UNIVERSITÀ DEGLI STUDI DI TORINO

Doctoral School in Life and Health Sciences

PhD Program in Complex System for Life Sciences

XXXIII cycle

Systematic evaluation of genomic landscape changes during engraftment and propagation of patient-derived cancer xenografts

Jessica Giordano

Advisor: Prof. Enzo Medico

To eyes, sparkling with life

Zeus, who leads onward mortals to be wise,
Appoints that suffering masterfully teach.
Aeschylus, Agamemnon (176-177)

Contents

Abstract

Patient-Derived Xenografts (PDXs) are preclinical models extensively used to characterize tumor biology and response to treatments. The relevance of PDXs as models depends on their ability to recapitulate the human tumor of origin. Several independent studies reported that PDXs retain the morphological, pharmacological and genomic features of their originating tumors during engraftment and propagation. On the contrary, a recent study, focused on the characterization of Copy Number Alteration (CNA) alterations, reported systematic divergence of PDX profiles compared to patient tumor samples (PT), supposedly originating from selective pressures imposed by the mouse host. However, the limited number of matched PT/PDX samples analyzed combined with the small cohort size per tumor type highlighted the need for a larger scale and more systematic analysis.

To systematically explore CNA dynamics during PDX engraftment and propagations, in a joint international effort of the EurOPDX and PDXNet consortia, we exhaustively analyzed CNA profiles of 1451 PDX and PT samples from 509 PDX models. Overall, we observed strong concordance between matched PT-PDX and PDX-PDX pairs, and no apparent downward trend over tumor engraftment and passaging. Nonetheless, some PDX models displayed CNA profile variations. However, it was unclear whether such changes arose from selective pressure imposed by the mouse host or spontaneous tumor evolution and intratumor heterogeneity. Hence, here, we focused on two large colorectal and breast cancer series, composed of 87 and 43 matched triplets of PT, PDX at early passage (PDX-early) and PDX at later passage (PDX-late), respectively. For both tumor types, we assembled genomic data from matched PT, PDX-early, and PDX-late cohorts. And, we estimated CNA recurrence by GISTIC separately for each cohort. We assumed that if mouse-specific selective pressure was occurring, recurrent changes in the

CNA profile would emerge in the PDX early cohort compared to the PT cohort and further increase in the PDX late cohort. However, GISTIC CNA profiles of the PT, PDX-early, and PDX-late cohorts were virtually indistinguishable, with minor changes, not functionally related. The GISTIC profiles of our cohorts recapitulated at large those generated by the TCGA for colorectal and breast cancer. Therefore, we were confident that our results were not affected by the lack of representativeness of CNA lesions per tumor type.

In summary, our analyses excluded a systematic mouse-driven genetic selection during PDX engraftment and propagation, supporting the assumption of a high degree of molecular fidelity of PDX models compared to patient tumor samples. Consequently, PDX models can be reliably implemented for anticancer drug testing.

Chapter 1

Introduction

One of the main advantages of PDXs is the possibility of studying the behavior of human cancer cells in a natural microenvironment, where they interact with the stromal components contributed by the murine host, typically absent in other experimental models, such as cancer cell lines, or tumor derived organoids.[1,2] However, as any patient-derived cancer model is interrogated to make decisions on how best to treat cancer patients and to explore the biology of human cancers, it is crucial to assess that PDXs faithfully recapitulate the genomic features of the originating human tumors.

1.1 Cancer as an evolutionary process

Cancer refers to a collection of more than 100 diseases that can develop almost everywhere in the human body.[3] In 2018, cancer accounted for an estimated 9.6 million deaths, proving as a leading cause of death worldwide.[4] Although each cancer type and, even, each patient's cancer has its distinct features, all cancers arise when some of the body's cells start to abnormally growing, generating a mass of cells named tumor, potentially invading surrounding tissues and spreading into distant body sites.[5]

1.1.1 Genomic instability triggers the conversion of normal cells into tumor cells

Tumor initiation and proliferation is a multi-step process, in which normal human cells gradually become malignant through the successive acquisition of genetic alterations in key regulatory genes.[6]

During malignant transformation, cancer cells start growing out of control and becoming invasive, uncaring about the maintenance of tissue functions and stability. More specifically, as tissue growth is regulated both by the rate of cell division and cell death or apoptosis, the uncontrolled growth, peculiar to cancer cells, is in many cancer types a result of higher cell proliferation and a reduced cell death rate than in normal cells.

As mentioned above, abnormal growth in malignant cells is a consequence of specifically mutated genes. In detail, these genes can be divided into three groups according to their role in tumor formation: oncogenes, tumor suppressor genes, and DNA repair genes.

DNA repair genes are involved in the normal repair of DNA damage. Hence, they play an essential role in the maintenance of genome integrity. Loss of their function causes genomic instability which increases the frequency of mutations in oncogenes and tumor suppressor genes[7].
Oncogenes activate mutations such as amplification, small mutations, or translocations and, since mutations in these genes are usually dominant, only one allele of the gene needs to be affected to cause that cells divide out of control.
On the other hand, tumor suppressor genes protect the normal cells from turning into cancer cells. Therefore, a loss of their function determines malignant transformation. Importantly, to lose their tumor-suppressing activity, both alleles usually have to be inactivated.

1.1.2 Selection of the fittest : the clonal evolution model

We described the key features of cancer cells versus normal cells and the mechanisms that can lead to tumor formation. Now we clarify why tumor initiation and progression is defined as a multi-step evolutionary process.

In this respect, P. Nowell, in 1976, proposed that most tumors arise from

a single normal cell affected by a genetic change, which provides it with a selective growth advantage over surrounding normal cells. Then, during tumor proliferation, as a result of genetic instability, tumor cells acquire new genetic alterations in the expanding population. Nonetheless, the majority of these genetic variants have no phenotypic advantages; thus they are eliminated. However, sporadically, a variant becomes the founder of a new predominant subpopulation of cells, since it has an additional selective advantage to the original cancer cells as well as to normal cells. As a result, over time, cancer cells undergo an evolutionary process in which increasingly aggressive subpopulations of cells are sequentially selected.[8,9]
Notably, this hypothesis has been demonstrated by the discovery of intratumor subclonal heterogeneity and clonal selection in multiple cancer types.[10–15]

1.1.3 The tumor evolution paradigm

Studies in microbial experimental evolution have brought insights that the cancer evolution is quite similar to the evolution of asexual microorganisms.[16] Therefore, as well as occurs for every evolutionary system, even cancer evolution should be regulated by the dynamic interplay of the same three fundamental processes (Fig. 1.1):

- the **mutation process**, i.e. the random generation of inheritable new variations in the population;
- the **genetic drift**, which changes the frequency of genotypes in the population due to random birth and death events;
- the **Darwinian selection**, which changes the frequency of genotypes in the population, based on their relative fitness advantage.

Note that the acquisition of inheritable alterations and genetic drift are both stochastic processes continuously occurring. Instead, the Darwinian selection is a deterministic force that depends on the environmental context.

Thus, mutations increase heterogeneity, while drift and selection generally reduce heterogeneity. Indeed genetic drift and selection processes modify the frequency of alleles in a population, rendering some larger or even dominant, and others to go extinct.

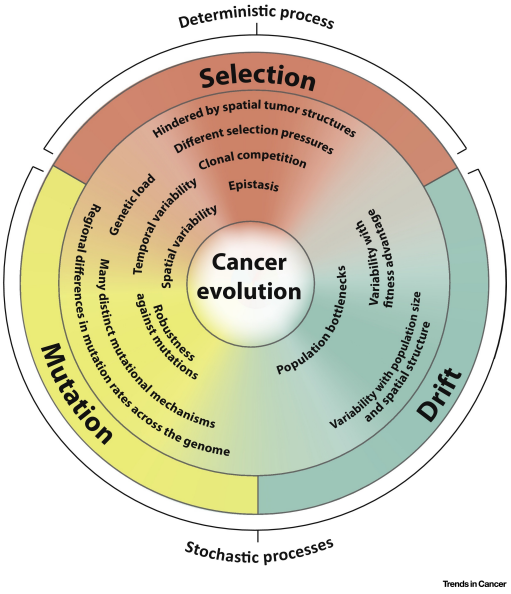

Figure 1.1: **Mutation, Selection, and Drift are the three basic processes shaping cancer evolution.** Interdependencies and spatial and temporal variability of mutation, selection, and drift produce additional levels of complexity (middle section), which together influence cancer evolution (center) [17].

1.1.3.1 Mutation (stochastic)

The mutation process is essential for evolution, as the variations introduced by mutational forces are the substrate on which selection can act. Typically, mutation may shape the cancer genome in highly different ways. In detail, inheritable somatic variations encompass multiple genetic alterations such as point mutations, insertions or deletions of base pairs, and larger structural variations[17,18].

Any cell has a baseline mutation rate, however, elevated mutation rates are hallmarks of cancer cells. Respectively, different genomic regions may have variable mutation rates depending on the DNA replication timing and chromatin accessibility.[19–21] Specifically, mutation frequencies are higher in late DNA replication timing regions and inaccessible heterochromatin-like domains.

Point mutations are single-base-pair changes that can modify the protein-coding region, making it not functional, as happens in the case of tumor-suppressive mutations, or altering its function, as occurs for oncogenic mutations. Instead, INsertions and DELetions of base pairs, generally called INDELs, are slightly larger changes, which can lead to similar consequences. Moreover, even larger structural variations, which include whole-genome doubling, chromosomal loss or gains, and translocations, often affect the cancer genome.[22]

1.1.3.2 Genetic drift (stochastic)

Genetic drift is the change in the frequency of an allele in a population caused by random birth and death events. In detail, each cell in a cancer subclone has a specific probability of dying due to random factors, and sometimes all cells of a subclone die, although this subclone holds highly beneficial mutations. Notably, the impact of drift is bigger in smaller populations and is more relevant after population bottlenecks. Therefore, as a result of drift, even the expansion of a clone with high fitness is not predictable with certainty, unless the abundance of this clone overcomes a certain amount such that it eludes possible extinction via drift. Moreover, experimental data demonstrated that drift affects cancer initiation more than cancer progression.[23,24]

1.1.3.3 Selection (deterministic)

When a cell acquires a new mutation able to enhance its ability to survive and to reproduce in certain environmental conditions, eluding potential extinction through drift, this cell gradually increases in the amount within the population.

In many cancer types, were identified different intratumoral subclones, carrying distinct driver mutations, showing different phenotypes, and growing with branched phylogenies.[10,12,25–27] As a consequence, the presence of different subclones within the same tumor can potentially result in competition among these multiple subclonal populations. Hence, the fitness of each subclone within the tumor depends on the fitness of the other competing subclones.[28] Therefore, beneficial mutations escaping the extinction via drift can still be eliminated by competing subclones. Thus, in this scenario, predicting evolutionary outcomes becomes more challenging.

However, subclonal competition is arguably limited to nearby subclones because of spatial constraints typical of solid tumors. Consequently, the 3D spatial structures of the solid tumors may enhance the formation and preservation of subclonal heterogeneity and drive the system towards a more stochastic behavior. As a result, solid tumors may be considered ecological systems composed of multiple small and localized subpopulations, each competing only with neighboring subpopulations.[17]

1.1.3.4 Intra-tumor heterogeneity fuels tumor evolution

As referred above, the interaction between the mutational, the genetic drift, and the selection processes lead to tumor masses composed of highly diverse populations of cells. Remarkably, this spatial and temporal heterogeneity within a tumor, briefly named intra-tumor heterogeneity (ITH), has been experimentally observed in many studies, across multiple tumor types, as extensively reported.[29]

From an evolutionary perspective, the diversity present in a population promotes its *evolvability*. In this respect, consider two scenarios: in the first set, the tumor includes all identical cells (*homogeneous tumor*), while, in the other set, it comprises highly phenotypically diverse cells (*heterogeneous tumor*). Suppose, then, that the tumor undergoes new selective pressures. In the case of a homogeneous tumor, all tumor cells adapt to the new condition,

as nothing happens, or they become extinct. On the contrary, in the case of a heterogeneous tumor, the cell population likely encompasses cells sensitive to the new pressure, that will die, and also subclone resistant to the pressure, that will survive and keep to grow to the point of becoming dominant in the tumor. Consequently, more heterogeneous tumors are more likely to evolve and generate metastases and/or a clone resistant to therapy[29] (Fig. 1.2).

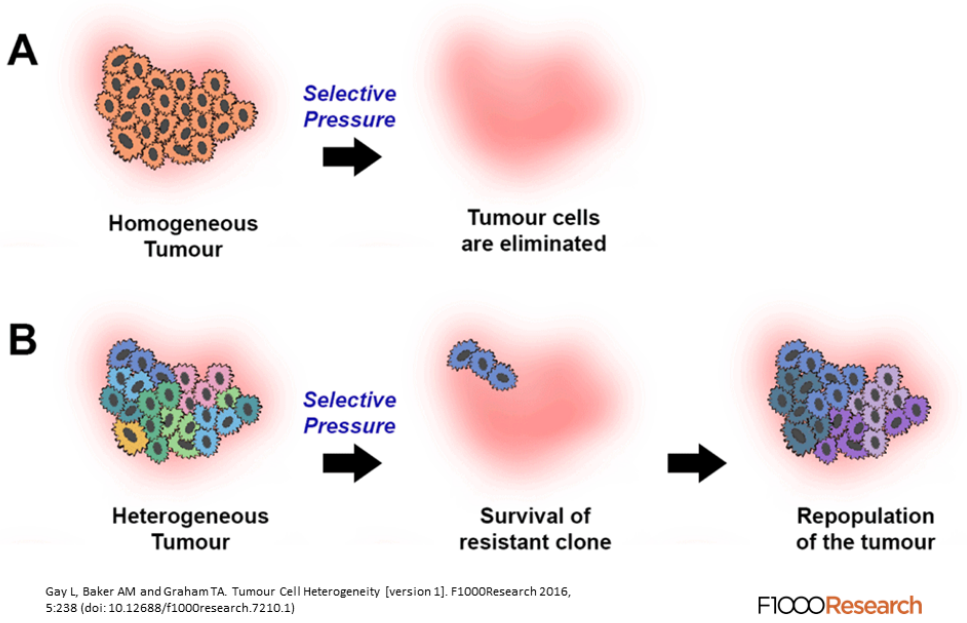

Figure 1.2: **Intratumor heterogeneity promotes tumor evolution**. A homogeneous tumor (**A**) is eradicated under a selective pressure, whereas a heterogeneous tumor (**B**) survives, although a selective pressure is imposed, as it more likely contains a resistant clone, that repopulates the tumor [29].

1.1.4 Implications of intra-tumor heterogeneity

The intra-tumor heterogeneity observed in cancer poses challenges to clinical diagnoses and therapeutic decisions Specifically, as cancer is a disease that can affect multiple organs or tissues in the body through metastatic lesions, the intra-tumor heterogeneity phenomenon should be considered compared to both primary tumors and metastases. Therefore, to provide a comprehensive assessment of the molecular features of the disease of a cancer patient,

tumor specimens from multiple spatially distinct regions of the same tumor, collected at different time points and from both primary tumors and metastases, if any, are needed. Accordingly, given these overall considerations, modern oncology aims to accurately determine and efficiently control the heterogeneity within tumors, to translate these insights into personalized-therapeutic strategies.

1.1.4.1 Genetic discrepancies between primary tumors and metastasis and between distinct metastatic lesions

Metastasis arises when one or more clones from a primary tumor form a new tumor in a distant site, in other organs or tissues of the body. Therefore, a *population bottleneck* occurs during the seeding of metastasis. As a consequence, the genetic heterogeneity of metastasis can decrease to the relative primary tumors.[30] In this respect, many studies have reported reduced heterogeneity or monoclonality in metastases, in multiple tumor types.[31–33] Nevertheless, more than one clone may also seed metastasis.[33–35] Thus, in a minority of cases, increased heterogeneity in metastases compared to matched primary tumors has been also shown.[36] As a result, genetic discrepancies between primary tumors and metastasis are direct consequences of extensively reported intratumor heterogeneity.[37,38]

It has been also observed that distinct metastasis types display different heterogeneity levels. Specifically, if inter-metastatic diversity reflects the diversity of primary tumor, this would indicate that many if not all subclones present in the primary tumor have similar metastatic potential. Conversely, evidence of homogeneous metastases would suggest that they are formed by a single clone provided with superior metastatic ability.
Concerning human colorectal cancer, considerably different inter- and intra-lesion heterogeneity has been observed for lymph node and distant organ metastases.[39] In detail, it has been reported that distant metastases tend to be genetically similar to each other. Conversely, lymph node metastases exhibit high levels of inter-lesion diversity. Moreover, metastases in lymph nodes display higher levels of intra-lesion heterogeneity compared to those in distant organs. Therefore, altogether, these pieces of evidence demonstrate that a reduced number of primary tumor clones seed distant metastases in respect to lymph node metastases. As many cells from the primary tumor seem to be able of migrating to and growing in lymph nodes, weaker selection

acts in correspondence of lymph nodes. In contrast, as distant metastases are typically composed of homogeneous cellular groups, a stricter evolutionary bottleneck is present in this case. Hence, it is clear that the evolutionary forces shaping the seeding of lymph nodes and distant metastases are fundamentally different.

