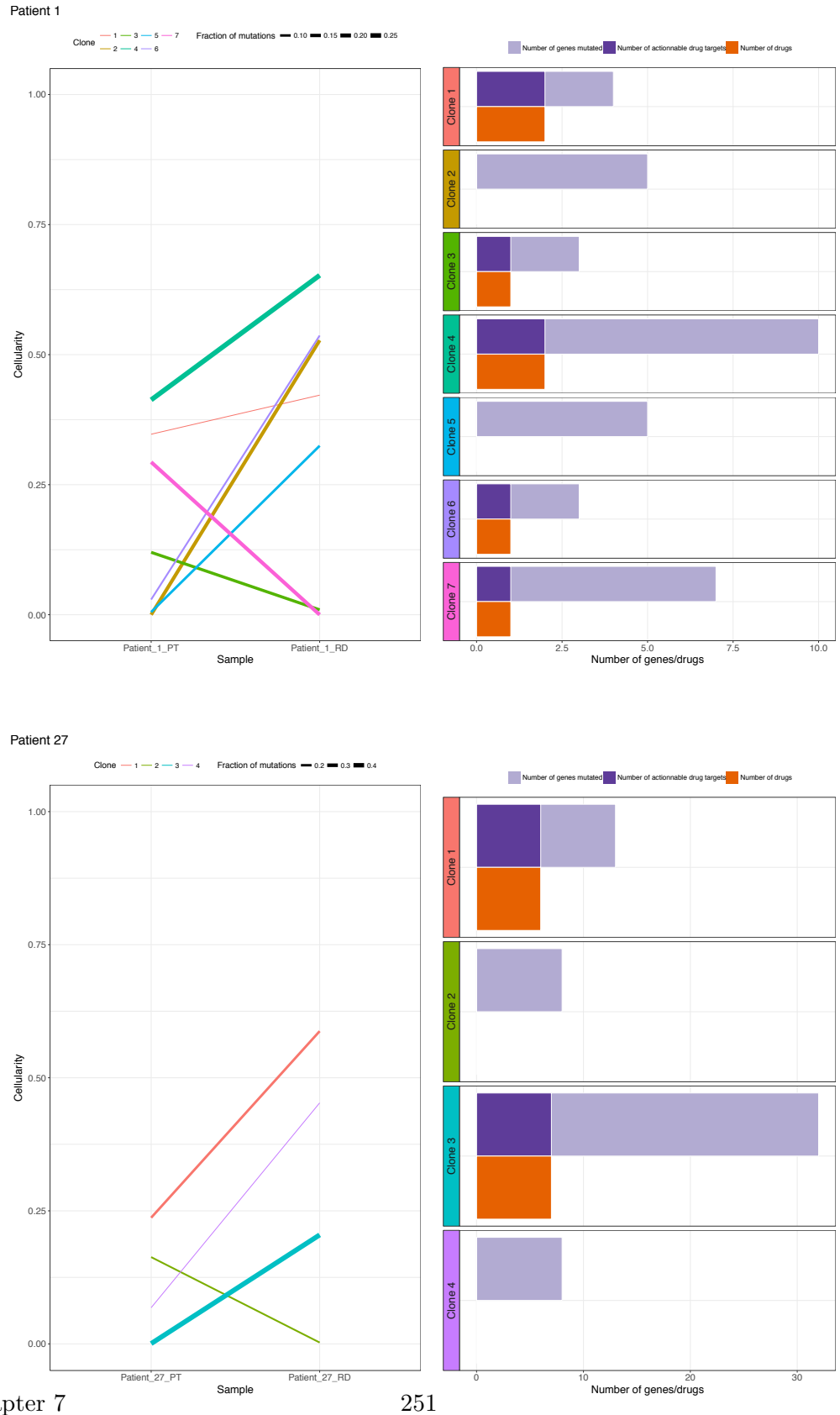


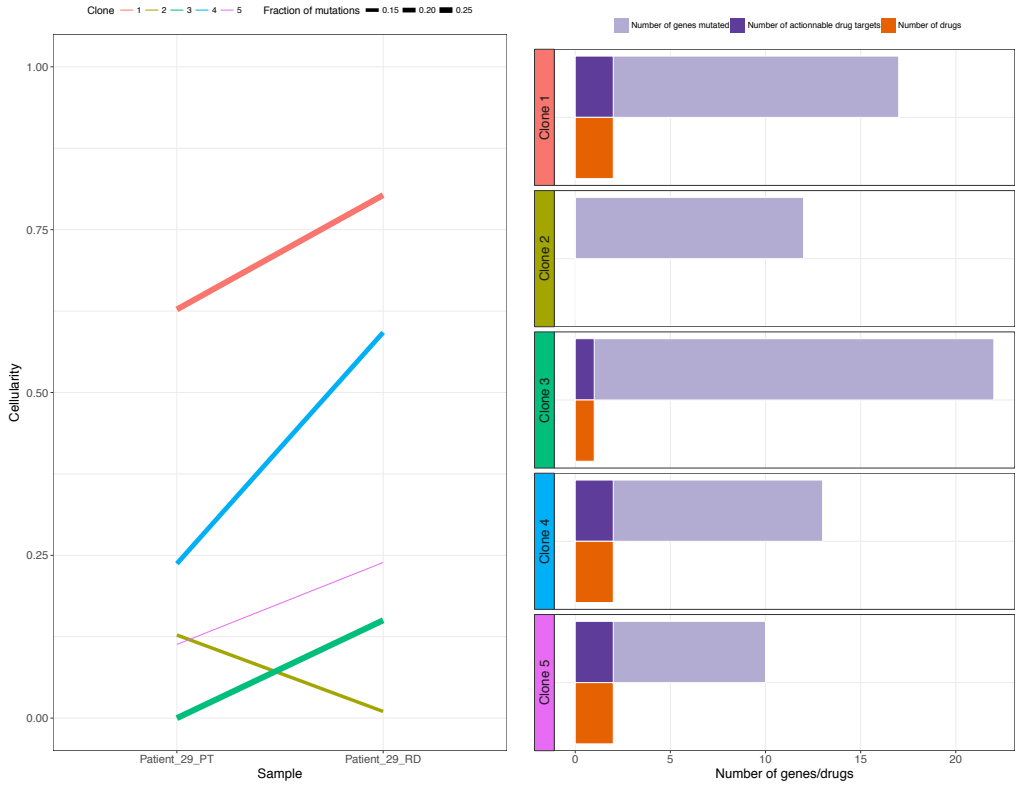
Figure B.6: Mutation signatures of SNVs in primary tumor and residual disease. Left: Pie charts display the pre and post-neoadjuvant mutational signatures in all patients. Right: the 96 trinucleotide mutational spectra of pre and post-neoadjuvant mutations in all patients was inferred by deconstructSigs.

Clonal evolution between primary tumors and residual diseases. The left panel displays the estimated clonal evolution between the PT and RD of the patient. The width of the segment corresponds to the fraction of mutation carried by the sub-clone. The right panel displays for each estimated sub-clone the number of genes with non-silent mutation (light purple), the number of genes with non-silent mutation that interact with the drugs (dark violet), and the number of associated drugs (orange).

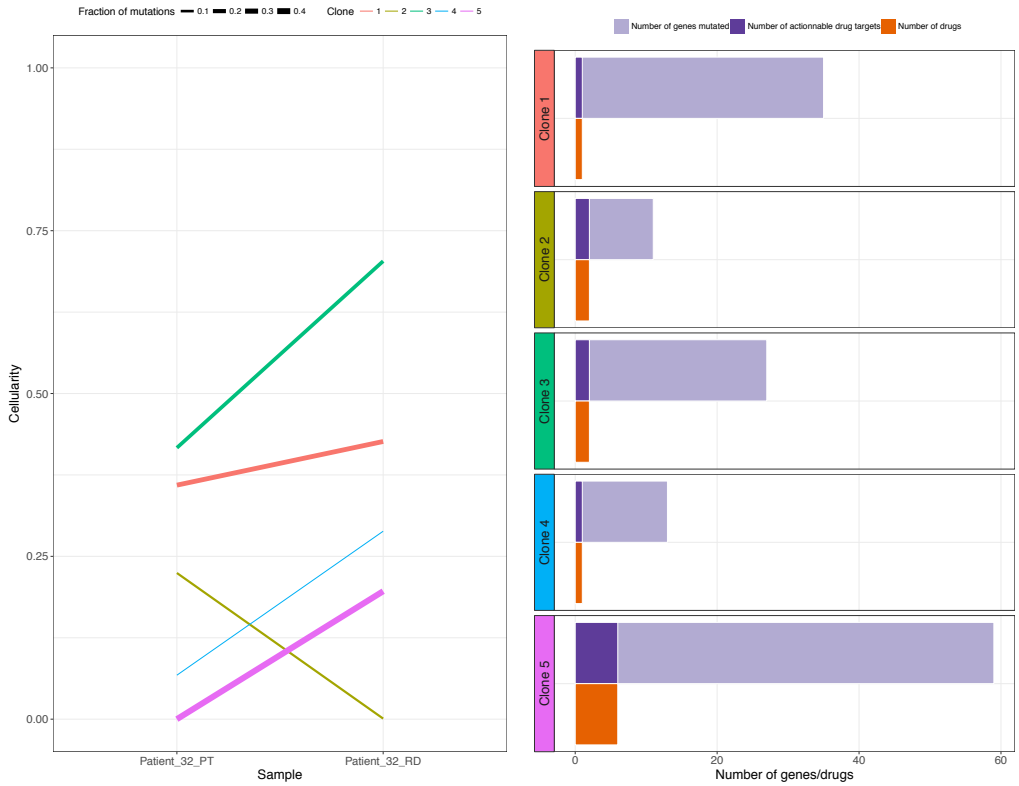
Figure B.7


Patient 1
Clone 1 2 3 4 5 6 7
Fraction of mutations 0.10 0.15 0.20 0.25
Number of genes mutated
Number of actionnable drug targets
Number of drugs
Cellularity
1.00
0.75
0.50
0.25
0.00
Patient_1_PT
Patient_1_RD
Sample
Clone 1
Clone 2
Clone 3
Clone 4
Clone 5
Clone 6
Clone 7
0.0
2.5
5.0
7.5
10.0
Number of genes/drugs
Patient 27
Clone 1 2 3 4
Fraction of mutations 0.2 0.3 0.4
Number of genes mutated
Number of actionnable drug targets
Number of drugs
Cellularity
1.00
0.75
0.50
0.25
0.00
Patient_27_PT
Patient_27_RD
Sample
Clone 1
Clone 2
Clone 3
Clone 4
0
10
20
30
Number of genes/drugs