1.1.4.2 The importance of ITH in clinical diagnosis and therapeutic responses

Understanding ITH is particularly relevant since genetic diversity within tumors complicates definitive clinical diagnosis and causes targeted therapy failure and resistance. Indeed, in presence of spatial intratumor heterogeneity, a targeted biopsy is not representative of the whole tumor. Hence, also clinical decision-making based on a biopsy including the dominant clone in a given sample might not be sufficiently accurate to eradicate the tumor mass. As even minor subpopulations of tumor cells could generate resistance to treatment, to decipher and assess the extent of heterogeneity within tumors, many studies performed genetic analysis of multi-region samplings revealing the genomic architecture, subclonal diversification, and evolution of multiple tumor types.[12,26,27,40]
Moreover, as, in principle, primary tumors and relative metastases may display genetic divergence even because of ITH, accurate treatment decision-making should assess biological features of both primary tumors and any metastases to evaluate the clinical relevance of potential discrepancies.[37] In this respect, different studies recommend the acquisition of biopsy of both primary tumors and metastases, demonstrating that substantial discordance in receptor status between primary tumor and metastases of breast and colorectal cancer could occur[41–46].

1.2 Uniqueness of PDXs over cancer models

Most of our understanding of the molecular basis of cancer results from the study of model systems derived from human tumor specimens. Therefore, it is fundamental that such model systems, termed *patient-derived cancer models*, on which cancer research depends, are representative of the originating tumors.

For many decades, cell lines derived from patients' samples and then modi-

fied to grow in artificial culture conditions have been essential tools in basic and preclinical cancer research, both cultured *in vitro* or grown as xenografts. However, although these model systems have proven to be particularly useful for deciphering cancer cell biology, in many cases, they have also shown limitations in accurately recapitulating the original tumor. Specifically, in several cases, cell lines have turned out to be inappropriate for biomarker discovery, drug screening, and therapeutic preclinical testing.[50] Therefore, many efforts have been made to find preclinical models able to more accurately predict the clinical outcome. This deal of energy has resulted in the generation of patient-derived cancer models, established either by engrafting fresh tumor tissues in experimental animals, e.g., patient-derived xenograft models (PDXs), or deriving 3D structures from human cancer tissues, i.e., organoids, or growing tumor cells *in vitro* 2D tissue culture conditions for a limited period.[47]

Among these categories of preclinical models, those providing the possibility of studying the growth of cancer cells in a more natural microenvironment, have been proven particularly relevant.[47] They includes the so called *patient-derived xenografts* (PDXs). Specifically, PDXs are obtained by direct implantation of fresh, surgically derived, clinical tumor samples in immunodeficient mice. Upon engraftment and adaptation to the murine host, PDX tumors are then grown and propagated across multiple generations of mice, a process called *passage*, to generate cohorts of PDX samples derived from the same patient tumor (Fig. 1.3). Unlike cancer cell lines, PDX models are established by the engraftment of intact tissue. Hence, the tumor architecture and the relative proportion of cancer and stromal cells are both maintained, enhancing the capability of PDXs in representing the human tumor of origin. Moreover, tumor samples can be transplanted subcutaneously or orthotopically to better recapitulate the microenvironmental interactions occurring within patients, depending on the original tumor type.[47] As a result, PDX models have been successfully derived from multiple solid or hematologic primary and metastatic tumors, emerging as a platform which provides a unique opportunity for investigating tumor biology and therapeutic response and resistance.[48,49]

It is largely accepted that the growth and spreading of solid tumors can be affected by the vascular, mesenchymal, and immune cells surrounding and feeding it and collectively constituting the so-called *tumor microenvironment* or *tumor stroma*. Therefore, as cancer cells of PDXs preserve their ability

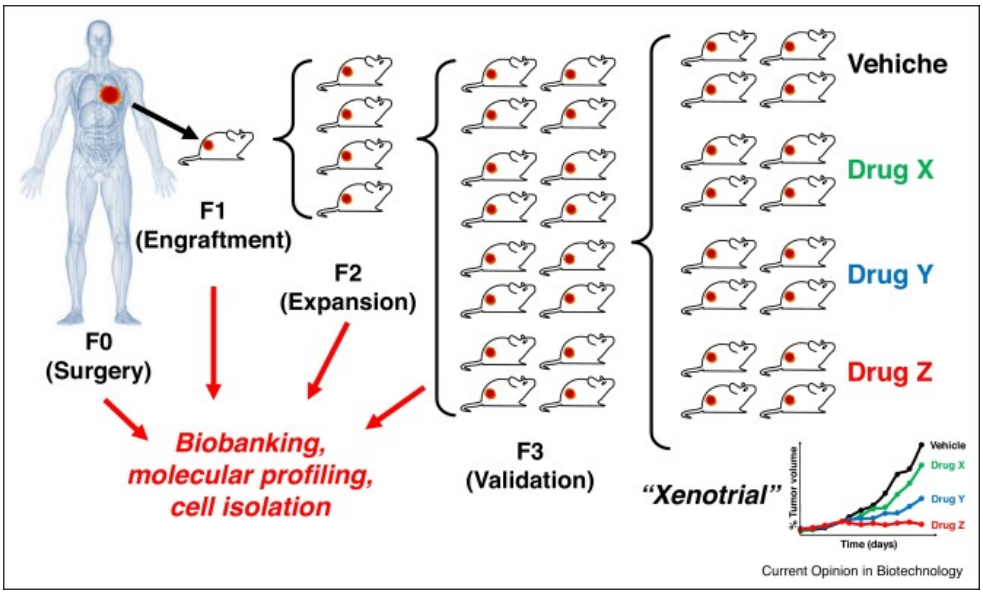

Figure 1.3: **Engraftment and propagation of PDXs**. The patient-derived tumour is engrafted into an immunodeficient mouse and propagated in multiple generation of mice. At each passage, PDX-derived tumour samples can be collected to develop a tissue biobank for molecular profiles and ex vivo experiments [2].

to interact with stromal cells, PDXs are an excellent experimental model to characterize tumor-stroma interactions, a possibility that is completely lacking in *in vitro* models.

Moreover, PDX models represent a virtually unlimited source of tissue availability for generating a *living tissue biobank*, as tumor tissue growth in the mouse host can be extracted and then cryopreserved. Hence, they are used to perform experiments and molecular profiles to identify associations between genotypes and drug response.

1.2.1 Powerful models of drug response and resistance

Concerning *in vitro* derivatives, specific protocols for cell isolation and derivation of 2D and 3D cultures from various tumor types have been developed. Remarkably, the genotype-driven responses of these *in vitro* models have been recapitulated *in vivo* in matched PDXs.[56,57] Therefore, 2D models are still largely used for high-throughput screenings thanks to their simplicity and low cost, although they usually display a low capacity of proliferating in culture. On the other hand, three-dimensional cultures better resemble the physical features and the architecture of the original tumors, but they lack the stromal component unlike PDXs.[58]
For these reasons, nowadays, PDXs models are widely interrogated to elucidate drug sensibility and resistance that are observed in human tumors. In this respect, large-scale pharmacogenomic *in vivo* screens performed on more than 1000 PDX models from different tissues shown the value of these preclinical models in terms of reproducibility and clinical translatability to identify associations between genotypes and treatment response and resistance.[51–53] Alternatively, PDX samples can be employed as *avatars* of the patient to test multiple treatments.[54,55] Thus, only the most effective drug is administered to the patient, revealing that PDXs can play a relevant role also in personalized medicine.

Nonetheless, since all patient-derived experimental models have their strengths and weaknesses, the best strategy to address many scientific questions may be to use them in a complementary rather than in an alternative way.

1.2.2 Challenges and opportunities

Certainly, PDXs have played a key role in bringing advancement in cancer research. However, also PDXs, like any model system, present some limitations. Indeed, currently, PDX models are established by transplanting human tumor tissue into immunodeficient mice to prevent that the xenotransplants would be rejected by the immune system of the host.[59] Nevertheless, it is now clear that immune cells are virtually present in every neoplastic lesion, from subtle infiltrations to gross inflammations, and that the immune system has a relevant role in tumor evolution.[60]

In this respect, multiple pieces of evidence indicate that an inflammatory microenvironment is an essential component of all tumors and that immune-inflammatory cells can actively promote tumor progression, stimulating angiogenesis, cancer cell proliferation, and invasiveness.[61] Coupled with its protumorigenic effects, inflammatory conditions also influence the host immune response to tumors. Thus, inflammation can be also used in cancer immunotherapy[62] and to increase the response to chemotherapy.[63]
As a result of inflammation, the tumor microenvironment consist of innate immune cells, adaptive immune cells in addition to the cancer cells and their surrounding stroma components. These different cells communicate with each other, controlling and shaping tumor growth. Specifically, the expression of multiple immune mediators and modulators as well as the abundance and activation state of different cell types in the tumor microenvironment determine whether tumor-promoting inflammation or antitumor immunity will dominate.[64,65]
As a consequence, the absence of the immune system components in PDXs hinders the possibility of employing these models for studying the roles of the immune system in tumor development and in immune-based therapy response.[66–68] Moreover, the lack of the constraints imposed by the human immune system may explain the observations that, at the molecular level, serially transplanted PDX tumors are more aggressive than parental tumors and are more similar to metastatic or recurrent tumors.[59,69,70]

Interestingly, some issues affecting PDX models may be resolved through the use of other model systems, called genetically engineered mouse models (GEMMs) of cancer.
GEMMs are mainly generated by manipulating a single gene or a handful

of genes of interest to lead a tumorigenic response. On a hand, in contrast to models based on cancer cell inoculation or tumor tissue implantation, GEMMs develop de novo tumors in a natural immune-proficient microenvironment.[71] On the other hand, similarly to PDXs models, it has been demonstrated that GEMMs closely mimic the histopathological, molecular, and clinical features of the originating human tumors, supporting their adequacy as models of human cancers.[72,73] Since GEMMs reliably capture both tumor cell-intrinsic and cell-extrinsic factors driving de novo tumor initiation and progression toward metastatic disease, these models have proven to be essential for preclinical research. Specifically, in the last decades, GEMMs have successfully been employed to validate candidate cancer genes and drug targets, assess therapy efficacy, study the contribution of the tumor microenvironment, and unravel the mechanisms of drug resistance.[71]
Consequently, these cancer models are valuable tools to elucidate the role of individual genes and their mutated counterparts in tumorigenesis, as well as the cooperation of individual mutations in tumor development, and to model known cancer predisposition syndromes.[74]
Conversely, PDX models allow to study tumor progression on large size cohorts of tumors and to develop novel combinatorial treatment strategies, increasing anti-cancer drugs efficacy. Furthermore, xenografts have a high degree of predictability and rapidity of tumor formation, which makes them easy to use[74].

An other strength of PDX models is that they can be generated also with a limited quantity of biological material. Nonetheless, when the studied tumor type is particularly heterogeneous, this procedure of PDX derivation may be confounding. Indeed, in such a case, a single biopsy, and correspondingly the derived PDX, could not be representative of the heterogeneity of the patient's tumor.[75,76] Therefore, owing to spatial genetic variability, patient tumor and derived PDXs may display distinct responses to the same treatment. Nevertheless, the sampling population bottleneck that occurred during PDX establishment rather than the intrinsic weakness of the model in recapitulating the tumor of origin would be responsible for this divergence in drug response between patients and PDXs.
Thus, to reduce this phenomenon, it is recommended to carry on standardized preclinical designs. Moreover, the disgregation and mixing of heterogeneous

tumor masses before implantation could increase the heterogeneity of clones' representation in xenografts. However, the cost is the loss of the original tumor architecture.

1.2.3 Genetic fidelity of PDXs

Despite few studies have reported possible population bottleneck during PDX engraftment of breast cancer,[77,78] a large growing body of literature has documented that PDX models mostly preserve the clonal architecture of the original human tumor and recapitulate the transcriptomic, epigenomic, and histological landscapes of the patient tumor (Fig. 1.4).[79–81] Importantly, in PDXs of breast cancer reporting genetic variations compared to the original tumor, it has been noted that the genetic changes identified after engraftment do not affect known breast cancer oncogenic drivers.[78] Therefore, this result suggests that evolution, occurring in PDXs upon engraftment, is essentially neutral and that the representation of relevant genes is preserved.

Furthermore, in support of the robustness of these preclinical models, it has been reported that the mechanisms of resistance detected in PDXs mirror those found in their original patient tumors and that tumors clinically displaying resistance resulted also refractory to treatment in PDXs.[75,82] Hence, these studies demonstrated that PDX models can predict clinical outcomes with accuracy.

1.2.4 Mouse-driven selective pressures or genetic drift in PDXs?

Although conservation of the genomic landscape during PDX engraftment and passaging has been extensively reported in the literature, recent studies highlighted the possibility that human tumors grown in a murine microenvironment undergo a mouse-driven selection, which may affect their reliability as models of human cancer.[83–85]

In this respect, these studies have characterized Copy Number Alteration (CNA) dynamics in PDXs and have reported that a median of ~10% of the genome is differentially altered between human tumors and PDXs.[83,85] Moreover, on one hand, according to previous evidence in PDXs from breast cancer, they have suggested that most of the genomic divergence observed in

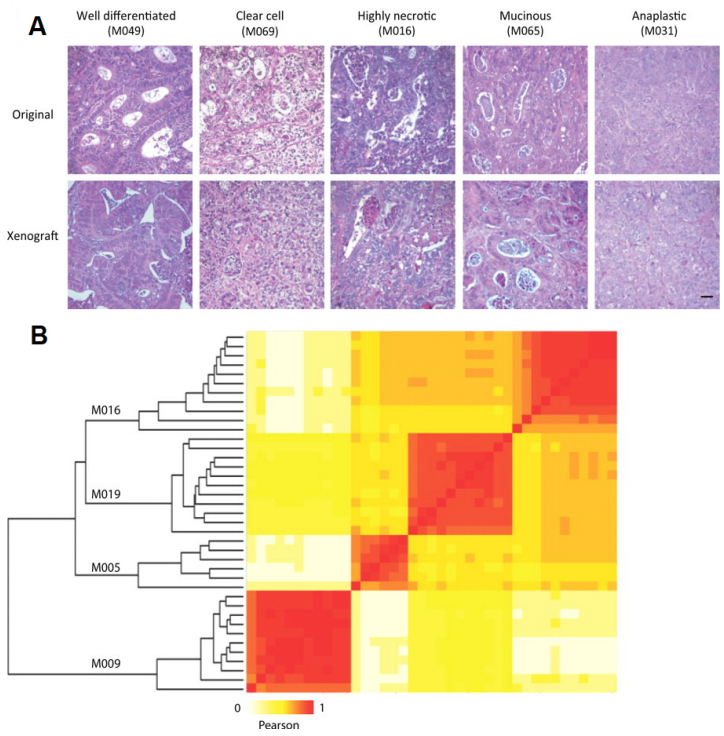

Figure 1.4: **Subclones selection in PDX models**. **A**, xenografted tumors retained the histopathologic characteristics of original samples. **C**, genetic concordance between xenografts and their original counterparts[87].

individual PDX models compared to the relative patient tumors is the result of population bottleneck of subclones present at a reduced frequency in the original tumors, occurring during PDX engraftment (Fig. 1.5). Thus, once again, intratumor heterogeneity played a crucial role in determining genetic differences between patient tumors and derived PDXs.

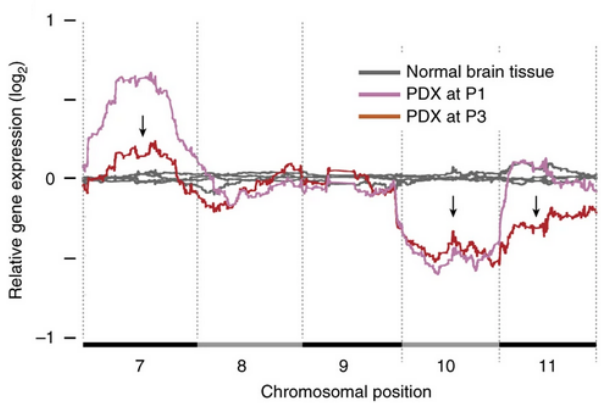

Figure 1.5: **Subclones selection in PDX models**. Gene expression moving-average plots for normal brain tissue (gray), a GBM PDX model at P1 (pink) and a GBM PDX model at P3 (red). Trisomy 7 disappears, monosomy 10 is retained, monosomy 11 emerges within two *in vivo* passages [83].

On the other hand, they have found that, in five cancer types, 12 arm-level genetic events, recurrently observed in TCGA samples, are lost when the tumors are transplanted into mice (Fig. 1.6). As a consequence, they have claimed that the selective pressures occurring in mouse hosts are different compared to those of humans, questioning the robustness of PDXs as models of human cancers. Therefore, if PDXs undergo a mouse-specific tumor evolution, their capacity of faithfully recapitulating patient treatment response is strongly impacted. Noteworthy, several limitations affect the experimental design of this work. Specifically, unmatched patient tumor and PDX sample cohorts and small size PDX cohorts per tumor type were available in this study. Moreover, most of their copy number profiles were inferred on RNA rather than DNA data and it is known that gene expression data allow low-resolution CNA estimates only.[86] Therefore, further investigations inferred on a larger scale, more systematic and DNA-based analysis are needed, to

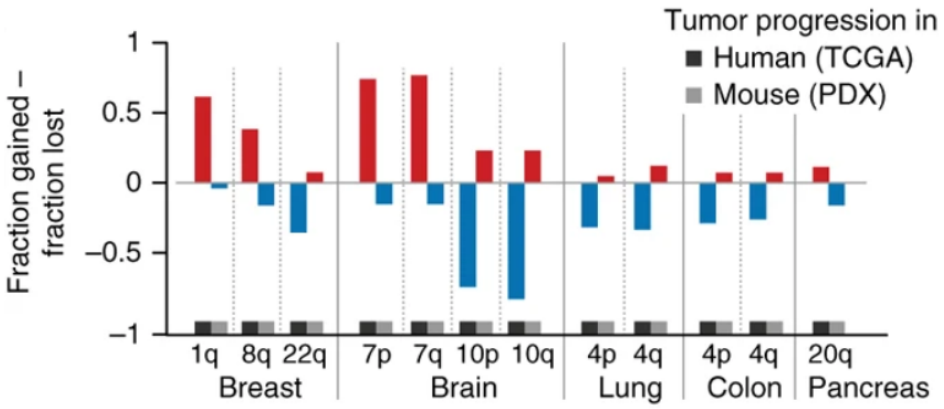

Figure 1.6: **Recurrent arm-level TCGA CNAs tend to disappear throughout PDX passaging**. Bar plots represent the difference between the fraction with gain and the fraction with loss for 12 recurrent TCGA arm-level CNAs [83].

discern whether genetic variations between human tumors and PDXs are the result of mouse-driven selective pressures or simply neutral evolution and genetic drift.