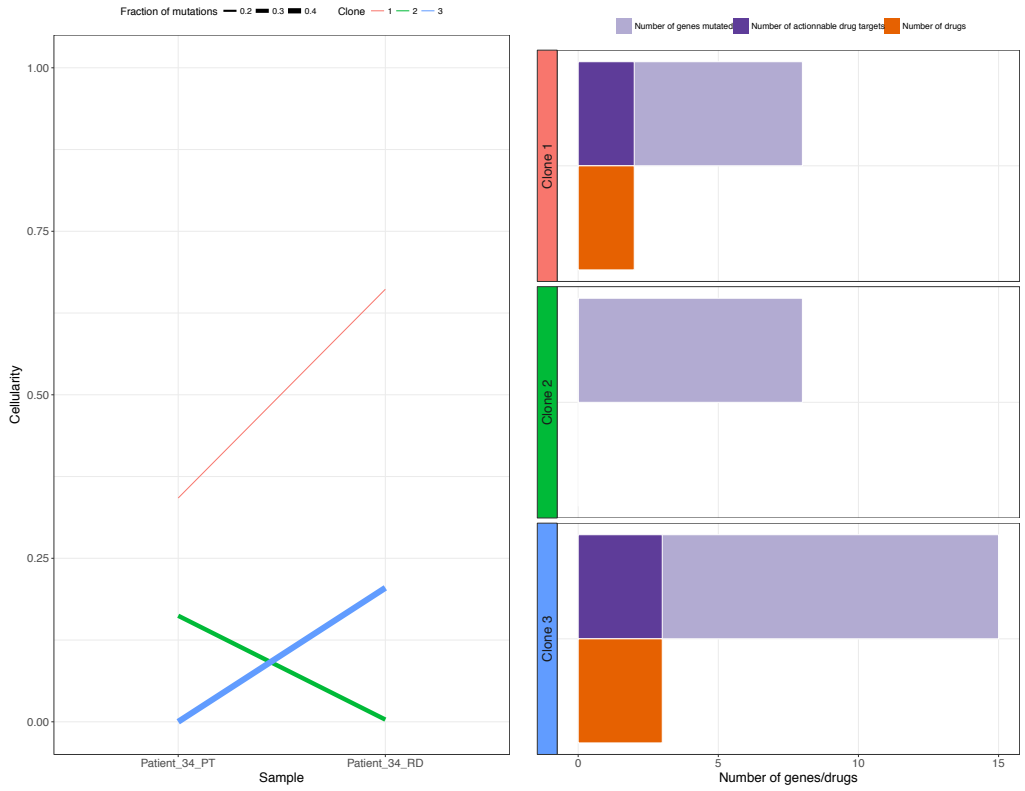
Patient 29



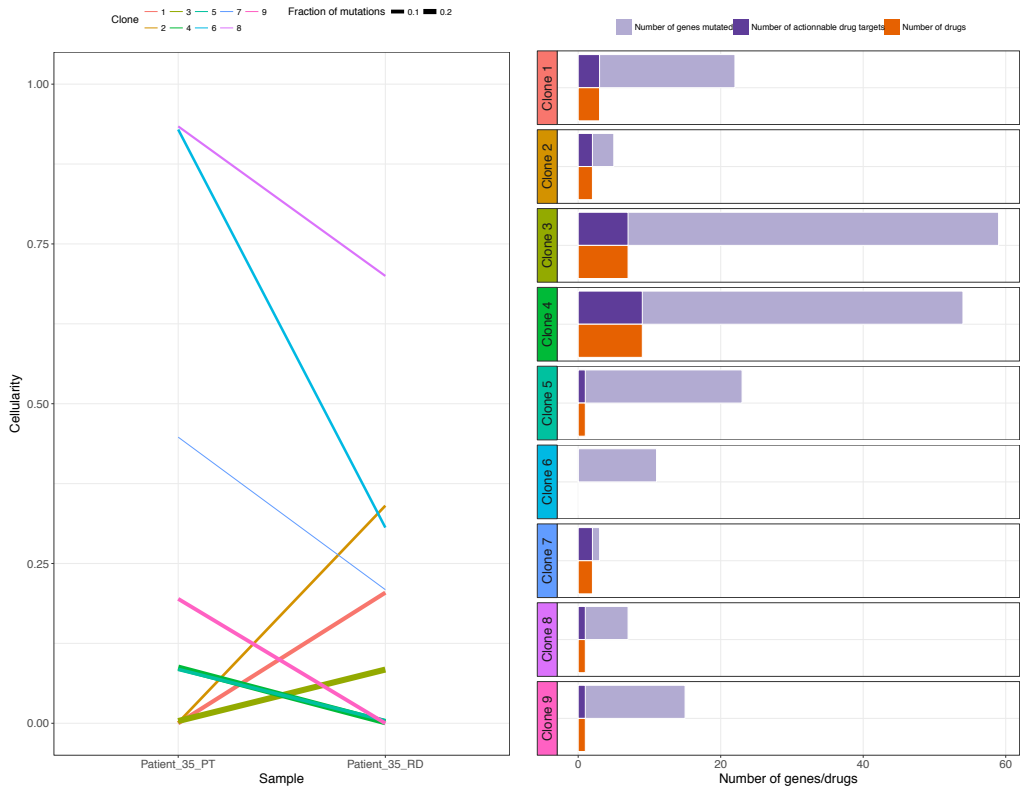
Patient 32



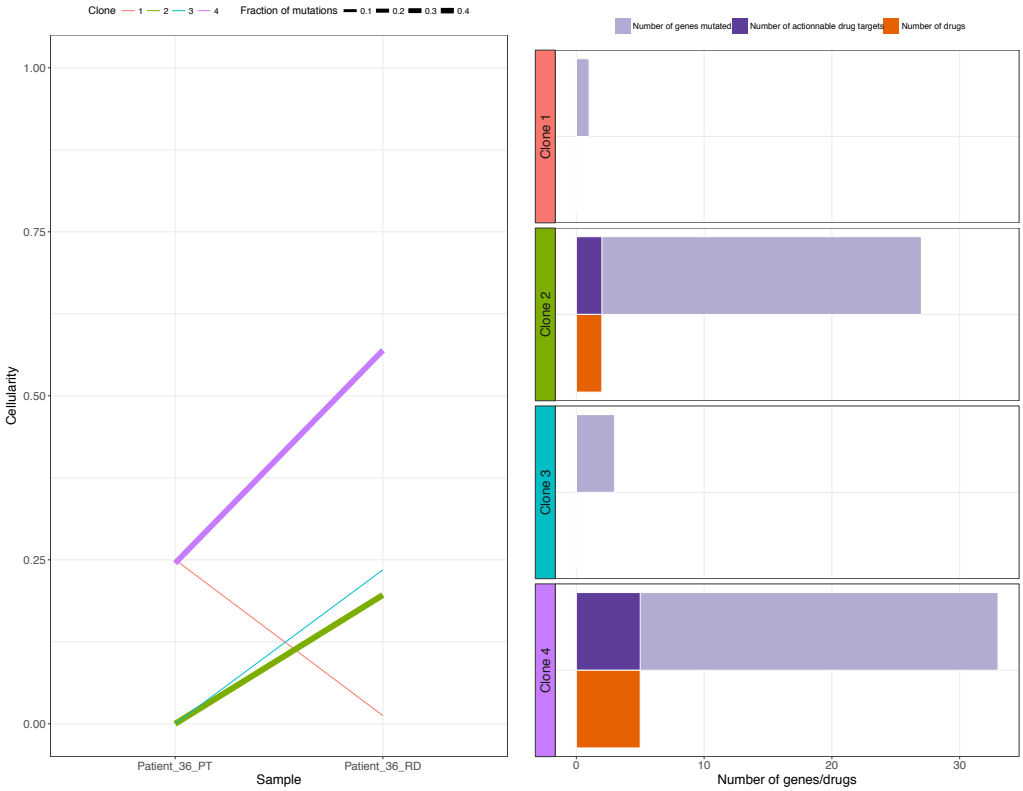
Patient 34



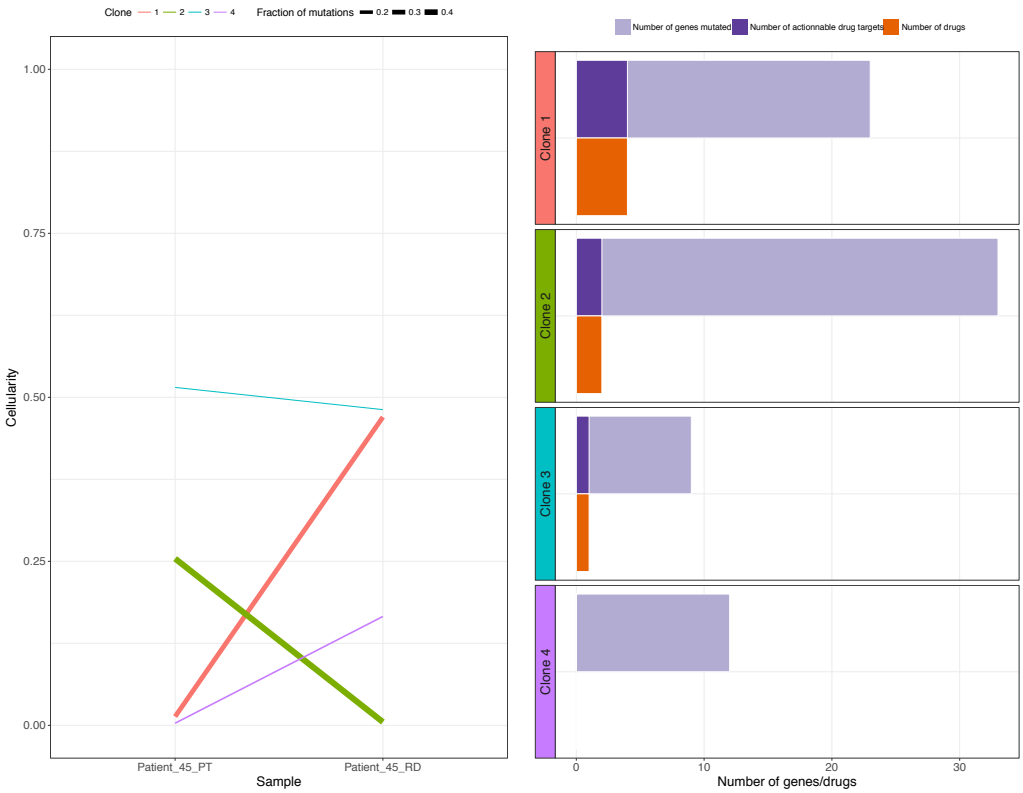
Patient 35



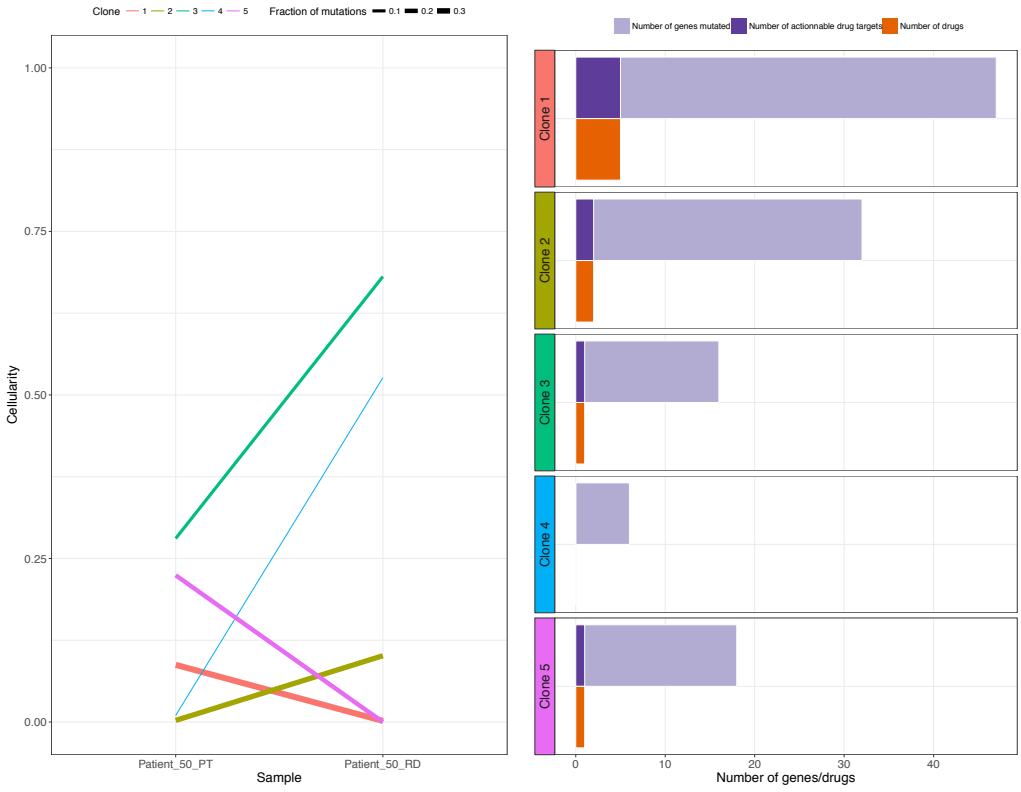
Patient 36



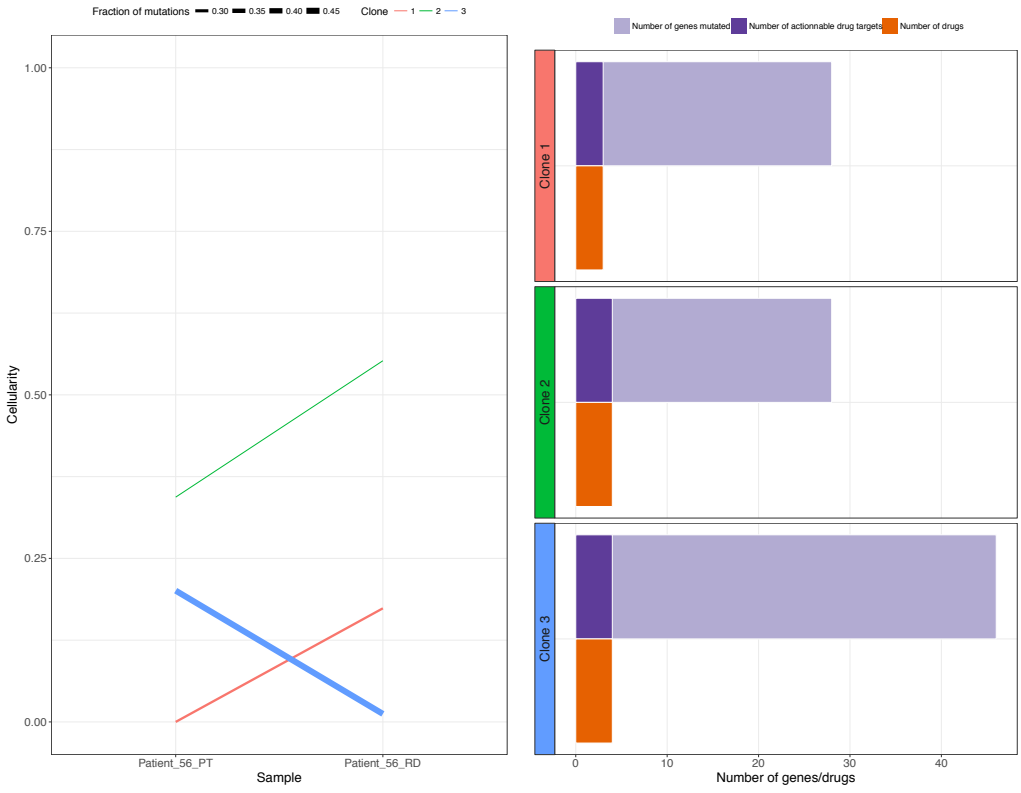
Patient 45



Patient 50



Patient 56



Glossary

IC_{50} The concentration at which the drug inhibited 50% of maximal cell growth

A3SS Alternative 3' splice site

A5SS Alternative 5' splice site

AR Androgen Receptor

AS Alternative splicing

AUC area under the dose response curve

BLBC Basal-Like Breast Cancer

CNA Copy Number Alterations

CNV copy number variation

CNV copy number variations

DCIS Ductal Carcinoma In Situ

DEG Differentially Expressed Genes

DEG Differentially Expressed Genes

ER Estrogen receptor

FDP False Discovery Proportion

FDR False Discovery Rate

FWER Family-Wise Error Rate

GEM Genetically Modified Mice

HER2 Human Epidermal Growth Factor Receptor 2

HTS High-Throughput Screening

IDC Invasive Ductal Carcinoma

ILC Invasive Lobular Carcinoma

indels Insertion and Deletion of one or more nucleotides

JR Joint Risk

LN Lymph Node

LOH Loss Of Heterozygosity

MXE Mutually exclusive exon

NAC Neoadjuvant Chemotherapy

NGS Next Generation Sequencing

PCR Pathological Complete Response

ph-FDP Post-Hoc False Discovery Proportion

PR Progesterone Receptor

PT Primary Tumor

RBP RNA-binding proteins

RD Residual Disease

RI Retention of intron

SDSE Significant Differential Splicing Events

SE Skipping exon

SNP Single Nucleotide Polymorphism

SNV Somatic Nucleic Variant

TILs Tumor-infiltrating lymphocytes

TNBC Triple Negative Breast Cancer

VAF Variant Allele Frequency

WES Whole Exome Sequencing

WGS Whole Genome Sequencing

References

[1] G. Abraham, A. Kowalczyk, S. Loi, I. Haviv, and J. Zobel. Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC bioinformatics*, 11:277, 2010. 120