Chapter 2

Project aim and plan

This thesis aims to provide a systematic assessment of genetic changes likely occurring during the establishment and propagation of patient-derived xenograft (PDX) cancer models.
In details, my project is focused on three main aims:

1. to evaluate the genetic stability of PDX models from PTs through late-passage PDXs
2. to investigate whether the mouse host imposes selective pressures during PDX engraftment and propagation, affecting the accuracy of PDXs in modeling human cancer
3. to compare PDX-associated genetic evolution to what patients experienced naturally in their tumors

To achieve these objectives, we assembled a large size and international collection of PDX models and matched patient tumors, in collaboration with the EurOPDX and PDXNet Consortia, and we estimated their CNA profiles. The combined PDX data include 1,451 unique samples, comprise 509 PDX models, represent 16 tumor types and encompass samples profiled with multiple genomic platforms.
On a hand, our collection includes 324 models with matched PDX samples and corresponding patient tumors. On the other hand, it encompasses 328 models with multiple PDX samples assayed at either different passages (ranging from P0-P21) or different lineages of propagation into distinct mice.

We took advantage of this dataset :

- to benchmark CNA profiles inferred from DNA and RNA-based platforms;
- to elucidate potential copy number evolution driven by the mouse host in PDXs derived from different tumor types;
- to investigate possible recurrent CNA changes occurring in PDXs in specific tumor types;
- to compare PDX-specific evolution with the levels of copy number variation in multi-region samples of PTs.

Accordingly, we firstly evaluated copy number alteration (CNA) changes during engraftment and passaging, in a large collection of more than 500 PDX models, comparing both DNA and RNA-based approaches across a variety of tumor types.
Then, to understand whether the potential CNA changes observed between patient tumors and PDXs result from spontaneous tumor evolution and intratumor heterogeneity or mouse-driven selective pressures, we searched for recurrent genetic shifts progressively arising in PDXs. Indeed, whether the mouse host imposed specific selective pressures, genomic changes should emerge, systematically and reproducibly, during the establishment and propagation of PDXs.

Chapter 3

Materials and Methods

3.1 Experimental details for sample collection, PDX engraftment and passaging, and array or sequencing

3.1.1 EurOPDX colorectal cancer (EurOPDX CRC)

The copy number stability of colorectal cancer was studied based not directly on primary tumors, but rather on metastasis located in the liver. Therefore, as discussed in (1.1.4), the intra-tumor heterogeneity of liver metastatic colorectal cancer may not recapitulate that of the relative primary tumor. However, as we aim to evaluate the genetic robustness of PDXs, the absolute requirement is having matched PDX samples and patient tumors, irrespective of the primary or metastatic origin.
Liver-metastatic colorectal cancer samples were obtained from surgical resection of liver metastases at the Candiolo Cancer Institute, the Mauriziano Umberto I Hospital, and the San Giovanni Battista Hospital. Informed consent for research use was obtained from all patients at the enrolling institution before tissue banking, and study approval was obtained from the ethics committees of the three centers.

Tissue from hepatic metastasectomy in affected individuals was fragmented and either frozen or prepared for implantation as described previously.[87,88] Non-obese diabetic/severe combined immunodeficient (NOD/SCID) female

mice (4–6 weeks old) were used for tumor implantation.

Whole-genome sequencing was conducted as follows: DNA was extracted using Maxwell RSC Blood DNA kit (Promega AS1400) from colorectal cancer liver metastasis and corresponding tumor grafts at different passages. Genomic DNA was fragmented and used for Illumina TruSeq library construction (Illumina) according to the manufacturer's instructions. Libraries were then purified with Qiagen MinElute column purification kit and eluted in 17 µl of 70°C EB to obtain 15 µl of DNA library. The libraries were sequenced on HiSeq4000 (Illumina) with single-end reads of 51bp at low coverage (~0.1x genome coverage on average).

3.1.2 EurOPDX breast cancer (EurOPDX BRCA)

Human breast tumors were obtained from surgical resections at the Netherlands Cancer Institute (NKI), Institut Curie (IC), and Vall d'Hebron Institute of Oncology (VHIO). Engraftment was conducted with different procedures at each center.

NKI: Small tumor fragments (2mm diameter) were implanted into the 4th mammary fat pad of 8-week-old Swiss female nude mice. Mice were checked for tumor appearance once a week and supplemented with estrogen if the tumor was ER-positive. After palpable tumor detection, tumor size was measured twice a week. When tumors reached a size of 700-1000 mm3, animals were sacrificed and tumors were explanted and subdivided into fragments for serial transplantation as described above, or for frozen vital storage in liquid nitrogen.

IC: Breast cancer fragments were obtained from patients at the time of surgery, with informed written patient consent. Fragments of 30 to 60 mm3 were grafted into the interscapular fat pad of 8 to 12-week-old female Swiss nude mice. Mice were supplemented with estrogen. Xenografts appeared at the graft site 2 to 8 months after grafting. When tumors were close to 1500 mm3, they were subsequently transplanted from mouse to mouse and stocked frozen in DMSO-fetal calf serum (FCS) solution or frozen dried in nitrogen. Fragment fixed tissues in phosphate-buffered saline (PBS) 10% formol for histologic studies were also stored. The experimental protocol and animal housing were under institutional guidelines as proposed by the French Ethics Committee (Agreement B75-05-18, France).

VHIO: Fresh tumor samples from patients with breast cancer were collected for implantation following an institutional IRB-approved protocol and the associated informed consent, or by the National Research Ethics Service, Cambridgeshire 2 REC (REC reference number: 08/H0308/178). Experiments were conducted following the European Union's animal care directive (2010/63/EU) and were approved by the Ethical Committee of Animal Experimentation of the Vall d'Hebron Research Institute. Surgical or biopsy specimens from primary tumors or metastatic lesions were immediately implanted in mice. Fragments of 30 to 60 mm3 were implanted into the mammary fat pad (surgery samples) or the lower flank (metastatic samples) of a 6-week-old female athymic HsdCpb: NMRI-Foxn1nu mice (Harlan Laboratories). Animals were continuously supplemented with estradiol. Upon growth of the engrafted tumors, the model was perpetuated by serial transplantation onto the lower flank. Tumor growth was measured with a caliper bi-weekly. In all experiments, mouse weight was recorded twice weekly. When tumors reached 1500 mm3, mice were euthanized and tumors were explanted.

Whole-genome sequencing was conducted as follows: genomic DNA was extracted from breast cancers and corresponding PDXs using (i) QIAamp DNA Mini Kit s(50) (#51304, Qiagen) (IC) or (ii) according to Laird PW's protocol16 (NKI and VHIO). The amount of double-stranded DNA in the genomic DNA samples was quantified by using the Qubit® dsDNA HS Assay Kit (Invitrogen, cat no Q32851). Up to 2000 ng of double-stranded genomic DNA were fragmented by Covaris shearing to obtain fragment sizes of 160-180bp. Samples were purified using 1.6X Agencourt AMPure XP PCR Purification beads according to the manufacturer's instructions (Beckman Coulter, cat no A63881). The sheared DNA samples were quantified and qualified on a BioAnalyzer system using the DNA7500 assay kit (Agilent Technologies cat no. 5067-1506). With an input of a maximum of 1 µg sheared DNA, library preparation for Illumina sequencing was performed using the KAPA HTP Library Preparation Kit (KAPA Biosystems, KK8234). During library enrichment, 4-6 PCR cycles were used to obtain enough yield for sequencing. After library preparation, the libraries were cleaned up using 1X AMPure XP beads. All DNA libraries were analyzed on the GX Caliper (a PerkinElmer company) using the HT DNA High Sensitivity LabChip, for determining the molarity. Up to two pools of 24, uniquely indexed samples and one pool of 81 uniquely indexed samples were mixed by equimolar pooling in a final concentration of 10nM, and subjected to sequencing on an Illumina HiSeq2500

machine in a total of 12 lanes of a single read 65bp run at low coverage (~0.4x genome coverage on average), according to manufacturer's instructions.

3.1.3 Seoul National University-Jackson Laboratory (SNU-JAX)

Gastric cancer tissues paired with normal gastric tissues and blood samples were obtained from individuals who underwent gastrectomies at the Hospital of Seoul National University from 2014 to 2016. All samples were obtained with informed consent at the Hospital of Seoul National University.

Gastric cancer samples were divided into several small pieces (2mm x 2mm) and used to generate PDX models and for genomic analysis. Mice were cared for according to guidelines of the Institutional Animal Care and Use Committee of the Seoul National University. For PDX models, surgically resected tissues were minced into pieces approximately ~2 mm in size and injected into the subcutaneous area in the flanks of 6-week-old NOD/SCID/IL-2 γ-receptor null female mice. When a tumor reached >700~1000 mm^3, the mouse was sacrificed, and tumor tissues were stored. Tumor tissues were divided and stored for several purposes: (1) Tumor tissues were cryopreserved in liquid nitrogen and stored at -80 °C for generating next passage PDXs. (2) Tumor tissues were frozen in liquid nitrogen for genomic analysis.

Whole-exome sequencing was conducted as follows: Genomic DNA was extracted from blood and tissues using DNeasy blood and tissue kit (QIAGEN) and checked for purity, concentration, and integrity by OD260/280 ratio using NanoDrop Instruments and agarose gel electrophoresis. DNA was sheared by fragmentation by Bioruptor and purified using Agencourt AMPure XP beads. DNA samples were then tested for size distribution and concentration using an Agilent Bioanalyzer 2100. Standard protocols were utilized for adaptor ligation, indexing, high-fidelity PCR amplification. Subsequently, exome enrichment was performed by hybrid capture with the All Exon v5 capture library. Capture libraries were amplified, pooled, and submitted to the commercial sequencing company (Macrogen) for 100bp paired-end, multiplex sequencing on a HiSeq 2000 sequencing system. Average coverage for normal samples was 62.67x (38.97 min – 108.77 max) and was 102.35x for tumor samples (36.02 min – 150.49 max).

RNA-Sequencing data was generated as follows: RNA was extracted from

tissues using the RNeasy Mini Kit the TruSeq RNA Sample Preparation v2 Kit (Illumina, San Diego, CA) according to the manufacturer's protocol. Libraries were submitted to the commercial sequencing company (Macrogen) for 100bp paired-end, multiplex sequencing on a HiSeq 2000 sequencer.

3.1.4 Shanghai Institute for Biological Sciences (SIBS)

Gene expression and copy number data, generated by the Affymetrix Human Genome U133 Plus 2.0 Array and Affymetrix Human SNP 6.0 platforms respectively, of hepatocellular carcinoma (HCC) PDX models, were retrieved from the Gene Expression Omnibus (GEO) accession ID GSE90653.[89] Expression microarray data generated by the Affymetrix Human Genome U133 Plus 2.0 Array for normal liver were downloaded from the GEO and ArrayExpress: GSE3526,[90] GSE33006[91] and E-MTAB-1503-3.[92]

3.2 Copy number alteration (CNA) estimation methods

3.2.1 SNP array

For Affymetrix Human SNP 6.0 arrays, PennCNV-Affy and Affymetrix Power Tools[93] were used to extract the B-allele frequency (BAF) and Log R Ratio (LRR) from the CEL files. Due to the absence of paired-normal samples, the allele-specific signal intensity for each PDX tumor was normalized relative to 300 randomly selected sex-matched Affymetrix Human SNP 6.0 array CEL files obtained from the International HapMap project.[94] For Illumina Infinium Omni2.5Exome-8 SNP arrays (v1.3 and v1.4 kit), the Illumina GenomeStudio software was used to extract the B-allele frequency (BAF) and Log R Ratio (LRR) from the signal intensity of each probe. The single sample mode of the Illumina GenomeStudio was used, which normalizes the signal intensities of the probes with an Illumina in-house dataset. The single tumor version of ASCAT[95] (v2.4.3 for JAX SNP data, v2.5.1 for SIBS SNP data) was used for GC correction, predictions of the heterozygous germline SNPs based on the SNP array platform, and estimation of ploidy, tumor content, and allele-specific copy number segments. The resultant copy number segments were annotated with $\log_2$ ratio of total copy number relative to predicted ploidy from ASCAT.

3.2.2 Low-pass whole-genome sequencing (WGS) data

Whole-genome sequence reads from EurOPDX CRC liver metastasis and corresponding tumor grafts at different passages were mapped to the reference human genome (GRCh37) using Burrows-Wheeler Aligner[96] (BWA) v0.7.12. SAMTools[97] v0.1.18 was used to convert SAM files into BAM files and Picard v1.43 to remove PCR duplicates (http://broadinstitute.github.io/picard/). Raw copy number profiles for each sample were estimated by QDNAseq[98] R package v1.20 by dividing the human reference genome in non-overlapping 50 kb windows and counting the number of reads in each bin. Bins in problematic regions were removed.[99] Read counts were corrected for GC content and mappability by a LOESS regression, median-normalized, and $\log_2$-transformed. Values below –1000 in each chromosome were floored to the first value greater than –1000 in the same chromosome. Raw $\log_2$ ratio values were then segmented using the ASCAT[95] algorithm implemented in the ASCAT R package v2.0.7.

Whole-genome sequence reads from EurOPDX BRCA tumors and corresponding tumor grafts at different passages were mapped to the reference human genome (GRCh38) and mouse genome (GRCm38/mm10, Ensembl 76) using Burrows-Wheeler Aligner (BWA) v0.7.15. Subsequently, mouse reads were excluded with XenofilteR.[100] Raw copy number profiles were estimated for each sample by dividing the human reference genome in non-overlapping 20 kb windows and counting the number of reads in each bin. Only reads with at least mapping quality 37 were considered. Bins within problematic regions (i.e. multi mapper regions) were excluded. Downstream analysis to estimate copy number was conducted as described above.

3.2.3 Whole-exome sequencing (WES) data

All the samples were subjected to quality control (filtering and trimming of poor-quality reads and bases) using an in-house QC script with the cut-off that half of the read length should be 20 in base quality at the Phred scale. We further removed the known adaptors using cut-adapt[101] v1.15 11 at -m 36. Afterward, we aligned the reads to the human genome (GRCh38.p5) using bwakit[96] v0.7.15. Engrafted tumor samples were subjected to the additional step of mouse read removal using Xenome[102] v1.0.0, with default parameters. The alignment was converted to BAM format using Picard SortSam v2.8.1 (https://broadinstitute.github.io/picard/), and dupli-

cates were removed by Picard MarkDuplicates utility. BaseRecalibrator from the Genome Analysis Tool Kit[103,104] (GATK) v4.0.5.1 was used to adjust the quality of raw reads. Training files for the base quality scale recalibration were Mills_and_1000G_gold_standard.indels.hg38.vcf.gz, Homo_sapiens_assembly38.known_indels.vcf.gz, and dbSNP v151. Mean target coverage was determined for each sample by Picard CollectHsMetrics. Aligned bams were subset to target region by GATK and SAMTools[97] v0.1.18 was used to generate the pileup for each sample. Pileup data were used for CNA estimation as calculated with Sequenza[105] v2.1.2. Both tumor and normal data, that utilized the same capture array, were used as input. pileup2seqz and GC-windows (-w 50) modules from sequenza-utils.py utility were used to create the native seqz format file for Sequenza and compute the average GC content in sliding windows from hg38 genome, respectively. Finally, we ran the three Sequenza modules with these modified parameters (sequenza.extract: assembly = "hg38", sequenza.fit: chromosome.list = 1:23, and sequenza.results: chromosome.list = 1:23) to estimate the segments of copy number gains/losses. Finally, segments lacking read counts, in which 50% of the segment with zero read coverage, were removed. A reference implementation of this workflow is developed and deployed in the cancer genomics cloud at SevenBridges (https://cgc.sbgenomics.com/public/apps#pdxnet/pdx-wf-commit2/wes-cnv-tumor-normal-workflow/, https://cgc.sbgenomics.com/public/apps#pdxnet/pdx-wf-commit2/pdx-wes-cnv-xenome-tumor-normal-workflow/).

3.2.4 RNA-sequencing (RNA-seq) and gene expression microarray (EXPARR) data

For SNU-JAX RNA-Seq data, simultaneous read alignment was performed to both the mouse (mm10) and the human genome (GRCh38.p5) and only human-specific reads were used for the expression quantification. Moreover, to be able to compare the mRNA expression values between samples independently of the dataset origin, the Transcripts Per Million (TPM) normalization method was carried on using RNA-Seq by Expectation Maximization[106] (RSEM) with ensemble GTF reference GRCh38.92.
For gene expression microarray data for SIBS HCC and normal liver samples from GEO and ArrayExpress databases were profiled as follows. After initial quality control and outlier removal, CEL files were normalized according

to the RMA algorithm and probesets were annotated according to the Affymetrix annotation file for HG-U133 Plus 2, released on 2016-03-15 build 36.

For expression-based copy number inference, we referred to the previous protocols of e-karyotyping and CGH-Explorer for both RNA-seq and gene expression array data.[86,107–109]
In detail, for each cancer type, expression values of the tumor and corresponding normal samples were merged in a single table, and gene identifiers were annotated with chromosomal nucleotide positions. Then multiple criteria for data cleaning were implemented. At first, genes located on sex chromosomes were excluded. Then genes with expression values below 1 TPM (RNAseq) or probeset $\log_2$-values below 6 (microarray) in more than 20% of the analyzed dataset were removed. At this point, the remaining gene expression values below the thresholds were respectively raised to 1 TPM or $\log_2$-value of 6. Furthermore, in the case of multiple transcripts (RNA-seq) or probesets (microarray) per gene, the one with the highest median value across the entire dataset was selected. Finally, the sum of squares of the expression values relative to their median expression across all samples was calculated for each gene, and the 10% most highly variable genes were removed.
To produce relative copy number values, for each gene, the median $\log_2$ expression value in normal samples was subtracted from the $\log_2$ expression value in each tumor sample and subsequently input in CGH-explorer. Instead, for tumor-only datasets, the median $\log_2$ expression value in the same set of tumor samples was subtracted. At this stage, the preprocessed relative expression profiles of each sample were individually analyzed using CGH-Explorer.[109]
Specifically, as part of the CGH-Explorer program, the Piecewise Constant Fitting (PCF) algorithm was applied to convert gene-level into segment-level copy number values. At last, parameters previously reported were used in the CGH-Explorer program to carry on copy number calling:[83] least allowed deviation = 0.25; least allowed aberration size = 30; winsorize at quantile = 0.001; penalty = 12; threshold = 0.01.

3.3 Filtering and gene annotation of copy number segments

Copy number (CN) segments with $\log_2$ copy number ratio estimated from the various platforms were processed in the following steps. Segments <1kb were filtered based on the definition of CNA.[110] Also, SNP array segments had to be covered by >10 probes, with an average probe density of 1 probe per 5kb. The copy number segments were then probes, with an average probe density of 1 probe per 5kb. The copy number segments were then re-center copy number segments. Median-centered copy number segments were visualized using IGV[111] v2.4.13 and GenVisR[112] v1.16.1. Median-centered copy number values of genes were calculated by intersecting the genome coordinates of copy number segments with the genome coordinates of genes (Ensembl Genes 93 for human genome assembly GRCh38, Ensembl Genes 96 for human genome assembly GRCh37). In the case where a gene overlaps multiple segments, the most conservative (lowest) estimate of copy number was used to represent the copy number of the entire intact gene.

3.4 Correlation of CNA profiles

The similarity of two CNA profiles is quantified by the Pearson correlation coefficient of $\log_2$(CN ratio) of 100kb-windows binned from segments or genes between 2 samples. Using correlation avoided the issue of making copy number gain and loss calls based on thresholds. Median-centering of each CNA profile approximates normalization by the sample ploidy. One caveat of our approach, however, is that it cannot distinguish genome-wide multiplication of ploidy between samples, as the correlation statistic is invariant to such genome-wide transformations. As such we cannot assess whether ploidy changes occur between samples of a given model.

3.5 Comparison of CNA profiles between different platforms

The copy number segments of each pair of samples were intersected and binned into 100kb-windows using Bedtools. Then the Pearson correlation coefficient was calculated for the $\log_2$(CN ratio) of the windows.

3.6 Association of mutations with copy number correlations

For WGS data, we collected the mutational status (wild-type or mutated) of TP53, BRCA1, and BRCA2 per model where available, which may or may not be obtained from the same tumor samples used in this study. For WGS data, mutations were obtained from the whole-exome or targeted panel sequencing[113] (unpublished data), and high-quality and likely functional mutations were retained. For each sample pair with copy number correlations, the mutational status was available on a per model basis. BRCA is labeled as mutated when either BRCA1 or BRCA2 is mutated.

3.7 Annotation with gene sets with known cancer or treatment-related functions

Copy number genes were annotated by various gene sets with cancer of treatment-related functions gathered from various databases and publications. 1. Genes in 10 oncogenic signaling pathways curated by TCGA and genes found to be frequently altered in different cancer types[114]. 2. Genes with gain in copy number or expression, or loss in copy number or expression that conferred therapeutic sensitivity, resistance or increase/decrease in drug response from the JAX Clinical Knowledgebase[115] (JAX-CKB) based on literature curation (https://ckbhome.jax.org/, as of 06-18-2019) 3. Genes with evidence of promoting oncogenic transformation by amplification or deletion from the Cancer Gene Census[116] (COSMIC v89). 4. Significantly amplified or deleted genes in TCGA cohorts of breast cancer,[117] colorectal cancer,[118] lung adenocarcinoma,[119] and lung squamous cell carcinoma[120] by GISTIC analysis.

3.8 GISTIC analysis of WGS data

The GISTIC[121] algorithm (GISTIC 2 v6.15.28) was applied on the segmented profiles using the GISTIC GenePattern module (https://cloud.genepattern.org/), with default parameters and genome reference files Human_Hg19.mat for EuroPDX CRC data and hg38.UCSC.add_miR.160920.refgene.mat for EuroPDX BRCA data. For each dataset, GISTIC provides separate results

(including segments, G-scores, and FDR q-values) separately for recurrent amplifications and recurrent deletions. Deletion G-scores were assigned negative values for visualization. We observed that the G-Score range was systematically lower in PT cohorts, which is likely the result of the dilution of CNA by normal stromal DNA. In contrast, human stromal DNA in PDX samples was lower or negligible. To account for this difference in gene-level G-scores, PDXs at early and late passages were scaled to PT gene-level G-score values using global linear regression, separately for amplification and deletion outputs.

3.9 Gene set enrichment analysis (GSEA) of WGS data

To assess the biological functions associated with the recurrent alterations detected by the GISTIC analysis, we performed GSEAPreranked analysis[122] on gene-level GISTIC G-score profiles, for both amplifications and deletions. In particular, we applied the algorithm with 1000 permutations on various gene set collections from the Molecular Signatures Database[123,124] (MSigDB): H (Hallmark), C2 (Curated: CGP chemical and genetic perturbations, CP canonical pathways), C5 (Gene Ontology: BP biological process, MF molecular function, CC cellular component) and C6 (Oncogenic Signatures) composed of 50, 4762, 5917 and 189 gene sets respectively. We also included gene sets with known cancer or treatment-related functions described in an earlier section. We noted that multiple genes with contiguous chromosomal locations, typically in recurrent amplicons, generated spurious enrichment for gene sets that consists of multiple genes of adjacent positions, while very few or none of them had a significant GISTIC G-score. To avoid this confounding issue, we only considered the "leading edge genes", i.e. those genes with increasing Normalized Enrichment Score (NES) up to its maximum value that contribute to the GSEA significance for a given gene set. The leading-edge subset can be interpreted as the core that accounts for the gene set's enrichment signal (http://software.broadinstitute.org/gsea). We included a requirement that the leading edge genes passing the GISTIC G-score significant thresholds based on GISTIC q-value 0.25 make up at least 20% of the gene set. This 20% threshold was chosen as the minimum threshold at which gene sets assembled from TCGA-generated lists of genes with recur-

rent CNA in CRC or BRCA were identified as significant in GSEA. Finally, gene sets with a NES greater than 1.5 and an FDR q-value of less than 0.05, which passed the leading-edge criteria, were considered significantly enriched in genes affected by recurrent CNAs.

Chapter 4

Results

4.1 Catalog of copy number alterations (CNAs) in PDXs

To ensure a comprehensive and systematic analysis of Copy Number Alterations in PDXs, we assembled the CNA profiles of 1451 unique samples:

- 324 patient tumor (PT)
- 1127 PDX samples.

Notably, this collection included 509 PDX models and covered 16 broad tumor types. Moreover, it was collected by the contribution of the EurOPDX, the PDXNET Consortium and by other published datasets.[125,126]

Notably, each model of the assembled PDX data encompasses matched samples. In detail, 324 PDX models include their corresponding patient tumors. Instead, 328 PDX models include multiple PDX samples assayed at either different passages (ranging from P0-P21) or different lineages of propagation in mice (Fig. 4.1a).

We estimated copy number measurements on both DNA and RNA data obtained by five data types :

- single nucleotide polymorphism (SNP) array
- whole-exome sequencing (WES)
- low-pass whole-genome sequencing (WGS)

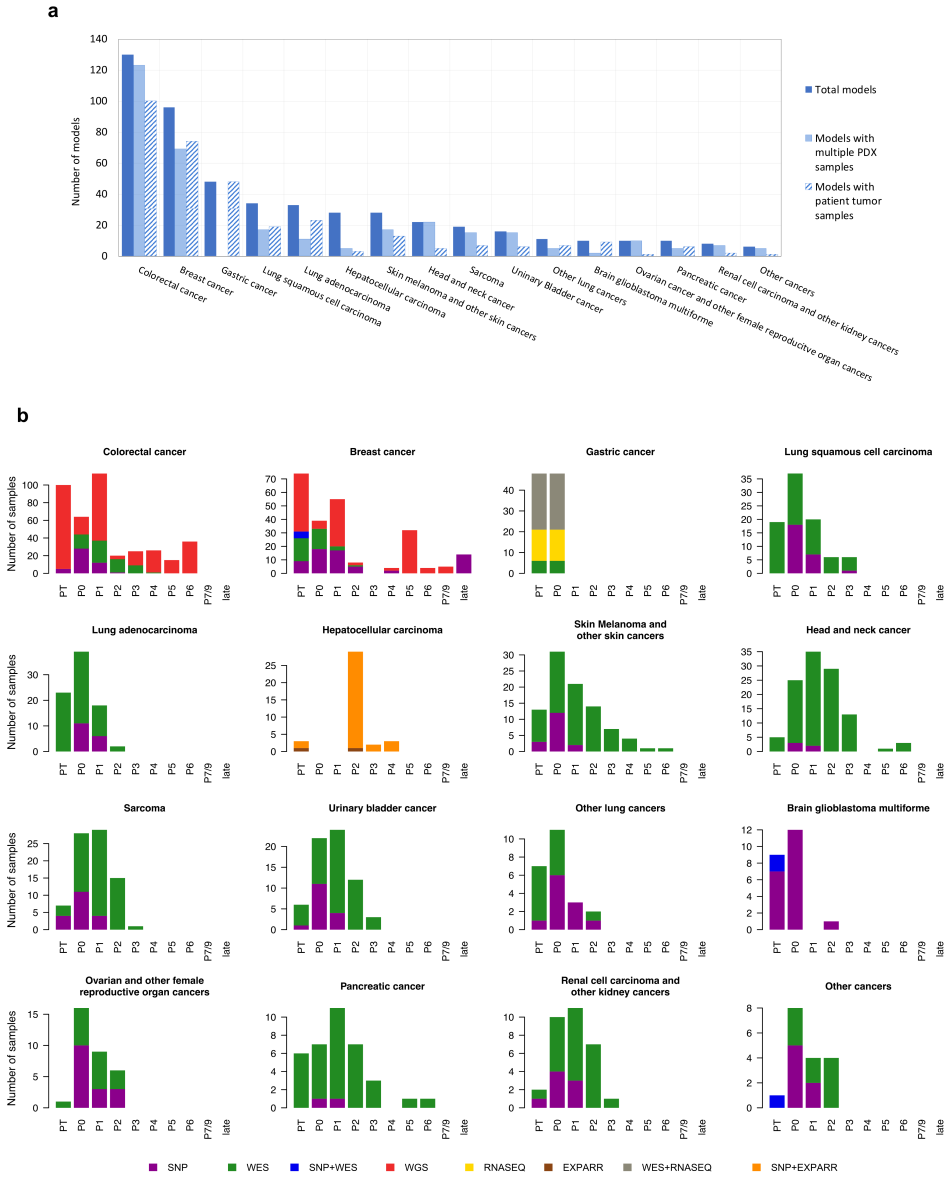

Figure 4.1: **PDX datasets used for copy number profiling across 16 tumor types. a**, Numbers of PDX models for each tumor type, with models including matched PT samples or multiple PDX samples. **b**, Distributions of datasets by passage number and assay platform for PTs and PDX samples, by tumor type [145].

- RNA sequencing (RNA-seq)
- gene expression (GEP) array.

To provide a benchmark of CNAs inferred from different genomic platforms, we assembled a dataset with matched measurements across multiple platforms (Fig. 4.1b). Precisely, we combined:

- 8 patient tumor samples with matched WES and SNP array data;
- 27 patient tumor and 27 PDX samples with matched WES and RNA-seq data;
- 2 patient tumor and 33 PDX samples with matched SNP and GEP array data.

4.2 A benchmark of copy number profiles inferred on DNA and RNA data

We took advantage of our dataset with matched measurements across multiple platforms, to assess the accuracy of CNA profiles estimated on DNA and RNA data.

It has been reported that copy number calling could be noisy for several data types.[127,128] In this respect, in our dataset, we observed that quantitative comparisons between CNA profiles are sensitive to:

1. the thresholds and baselines used to define gains and losses;
2. the dynamic range of copy number values from each platform;
3. the differential impacts of normal cell contamination for different measurements.

Therefore, to control for such systematic biases, we assessed the similarity between two CNA profiles by the Pearson correlation of their $\log_2$(CN ratio) values across the genome in 100-kb windows. Moreover, we identified regions with discrepant copy number as those with outlier values from the linear regression model.

Table 4.1: Resolution and dynamic range of CN segments from SNP arrays and WES.

	Median/Mean segment size (Mb)	Range of log2(CN)
SNP	1.49/4.05	[-8.62,2.84]
WES	4.70/14.6	[-3.04,1.85]

4.2.1 CNAs consistency of CNAs from WES and SNP array data

On one hand, SNP arrays are largely accepted for CNA profiles estimation.[129,130] On the other hand, WES data are reported to have more uncertainty.[105,131] Therefore, we validated our CNA estimates based on WES against those based on SNP array for matched samples.

We observed that copy number segments from SNP arrays had higher resolution and broader dynamic range than those estimated from WES (Fig. 4.2a-b and Table 4.1). Interestingly, the differences in resolution and dynamic range across these two platforms were statistically significant ($P <$ 2.2 x 10-16). Notably, the difference in resolution and dynamic range was apparent in the linear regressions between platforms (Fig. 4.3a).

These observations were consistent with the broad factors affecting CNA estimates across platforms such as :

- the positional distribution of sequencing loci
- the sequencing depth of WES
- the superior removal of normal cell contamination by SNP array CNA analysis workflows using SNP allele frequencies.[95]

However, despite of SNP arrays superiority in copy number inference, we observed strong agreement between SNP arrays and WES for matched samples assayed on both platforms (Fig. 4.4). Notably, Pearson correlation coefficients of matched samples were significantly higher than those of unmatched samples (range: 0.913 – 0.957 for matched samples, 0.0366 – 0.354 for unmatched samples, p = 1.02e-06) except for two samples that lacked CNA aberrations and were removed (Fig. 4.2c and Fig. 4.5).

Furthermore, the discordant copy number regions largely correspond to small

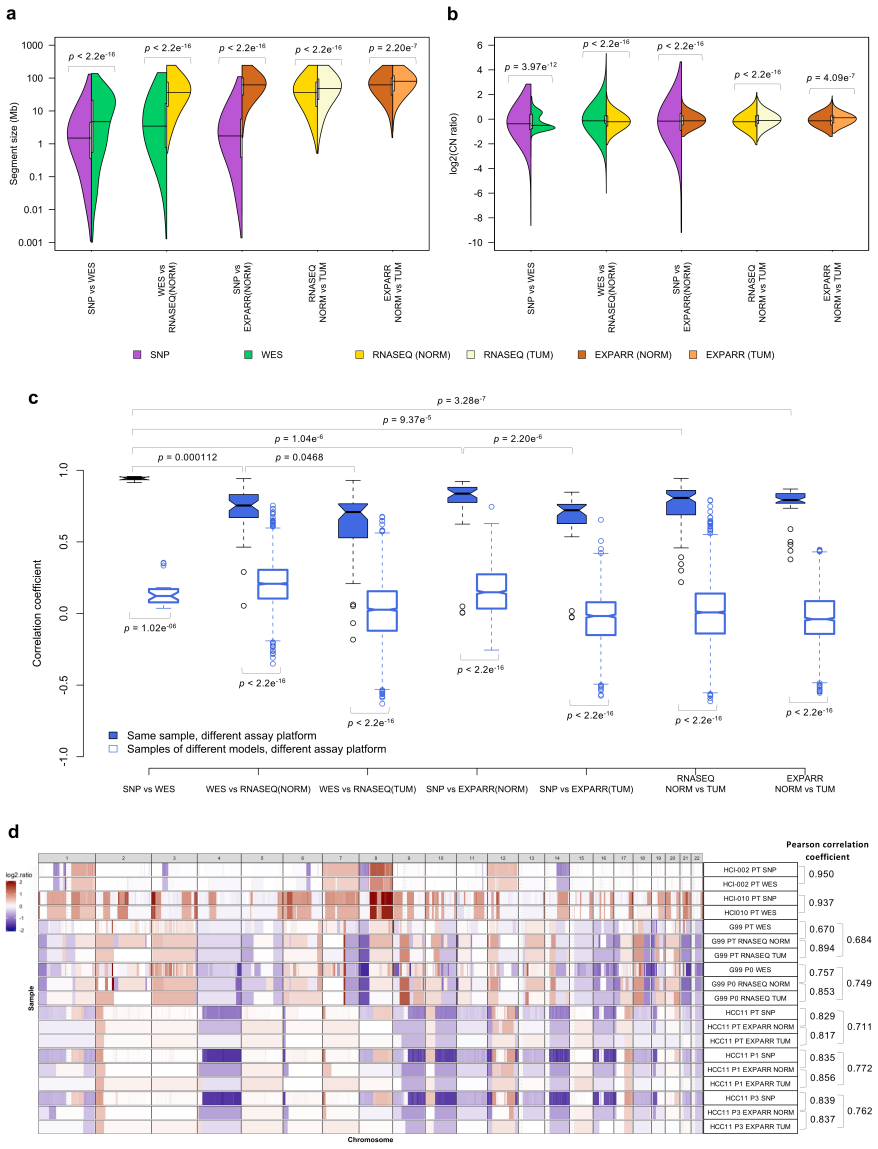

Figure 4.2: **Comparisons of resolution and accuracy for CNAs estimated using DNA and expression-based methods across different measurement platforms.** Copy Number Alterations (CNAs) analysis of profiles inferred on different platforms. Pairwise comparisons of the distributions of CNA segment sizes estimated (**a**) and $\log_2$(copy number ratio) values (**b**). **c**, Distributions of Pearson correlation coefficients of median-centered $\log_2$(copy number ratio) values in 100-kb windows from CNA segments between pairs of samples. **d**, Examples of matched CNA profiles [145].