[2] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, apr 2010. 29, 82

[3] R. Akbani, P. K. S. Ng, H. M. J. Werner, M. Shahmoradgoli, F. Zhang, Z. Ju, W. Liu, J.-Y. Yang, K. Yoshihara, J. Li, S. Ling, E. G. Seviour, P. T. Ram, J. D. Minna, L. Diao, P. Tong, J. V. Heymach, S. M. Hill, F. Dondelinger, N. Städler, L. A. Byers, F. Meric-Bernstam, J. N. Weinstein, B. M. Broom, R. G. W. Verhaak, H. Liang, S. Mukherjee, Y. Lu, and G. B. Mills. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature Communications*, 5:ncomms4887, may 2014. 86

[4] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. a. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. a. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, M. Imielinsk, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. a. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500:415–21, 2013. 30, 31, 78, 80

[5] American Cancer Society. Global Cancer Facts & Figures 3rd Edition. *American Cancer Society*, (800):1–64, 2015. 1, 8, 9

[6] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010. 24

[7] S. Anders, P. T. Pyl, and W. Huber. HTSeq A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, jan 2014. 89

[8] S. Anders, a. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Res*, 22(10):2008–2017, 2012. 26

[9] E. S. Anderson, C.-H. Lin, X. Xiao, P. Stoilov, C. B. Burge, and D. L. Black. The cardiotonic steroid digitoxin regulates alternative splicing through depletion of the splicing factors SRSF3 and TRA2B. *RNA (New York, N.Y.)*, 18(5):1041–9, may 2012. 72

[10] F. André, T. Bachelot, F. Commo, M. Campone, M. Arnedos, V. Dieras, M. Lacroix-Triki, L. Lacroix, P. Cohen, D. Gentien, J. Adélaide, F. Dalenc, A. Goncalves, C. Levy, J. M. Ferrero, J. Bonneterre, C. Lefeuvre, M. Jimenez, T. Filleron, and H. Bonnefoi. Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: A multicentre, prospective trial (SAFIR01/UNICANCER). *The Lancet Oncology*, 15(3):267–274, 2014. 84

[11] T. T. Ashburn and K. B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews. Drug discovery*, 3(8):673–683, 2004. 46

[12] W. K. Bae, K. H. Yoo, J. S. Lee, Y. Kim, I. J. Chung, M. H. Park, J. H. Yoon, P. A. Furth, and L. Hennighausen. The methyltransferase EZH2 is not required for mammary cancer development, although high EZH2 and low H3K27me3 correlate with poor prognosis of ER-positive breast cancers. *Molecular Carcinogenesis*, 54(10):1172–1180, 2015. 168

[13] J. M. Balko, R. S. Cook, D. B. Vaught, M. G. Kuba, T. W. Miller, N. E. Bhola, M. E. Sanders, N. M. Granja-Ingram, J. J. Smith, I. M. Meszoely, J. Salter, M. Dowsett, K. Stemke-Hale, A. M. González-Angulo, G. B. Mills, J. a. Pinto, H. L. Gómez, and C. L. Arteaga. Profiling of residual breast cancers after neoadjuvant chemotherapy identifies DUSP4 deficiency as a mechanism of drug resistance. *Nature medicine*, 18(7):1052–9, 2012. 64, 86

[14] J. M. Balko, J. M. Giltnane, K. Wang, L. J. Schwarz, C. D. Young, R. S. Cook, P. Owens, M. E. Sanders, M. G. Kuba, V. Sánchez, R. Kurupi, P. D. Moore, J. A. Pinto, F. D. Doimi, H. Gómez, D. Horiuchi, A. Goga, B. D. Lehmann, J. A. Bauer, J. A. Pietenpol, J. S. Ross, G. A. Palmer, R. Yelensky, M. Cronin, V. A. Miller, P. J. Stephens, and C. L. Arteaga. Molecular profiling of the residual disease of triple-negative breast cancers after neoadjuvant chemotherapy identifies actionable therapeutic targets. *Cancer Discovery*, 4(2):232–245, 2014. 14, 16, 17, 81, 83

[15] Y. J. Bang, E. Van Cutsem, A. Feyereislova, H. C. Chung, L. Shen, A. Sawaki, F. Lordick, A. Ohtsu, Y. Omuro, T. Satoh, G. Aprile, E. Kulikov, J. Hill, M. Lehle, J. Rüschoff, and Y. K. Kang. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): A phase 3, open-label, randomised controlled trial. *The Lancet*, 376(9742):687–697, aug 2010. 7

[16] J. Barretina, G. Caponigro, and N. Stransky. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012. 8, 33, 56, 57, 158

[17] K. R. Bauer, M. Brown, R. D. Cress, C. A. Parise, and V. Caggiano. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: A population-based study from the California Cancer Registry. *Cancer*, 109(9):1721–1728, may 2007. 12

[18] E. Becht, A. De Reyniès, N. A. Giraldo, C. Pilati, B. Buttard, L. Lacroix, J. Selves, C. Sautès-Fridman, P. Laurent-Puig, and W. H. Fridman. Immune and stromal classification of Colorectal cancer is associated with molecular subtypes and relevant for precision immunotherapy. *Clinical Cancer Research*, 22(16):4057–4066, 2016. 66

[19] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing, 1995. 28

[20] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under depencency. *The Annals of Statistics*, 29(4):1165–1188, 2001. 28, 106, 115

[21] V. Beral, D. Bull, R. Doll, R. Peto, and G. Reeves. Breast cancer and breastfeeding: Collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. *Lancet*, 360(9328):187–195, jul 2002. 9

[22] E. Bernard, L. Jacob, J. Mairal, and J. P. Vert. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics*, 30(17):2447–2455, sep 2014. 25

[23] N. E. Bhola, V. M. Jansen, J. P. Koch, H. Li, L. Formisano, J. A. Williams, J. R. Grandis, and C. L. Arteaga. Treatment of triple-negative breast cancer with TORC1/2 inhibitors sustains a drug-resistant and notch-dependent cancer stem cell population. *Cancer Research*, 76(2):440–452, jan 2016. 87

[24] G. Blanchard, P. Neuvial, and E. Roquain. Post hoc inference via joint family-wise error rate control. (2014):1–34, 2017. 28, 29, 106, 108, 109

[25] V. Boeva, T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher, I. Janoueix-Lerosey, O. Delattre, and E. Barillot. Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28(3):423–425, feb 2012. 30

[26] M. A. Bollet, N. Servant, P. Neuvial, C. Decraene, I. Lebigot, J. P. Meyniel, Y. De Rycke, A. Savignoni, G. Rigaill, P. Hupé, A. Fourquet, B. Sigal-Zafrani, E. Barillot, and J. P. Thiery. High-resolution mapping of DNA breakpoints to define true recurrences among ipsilateral breast cancers. *Journal of the National Cancer Institute*, 100(1):48–58, 2008. 74

[27] P. Bouwman and J. Jonkers. The effects of deregulated DNA damage signalling on cancer chemotherapy response and resistance. *Nature Reviews Cancer*, 12(9):587–598, 2012. 16

[28] G. Box. Robustness in the strategy of scientific model building. *Robustness in statistics*, 1979. 7

[29] A. P. Bracken and K. Helin. Polycomb group proteins: navigators of lineage pathways led astray in cancer. *Nature Reviews Cancer*, 9(11):773–784, 2009. 168

[30] I. Buchhalter, B. Hutter, T. S. Alioto, T. A. Beck, P. C. Boutros, B. Brors, P. Adam, S. Chotewutmontri, R. E. Denroche, S. Derdak, N. Diessl, L. Feuerbach, A. Fujimoto, S. Gröbner, M. Gut, N. J. Harding, M. Heinold, L. E. Heisler, J. Hinton, N. Jäger, D. Jones, R. Kabbe, A. Korshunov, J. D. Mcpherson, H. Nakagawa, C. Previti, K. Raine, P. Ribeca, S. Schmidt, L. Stebbings, P. S. Tarpey, J. W. Teague, L. Tonon, D. A. Wheeler, L. Xi, T. N. Yamaguchi, A.-S. Sertier, S. M. Pfister, P. Lichter, and R. Eils. A comprehensive multicenter comparison of whole genome sequencing pipelines using a uniform tumor - normal sample pair. *bioRxiv*, 2014. 31

[31] M. D. Burstein, A. Tsimelzon, G. M. Poage, K. R. Covington, A. Contreras, S. A. W. Fuqua, M. I. Savage, C. K. Osborne, S. G. Hilsenbeck, J. C. Chang, G. B. Mills, C. C. Lau, and P. H. Brown. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clinical Cancer Research*, 21(7):1688–1698, sep 2015. 120, 187

[32] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27, 1974. 41, 50

[33] L. A. Carey, C. M. Perou, C. A. Livasy, L. G. Dressler, D. Cowan, K. Conway, G. Karaca, M. A. Troester, C. K. Tse, S. Edmiston, S. L. Deming, J. Geradts, M. C. U. Cheang, T. O. Nielsen, P. G. Moorman, H. S. Earp, and R. C. Millikan. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA : the journal of the American Medical Association*, 295(21):2492–502, jun 2006. 12

[34] P. B. Chapman, A. Hauschild, C. Robert, J. B. Haanen, P. Ascierto, J. Larkin, R. Dummer, C. Garbe, A. Testori, M. Maio, D. Hogg, P. Lorigan, C. Lebbe, T. Jouary, D. Schadendorf, A. Ribas, S. J. O'Day, J. A. Sosman, J. M. Kirkwood, A. M. M. Eggermont, B. Dreno, K. Nolop, J. Li, B. Nelson, J. Hou, R. J. Lee, K. T. Flaherty, G. A. McArthur, and BRIM-3 Study Group. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *The New England journal of medicine*, 364(26):2507–16, jun 2011. 6, 8, 34

[35] M. Chen and J. L. Manley. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature Reviews Molecular Cell Biology*, 10(11):741–54, nov 2009. 71

[36] S. Cho, M. Lu, X. He, P.-L. R. Ee, U. Bhat, E. Schneider, L. Miele, and W. T. Beck. Notch1 regulates the expression of the multidrug resistance gene ABCC1/MRP1 in cultured cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, 108(51):20778–20783, dec 2011. 87

[37] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80–92, 2014. 30, 82

[38] E. E. Cohen, D. W. Davis, T. G. Karrison, T. Y. Seiwert, S. J. Wong, S. Nattam, M. F. Kozloff, J. I. Clark, D. H. Yan, W. Liu, C. Pierce, J. E. Dancey, K. Stenson, E. Blair, A. Dekker, and E. E. Vokes. Erlotinib and bevacizumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck: a phase I/II study. *The Lancet Oncology*, 10(3):247–257, mar 2009. 47

[39] M. Colleoni, Z. Sun, K. N. Price, P. Karlsson, J. F. Forbes, B. Thurlimann, L. Gianni, M. Castiglione, R. D. Gelber, A. S. Coates, and A. Goldhirsch. Annual hazard rates of recurrence for breast cancer during 24 years of follow-up: Results from the international breast cancer study group trials I to V. *Journal of Clinical Oncology*, 34(9):927–935, 2016. 15

[40] E. M. Coonrod, J. D. Durtschi, R. L. Margraf, and K. V. Voelkerding. Developing genome and exome sequencing for candidate gene identification in inherited disorders: An integrated technical and bioinformatics approach. *Archives of Pathology and Laboratory Medicine*, 137(3):415–433, 2013. 21, 30

[41] P. Cortazar, L. Zhang, M. Untch, K. Mehta, J. P. Costantino, N. Wolmark, H. Bonnefoi, D. Cameron, L. Gianni, P. Valagussa, S. M. Swain, T. Prowell, S. Loibl, D. L. Wickerham, J. Bogaerts, J. Baselga, C. Perou, G. Blumenthal, J. Blohmer, E. P. Mamounas, J. Bergh, V. Semiglazov, R. Justice, H. Eidtmann, S. Paik, M. Piccart, R. Sridhara, P. a. Fasching, L. Slaets, S. Tang, B. Gerber, C. E. Geyer, R. Pazdur, N. Ditsch, P. Rastogi, W. Eiermann, and G. Von Minckwitz. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *The Lancet*, 384(9938):164–172, jul 2014. 15, 16, 46

[42] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-Ud-Din, P. Hintsanen, S. a. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J. W. Gray, and G. Stolovitzky. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, pages 1–103, 2014. 22

[43] F. Crick. On Protein Synthesis. *The Symposia of the Society for Experimental Biology*, pages 138–166, 1958. 3

[44] C. M. Croce. Oncogenes and Cancer. *New England Journal of Medicine*, 358(5):502–511, jan 2008. 3

[45] Y. Cun and H. Fröhlich. Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics*, 13:69, 2012. 45, 120

[46] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–52, jun 2012. 11, 14, 74, 86, 120, 135

[47] M. Danan-gotthold, R. Golan-gerstl, E. Eisenberg, K. Meir, R. Karni, and E. Y. Levanon. Identification of recurrent regulated alternative splicing events across human solid tumors. 43(10):1–15, 2015. 25

[48] P. Deveau, L. Colmet Daage, D. Oldridge, V. Bernard, A. Bellini, M. Chicard, N. Clement, E. Lapouble, V. Combaret, A. Boland, V. Meyer, J.-F. Deleuze, I. Janoueix-Lerosey, E. Barillot, O. Delattre, J. Maris, G. Schleiermacher, and V. Boeva. Clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction. *bioRxiv*, pages 1–32, 2016. 82

[49] A. Di Leo and E. Moretti. Anthracyclines: The first generation of cytotoxic targeted agents? A possible dream. *Journal of Clinical Oncology*, 26(31):5011–5013, 2008. 87

[50] M. V. Dieci, C. Criscitiello, A. Goubar, G. Viale, P. Conte, V. Guarneri, G. Ficarra, M. C. Mathieu, S. Delaloge, G. Curigliano, and F. Andre. Prognostic value of tumor-infiltrating lymphocytes on residual disease after primary chemotherapy for triple-negative breast cancer: A retrospective multicenter study. *Annals of Oncology*, 25(3):611–618, 2014. 66

[51] K. A. Dittmar, P. Jiang, J. W. Park, K. Amirikian, J. Wan, S. Shen, Y. Xing, and R. P. Carstens. Genome-Wide Determination of a Broad ESRP-Regulated Posttranscriptional Network by High-Throughput Sequencing. *Molecular and Cellular Biology*, 32(8):1468–1482, apr 2012. 72

[52] S. Domcke, R. Sinha, D. a. Levine, C. Sander, and N. Schultz. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature communications*, 4:2126, jan 2013. 8

[53] DREAM Challenge. AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge - syn4231880, 2015. 22

[54] D. Eddelbuettel and R. François. Rcpp : Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8):1–18, apr 2011. 109

[55] J. Eswaran, A. Horvath, S. Godbole, S. D. Reddy, P. Mudvari, K. Ohshiro, D. Cyanam, S. Nair, S. A. W. Fuqua, K. Polyak, L. D. Florea, and R. Kumar. RNA sequencing of cancer reveals novel splicing alterations. *Scientific reports*, 3:1689, 2013. 25, 70

[56] S. Falgreen, K. Dybkær, K. H. Young, Z. Y. Xu-Monette, T. C. El-Galaly, M. B. Laursen, J. S. Bødker, M. K. Kjeldsen, A. Schmitz, M. Nyegaard, H. E. Johnsen, and M. Bøgsted. Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. *BMC cancer*, 15:235, 2015. 8

[57] M. Fallahi-Sichani, S. Honarnejad, L. M. Heiser, J. W. Gray, and P. K. Sorger. Metrics other than potency reveal systematic variation in responses to cancer drugs. *Nature chemical biology*, 9(11):708–14, 2013. 42

[58] F. Favero, T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund. Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology*, 26(1):64–70, jan 2015. 76

[59] P. Fici, G. Gallerani, A.-P. Morel, L. Mercatali, T. Ibrahim, E. Scarpi, D. Amadori, A. Puisieux, M. Rigaud, and F. Fabbri. Splicing factor ratio as an index of epithelial-mesenchymal transition and tumor aggressiveness in breast cancer. *Oncotarget*, 8(2):2423–2436, nov 2016. 72

[60] W. D. Foulkes, I. E. Smith, and J. S. Reis-Filho. Triple-Negative Breast Cancer. *New England Journal of Medicine*, 363(20):1938–1948, nov 2010. 63, 64

[61] W. H. Fridman, F. Pagès, C. Sautès-Fridman, and J. Galon. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer*, 12(4):298–306, 2012. 120

[62] H. Fröhlich. Network Based Consensus Gene Signatures for Biomarker Discovery in Breast Cancer. *PLoS ONE*, 6(10):e25364, oct 2011. 120

[63] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, Q. Liu, F. Iorio, D. Surdez, L. Chen, R. J. Milano, G. R. Bignell, A. T. Tam, H. Davies, J. A. Stevenson, S. Barthorpe, S. R. Lutz, F. Kogera, K. Lawrence, A. Mclaren-douglas, T. Mironenko, H. Thi, L. Richardson, W. Zhou, W. Hur, W. Yang, X. Deng, A. Butler, H. G. Choi, J. W. Chang, J. Baselga, I. Stamenkovic, J. A. Engelman, V. Sreenath, O. Delattre, J. Saez-rodriguez, N. S. Gray, P. A. Futreal, D. A. Haber, and M. R. Stratton. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012. 8, 33

[64] Y. Ge, T. P. Speed, and S. Dudoit. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003. 115

[65] P. Geeleher, N. J. Cox, and R. S. Huang. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome biology*, 15(3):R47, mar 2014. 8

[66] C. A. Geisberg and D. B. Sawyer. Mechanisms of anthracycline cardiotoxicity and strategies to decrease cardiac damage. *Current Hypertension Reports*, 12(6):404–410, 2010. 70

[67] J. J. Goeman and A. Solari. Multiple Testing for Exploratory Research. *Institute of Mathematical Statistics*, 26(4):584–597, 2011. 105

[68] A. Gonçalves, O. Trédan, C. Villanueva, and C. Dumontet. Les anticorps conjugués en oncologie: Du concept au trastuzumab emtansine (T-DM1), 2012. 12

[69] I. Gonzalez. Tutorial : Statistical analysis of RNA-Seq data. 2014. 24

[70] J. Greshock, K. E. Bachman, Y. Y. Degenhardt, J. Jing, Y. H. Wen, S. Eastman, E. McNeil, C. Moy, R. Wegrzyn, K. Auger, M. A. Hardwicke, and R. Wooster. Molecular target class is predictive of in vitro response profile. *Cancer Research*, 70(9):3677–3686, 2010. 43, 56

[71] B. G. Haffty, Q. Yang, M. Reiss, T. Kearney, S. A. Higgins, J. Weidhaas, L. Harris, W. Hait, and D. Toppmeyer. Locoregional relapse and distant metastasis in conservatively managed triple negative early-stage breast cancer. *Journal of Clinical Oncology*, 24(36):5652–5657, dec 2006. 12

[72] B. Haibe-Kains, N. El-Hachem, N. J. Birkbak, A. C. Jin, A. H. Beck, H. J. W. L. Aerts, and J. Quackenbush. Inconsistency in large pharmacogenomic studies. *Nature*, 504(7480):389–93, dec 2013. 8, 34, 36, 42, 46, 47, 48, 53, 186

[73] N. Hamajima, K. Hirose, K. Tajima, T. Rohan, C. M. Friedenreich, E. E. Calle, S. M. Gapstur, A. V. Patel, R. J. Coates, J. M. Liff, R. Talamini, N. Chantarakul, S. Koetsawang, D. Rachawat, Y. Marcou, E. Kakouri, S. W. Duffy, A. Morabia, L. Schuman, W. Stewart, M. Szklo, P. F. Coogan, J. R. Palmer, L. Rosenberg, P. Band, A. J. Coldman, R. P. Gallagher, T. G. Hislop, P. Yang, S. R. Cummings, K. Canfell, F. Sitas, P. Chao, J. Lissowska, P. L. Horn-Ross, E. M. John, L. M. Kolonel, A. M. Nomura, R. Ghiasvand, J. Hu, K. C. Johnson, Y. Mao, V. Beral, D. Bull, K. Callaghan, B. Crossley, A. Goodill, J. Green, C. Hermon, T. Key, I. Lindgard, B. Liu, K. Pirie, G. Reeves, R. Collins, R. Doll, R. Peto, T. Bishop, I. S. Fentiman, S. De Sanjosé, C. A. Gonzalez, N. Lee, P. Marchbanks, H. W. Ory, H. B. Peterson, P. Wingo, K. Ebeling, D. Kunde, P. Nishan, J. L. Hopper, H. Eliassen, S. Hankinson, V. Gajalakshmi, N. Martin, T. Pardthaisong, S. Silpisornkosol, C. Theetranont, B. Boosiri, S. Chutivongse, P. Jimakorn, P. Virutamasen, C. Wongsrichanalai, A. Neugut, R. Santella, C. J. Baines, N. Kreiger, A. B. Miller, C. Wall, A. Tjonneland, T. Jorgensen, C. Stahlberg, A. T. Pedersen, D. Flesch-Janys, N. Hakansson, J. Cauley, I. Heuch, H. O. Adami, I. Persson, E. Weiderpass, C. Magnusson, J. Chang-Claude, R. Kaaks, M. McCredie, C. Paul, D. C. Skegg, G. F. Spears, M. Iwasaki, S. Tsugane, G. Anderson, J. R. Daling, J. Hampton, W. B. Hutchinson, C. I. Li, K. Malone, M. Mandelson, P. Newcomb, E. A. Noonan, R. M. Ray, J. L. Stanford, M. T. Tang, D. B. Thomas, N. S. Weiss, E. White, A. Izquierdo, P. Viladiu, E. O. Fourkala, I. Jacobs, U. Menon, A. Ryan, H. R. Cuevas, P. Ontiveros, A. Palet, S. B. Salazar, N. Aristizabal, A. Cuadros, L. Tryggvadottir, H. Tulinius, E. Riboli, N. Andrieu, A. Bachelot, M. G. Lê, A. Brêmond, B. Gairard, J. Lansac, L. Piana, R. Renaud, F. Clavel-Chapelon, A. Fournier, M. Touillaud, S. Mesrine, N. Chabbert-Buffet, M. C. Boutron-Ruault, A. Wolk, G. Torres-Mejia, S. Franceschi, I. Romieu, P. Boyle, F. Lubin, B. Modan, E. Ron, Y. Wax, G. D. Friedman, R. A. Hiatt, F. Levi, K. Kosmelj, M. Primic-Zakelj, B. Ravnihar, J. Stare, A. Ekbom, G. Erlandsson, I. Persson, W. L. Beeson, G. Fraser, J. Peto, R. L. Hanson, M. C. Leske, M. C. Mahoney, P. C. Nasca, A. O. Varma, A. L. Weinstein, M. L. Hartman, H. Olsson, R. A. Goldbohm, P. A. van den Brandt, D. Palli, S. Teitelbaum, R. A. Apelo, J. Baens, J. R. de la Cruz, B. Javier, L. B. Lacaya, C. A. Ngelangel, C. La Vecchia, E. Negri, E. Marubini, M. Ferraroni, M. C. Pike, M. Gerber, S. Richardson, C. Segala, D. Gatei, P. Kenya, A. Kungu, J. G. Mati, L. A. Brinton, M. Freedman, R. Hoover, C. Schairer, R. Ziegler, E. Banks, R. Spirtas, H. P. Lee, M. A. Rookus, F. E. van Leeuwen, J. A. Schoenberg, S. Graff-Iversen, R. Selmer, L. Jones, K. McPherson, A. Neil, M. Vessey, D. Yeates, K. Mabuchi, D. Preston, P. Hannaford, C. Kay, S. E. McCann, L. Rosero-Bixby, Y. T. Gao, F. Jin, J. M. Yuan, H. Y. Wei, T. Yun, C. Zhiheng, G. Berry, J. C. Booth, T. Jelihovsky, R. MacLennan, R. Shearman, A. Hadjisavvas, K. Kyriacou, M. Loisidou, X. Zhou, Q. S. Wang, M. Kawai, Y. Minami, I. Tsuji, E. Lund, M. Kumle, H. Stalsberg, X. O. Shu, W. Zheng, E. M. Monninkhof, N. C. Onland-Moret, P. H. Peeters, K. Katsouyanni, A. Trichopoulou,

D. Trichopoulos, A. Tzonou, K. A. Baltzell, A. Dabancens, L. Martinez, R. Molina, O. Salas, F. E. Alexander, K. Anderson, A. R. Folsom, M. D. Gammon, B. S. Hulka, R. Millikan, C. E. Chilvers, F. Lumachi, C. Bain, F. Schofield, V. Siskind, T. R. Rebbeck, L. R. Bernstein, S. Enger, R. W. Haile, A. Paganini-Hill, R. K. Ross, G. Ursin, A. H. Wu, M. C. Yu, M. Ewertz, E. A. Clarke, L. Bergkvist, G. L. Anderson, M. Gass, M. J. O'Sullivan, A. Kalache, T. M. Farley, S. Holck, O. Meirik, and A. Fukao. Menarche, menopause, and breast cancer risk: Individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *The Lancet Oncology*, 13(11):1141–1151, nov 2012. 9

[74] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000. 2, 6

[75] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011. 2, 3, 6, 8

[76] J. M. Harvey, G. M. Clark, C. K. Osborne, and D. C. Allred. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. *Journal of Clinical Oncology*, 17(5):1474–1481, may 1999. 88