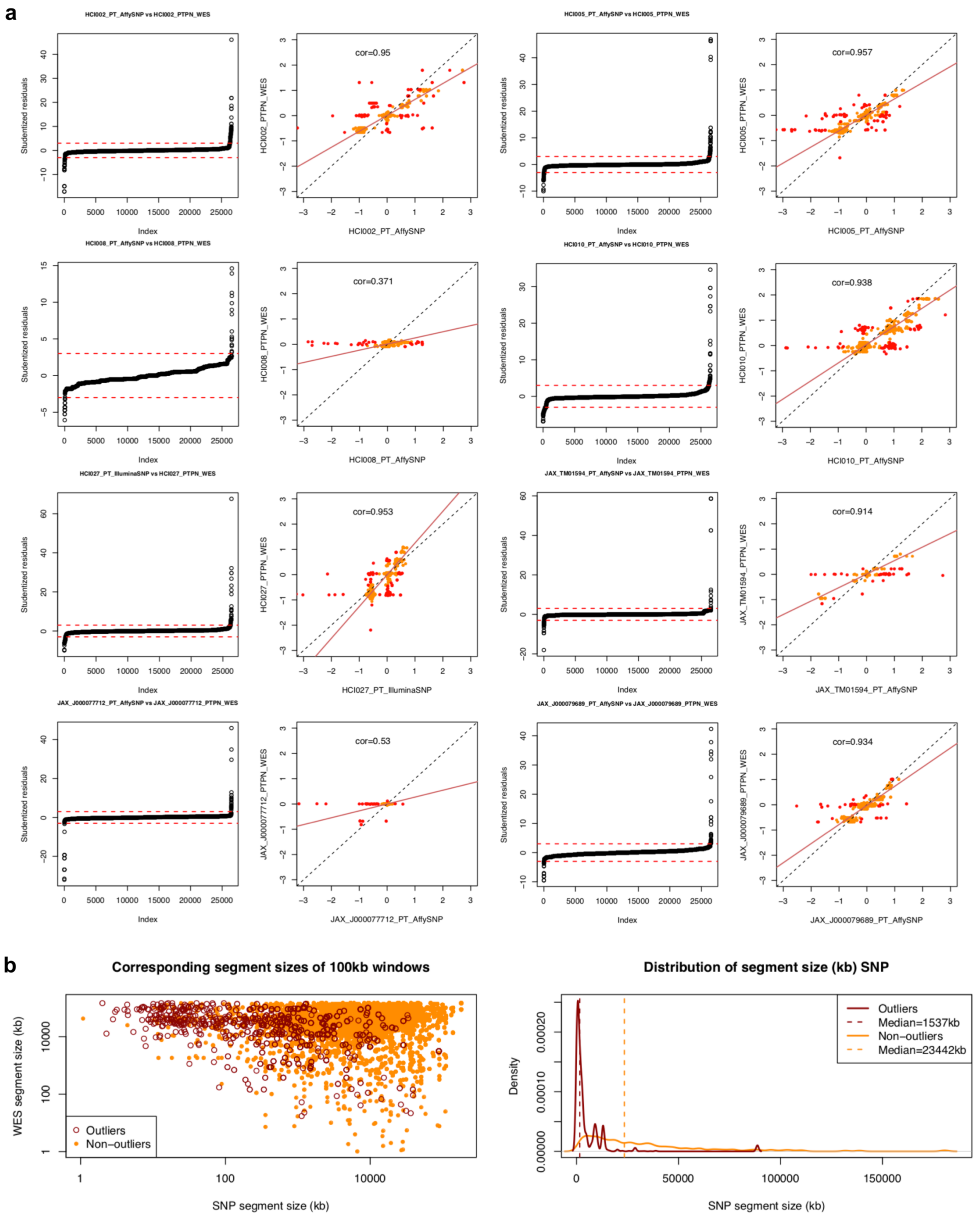

Figure 4.3: **Difference in range in the copy number values of matched CNA profiles estimated from SNP array and WES. a**, Pearson correlation and linear regression of the $\log_2$(CN ratio) of 100kb-windows binned from copy number segments of CNA profiles. Outliers of the linear regression (red points) are identified by studentized residuals $>$ 3 and $<$ -3. **b**, Comparison of segment sizes between the combined outlier and non-outliers in (a) [145].

focal events (average size 1.53 Mb) detectable by SNP arrays but missed by WES (Fig. 4.3b). Hence, CNA profiling by WES was reliable in most regions in this small dataset.

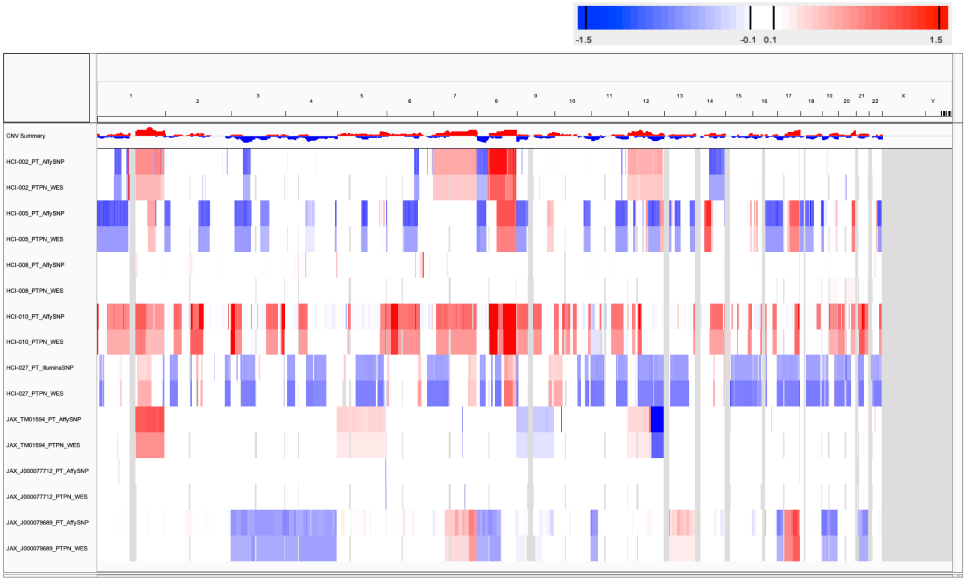

Figure 4.4: **Benchmark of CNA inferences based on SNP vs WES.** CNA profiles for matched patient tumor samples estimated from SNP array and WES for *SNP vs WES* benchmarking.

4.2.2 Low accuracy for gene expression-derived CNA profiles

We also assessed the suitability of gene expression for copy number quantification. Therefore, we adapted the e-karyotyping methods used in Ben-David et al. [83,86,108] for RNA-seq and gene expression array data.

We applied e-karyotyping to our datasets with matched samples profiled on SNP and GEP array or WES and RNA-seq platforms.

Concerning RNA-based data, for each tumor type, we centered the expression values on the median expression of normal or tumor RNA samples, when normal profiles were not available.

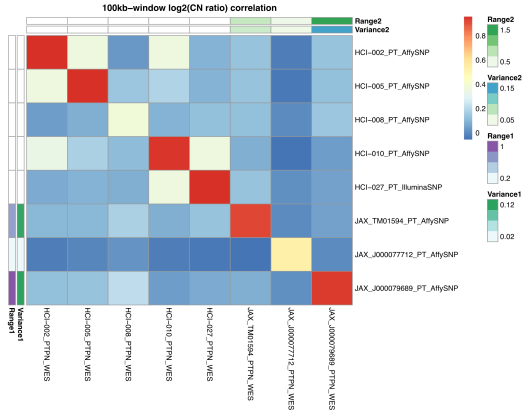

Figure 4.5: **Similarity of matched CNA profiles inferred on SNP array and WES.** Heatmap representing the Pearson correlation coefficients of the $\log_2$(CNratio) of 100kb-windows binned from copy number segments of CNA profiles between matched samples estimated from SNP array and WES.

We observed that copy number segments centered using normal expression were of higher resolution and broader dynamic range compared to those inferred by calibration with tumor samples (Table 4.2). Interestingly, these results were consistent for both RNA-seq and gene expression arrays platforms.

Moreover, we noticed that alternative expression calibrations determine high variability in CNA inferences, especially for regions frequently called gains or losses in specific tumor types, as identified by GISTIC analysis in other

Table 4.2: Resolution and dynamic range of CN segments from RNA-seq and GEP-array centered on NORM and TUM median expression.

	Median/Mean segment size (Mb)	Range of log2(CN)
RNASEQ NORM	36.0/51.9	[-2.07,2.17]
RNASEQ TUM	48.2/65.3	[-1.79,1.81]
EXPARR NORM	62.0/72.4	[-1.40,1.89]
EXPARR TUM	80.1/85.2	[-1.13,1.59]

Table 4.3: Resolution and dynamic range of CN segments from DNA and RNA based methods.

	Median/Mean segment size (Mb)	Range of log2(CN)
WES	3.45/14.0	[-6.00,5.33]
RNASEQ NORM	36.0/51.9	[-2.07,2.17]
SNP	1.73/5.18	[-9.19,4.65]
EXPARR NORM	62.0/72.4	[-1.40,1.89]

studies.[132–134] Specifically, when the median expression level accounts for the aneuploid state at a given locus, aberrations are missed in the samples of interest, if calibration with tumor samples is implemented. In this respect, we observed that chromosomes 8q and 13 were almost exclusively identified as gains, and chromosomes 21 and 22 were almost exclusively as losses in the gastric cancer RNA-Seq dataset when normal samples were used for calibration. Similarly, we called exclusive gains in chromosomes 7q and 20 and losses in chromosomes 4q31-35, 8p,16q, and 21 using normal samples for calibration for the hepatocellular carcinoma expression array dataset. However, using the calibration based on tumor samples, these regions resulted erroneously called with approximately equal frequencies of gains and losses compared to previous reports (Fig. 4.6).[132–134] Therefore, these observations indicate that RNA-based CNA profiles calibrated by tumor samples are problematic.

We then compared RNA-based copy number profiles to those inferred on DNA data.

Gene expression based copy number segments had segmental resolution an order of magnitude worse than the DNA-based methods (Table 4.3). Furthermore, the range of detectable copy number values was also superior for DNA-based methods (Table 4.3).

In addition, we reported lack of correlation between the expression-based and DNA-based methods (range: 0.0541-0.942 for WES vs RNASEQ (NORM); 0.00517-0.921 for SNP vs EXPARR (NORM)) (Fig. 4.2c, 4.7a, 4.8a). Moreover, CNA estimates after tumor-based expression normalization resulted in further discordance with DNA-based copy number results (range: –0.182-0.929, P = 0.0468 for WES 202 vs RNASEQ (TUM); –0.0274-0.847, P = 2.20

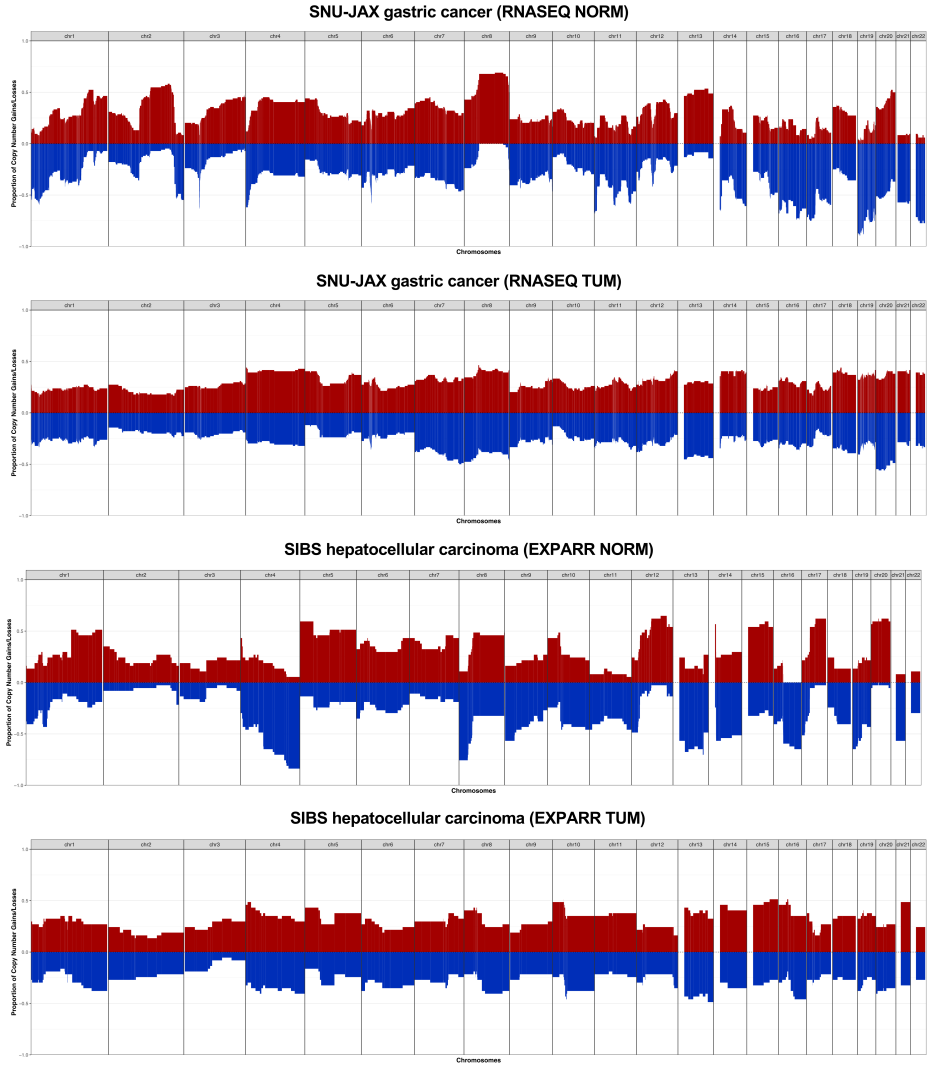

Figure 4.6: **High variability in CNA inferences calibrated on normal vs tumoral samples.** Frequencies of copy number gains ($\log_2$(CN ratio) > 0.1) and losses ($\log_2$(CN ratio) < -0.1) estimated from RNA-Seq and gene expression array normalized by median expression of normal samples of the same tumor type (RNASEQ NORM, EXPARR NORM) or median expression of same set of patient tumors (RNASEQ TUM, EXPARR TUM).

x 10-6 for SNP vs EXPARR (TUM)) (Fig. 4.2c, 4.7b, 4.8b). Representative examples illustrating the superior resolution and accuracy from DNA-based estimates are given in Fig. 4.2d.

We have shown that copy number profiles inferred on DNA data have higher resolution and broader dynamic range than those estimated on RNA platforms. Hence, DNA-based CNA profiles are generally more reliable than those RNA-based. Moreover, concerning CNA profiles based on RNA, we have observed that calibrating the expression on normal or, alternatively, on tumor tissue expression levels, strongly impacts the accuracy of copy number calling. Thus, if copy number estimates based on gene expression data would be the only one available, it is important to be aware that they are gross quantification of copy number values and that they may be highly sensitive to the procedure implemented for signal calibration.
Therefore, for the reasons mentioned, we decide to evaluate the genetic robustness of PDX models separately for DNA and RNA data.

4.3 Concordance of PDXs with patient tumors and during passaging

To track the similarity of CNA profiles during engraftment and passaging, we calculated the Pearson correlation of gene-level copy number for samples measured on the same platform.

We considered only pairs of either patient tumor-PDX (PT-PDX) or PDX-PDX derived from the same PDX model, yielding 501 PT-PDX and 1257 PDX-PDX pairs.