[77] C. Hatzis, P. L. Bedard, N. J. Birkbak, A. H. Beck, H. J. W. L. Aerts, D. F. Stern, L. Shi, R. Clarke, J. Quackenbush, and B. Haibe-Kains. Enhancing reproducibility in cancer drug screening: how do we move forward? *Cancer research*, 74(15):4016–23, aug 2014. 8

[78] P. M. Haverty, E. Lin, J. Tan, Y. Yu, B. Lam, S. Lianoglou, R. M. Neve, S. Martin, J. Settleman, R. L. Yauch, and R. Bourgon. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature*, 533(7603):333–337, 2016. 43, 48, 57

[79] H.Bonsang-Kitzis, B.Sadacca, AS.Hamy-Petit, M.Moarii, A.Pinheiro, C.Laurent, and F.Reyal. Biological network-driven gene selection identifies a stromal immune module as a key determinant of triple-negative breast carcinoma prognosis. *OncoImmunology*, (June):37–41, 2015. 46, 135

[80] M. A. Healey, R. Hu, A. H. Beck, L. C. Collins, S. J. Schnitt, R. M. Tamimi, and A. Hazra. Association of H3K9me3 and H3K27me3 repressive histone marks with breast cancer subtypes in the Nurses??? Health Study. *Breast Cancer Research and Treatment*, 147(3):639–651, 2014. 168

[81] L. M. Heiser, A. Sadanandam, W.-L. Kuo, S. C. Benz, T. C. Goldstein, S. Ng, W. J. Gibb, N. J. Wang, S. Ziyad, F. Tong, N. Bayani, Z. Hu, J. I. Billig, A. Dueregger, S. Lewis, L. Jakkula, J. E. Korkola, S. Durinck, F. Pepin, Y. Guan, E. Purdom, P. Neuvial, H. Bengtsson, K. W. Wood, P. G. Smith, L. T. Vassilev, B. T. Hennessy, J. Greshock, K. E. Bachman, M. A. Hardwicke, J. W. Park, L. J. Marton, D. M. Wolf, E. A. Collisson, R. M. Neve, G. B. Mills, T. P. Speed, H. S. Feiler, R. F. Wooster, D. Haussler, J. M. Stuart, J. W. Gray, and P. T. Spellman. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences*, 109(8):2724–2729, 2012. 34, 42

[82] G. H. Heppner and B. E. Miller. Tumor heterogeneity: biological implications and therapeutic consequences. *CANCER AND METASTASIS REVIEW*, 2(1):5–23, 1983. 4

[83] K. P. Hoeflich, C. O'Brien, Z. Boyd, G. Cavet, S. Guerrero, K. Jung, T. Januario, H. Savage, E. Punnoose, T. Truong, W. Zhou, L. Berry, L. Murray, L. Amler, M. Belvin, L. S. Friedman, and M. R. Lackner. In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models. *Clin Cancer Res*, 15(14):4649–4664, 2009. 57

[84] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer*, 13(10):714–726, 2013. 16

[85] D. Horiuchi, L. Kusdra, N. E. Huskey, S. Chandriani, M. E. Lenburg, A. M. Gonzalez-Angulo, K. J. Creasman, A. V. Bazarov, J. W. Smyth, S. E. Davis, P. Yaswen, G. B. Mills, L. J. Esserman, and A. Goga. MYC pathway activation in triple-negative breast cancer is synthetic lethal with CDK inhibition. *The Journal of Experimental Medicine*, 209(4):679–696, apr 2012. 74

[86] G. N. Hortobagyi. Trastuzumab in the Treatment of Breast Cancer. *New England Journal of Medicine*, 353(16):1734–1736, oct 2005. 12

[87] N. Howlader, A. Noone, M. Krpcho, D. Miller, K. Bishop, C. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. Lewis, H. Chen, E. Feuer, and K. Cronin. SEER Cancer Statistics Review, 1975-2014, National Cancer Institute. Technical report, 2016. 14

[88] D. M. Hyman, I. Puzanov, V. Subbiah, J. E. Faris, I. Chau, J.-Y. Blay, J. Wolf, N. S. Raje, E. L. Diamond, A. Hollebecque, R. Gervais, M. E. Elez-Fernandez, A. Italiano, R.-D. Hofheinz, M. Hidalgo, E. Chan, M. Schuler, S. F. Lasserre, M. Makrutzki, F. Sirzen, M. L. Veronese, J. Tabernero, and J. Baselga. Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations. *The New England journal of medicine*, 373(8):726–36, aug 2015. 34

[89] M. Ignatiadis, S. K. Singhal, C. Desmedt, B. Haibe-Kains, C. Criscitiello, F. Andre, S. Loi, M. Piccart, S. Michiels, and C. Sotiriou. Gene Modules and Response to Neoadjuvant Chemotherapy in Breast Cancer Subtypes: A Pooled Analysis. *Journal of Clinical Oncology*, 30(16):1996–2004, 2012. 120, 135

[90] T. I. C. G. International Cancer Genome Consortium, T. J. Hudson, W. Anderson, A. Artez, A. D. Barker, C. Bell, R. R. Bernabé, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, A. Guttmacher, M. Guyer, F. M. Hemsley, J. L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusada, D. P. Lane, F. Laplace, L. Youyong, G. Nettekoven, B. Ozenberger, J. Peterson, T. S. Rao, J. Remacle, A. J. Schafer, T. Shibata, M. R. Stratton, J. G. Vockley, K. Watanabe, H. Yang, M. M. F. Yuen, B. M. Knoppers, M. Bobrow, A. Cambon-Thomsen, L. G. Dressler, S. O. M. Dyke, Y. Joly, K. Kato, K. L. Kennedy, P. Nicolás, M. J. Parker, E. Rial-Sebbag, C. M. Romeo-Casabona, K. M. Shaw, S. Wallace, G. L. Wiesner, N. Zeps, P. Lichter, A. V. Biankin, C. Chabannon, L. Chin, B. Clément, E. de Alava, F. Degos, M. L. Ferguson, P. Geary, D. N. Hayes, T. J. Hudson, A. L. Johns, A. Kasprzyk, H. Nakagawa, R. Penny, M. A. Piris, R. Sarin, A. Scarpa, T. Shibata, M. van de Vijver,

P. A. Futreal, H. Aburatani, M. Bayés, D. D. L. Botwell, P. J. Campbell, X. Estivill, D. S. Gerhard, S. M. Grimmond, I. Gut, M. Hirst, C. López-Otín, P. Majumder, M. Marra, J. D. McPherson, H. Nakagawa, Z. Ning, X. S. Puente, Y. Ruan, T. Shibata, M. R. Stratton, H. G. Stunnenberg, H. Swerdlow, V. E. Velculescu, R. K. Wilson, H. H. Xue, L. Yang, P. T. Spellman, G. D. Bader, P. C. Boutros, P. J. Campbell, P. Flicek, G. Getz, R. Guigó, G. Guo, D. Haussler, S. Heath, T. J. Hubbard, T. Jiang, S. M. Jones, Q. Li, N. López-Bigas, R. Luo, L. Muthuswamy, B. F. F. Ouellette, J. V. Pearson, X. S. Puente, V. Quesada, B. J. Raphael, C. Sander, T. Shibata, T. P. Speed, L. D. Stein, J. M. Stuart, J. W. Teague, Y. Totoki, T. Tsunoda, A. Valencia, D. A. Wheeler, H. Wu, S. Zhao, G. Zhou, L. D. Stein, R. Guigó, T. J. Hubbard, Y. Joly, S. M. Jones, A. Kasprzyk, M. Lathrop, N. López-Bigas, B. F. F. Ouellette, P. T. Spellman, J. W. Teague, G. Thomas, A. Valencia, T. Yoshida, K. L. Kennedy, M. Axton, S. O. M. Dyke, P. A. Futreal, D. S. Gerhard, C. Gunter, M. Guyer, T. J. Hudson, J. D. McPherson, L. J. Miller, B. Ozenberger, K. M. Shaw, A. Kasprzyk, L. D. Stein, J. Zhang, S. A. Haider, J. Wang, C. K. Yung, A. Cros, A. Cross, Y. Liang, S. Gnaneshan, J. Guberman, J. Hsu, M. Bobrow, D. R. C. Chalmers, K. W. Hasel, Y. Joly, T. S. H. Kaan, K. L. Kennedy, B. M. Knoppers, W. W. Lowrance, T. Masui, P. Nicolás, E. Rial-Sebbag, L. L. Rodriguez, C. Vergely, T. Yoshida, S. M. Grimmond, A. V. Biankin, D. D. L. Bowtell, N. Cloonan, A. DeFazio, J. R. Eshleman, D. Etemadmoghadam, B. B. Gardiner, B. A. Gardiner, J. G. Kench, A. Scarpa, R. L. Sutherland, M. A. Tempero, N. J. Waddell, P. J. Wilson, J. D. McPherson, S. Gallinger, M.-S. Tsao, P. A. Shaw, G. M. Petersen, D. Mukhopadhyay, L. Chin, R. A. DePinho, S. Thayer, L. Muthuswamy, K. Shazand, T. Beck, M. Sam, L. Timms, V. Ballin, Y. Lu, J. Ji, X. Zhang, F. Chen, X. Hu, G. Zhou, Q. Yang, G. Tian, L. Zhang, X. Xing, X. Li, Z. Zhu, Y. Yu, J. Yu, H. Yang, M. Lathrop, J. Tost, P. Brennan, I. Holcatova, D. Zaridze, A. Brazma, L. Egevard, E. Prokhortchouk, R. E. Banks, M. Uhlén, A. Cambon-Thomsen, J. Viksna, F. Ponten, K. Skryabin, M. R. Stratton, P. A. Futreal, E. Birney, A. Borg, A.-L. Børresen-Dale, C. Caldas, J. A. Foekens, S. Martin, J. S. Reis-Filho, A. L. Richardson, C. Sotiriou, H. G. Stunnenberg, G. Thoms, M. van de Vijver, L. van't Veer, F. Calvo, D. Birnbaum, H. Blanche, P. Boucher, S. Boyault, C. Chabannon, I. Gut, J. D. Masson-Jacquemier, M. Lathrop, I. Pauporté, X. Pivot, A. Vincent-Salomon, E. Tabone, C. Theillet, G. Thomas, J. Tost, I. Treilleux, F. Calvo, P. Bioulac-Sage, B. Clément, T. Decaens, F. Degos, D. Franco, I. Gut, M. Gut, S. Heath, M. Lathrop, D. Samuel, G. Thomas, J. Zucman-Rossi, P. Lichter, R. Eils, B. Brors, J. O. Korbel, A. Korshunov, P. Landgraf, H. Lehrach, S. Pfister, B. Radlwimmer, G. Reifenberger, M. D. Taylor, C. von Kalle, P. P. Majumder, R. Sarin, T. S. Rao, M. K. Bhan, A. Scarpa, P. Pederzoli, R. A. Lawlor, M. Delledonne, A. Bardelli, A. V. Biankin, S. M. Grimmond, T. Gress, D. Klimstra, G. Zamboni, T. Shibata, Y. Nakamura, H. Nakagawa, J. Kusada, T. Tsunoda, S. Miyano, H. Aburatani, K. Kato, A. Fujimoto, T. Yoshida, E. Campo, C. López-Otín, X. Estivill, R. Guigó, S. de Sanjosé, M. A. Piris, E. Montserrat, M. González-Díaz, X. S. Puente, P. Jares, A. Valencia, H. Himmelbauer, H. Himmelbaue, V. Quesada, S. Bea, M. R. Stratton, P. A. Futreal, P. J. Campbell, A. Vincent-Salomon, A. L. Richardson, J. S. Reis-Filho, M. van de Vijver, G. Thomas, J. D. Masson-Jacquemier, S. Aparicio, A. Borg, A.-L. Børresen-Dale, C. Caldas, J. A. Foekens, H. G. Stunnenberg, L. van't Veer, D. F. Easton, P. T. Spellman, S. Martin, A. D. Barker, L. Chin, F. S. Collins, C. C. Compton, M. L. Ferguson, D. S. Gerhard, G. Getz, C. Gunter, A. Guttmacher, M. Guyer, D. N. Hayes, E. S. Lander, B. Ozenberger, R. Penny, J. Peterson, C. Sander, K. M. Shaw, T. P. Speed, P. T. Spellman, J. G. Vockley, D. A. Wheeler, R. K. Wilson, T. J. Hudson, L. Chin, B. M. Knoppers, E. S. Lander, P. Lichter, L. D. Stein, M. R. Stratton, W. Anderson, A. D. Barker, C. Bell, M. Bobrow, W. Burke, F. S. Collins, C. C. Compton, R. A. DePinho, D. F. Easton, P. A. Futreal, D. S. Gerhard, A. R. Green, M. Guyer, S. R. Hamilton, T. J. Hubbard, O. P. Kallioniemi, K. L. Kennedy, T. J. Ley, E. T. Liu, Y. Lu, P. Majumder, M. Marra, B. Ozenberger, J. Peterson, A. J. Schafer, P. T. Spellman, H. G. Stunnenberg, B. J. Wainwright, R. K. Wilson, and H. Yang. International network of cancer genome projects. *Nature*, 464(7291):993–8, apr 2010. 30

[91] J. P. A. Ioannidis. Why most published research findings are false, aug 2005. 189

[92] R. A. Irizarry. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, apr 2003. 47

[93] T. Jiang, W. Shi, V. B. Wali, L. S. Pongor, C. Li, R. Lau, B. Győrffy, R. P. Lifton, W. F. Symmans, L. Pusztai, and C. Hatzis. Predictors of Chemosensitivity in Triple Negative Breast Cancer: An Integrated Genomic Analysis. *PLoS Medicine*, 13(12):1–23, 2016. 64, 86