Firstly, we carried on similarity analysis of the PT-PDX pairs to quantify the extent of CNA conservation in PDXs relative to their originating tumors. Then, we performed a similarity analysis of the PDX-PDX pairs to evaluate the amount of copy number changes occurring during PDX expansion and passaging. Moreover, we adopted a pan-cancer approach, since we were interested in elucidating potential tumor type-independent copy number evolution in PDXs driven by the mouse host.

For all DNA-based platforms, we reported high similarity between matched PT-PDX and PDX-PDX pairs. Interestingly, the similarity scores of PT-

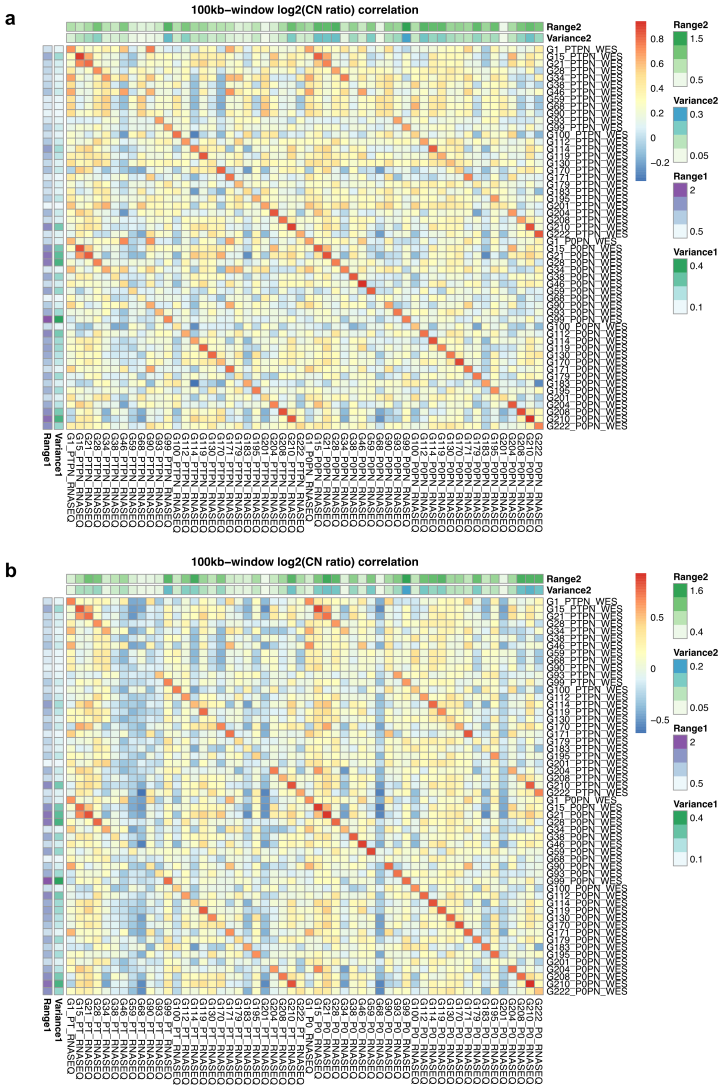

Figure 4.7: **Similarity of matched CNA profiles inferred on WES and RNA-seq differentiating calibrations of gene expression data on normal and tumoral samples.** Heatmap representing the Pearson correlation coefficients of the $\log_2$(CN ratio) of 100kb-windows binned from copy number segments of CNA profiles between matched samples estimated from WES and RNA-Seq, (**a**) normalized by median expression of normal samples of the same tumor type *WES vs RNASEQ (NORM)* or (**b**) median expression of same set of patient tumors *WES vs RNASEQ (TUM)*.

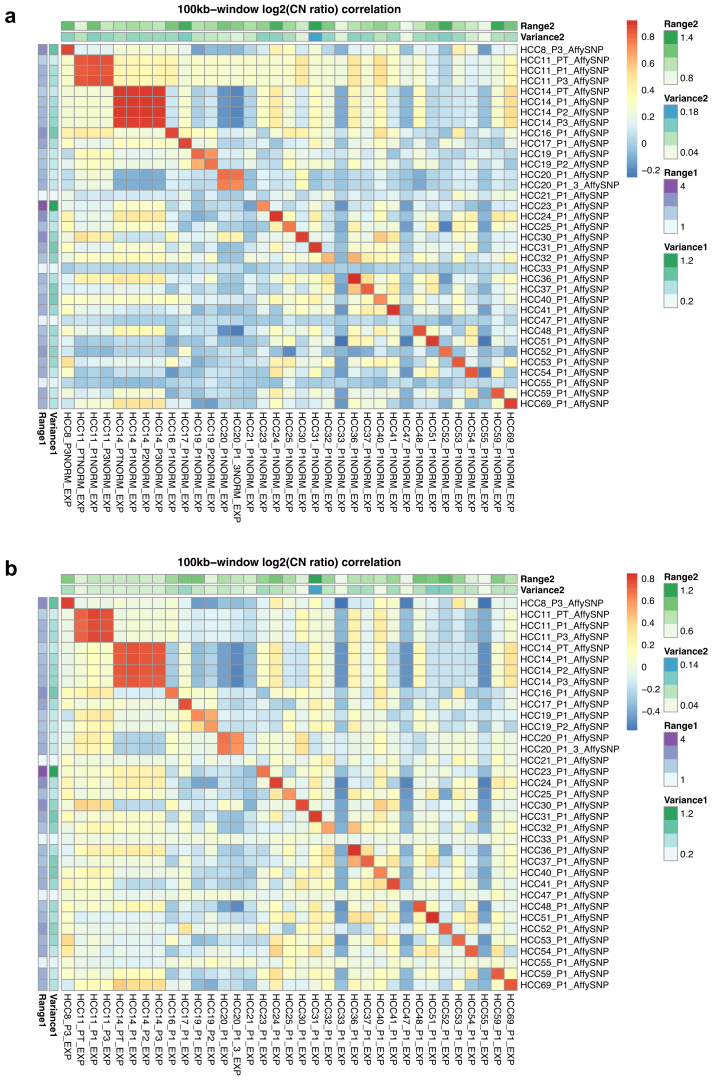

Figure 4.8: **Similarity of matched CNA profiles inferred on SNP array and gene expression microarray differentiating calibrations of gene expression data on normal and tumoral samples.** Heatmap representing the Pearson correlation coefficients of the $\log_2$(CN ratio) of 100kb-windows binned from copy number segments of CNA profiles between matched samples estimated from SNP array and gene expression microarray, (**a**) normalized by the median expression of normal samples of the same tumor type *SNP vs EXPARR (NORM)* or **(b)** median expression of same set of patient tumors *SNP vs EXPARR (TUM)*.

PDX and PDX-PDX pairs derived from the same PDX models were significantly higher than those of PT-PDX and PDX-PDX pairs achieved from different models from the same tumor type and center($p < 2.2e\text{-}16$)(Fig. 4.9a-c).

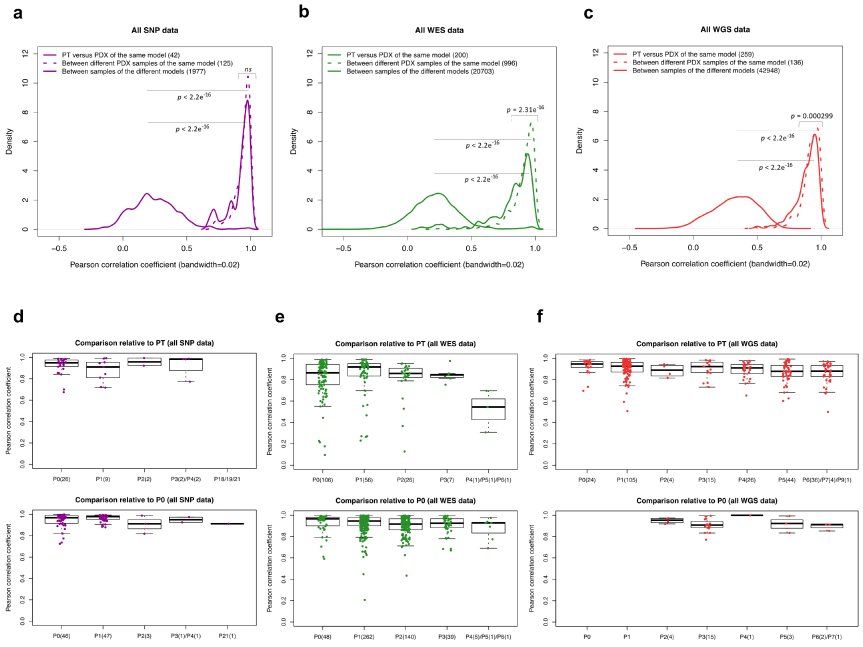

Figure 4.9: **Comparisons of CNAs from PTs with early and late PDX passages.** **a-c**, Distributions of Pearson correlation coefficients of gene-based copy number, estimated by SNP array (**a**), WES (**b**) and WGS (**c**) between: PT–PDX and PDX–PDX samples of the same model; and pair of samples of different models from a common tumor type and contributing center. **d-f**, Distributions of Pearson correlation coefficients of gene-based copy number, estimated by SNP array (**d**), WES (**e**) and WGS (**f**) among PT and PDX passages of the same model, relative to PT (top) and P0 (bottom) are shown.[145].

For SNP array data, the difference in the correlation values between PT-PDX and PDX-PDX pairs was not significant (median correlation = 0.950 for PT–PDX and 0.964 for PDX–PDX; P>0.05). Conversely, there were small but statistically significant shifts of WES (PT–PDX = 0.874; PDX–PDX = 0.936;

P = 2.31×10−16) and WGS data (PT–PDX = 0.914; PDX–PDX = 0.931; P = 0.000299).

We compared the spectrum of copy number alteration values assumed by PT and PDX sample profiles. We observed that PT samples had a narrower CNA range than their derived PDXs, whatever platform used. However, this behavior was particularly apparent for CNA profiles based on WES and WGS (median ratios for PT/PDX and PDX/PDX, respectively=0.832 and 0.982 (P=0.000120) for SNP, 0.626 and 0.996 (P<2.2×10−16) for WES and 0.667 and 1.00 (P<2.2×10−16) for WGS). We expected that stromal DNA in PT samples dilutes the CNA signal. On the other hand, human stromal DNA is reduced in PDXs, because replaced by the murine counterpart. Moreover, WES and WGS platforms have higher uncertainties than SNP array data in estimating stromal DNA contributions. Hence, technical limitations, rather than biological variations between PT and PDXs, more reliably explain the slightly significant differences in the correlation values between PT-PDX and PDX-PDX matched pairs inferred on WES and WGS data.

Finally, we performed a similarity analysis between matched PT-PDX and PDX-PDX pairs using RNA-based data. We reported that expression-based CNA profiles could overestimate copy number changes during engraftment and passages (Fig. 4.10, 4.11). Notably, the similarity scores of matched pairs of PT-PDX and PDX-PDX tended to be systematically lower in gene-expression based CN profiles than in those estimated from DNA data in the SIBS hepatocellular carcinoma (HCC) dataset (Fig. 4.12).

4.3.1 PDX samples at late passages maintain CNA profiles similar to early passages

Next, we investigated whether any systematic evolution of copy number profiles occurred during PDX engraftment and passaging. We hypothesized that systematic mouse environment-driven evolution, if present, should reduce the correlations between the copy number profiles of matched PT and PDX or matched PDX pairs. Moreover, such reduction should be systematically observable at each subsequent passage in the mouse.

We observed no apparent effect during PDX passaging of CN profiles estimated on SNP, WES, and WGS platforms (Fig. 4.9d-f and 4.13).

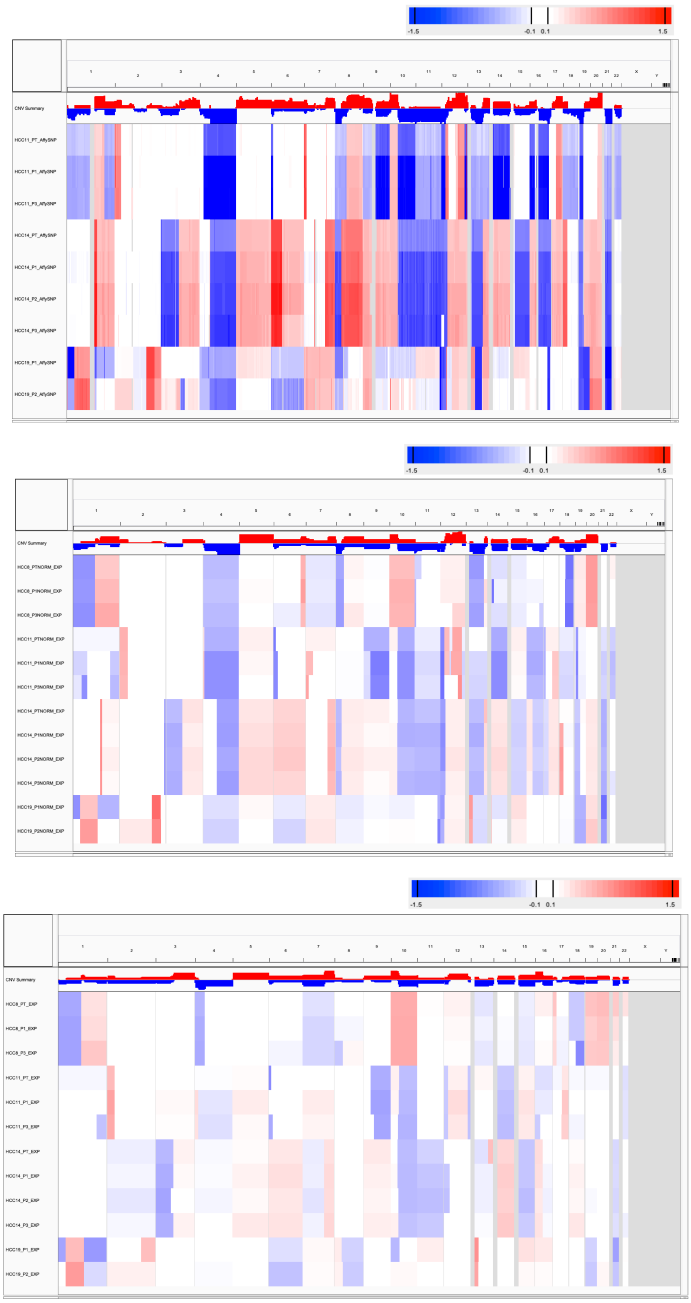

Figure 4.10: **CNA profiles of samples from SIBS hepatocellular carcinoma (HCC) dataset.** Copy number profiles are inferred on SNP array and gene expression array normalized by the median expression of normal liver tissue samples and of tumor samples of the same dataset.

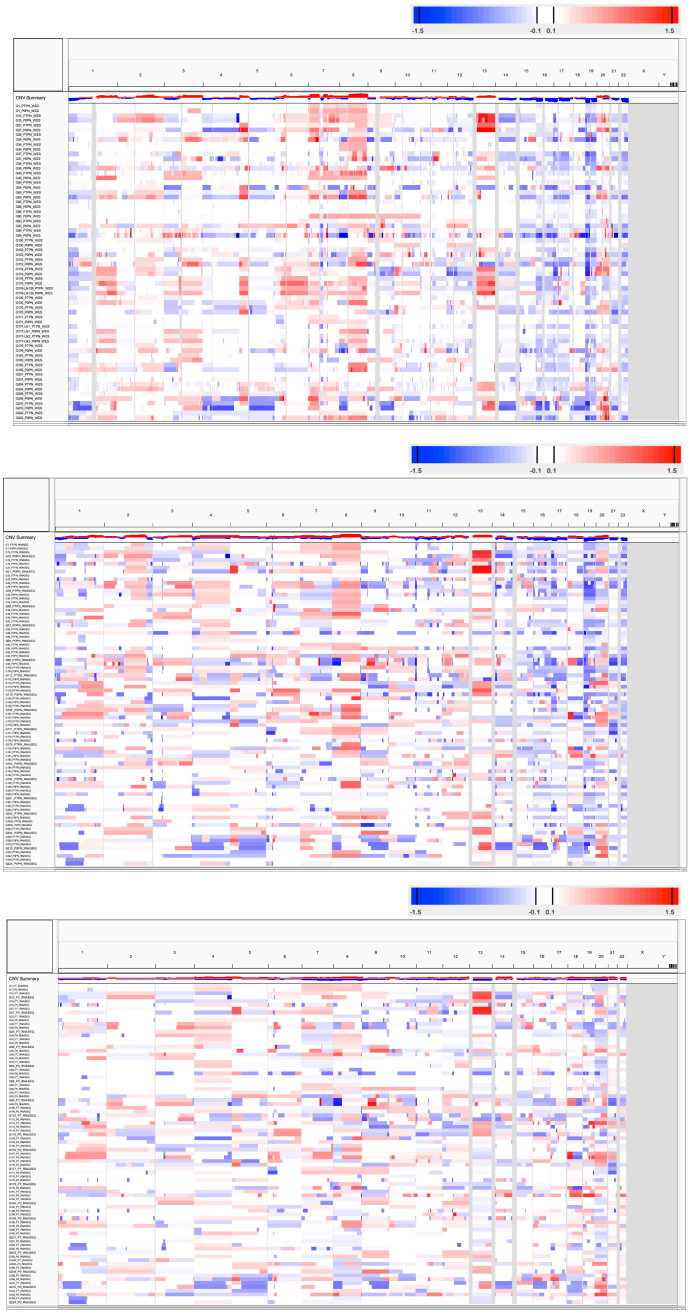

Figure 4.11: **CNA profiles of samples from SNU-JAX gastric cancer dataset.** Copy number profiles are inferred on WES and RNA-Seq normalized by the median expression of normal gastric tissue samples from the same patients and of tumor samples of the same dataset.