[94] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, M. D. M. Leiserson, C. A. Miller, J. S. Welch, M. J. Walter, M. C. Wendl, T. J. Ley, R. K. Wilson, B. J. Raphael, and L. Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 2013. 78

[95] C. G. Kleer, Q. Cao, S. Varambally, R. Shen, I. Ota, S. A. Tomlins, D. Ghosh, R. G. A. B. Sewalt, A. P. Otte, D. F. Hayes, M. S. Sabel, D. Livant, S. J. Weiss, M. A. Rubin, and A. M. Chinnaiyan. EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proceedings of the National Academy of Sciences*, 100(20):11606–11611, 2003. 168

[96] C. Klijn, S. Durinck, E. W. Stawiski, P. M. Haverty, Z. Jiang, H. Liu, J. Degenhardt, O. Mayba, F. Gnad, J. Liu, G. Pau, J. Reeder, Y. Cao, K. Mukhyala, S. K. Selvaraj, M. Yu, G. J. Zynda, M. J. Brauer, T. D. Wu, R. C. Gentleman, G. Manning, R. L. Yauch, R. Bourgon, D. Stokoe, Z. Modrusan, R. M. Neve, F. J. D. Sauvage, J. Settleman, S. Seshagiri, and Z. Zhang. A comprehensive transcriptional portrait of human cancer cell lines. *Nature Biotechnology*, (December), 2014. 40

[97] G. E. Konecny, M. D. Pegram, N. Venkatesan, R. Finn, G. Yang, M. Rahmeh, M. Untch, D. W. Rusnak, G. Spehar, R. J. Mullin, B. R. Keith, T. M. Gilmer, M. Berger, K. C. Podratz, and D. J. Slamon. Activity of the dual kinase inhibitor lapatinib (GW572016) against HER-2-overexpressing and trastuzumab-treated breast cancer cells. *Cancer research*, 66(3):1630–9, feb 2006. 8, 34

[98] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29, 2014. 24, 25, 90

[99] C. Le Tourneau, J.-P. Delord, A. Gonçalves, C. Gavoille, C. Dubot, N. Isambert, M. Campone, O. Trédan, M.-A. Massiani, C. Mauborgne, S. Armanet, N. Servant, I. Bièche, V. Bernard, D. Gentien, P. Jezequel, V. Attignon, S. Boyault, A. Vincent-Salomon, V. Servois, M.-P. Sablin, M. Kamal, and X. Paoletti. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *The Lancet Oncology*, 16(13):1324–1334, 2015. 84

[100] B. Lehmann and J. Bauer. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of of Clinical Investigation*, 2011. 13, 46, 86, 187

[101] G. Li, W. Hu, J. Wang, X. Deng, P. Zhang, X. Zhang, C. Xie, and S. Wu. Phase II study of concurrent chemoradiation in combination with erlotinib for locally advanced esophageal carcinoma. *International Journal of Radiation Oncology Biology Physics*, 78(5):1407–1412, dec 2010. 47

[102] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009. 109

[103] Q. Li, N. J. Birkbak, B. Gyorffy, Z. Szallasi, and A. C. Eklund. Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics*, 12(1):474, 2011. 48

[104] C. Liedtke, C. Mazouni, K. R. Hess, F. André, A. Tordai, J. A. Mejia, W. F. Symmans, A. M. Gonzalez-Angulo, B. Hennessy, M. Green, M. Cristofanilli, G. N. Hortobagyi, and L. Pusztai. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *Journal of Clinical Oncology*, 26(8):1275–1281, 2008. 16, 46, 64

[105] L. S. Lindström, E. Karlsson, U. M. Wilking, U. Johansson, J. Hartman, E. K. Lidbrink, T. Hatschek, L. Skoog, and J. Bergh. Clinically used breast cancer markers such as estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 are unstable throughout tumor progression. *Journal of Clinical Oncology*, 30(21):2601–2608, 2012. 87

[106] E. H. Lips, M. Michaut, M. Hoogstraat, L. Mulder, N. J. Besselink, M. J. Koudijs, E. Cuppen, E. E. Voest, R. Bernards, P. M. Nederlof, J. Wesseling, S. Rodenhuis, and L. F. Wessels. Next generation sequencing of triple negative breast cancer to find predictors for chemotherapy response. *Breast Cancer Res*, 17(1):134–43, 2015. 79, 86

[107] E. H. Lips, L. Mulder, A. Oonk, L. E. van der Kolk, F. B. L. Hogervorst, A. L. T. Imholz, J. Wesseling, S. Rodenhuis, and P. M. Nederlof. Triple-negative breast cancer: BRCAness and concordance of clinical features with BRCA1-mutation carriers. *British Journal of Cancer*, 108(10):2172–2177, 2013. 14

[108] Z. Liu, G. Zhu, R. H. Getzenberg, and R. W. Veltri. The upregulation of PI3K/Akt and MAP kinase pathways is associated with resistance of microtubule-targeting drugs in prostate cancer. *Journal of Cellular Biochemistry*, 116(7):1341–1349, 2015. 70

[109] D. B. Longley and P. G. Johnston. Molecular mechanisms of drug resistance. *Journal of Pathology*, 205(2):275–292, 2005. 17

[110] F. Lorio, T. A. Knijnenburg, D. J. Vis, J. Saez-Rodriguez, U. Mcdermott, M. J. G. Correspondence, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonç Alves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. Van Dyk, H. Chang, H. De Silva, H. Heyn, X. Deng, R. K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, and M. J. Garnett. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 16613616(421):1–15, 2016. 8, 79, 82, 83

[111] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*, pages 1–21, 2014. 90

[112] E. Marangoni, A. Vincent-Salomon, N. Auger, A. Degeorges, F. Assayag, P. de Cremoux, L. de Plater, C. Guyader, G. De Pinieux, J.-G. Judde, M. Rebucci, C. Tran-Perennou, X. Sastre-Garau, B. Sigal-Zafrani, O. Delattre, V. Dieras, and M.-F. Poupon. A New Model of Patient Tumor-Derived Breast Cancer Xenografts for Preclinical Assays. *Clinical Cancer Research*, 13(13):3989–3998, jul 2007. 7

[113] R. Marcus, E. Peritz, and K. E. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976. 106

[114] E. Martis. High-Throughput Screening: The Hits and Leads of Drug Discovery- An Overview — Elvis Martis - Academia.edu. *Journal of Applied Pharmaceutical Science*, 01(01):02–10, 2011. 22

[115] M. Marty, F. Cognetti, D. Maraninchi, R. Snyder, L. Mauriac, M. Tubiana-Hulin, S. Chan, D. Grimes, A. Antón, A. Lluch, J. Kennedy, K. O'Byrne, P. Conte, M. Green, C. Ward, K. Mayne, and J.-M. Extra. Randomized phase II trial of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer administered as first-line treatment: the M77001 study group. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 23(19):4265–74, jul 2005. 6

[116] N. Matosin, E. Frank, M. Engel, J. S. Lum, and K. A. Newell. Negativity towards negative results: a discussion of the disconnect between scientific worth and scientific culture. *Disease Models & Mechanisms*, 7(2):171–173, feb 2014. 187

[117] M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2):242–253, apr 2010. 47

[118] N. McGranahan and C. Swanton. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, 168(4):613–628, 2017. 4

[119] D. W. McMillin, J. M. Negri, and C. S. Mitsiades. The role of tumour-stromal interactions in modifying drug response: challenges and opportunities. *Nature reviews. Drug discovery*, 12(3):217–28, 2013. 8, 16, 18

[120] T. Minami, T. Kijima, Y. Otani, S. Kohmo, R. Takahashi, I. Nagatomo, H. Hirata, M. Suzuki, K. Inoue, Y. Takeda, H. Kida, I. Tachibana, and A. Kumanogoh. HER2 As Therapeutic Target for Overcoming ATP-Binding Cassette Transporter-Mediated Chemoresistance in Small Cell Lung Cancer. *Molecular Cancer Therapeutics*, 11(4):830–841, apr 2012. 46

[121] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus Clustering - A resampling-based method for class discovery and visualization of gene expression microarray data Stefano. *Machine Learning*, 52(1):1–34, 2003. 24, 34

[122] L. Moreno and A. D. J. Pearson. How can attrition rates be reduced in cancer drug discovery? *Expert opinion on drug discovery*, 8(4):363–8, 2013. 33

[123] E. Moretti, C. Desmedt, C. Biagioni, M. M. Regan, C. Oakman, D. Larsimont, F. Galardi, M. Piccart-Gebhart, C. Sotiriou, D. L. Rimm, and A. Di Leo. TOP2A protein by quantitative immunofluorescence as a predictor of response to epirubicin in the neoadjuvant treatment of breast cancer. *Future oncology (London, England)*, 9(10):1477–87, oct 2013. 87

[124] S. Myhre, O.-c. Lingjærde, B. T. Hennessy, M. R. Aure, M. S. Carey, J. Alsner, T. Tramm, J. Overgaard, G. B. Mills, A.-L. Børresen-Dale, and T. Sørlie. Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins. *Molecular Oncology*, 7(3):704–718, 2013. 86

[125] J. Y. Nam, N. K. Kim, S. C. Kim, J. G. Joung, R. Xi, S. Lee, P. J. Park, and W. Y. Park. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. *Briefings in Bioinformatics*, 17(2):185–192, 2016. 32

[126] I. national du cancer. Les cancers en france. *Institut national du cancer*, pages 1–240, 2016. 8

[127] P. C. Ng and S. Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, jul 2003. 82

[128] S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, A. Menzies, S. Martin, K. Leung, L. Chen, C. Leroy, M. Ramakrishna, R. Rance, K. W. Lau, L. J. Mudie, I. Varela, D. J. McBride, G. R. Bignell, S. L. Cooke, A. Shlien, J. Gamble, I. Whitmore, M. Maddison, P. S. Tarpey, H. R. Davies, E. Papaemmanuil, P. J. Stephens, S. McLaren, A. P. Butler, J. W. Teague, G. Jönsson, J. E. Garber, D. Silver, P. Miron, A. Fatima, S. Boyault, A. Langerod, A. Tutt, J. W. Martens, S. A. Aparicio, Å. Borg, A. V. Salomon, G. Thomas, A. L. Borresen-Dale, A. L. Richardson, M. S. Neuberger, P. A. Futreal, P. J. Campbell, and M. R. Stratton. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 2012. 80

[129] S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman, S. Morganella, M. R. Aure, O. C. Lingjærde, A. Langerød, M. Ringnér, S.-M. Ahn, S. Boyault, J. E. Brock, A. Broeks, A. Butler, C. Desmedt, L. Dirix, S. Dronov, A. Fatima, J. A. Foekens, M. Gerstung, G. K. J. Hooijer, S. J. Jang, D. R. Jones, H.-Y. Kim, T. A. King, S. Krishnamurthy, H. J. Lee, J.-y. Lee, Y. Li, S. McLaren, A. Menzies, V. Mustonen, S. O'Meara, I. Pauporté, X. Pivot, C. A. Purdie, K. Raine, K. Ramakrishnan, F. G. Rodríguez-González, G. Romieu, A. M. Sieuwerts, P. T. Simpson, R. Shepherd, L. Stebbings, O. A. Stefansson, J. Teague, S. Tommasi, I. Treilleux, G. G. Van den Eynden, P. Vermeulen, A. Vincent-Salomon, L. Yates, C. Caldas, L. van't Veer, A. Tutt, S. Knappskog, B. K. T. Tan, J. Jonkers, Å. Borg, N. T. Ueno, C. Sotiriou, A. Viari, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, J. W. M. Martens, A.-L. Børresen-Dale, A. L. Richardson, G. Kong, G. Thomas, and M. R. Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, pages 1–20, 2016. 14, 79, 80

[130] S. Nofech-Mozes, M. Trudeau, H. K. Kahn, R. Dent, E. Rawlinson, P. Sun, S. A. Narod, and W. M. Hanna. Patterns of recurrence in the basal and non-basal subtypes of triple-negative breast cancers. *Breast Cancer Research and Treatment*, 118(1):131–137, nov 2009. 63

[131] P. B. Olkhanud, Y. Rochman, M. Bodogai, E. Malchinkhuu, K. Wejksza, M. Xu, R. E. Gress, C. Hesdorffer, W. J. Leonard, and A. Biragyn. Thymic Stromal Lymphopoietin Is a Key Mediator of Breast Cancer Progression. *The Journal of Immunology*, 186(10):5656–5662, 2011. 157, 158

[132] E. A. O'Reilly, L. Gubbins, S. Sharma, R. Tully, M. H. Z. Guang, K. Weiner-Gorzel, J. McCaffrey, M. Harrison, F. Furlong, M. Kell, and A. McCann. The fate of chemoresistance in triple negative breast cancer (TNBC). *BBA Clinical*, 3:257–275, 2015. 16, 18

[133] J. W. Park, S. Jung, E. C. Rouchka, Y.-T. Tseng, and Y. Xing. rMAPS: RNA map analysis and plotting server for alternative exon regulation. *Nucleic Acids Research*, 44(W1):W333–W338, 2016. 26, 71, 90, 102

[134] S. Park, C. Shimizu, T. Shimoyama, M. Takeda, M. Ando, T. Kohno, N. Katsumata, Y. K. Kang, K. Nishio, and Y. Fujiwara. Gene expression profiling of ATP-binding cassette (ABC) transporters as a predictor of the pathologic response to neoadjuvant chemotherapy in breast cancer patients. *Breast Cancer Research and Treatment*, 99(1):9–17, 2006. 70

[135] A. Pedroza-Gonzalez, K. Xu, T.-C. Wu, C. Aspord, S. Tindle, F. Marches, M. Gallegos, E. C. Burton, D. Savino, T. Hori, Y. Tanaka, S. Zurawski, G. Zurawski, L. Bover, Y.-J. Liu, J. Banchereau, and A. K. Palucka. Thymic stromal lymphopoietin fosters human breast tumor growth by promoting type 2 inflammation. *The Journal of experimental medicine*, 208(3):479–90, 2011. 157, 158

[136] C. M. Perou, T. Sorlie, M. B. Eisen, M. V. D. Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, I. Fluge, A. Pergamenschikov,

C. Williams, S. X. Zhu, P. E. Lùnning, P. O. Brown, D. Botstein, and S. Grant. Molecular portraits of human breast tumours. *Nature*, 406(May):747–752, 2000. 11, 12, 23

[137] B. Pesch, B. Kendzia, P. Gustavsson, K.-H. Jöckel, G. Johnen, H. Pohlabeln, A. Olsson, W. Ahrens, I. M. Gross, I. Brüske, H.-E. Wichmann, F. Merletti, L. Richiardi, L. Simonato, C. Fortes, J. Siemiatycki, M.-E. Parent, D. Consonni, M. T. Landi, N. Caporaso, D. Zaridze, A. Cassidy, N. Szeszenia-Dabrowska, P. Rudnai, J. Lissowska, I. Stücker, E. Fabianova, R. S. Dumitru, V. Bencko, L. Foretova, V. Janout, C. M. Rudin, P. Brennan, P. Boffetta, K. Straif, and T. Brüning. Cigarette smoking and lung cancer–relative risk estimates for the major histological types from a pooled analysis of case-control studies. *International journal of cancer*, 131(5):1210–9, sep 2012. 9

[138] A. Phipps and C. Li. Breast Cancer Biology and Clinical Characteristics. In *Breast Cancer Epidemiology*, pages 47–72. 2010. 10, 11

[139] M. Pierre-Jean, G. Rigaill, and P. Neuvial. Performance evaluation of DNA copy number segmentation methods. *Briefings in Bioinformatics*, 16(4):600–615, 2014. 31

[140] K. Polyak and O. Metzger Filho. SnapShot: Breast Cancer. *Cancer Cell*, 22(4):562–562.e1, 2012. 14

[141] A. Prahallad, C. Sun, S. Huang, F. Di Nicolantonio, R. Salazar, D. Zecchin, R. L. Beijersbergen, A. Bardelli, and R. Bernards. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature*, 483(7387):100–3, 2012. 7

[142] A. Prat, B. Adamo, M. C. U. Cheang, C. K. Anders, L. A. Carey, and C. M. Perou. Molecular Characterization of Basal-Like and Non-Basal-Like Triple-Negative Breast Cancer. *The Oncologist*, 18(2):123–133, feb 2013. 13

[143] R. S. Pruthi, M. Nielsen, S. Heathcote, E. M. Wallen, W. K. Rathmell, P. Godley, Y. Whang, J. Fielding, H. Schultz, G. Grigson, A. Smith, and W. Kim. A phase II trial of neoadjuvant erlotinib in patients with muscle-invasive bladder cancer undergoing radical cystectomy: Clinical and pathological results. *BJU International*, 106(3):349–356, aug 2010. 47

[144] M. P. Quinlan, S. E. Quatela, M. R. Philips, and J. Settleman. Activated Kras, but Not Hras or Nras, May Initiate Tumors of Endodermal Origin via Stem Cell Expansion. *Molecular and Cellular Biology*, 28(8):2659–2674, apr 2008. 39

[145] J. J. Raizer, L. E. Abrey, A. B. Lassman, S. M. Chang, K. R. Lamborn, J. G. Kuhn, W. K. Yung, M. R. Gilbert, K. A. Aldape, P. Y. Wen, H. A. Fine, M. Mehta, L. M. DeAngelis, F. Lieberman, T. F. Cloughesy, H. I. Robins, J. Dancey, and M. D. Prados. A phase II trial of erlotinib in patients with recurrent malignant gliomas and nonprogressive glioblastoma multiforme postradiation therapy. *Neuro-Oncology*, 12(1):95–103, jan 2010. 57

[146] E. A. Rakha, M. E. El-Sayed, A. R. Green, E. C. Paish, A. H. Lee, and I. O. Ellis. Breast carcinoma with basal differentiation: A proposal for pathology definition based on basal cytokeratin expression. *Histopathology*, 50(4):434–438, mar 2007. 13

[147] E. A. Rakha, D. S. Tan, W. D. Foulkes, I. O. Ellis, A. Tutt, T. O. Nielsen, and J. S. Reis-Filho. Are triple-negative tumours and basal-like breast cancer synonymous? *Breast Cancer Research*, 9(6):404, 2007. 13

[148] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. F. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, jul 2013. 72

[149] M. G. Rees, B. Seashore-ludlow, J. H. Cheah, D. J. Adams, E. V. Price, S. Gill, S. Javaid, M. E. Coletti, V. L. Jones, N. E. Bodycombe, C. K. Soule, B. Alexander, A. Li, J. D. Kotz, C. S.-y. Hon, B. Munoz, T. Liefeld, D. A. Haber, C. B. Clish, J. A. Bittker, M. Palmer, K. Wagner, P. A. Clemons, A. F. Shamji, and S. L. Schreiber. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol*, 12(2):109–116, 2016. 6, 47

[150] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, apr 2015. 24, 49

[151] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, jan 2009. 24

[152] J. P. Romano and M. Wolf. Control of generalized error rates in multiple testing. *Annals of Statistics*, 35(4):1378–1408, 2007. 107

[153] R. Rosenthal, N. McGranahan, J. Herrero, B. S. Taylor, and C. Swanton. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1):31, dec 2016. 80

[154] Z. Safikhani, N. El-Hachem, R. Quevedo, P. Smirnov, A. Goldenberg, N. Juul Birkbak, C. Mason, C. Hatzis, L. Shi, H. J. W. L. Aerts, J. Quackenbush, and B. Haibe-Kains. Assessment of pharmacogenomic agreement. *F1000Research*, 5:825, 2016. 186

[155] T. Sanavia, F. Aiolli, G. Da San Martino, A. Bisognin, and B. Di Camillo. Improving biomarker list stability by integration of biological knowledge in the learning process. *BMC Bioinformatics*, 13(Suppl 4):S22, 2012. 45, 120

[156] J. F. Sathirapongsasuti, H. Lee, B. A. J. Horst, G. Brunner, A. J. Cochran, S. Binder, J. Quackenbush, and S. F. Nelson. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, 27(19):2648–2654, oct 2011. 30

[157] C. M. Sawai, J. Freund, P. Oh, D. Ndiaye-Lobry, J. C. Bretz, A. Strikoudis, L. Genesca, T. Trimarchi, M. A. Kelliher, M. Clark, J. Soulier, S. Chen-Kiang,

and I. Aifantis. Therapeutic Targeting of the Cyclin D3:CDK4/6 Complex in T Cell Leukemia. *Cancer Cell*, 22(4):452–465, oct 2012. 46

[158] M. Schwalbe. *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results*. National Academies Press (US), Washington (DC), 2016. 189

[159] S. P. Shah, A. Roth, R. Goya, A. Oloumi, G. Ha, Y. Zhao, G. Turashvili, J. Ding, K. Tse, G. Haffari, A. Bashashati, L. M. Prentice, J. Khattra, A. Burleigh, D. Yap, V. Bernard, A. McPherson, K. Shumansky, A. Crisan, R. Giuliany, A. Heravi-Moussavi, J. Rosner, D. Lai, I. Birol, R. Varhol, A. Tam, N. Dhalla, T. Zeng, K. Ma, S. K. Chan, M. Griffith, A. Moradian, S.-W. G. Cheng, G. B. Morin, P. Watson, K. Gelmon, S. Chia, S.-F. Chin, C. Curtis, O. M. Rueda, P. D. Pharoah, S. Damaraju, J. Mackey, K. Hoon, T. Harkins, V. Tadigotla, M. Sigaroudinia, P. Gascard, T. Tlsty, J. F. Costello, I. M. Meyer, C. J. Eaves, W. W. Wasserman, S. Jones, D. Huntsman, M. Hirst, C. Caldas, M. A. Marra, and S. Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–9, 2012. 14

[160] I. M. Shapiro, A. W. Cheng, N. C. Flytzanis, M. Balsamo, J. S. Condeelis, M. H. Oktay, C. B. Burge, and F. B. Gertler. An emt-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genetics*, 7(8):e1002218, aug 2011. 72

[161] R. Shen and V. E. Seshan. FACETS: Allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research*, 44(16):1–9, 2016. 20, 72, 91

[162] S. Shen, J. W. Park, Z.-x. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America*, 111(51):E5593–601, 2014. 26, 27, 69, 90

[163] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–11, jan 2001. 29

[164] H. Shigematsu, L. Lin, T. Takahashi, M. Nomura, M. Suzuki, I. I. Wistuba, K. M. Fong, H. Lee, S. Toyooka, N. Shimizu, T. Fujisawa, Z. Feng, J. A. Roth, J. Herz, J. D. Minna, and A. F. Gazdar. Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. *Journal of the National Cancer Institute*, 97(5):339–346, mar 2005. 57

[165] H. A. Shihab, J. Gough, M. Mort, D. N. Cooper, I. N. Day, and T. R. Gaunt. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Human Genomics*, 8(1):11, jun 2014. 82

[166] L. Song, L. Wang, Y. Li, H. Xiong, J. Wu, J. Li, and M. Li. Sam68 up-regulation correlates with, and its down-regulation inhibits, proliferation and tumourigenicity of breast cancer cells. *Journal of Pathology*, 222(3):227–237, jul 2010. 72

[167] H. M. Sontrop, P. D. Moerland, R. van den Ham, M. J. Reinders, and W. F. Verhaegh. A comprehensive sensitivity analysis of microarray breast cancer classification under feature variability. *BMC Bioinformatics*, 10(1):389, 2009. 45, 120

[168] T. Sørlie. Molecular portraits of breast cancer: Tumour subtypes as distinct disease entities. *European Journal of Cancer*, 40(18):2667–2675, 2004. 11, 12

[169] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A. L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–74, 2001. 11, 12, 23

[170] M. B. Sporn. The war on cancer, may 1996. 2

[171] T. Steijger, J. F. Abril, P. G. Engström, and F. Kokocinski. Europe PMC Funders Group Assessment of transcript reconstruction methods for RNA-seq. 10(12):1–20, 2014. 25, 29

[172] N. Stransky, M. Ghandi, G. V. Kryukov, L. A. Garraway, J. Lehár, M. Liu, D. Sonkin, A. Kauffmann, K. Venkatesan, E. J. Edelman, M. Riester, J. Barretina, G. Caponigro, R. Schlegel, W. R. Sellers, F. Stegmeier, M. Morrissey, A. Amzallag, I. Pruteanu-Malinici, D. A. Haber, S. Ramaswamy, C. H. Benes, M. P. Menden, F. Iorio, M. R. Stratton, U. McDermott, M. J. Garnett, and J. Saez-Rodriguez. Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, 2015. 186

[173] M. R. Stratton, P. J. Campbell, and P. Andrew F. The cancer genome. *Nature*, 458(7239):719–724, apr 2009. 4

[174] K. Strimmer. fdrtool: A versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24(12):1461–1462, jun 2008. 90

[175] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, oct 2005. 39

[176] A. Sveen, S. Kilpinen, A. Ruusulehto, R. Lothe, and R. Skotheim. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene*, 35(19):2413–2427, 2015. 3, 4

[177] W. Symmans, F. Peintinger, and C. Hatzis. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *Journal of Clinical*, 2007. 16, 64

[178] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. v. Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(Database):D561–D568, 2011. 34

[179] P. Szymański, M. Markowicz, and E. Mikiciuk-Olasik. Adaptation of high-throughput screening in drug discovery-toxicological screening tests. *International Journal of Molecular Sciences*, 13(1):427–452, 2012. 22

[180] J. H. Taube, J. I. Herschkowitz, K. Komurov, A. Y. Zhou, S. Gupta, J. Yang, K. Hartwell, T. T. Onder, P. B. Gupta, K. W. Evans, B. G. Hollier, P. T. Ram, E. S. Lander, J. M. Rosen, R. a. Weinberg, and S. a. Mani. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proceedings of the National Academy of Sciences of the United States of America*, 107(35):15449–15454, 2010. 40, 50

[181] B. Tauzin. More than 900 medicines and vaccines in clinical testing offer new hope in the fight against cancer. *DC: Medicines in Development for Cancer*, pages 1–2, 2009. 7, 22, 33

[182] B. Teicher. *Cancer drug resistance*. 2006. 70

[183] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. 14, 73, 74, 86, 158

[184] T. Thurnherr, F. Singer, D. J. Stekhoven, and N. Beerenwinkel. Genomic variant annotation workflow for clinical applications. *F1000Research*, 5:1963, oct 2016. 83, 84

[185] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–78, mar 2012. 25

[186] A.-M. Tsimberidou, N. G. Iskander, D. S. Hong, J. J. Wheler, G. S. Falchook, S. Fu, S. Piha-Paul, A. Naing, F. Janku, R. Luthra, Y. Ye, S. Wen, D. Berry, and R. Kurzrock. Personalized Medicine in a Phase I Clinical Trials Program: The MD Anderson Cancer Center Initiative. *Clinical Cancer Research*, 18(22):6373–6383, nov 2012. 34, 58