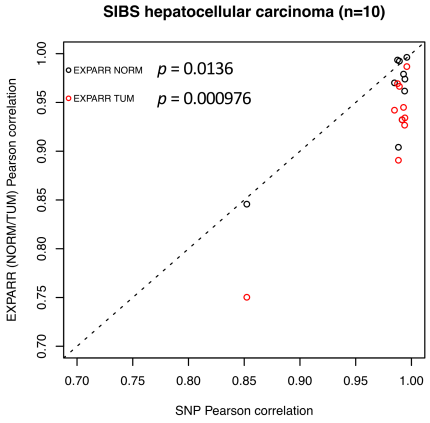

Figure 4.12: **Similarity of matched pairs of PT-PDX and PDX-PDX in gene-expression vs DNA based CNA profiles.** Scatter plots to compare the Pearson correlation coefficients of gene-based copy number using SNP array, gene expression array normalized using median expression of normal(RNASEQ/EXPARR NORM) and tumor (RNASEQ/EXPARR TUM) samples. P-values were computed by Wilcoxon signed-rank test.

The SNP data showed no significant difference between passages (Fig.4.9d). However, PDX models having very late passages exhibited a minor significant decrease in correlation compared to models with earlier passages (P < 8.98 x 10-5). Notably, this decrease in correlation indicated that some copy number changes could occur over long-term passaging (Fig. 4.13). Nonetheless, even at these late passages, the correlations to early passage PDXs remained high (median = 0.896). On the other hand, more variability in the correlation values could be observable for WES and WGS data (Fig. 4.9e-f and 4.13). However, the lack of a downward trend over passaging was also apparent in these sets of samples.

4.3.2 Lack of association between mutations in genome stability-related genes and PDX copy number stability

We explored whether the stability of copy number during engraftment and passaging is affected by mutations in genes known to impact genome stability. Specifically, as our larger size cohorts are given by CRC and BRCA derived

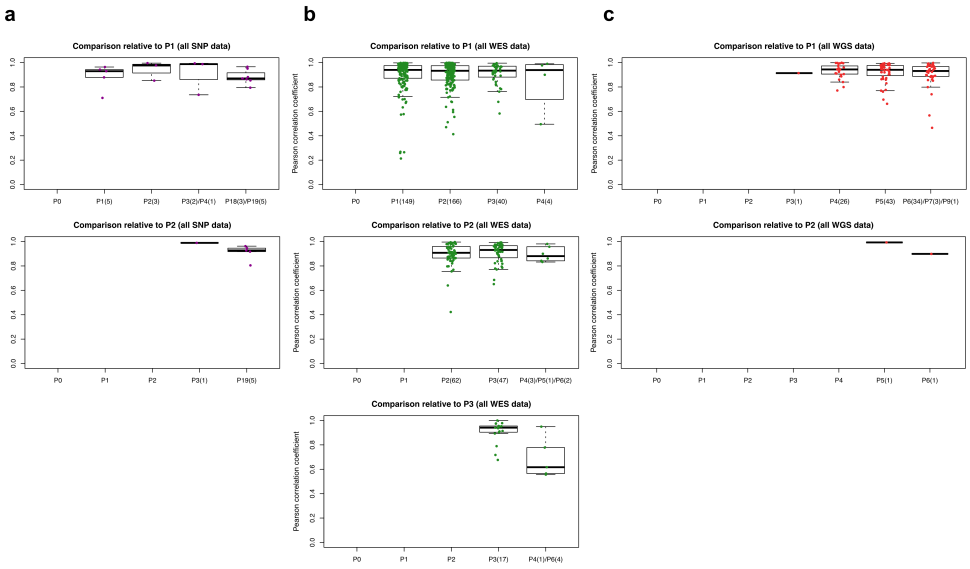

Figure 4.13: **Comparisons of CNAs relative to PDX samples at passages P1 or later.** Distribution of Pearson correlation coefficients of gene-based copy number, estimated by (a) SNP array, (b) WES, (c) WGS, between different combinations of patient tumor and PDX passages of the same model [145].

PDX models, we investigated whether the observed copy number instability is associated with the genetic mutations related to genetic instability and most frequently detected in CRC and BRCA tumor types, e. g. TP53 and BRCA.[135–138]

Hence, we compared the copy number correlations in models with wildtype versus mutated TP53 or BRCA, where available. However, we did not observe any consistent decrease in correlation associated with the mutational status of TP53 or BRCA (Fig. 4.14). Therefore, this indicated that mutations in such genes did not lead to copy number changes increased during PDX engraftment and passaging.

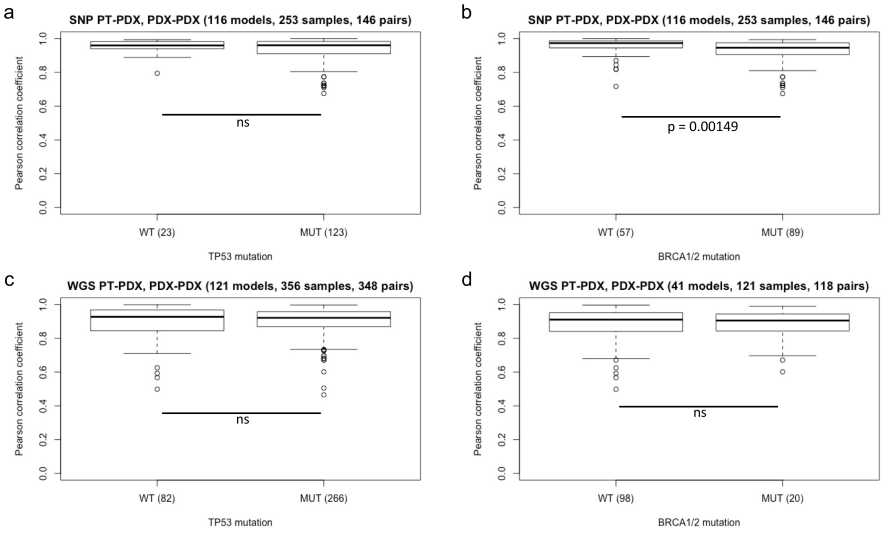

Figure 4.14: **Comparison of matched pairs of PT-PDX and PDX-PDX CNA profiles according to different mutational status.** Distribution of Pearson correlation coefficients of gene-based copy number for mutational status (WT: wildtype, MUT: mutant) of TP53 or BRCA1/BRCA2 of the samples or models for each correlation pair for, (**c**) and (**d**) SNP array and (**e**) and (**f**) WGS.

4.4 Spatial heterogeneity: a relevant source of genetic evolution in PDX models

We next compared the similarity between engrafted PDXs of the same model with the same passage number. Specifically, we defined PDX samples with the *same lineage* as those differing only by consecutive serial passages. For JAX SNP array and PDMR WES datasets, we defined samples with *different lineages* as those obtained by dividing and propagating a tumor into multiple mice (Fig. 4.15a). Instead, for the EurOPDX CRC and BRCA WGS datasets, we defined PDX samples with *different lineages* as they originated from distinct patient tumor fragments. Interestingly, we observed a lower correlation between PDX samples from different lineages compared to within a lineage (Fig. 4.15a-b, P = 0.0233 for SNP, P = 0.00119 for WES, P = 0.000232 for WGS), despite a majority of these pairwise comparisons exhibiting high correlation (>0.9).

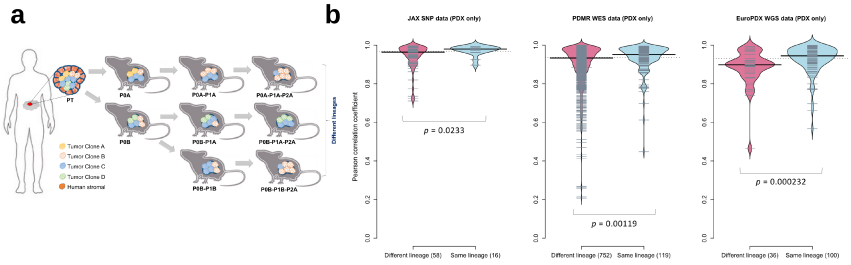

Figure 4.15: **Genetic divergence of PDX samples from different lineages compared to within a lineage.** **a**, Scheme of lineage splitting during passaging and expansion of tumors into multiple mice. **b**, Pearson correlation distributions for PDX sample pairs of different lineages and sample pairs within the same lineage, for (from left to right): JAX SNP array, PDMR WES and EuroPDX WGS datasets [145].

This indication suggested that lineage-splitting could be responsible for deviations in CNAs between samples. Therefore, copy number evolution during passaging mainly arises from evolved spatial heterogeneity.[139]

4.4.1 CNA evolution across PDXs is comparable to variation in multi-region samples

We then compared the CNA evolution occurring in PDXs with the levels of copy number variations in the multi-region samplings of non-small-cell lung cancer from the TRACERx Consortium.[140] Therefore, we performed analogous CNA correlation analyses between multi-region pairs. Interestingly, differences in correlation ($P > 0.05$) between multi-region patient and lung cancer PT-PDX pairs were not significant. The PDX-PDX pairs showed significantly better correlation than the multi-region pairs ($P < 0.05$, Fig. 4.16), across all lung cancer subtypes. Moreover, the correlations among intra-patient samples were lower median compared with those associated with CNA evolution during engraftment (PT-PDX) (Fig. 4.17). Nonetheless, the difference across variations in patient tumors versus PDX evolution was not statistically significant (Fig. 4.17), suggesting that CNA evolution across PDXs is no greater than variation in patient multiregion samples.

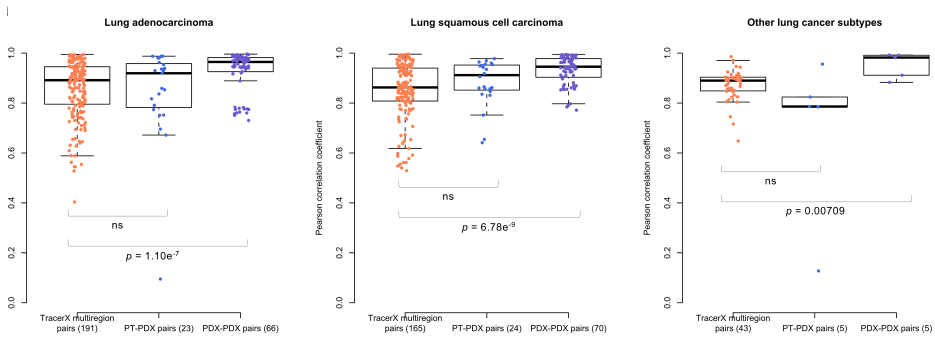

Figure 4.16: **Similarity of multi-region patient and lung cancer PT-PDX pairs.** Distributions of Pearson correlation coefficients of gene-based copy number for lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and other lung cancer subtypes, comparing different datasets. From left to right on the x-axis, these include: multiregion tumor samples of the same patient from TRACERx (n = 92 PTs; n = 295 multiregion samples); PT–PDX samples of the same model; and PDX–PDX samples of the same model [145]. P-values were computed by two-sided Wilcoxon rank-sum test ($P > 0.05$).

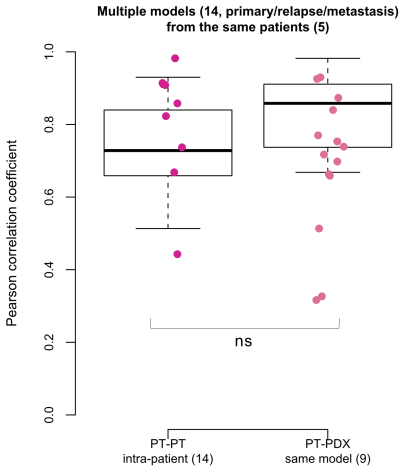

Figure 4.17: **Genetic variations in patient primary, relapse and metastasis samples vs PT-PDX pairs.** Distributions of Pearson correlation coefficients of gene-based copy number between intra-patient PT pairs (n = 14; primary, relapse or metastasis) from the same patient (n = 5) and corresponding PT–PDX pairs (derived from the same model; a different PT sample from the same patient generates a different model) for the same set of patients. P-values were computed by two-sided Wilcoxon rank-sum test (P $>$ 0.05). For all box and violin plots, the numbers of pairwise comparisons are indicated in the x-axis labels. In all box plots the center line represents the median, the box limits are the upper and lower quantiles, the whiskers extend to 1.5$\times$ the interquartile range and the dots represent all data points.

4.5 Absence of mouse-specific evolution in PDX models

Finally, we investigated whether recurrent CNA changes occur in PDXs in a tumor-type specific fashion. To this aim, we analyzed further the WGS-based CNA profiles of large metastatic colorectal (CRC) and breast cancer (BRCA) series. These datasets were respectively composed of 87 and 43 matched triplets of the patient tumor (PT), the PDX at early passage (PDX-early), and the PDX at later passage (PDX-late).

We carried out GISTIC analysis to identify recurrent CNAs by evaluating the frequency and amplitude of observed events.[141] In detail, GISTIC was applied separately for each PT, PDX-early (P0-P1 for CRC, P0-P2 for BRCA), and PDX-late (P2-P7 for CRC, P3-P9 for BRCA) cohorts of CRC and BRCA. As expected, CRCs and BRCAs generated different patterns of significant CNAs, with each similar to the GISTIC patterns in their respective TCGA series (Fig. 4.18). However, within each tumor type, the GISTIC profiles of the PT, PDX-early, and PDX-late cohorts were virtually indistinguishable (Fig. 4.19, 4.18), demonstrating no gross genomic alteration systematically acquired or lost in PDXs.

4.5.1 Absence of CNA shifts in 130 WGS patient tumor, early passage PDX and late passage PDX triplets

To perform an high-resolution analysis of recurrent genomic alterations, we carried out a gene-level analysis. Therefore, we attributed the GISTIC score (G-score) of the respective segment to each gene (Supplementary Table 7).

In both the CRC and BRCA cohorts, gene-level G-scores of the PTs highly correlated with the respective PDX-early and PDX-late cohorts (Fig. 4.19b-c). Moreover, PT versus PDX correlations was comparable to PDX-early versus PDX-late correlations. To search for progressive shifts, we compared the change in G-score (ΔG): (i) from tumor to PDX-early and (ii) from PDX-early to PDX-late. Correlations in these two ΔG values, as shown in the bottom-right panels of Fig. 4.19b and c, was absent or even slightly negative. Moreover, not a single gene had both ΔG concordant and passing the respective GISTIC threshold for significance (see Supplementary Table

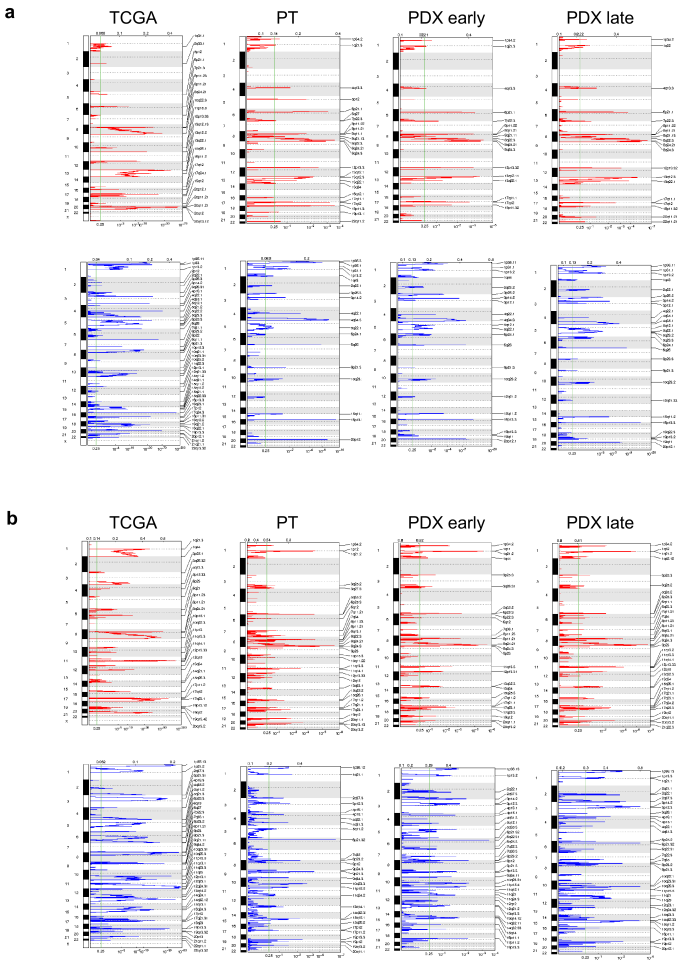

Figure 4.18: **GISTIC analysis of recurrent CNAs in TCGA primary tumors and EurOPDX collections of PTs and derived PDXs, at early and late passages, of (a) colorectal cancer and (b) breast cancer.** For each GISTIC plot the top axis reports the G-score and the bottom axis the q-value [145]. Red line plots: amplifications, blue line plots: deletions.