[187] N. Turner, A. Tutt, and A. Ashworth. Opinion: Hallmarks of 'BRCAness' in sporadic cancers. *Nature Reviews Cancer*, 4(10):814–819, 2004. 9, 14

[188] N. C. Turner and J. S. Reis-Filho. Tackling the Diversity of Triple-Negative Breast Cancer. *Clinical Cancer Research*, 19(23):6380–6388, dec 2013. 46

[189] E. M. Van Allen, V. W. Y. Lui, A. M. Egloff, E. M. Goetz, H. Li, J. T. Johnson, U. Duvvuri, J. E. Bauman, N. Stransky, Y. Zeng, B. R. Gilbert, K. P. Pendleton, L. Wang, S. Chiosea, C. Sougnez, N. Wagle, F. Zhang, Y. Du, D. Close, P. A. Johnston, A. McKenna, S. L. Carter, T. R. Golub, G. Getz, G. B. Mills, L. A. Garraway, and J. R. Grandis. Genomic Correlate of Exceptional Erlotinib Response in Head and Neck Squamous Cell Carcinoma. *JAMA oncology*, 1(2):238–44, may 2015. 47

[190] M. J. Van Den Bent, A. A. Brandes, R. Rampling, M. C. M. Kouwenhoven, J. M. Kros, A. F. Carpentier, P. M. Clement, M. Frenay, M. Campone, J. F. Baurain, J. P. Armand, M. J. B. Taphoorn, A. Tosoni, H. Kletzl, B. Klughammer, D. Lacombe, and T. Gorlia. Randomized phase II trial of erlotinib versus temozolomide or carmustine in recurrent glioblastoma: EORTC brain tumor group study 26034. *Journal of Clinical Oncology*, 27(8):1268–1274, mar 2009. 57

[191] R. Van Geel, E. Elez, J. C. Bendell, J. E. Faris, M. Lolkema, F. Eskens, P. Kavan, J.-P. Delord, M. H. Schuler, Z. A. Wainberg, Y. Yamada, T. Yoshino, E. Avsar, A. Chatterjee, P. Zhu, R. Bernards, J. Tabernero, and J. Schellens. Phase I study of the selective BRAF V600 inhibitor encorafenib (LGX818) combined with cetuximab and with or without the α -specific PI3K inhibitor BYL719 in patients with advanced BRAF -mutant colorectal cancer. *European Journal of Cancer*, 32:5s:(suppl; abstr 3514), 2014. 7

[192] S. Varambally, S. Dhanasekaran, and M. Zhou. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, 2002. 168

[193] J. P. Venables. Aberrant and Alternative Splicing in Cancer Aberrant and Alternative Splicing in Cancer. 64(21):7647–7654, 2004. 116

[194] A. R. Venkitaraman. Linking the Cellular Functions of BRCA Genes to Cancer Pathogenesis and Treatment. *Annual Review of Pathology: Mechanisms of Disease*, 4(1):461–487, 2009. 9

[195] V. D. Vijver. A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, 347(25):1999–2009, 2002. 11

[196] E. E. Vokes, E. E. W. Cohen, A. M. Mauer, T. G. Karrison, S. J. Wong, L. J. Skoog-Sluman, M. F. Kozloff, J. Dancey, and A. Dekker. A phase I study of erlotinib and bevacizumab for recurrent or metastatic squamous cell carcinoma of the head and neck (HNC). *Journal of Clinical Oncology*, 23(16_suppl):5504–5504, jun 2005. 47

[197] G. von Minckwitz, J. Blohmer, S. Costa, C. Denkert, H. Eidtmann, W. Eiermann, B. Gerber, C. Hanusch, J. Hilfrich, J. Huober, C. Jackisch, M. Kaufmann, S. Kümmel, S. Paepke, A. Schneeweiss, M. Untch, D. Zahm, K. Mehta, and S. Loibl. S3-2: Neoadjuvant Chemotherapy Adapted by Interim Response Improves Overall Survival of Primary Breast Cancer Patients – Results of the GeparTrio Trial. *Cancer Research*, 71(24 Supplement), 2014. 15

[198] G. von Minckwitz, J. U. Blohmer, S. D. Costa, C. Denkert, H. Eidtmann, W. Eiermann, B. Gerber, C. Hanusch, J. Hilfrich, J. Huober, C. Jackisch, M. Kaufmann, S. Kummel, S. Paepke, A. Schneeweiss, M. Untch, D. M. Zahm, K. Mehta, S. Loibl, S. Kümmel, S. Paepke, A. Schneeweiss, M. Untch, D. M. Zahm, K. Mehta, and S. Loibl. Response-Guided Neoadjuvant Chemotherapy for Breast Cancer. *Journal of Clinical Oncology*, 31(29):3623–3630, oct 2013. 15

[199] G. von Minckwitz, M. Untch, J.-U. J.-U. Blohmer, S. D. Costa, H. Eidtmann, P. A. Fasching, B. Gerber, W. Eiermann, J. Hilfrich, J. Huober, C. Jackisch, M. Kaufmann, G. E. Konecny, C. Denkert, V. Nekljudova, K. Mehta, and S. Loibl. Definition and Impact of Pathologic Complete Response on Prognosis After Neoadjuvant Chemotherapy in Various Intrinsic Breast Cancer Subtypes. *Journal of Clinical Oncology*, 30(15):1796–1804, may 2012. 15, 34, 42

[200] A. H. Wagner, A. C. Coffman, B. J. Ainscough, N. C. Spies, Z. L. Skidmore, K. M. Campbell, K. Krysiak, D. Pan, J. F. McMichael, J. M. Eldred, J. R. Walker, R. K. Wilson, E. R. Mardis, M. Griffith, and O. L. Griffith. DGIdb 2.0: Mining clinically relevant drug-gene interactions. *Nucleic Acids Research*, 44(D1):D1036–D1044, jan 2016. 83, 84

[201] Z. A. Wainberg, L.-S. Lin, B. DiCarlo, K. M. Dao, R. Patel, D. J. Park, H.-J. Wang, R. Elashoff, N. Ryba, and J. R. Hecht. Phase II trial of modified FOLFOX6 and erlotinib in patients with metastatic or advanced adenocarcinoma of the oesophagus and gastro-oesophageal junction. *British journal of cancer*, 105(6):760–5, sep 2011. 47

[202] D. Wang, S. A. Boerner, J. D. Winkler, and P. M. LoRusso. Clinical experience of MEK inhibitors in cancer therapy. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1773(8):1248–1255, aug 2007. 46

[203] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, sep 2010. 30

[204] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136–138, jan 2009. 24

[205] Y. Wang, J. Waters, M. L. Leung, A. Unruh, W. Roh, X. Shi, K. Chen, P. Scheet, S. Vattathil, H. Liang, A. Multani, H. Zhang, R. Zhao, F. Michor, F. Meric-Bernstam, and N. E. Navin. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, 2014. 15

[206] B. Weigelt, A. Mackay, R. A'Hern, R. Natrajan, D. S. Tan, M. Dowsett, A. Ashworth, and J. S. Reis-Filho. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol*, 11(4):339–349, 2010. 120

[207] J. Weinstein and P. Lorenzi. Cancer: Discrepancies in drug sensitivity. *Nature*, 504(7480):381–3, dec 2013. 187

[208] J. N. Weinstein. Cell Lines battle cancer. *Nature*, pages 9–10, 2012. 187

[209] J. N. Weinstein, T. G. Myers, P. M. O'Connor, S. H. Friend, A. J. Fornace, K. W. Kohn, T. Fojo, S. E. Bates, L. V. Rubinstein, N. L. Anderson, J. K. Buolamwini, W. W. van Osdol, A. P. Monks, D. A. Scudiero, E. A. Sausville, D. W. Zaharevitz, B. Bunow, V. N. Viswanadhan, G. S. Johnson, R. E. Wittes, and K. D. Paull. An information-intensive approach to the molecular pharmacology of cancer. *Science (New York, N.Y.)*, 275(5298):343–9, jan 1997. 33

[210] P. S. Weisman, C. K. Y. Ng, E. Brogi, R. E. Eisenberg, H. H. Won, S. Piscuoglio, M. R. De Filippo, R. Ioris, M. Akram, L. Norton, B. Weigelt, M. F. Berger, J. S. Reis-Filho, and H. Y. Wen. Genetic alterations of triple negative breast cancer by targeted next-generation sequencing and correlation with tumor morphology. *Modern Pathology*, 29(5):476–488, 2016. 73

[211] J. Wen, K. H. Toomer, Z. Chen, and X. Cai. Genome-wide analysis of alternative transcripts in human breast cancer. *Breast Cancer Research and Treatment*, 151(2):295–307, 2015. 72

[212] A. C. Wolff, M. E. H. Hammond, J. N. Schwartz, K. L. Hagerty, D. C. Allred, R. J. Cote, M. Dowsett, P. L. Fitzgibbons, W. M. Hanna, A. Langer, L. M. McShane, S. Paik, M. D. Pegram, E. A. Perez, M. F. Press, A. Rhodes, C. Sturgeon, S. E. Taube, R. Tubbs, G. H. Vance, M. Van De Vijver, T. M. Wheeler, and D. F. Hayes. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer, jan 2007. 88

[213] J. J. Yeh, E. D. Routh, T. Rubinas, J. Peacock, T. D. Martin, X. J. Shen, R. S. Sandler, H. J. Kim, T. O. Keku, and C. J. Der. KRAS/BRAF mutation status and ERK1/2 activation as biomarkers for MEK1/2 inhibitor therapy in colorectal cancer. *Molecular Cancer Therapeutics*, 8(4):834–843, apr 2009. 46

[214] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, 16(5):284–7, may 2012. 49

[215] Y. Zhong, S. Shen, Y. Zhou, F. Mao, Y. Lin, J. Guan, Y. Xu, S. Zhang, X. Liu, and Q. Sun. NOTCH1 is a poor prognostic factor for breast cancer and is associated with breast cancer stem cells. *OncoTargets and Therapy*, 9:6865–6871, nov 2016. 87

[216] F. Y. Zong, X. Fu, W. J. Wei, Y. G. Luo, M. Heiner, L. J. Cao, Z. Fang, R. Fang, D. Lu, H. Ji, and J. Hui. The RNA-Binding Protein QKI Suppresses Cancer-Associated Aberrant Splicing. *PLoS Genetics*, 10(4), 2014. 102

Titre : Analyse de données pharmacogénomiques et moléculaires pour comprendre la résistance aux traitements des cancers du sein triple négatif.

Mots clés : Cancer du sein triple négatif, lignées cellulaires, chimiothérapie néoadjuvante, RNAseq, génomique.

Résumé: Devant le grand nombre de tumeurs du sein triple négatif résistant aux traitements, il est essentiel de comprendre les mécanismes de résistance et de trouver de nouvelles molécules efficaces.

En premier lieu, nous analysons deux ensembles de données pharmacogénomiques à grande échelle. Nous proposons une nouvelle classification basée sur des profils transcriptomiques de lignées cellulaires, selon un processus de sélection de gènes basé sur des réseaux biologiques. Notre classification moléculaire montre une plus grande homogénéité dans la réponse aux médicaments que lorsque l'on regroupe les lignées cellulaires en fonction de leur tissu d'origine. Elle permet également d'identifier des profils similaires de réponse aux traitements.

Dans un second travail, nous étudions une cohorte de patients atteints d'un cancer du sein triple négatif ayant résisté à la chimiothérapie néoadjuvante. Nous effectuons des analyses moléculaires complètes basées sur du RNAseq et WES. Nous constatons une forte hétérogénéité moléculaire des tumeurs avant et après traitement. Bien que nous observons une évolution clonale sous traitement, aucun mécanisme récurrent de résistance n'a pu être identifié. Nos résultats suggèrent fortement que chaque tumeur a un profil moléculaire unique et qu'il est important d'étudier de grandes séries de tumeurs.

Enfin, nous améliorons une méthode pour tester la surreprésentation de motifs connus de protéines de liaison à l'ARN, dans un ensemble donné de séquences régulées. Cet outil utilise une approche innovante pour contrôler la proportion de faux positifs qui n'est pas réalisé par l'algorithme existant. Nous montrons l'efficacité de notre approche en utilisant deux séries de données différentes.

Title : Pharmacogenomic and high-throughput data analysis to overcome triple negative breast cancers drug resistance.

Keywords : Triple negative breast cancer, cancer cell lines, neoadjuvant chemotherapy, RNAseq, WES.

Abstract : Given the large number of treatment-resistant triple-negative breast cancers, it is essential to understand the mechanisms of resistance and to find new effective molecules.

First, we analyze two large-scale pharmacogenomic datasets. We propose a novel classification based on transcriptomic profiles of cell lines, according to a biological network-driven gene selection process. Our molecular classification shows greater homogeneity in drug response than when cell lines are grouped according to their original tissue. It also helps identify similar patterns of treatment response.

In a second analysis, we study a cohort of patients with triple-negative breast cancer who have resisted to neoadjuvant chemotherapy. We perform complete molecular analyzes based on RNAseq and WES. We observe a high molecular heterogeneity of tumors before and after treatment. Although we highlighted clonal evolution under treatment, no recurrent mechanism of resistance could be identified Our results strongly suggest that each tumor has a unique molecular profile and that that it is increasingly important to have large series of tumors.

Finally, we are improving a method for testing the overrepresentation of known RNA binding protein motifs in a given set of regulated sequences. This tool uses an innovative approach to control the proportion of false positives that is not realized by the existing algorithm. We show the effectiveness of our approach using two different datasets.