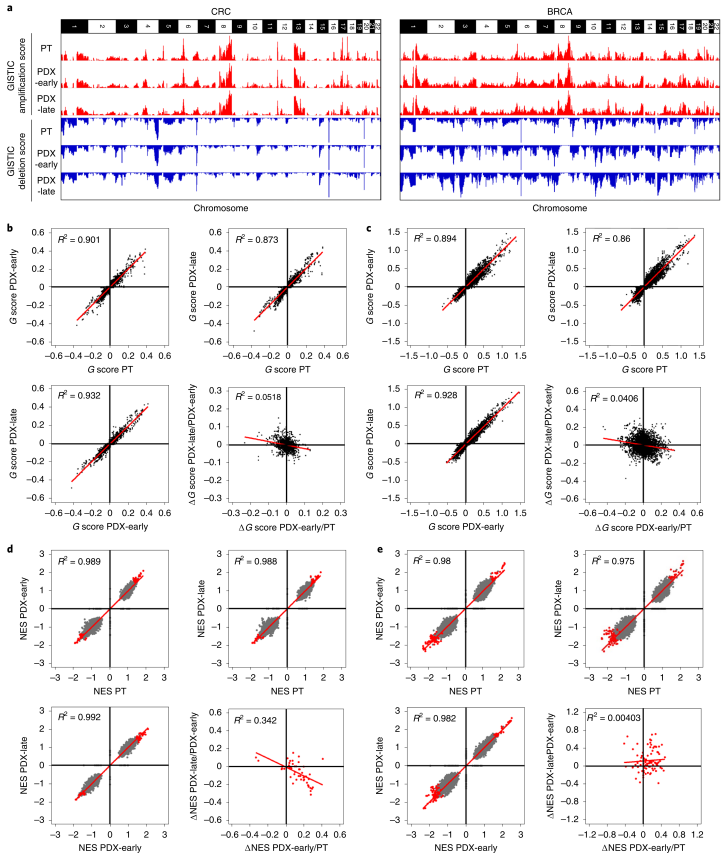

Figure 4.19: **Absence of mouse-driven recurrent CNAs during engraftment and propagation of CRC and BRCA PDXs. a**, Bar charts representing genome-wide G-scores for amplifications and deletions in each of the three cohorts of CRC (left; 87 triplets) and BRCA (right; 43 triplets): PT, PDX-early (P0–P1 for CRC; P0–P2 for BRCA) and PDX-late (P2–P7 for CRC; P3–P9 for BRCA). **b,c**, Scatter plots comparing gene-level G-scores between each of the three cohorts for CRC (**b**) and BRCA (**c**). The bottom-right panels of **b** and **c** show scatter plots comparing ΔG values from PT to PDX-early and from PDX-early to PDX-late. **d,e**, Scatter plots comparing GSEA NESs for gene sets between each of the three cohorts for CRC (**d**) and BRCA (**e**). The bottom-right panels of **d** and **e** show scatter plots comparing ΔNES from PT to PDX-early and from PDX-early to PDX-late. Gray data points represent all gene sets, whereas red data points represent gene sets significantly enriched in at least one of the three cohorts (that is, PT, PDX-early or PDX-late) [145].

8). Nonetheless, small segments of recurrent copy number gain or loss could be missed by this analysis due to the bin size imposed by the WGS coverage. However, overall, these results confirmed the absence of systematic CNA shifts in PDXs even under high resolution, gene-level analysis.

4.5.2 Lack of CNA-based functional shifts in triplets

We then considered the possibility of systematic copy number evolution at the pathway level in these triplets. In this regard, we performed Gene Set Enrichment Analysis (GSEA)[143] using G-scores to rank genes in each cohort. Multiple gene sets displayed significant enrichment in individual cohorts. Notably, these significant enrichments were consistent with the known recurrence of cancer CNAs at driver genes.

To avoid spurious apparent enrichment for sets of genes with adjacent chromosomal location, we implemented an additional filter based on G-score significance. Thus, after applying the Normalized Enrichment Score (NES), FDR q-value, and G-score filters, 49 gene sets were significant in at least one of the three CRC cohorts, and 89 gene sets in at least one of the three BRCA cohorts. Importantly, control gene sets composed of GISTIC hits identified in TCGA CRC and BRCA datasets were all significant. Therefore, our WGS cohorts properly recapitulated the major CNA features of these two cancer types. Moreover, differences associated with PDX engraftment and passage were negligible.

For both CRC and BRCA, the NES profiles for the ~8000 gene sets of PTs highly correlated with the respective PDX-early and PDX-late cohorts (Fig. 4.19d-e). Furthermore, PT versus PDX correlations was comparable to PDX-early versus PDX-late correlations. To search for progressive shifts, we calculated for each significant gene set ΔNES values between PT and PDX-early, as well as between early and late PDX. Similarly to what was observed for the ΔG-scores, as shown in the bottom-right panels of Fig. 4.19d and e, correlations were absent or at most slightly negative, confirming the absence of systematic CNA-based functional shifts in PDXs.

Chapter 5

Discussion

During the last decades, multiple model systems derived from patient tumors have been developed. Both *in vitro* and *in vivo* models have provided their contribution to study cancer biology and investigate drug responses, despite the intrinsic limitations affecting each of these systems.
Patient-derived xenografts (PDXs) are *in vivo* preclinical models generated by directly transplanting fragments of human tumors into immunodeficient mice. Therefore, the uniqueness of PDXs is the possibility of studying human cancer cells in a natural microenvironment, where they interact with the stromal components contributed by the murine host.[1,2] Nonetheless, the absence of the immune system components in PDXs hinders the possibility of employing these models for studying the roles of the immune system in tumor development and in immune-based therapy response.[66–68]

The importance of any cancer model relies on its ability to recapitulate the tumor of origin. Hence, regarding PDXs, they are considered robust model systems, provided that the murine microenvironment does not affect their biology and, potentially, their tumor evolution trajectories.
Indeed, cancer is an evolutionary process, in which, as expected, genomic alterations may emerge or disappear from patient tumors to PDXs. Thus, in this context, our goal is to discern whether potential genomics changes from patient tumors to PDXs are the result of a selective pressure imposed by the mouse host or a neutral tumor evolution as those occurring in patients. In presence of genetic drift and spatial heterogeneity, PDXs would mimic tumor evolution of patient tumors, confirming their suitability as models of

human cancers. Conversely, if the murine microenvironment would impose selective pressures, this would affect the entire reliability of the model system. Precisely, whether the mouse host would induce selective pressures affecting tumor evolution of PDXs, genetic changes would arise, systematically and progressively, from patient tumors to PDXs, during engraftment and propagation, impacting their reliability as human cancer models.

To this aim, we collected data which include the copy number profiles of matched patient tumor and PDXs, and we tracked the genomic stability of PDXs from the patient to long-term passages in the mice. In this respect, as tumor purity could affect the reliability of copy number events calls, we decided to avoid copy number calling. Thus, we assessed the similarity among the profiles of patient tumors and matched PDXs computing the pairwise correlation of their copy number ratio values.
Overall, this analysis showed strong concordance between matched PT-PDX and PDX-PDX pairs, and no apparent downward trend over tumor engraftment and passaging. We did observe larger deviations between PT–PDX than in PDX–PDX comparisons. However, this was probably due to the dilution of the PT signal by human stromal cells.
Some PDX models displayed CNA profile variations, but it was unclear whether such changes were the result of selective pressure imposed by the mouse host or of neutral tumor evolution and intratumor heterogeneity. Hence, to clarify the contribution of intratumor heterogeneity to the observed deviations in CNAs among pairs, we compared the similarity between PDXs derived from the same patient tumor sample versus those from distinct fragments of the same original tumor. Notably, we found that the splitting of tumors into fragments during PDX propagation was responsible for differences between PDX samples. Hence, spatial evolution within tumors seemed to produce variations among samples more than time or number of passage in the mouse. Moreover, we observed that the copy number shifts between PT and PDX were no more than the variations among multi-region tumor samplings. Importantly, this result corroborated the finding that spatial heterogeneity is a relevant source of genetic evolution in PDXs. Then, to investigate whether any selective pressures are imposed by the mouse microenvironment in individual tumor types, we focused on two large colorectal and breast cancer series, composed of 87 and 43 matched triplets of PT, PDX at early passage (PDX-early) and PDX at later passage (PDX-late), respectively. As a result, in this context, for both of these tumor types, genomic

data were assembled from matched PT, PDX-early, and PDX-late cohorts. Therefore, we performed an analysis of recurrent CNA events by GISTIC, separately for each cohort. Specifically, we assumed that if mouse-specific selective pressure was occurring, recurrent changes in the CNA profile would emerge in the PDX early cohort compared to the PT cohort and further increase in the PDX late cohort. However, we found that GISTIC CNA profiles of the PT, PDX-early, and PDX-late cohorts were virtually indistinguishable, with minor not functionally related changes only. Moreover, as the GISTIC profiles of our cohorts recapitulated at large those generated by the TCGA for colorectal and breast cancer, these results were not affected by the lack of representativeness of CNA lesions per tumor type.

Thus, on a hand, our analyses, excluding a systematic mouse-driven genetic selection during PDX engraftment and propagation, reinforced the finding that PDXs are prominent preclinical models, as they recapitulated the genomic landscapes of their original human tumors. On the other hand, our study challenged recent literature highlighting the possibility that PDXs undergo a mouse-induced copy number evolution.[83] In detail, the disagreement between our work and previous reports of copy number divergence in PDXs[83,85] strongly depends on the hypothesis tested for verifying the absence of mouse-driven selection in PDXs and on the methods and data types used for defining copy number profiles and discordance among pair of samples. These studies just reported that the percentage of altered genome varies from PT to PDXs. However, shifts in copy number values are expected during the evolution of tumor tissues. Therefore, a priori, the copy number changes reported between patient tumors and PDXs could simply be the result of population bottleneck during PDX engraftment and/or neutral evolution of tumors. Moreover, these reports almost exclusively estimated the copy number shifts from GEP array data, which, as we have shown, have low resolution and robustness. Furthermore, they called copy number gains and losses events and then measured the discordance among PT and PDX samples. Nevertheless, copy number events calling is still challenging without very high-depth sequencing data. Therefore, our results are not in real contradiction with recent reports, questioning the genomic fidelity of PDXs.

Our work, which has been published on Nature Genetics at the beginning of this year in collaboration with the PDXNET Consortium & EurOPDX Consortium teams,[145] reinforces the finding the PDXs are robust models of the genetic evolution of human tumors. Indeed, our in-depth tracking of

CNAs throughout PDX engraftment and passaging confirmed that tumors engrafted and passaged in PDX models maintain a high degree of molecular fidelity to the original PTs, thus verifying their suitability for preclinical drug testing.

Nonetheless, despite the volume and comprehensiveness of our CNA dataset, we do not rule out that future studies, hopefully including thousands of PDX models composed of matched patient tumors and PDX samples, might elucidate that, in some subpopulations of cancer types, the murine microenvironment might impact the tumor evolution of PDXs.
Moreover, it is reasonable that large amounts of higher-depth sequencing data will be available in the future. As a consequence, it will be possible to investigate subclonal dynamics of cancer cell populations with high accuracy, which, in turn, will lead to better clarify the contribution of spatial heterogeneity in the genomic differences between patient tumors and PDXs.

Overall, we do not exclude the PDXs will evolve in individual trajectories over time as a consequence of unavoidable evolutionary bottleneck occurring during PDX establishment and of spontaneous tumor evolution. Thus, for therapeutic dosing studies, we recommend confirming the existence of expected molecular targets and obtaining sequence characterizations in the cohorts used for testing as close to the time of the treatment study as is practical.

Acknowledgements

I wish to express my sincere appreciation to my advisor, Prof. Enzo Medico, for having provided me the possibility to join his team at the Candiolo Cancer Institute and for having guided me during my Ph.D. project with motivation and extensive knowledge.
I have my deepest thanks to my present and past lab colleagues for the scientific and moral support and the light-hearted moments of the day shared: Federica, Roberta, Ivan, Consalvo, and Erica.
A particular thanks go to Claudio, who has placed his trust in me from the very beginning of my project and has supported me throughout my Ph.D. Moreover, I would like to offer my special thanks to the core facilities colleagues and the office mates for their encouragement, their precious help, and laughter.
Of course, I would like to express my great gratitude to our collaborators, Prof. Jeffrey Chuang, and his team. In particular, I want to thanks Xingyi Woo, who has carried on the part of this project relative to the US National Cancer Institute (NCI) PDX Development and Trial Centers Research Network (PDXNet) Consortium. Furthermore, I express my thankfulness to our collaborators at the KU Leuven: Prof. Diether Lambrechts, who has allowed me to join his lab; Elodie Modave, and Bram Boeckx, who have trained me on the copy number quantification of Shallow Whole-Genome Sequencing data; Francesca Lodi, who has welcomed me warmly in Leuven and at the Campus Gasthuisberg. Additionally, I want to thank Prof. Michele De Bortoli for the excellent management of my Ph.D. course and the Ph.D. facilitator Danilo Lombardi for his helpfulness in clarifying any doubts arisen during the Ph.D. path.

Last but not the least, I would like to thank my parents, Angelina and Bruno, and my brother, Luca, for their constant patience, tolerance, and encouragement. A thank from deep in my heart goes to my grandparents, who, thanks to their eyes still capable of shining and being surprised, have the superpower of feeling me up, and giving me the strength of dealing with whatever difficulties.
The last few times have been quite hard. However, life has put on my way an incredible number of people who have taught me how to deal with troubles and pull out my strength. Therefore, I would like to thank: my friends at the Candiolo Cancer Institute, everyday donor of laughs and deep thought; Valentina, who I have met relatively recently, but I have the impression to know forever; Maura, who, with her life experience, is my look at the future; my colleagues at Giulio Natta high school, who have rubbed off on me their enthusiasm for their job and have trusted me from the very beginning, encouraging my initiatives. Finally, I should be grateful to my students, who, with their curiosity and vitality, with their pertinent, inappropriate, and, sometimes, even embarrassing questions, have made me grow up a lot during this last year.

References

1. Isella, C. *et al.* Stromal contribution to the colorectal cancer transcriptome. *Nature Genetics* **47**, 312–319 (2015).

2. Invrea, F. *et al.* Patient-derived xenografts (PDXs) as model systems for human cancer. vol. 63 151–156 (2020).

3. (US), N. I. of H. & Study, B. S. C. Understanding Cancer. (2007).

4. WHO2018 - Cancer.

5. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. vol. 458 719–724 (2009).

6. Cooper, G. M. The Development and Causes of Cancer. (2000).

7. Grandér, D. & Grandér, G. How do mutated oncogenes and tumor suppressor genes cause cancer? vol. 15 20–26 (1998).

8. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).

9. Tysnes, B. B. & Bjerkvig, R. Cancer initiation and progression: Involvement of stem cells and the microenvironment. vol. 1775 283–297 (2007).

10. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469**, 356–361 (2011).

11. Diaz, L. A. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537–540 (2012).

12. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* **366**, 883–892 (2012).

13. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).

14. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. vol. 501 338–345 (2013).

15. Polyak, K. Tumor Heterogeneity Confounds and Illuminates: A case for Darwinian tumor evolution. vol. 20 344–346 (2014).

16. Sprouffske, K., Merlo, L. M. F., Gerrish, P. J., Maley, C. C. & Sniegowski, P. D. Cancer in light of experimental evolution. vol. 22 (2012).

17. Lipinski, K. A. *et al.* Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. vol. 2 49–63 (2016).

18. Williams, M. J., Sottoriva, A. & Graham, T. A. Measuring Clonal Evolution in Cancer with Genomics. vol. 20 309–329 (2019).

19. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nature Genetics* **41**, 393–395 (2009).

20. Liu, L., De, S. & Michor, F. ARTICLE DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. (2013) doi:10.1038/ncomms2502.

21. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).

22. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).

23. Vermeulen, L. *et al.* Defining stem cell dynamics in models of intestinal tumor initiation. *Science* **342**, 995–998 (2013).

24. Kozar, S. *et al.* Continuous clonal labeling reveals small numbers of functional stem cells in intestinal crypts and adenomas. *Cell stem cell* **13**, 626–633 (2013).

25. Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 4009–14 (2013).

26. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics* **46**, 225–233 (2014).

27. Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine* **21**, 751–759 (2015).

28. Gerrish, P. & Lenski, R. The fate of competing beneficial mutations in an asexual population. *Genetica* **102**, 127–144 (1998).

29. Gay, L., Baker, A. M. & Graham, T. A. Tumour Cell Heterogeneity. vol. 5 (2016).

30. Turajlic, S. & Swanton, C. Metastasis as an evolutionary process. vol. 352 169–175 (2016).

31. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–95 (2011).

32. Turajlic, S. *et al.* Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. *Cell* **173**, 581–594.e12 (2018).

33. McPherson, A. *et al.* Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics* **48**, 758–767 (2016).

34. Reeves, M. Q., Kandyba, E., Harris, S., Del Rosario, R. & Balmain, A. Multicolour lineage tracing reveals clonal dynamics of squamous carcinoma evolution from initiation to metastasis. *Nature Cell Biology* **20**, 699–709 (2018).

35. Liu, X. *et al.* Homophilic CD44 interactions mediate tumor cell aggregation and polyclonal metastasis in patient-derived breast cancer models. *Cancer Discovery* **9**, 96–113 (2019).

36. Walter, D. *et al.* Genetic heterogeneity of primary lesion and metastasis in small intestine neuroendocrine tumors. *Scientific Reports* **8**, 1–9 (2018).

37. Vignot, S., Besse, B., André, F., Spano, J. P. & Soria, J. C. Discrepancies between primary tumor and metastasis: A literature review on clinically established biomarkers. vol. 84 301–313 (2012).

38. Sprouffske, K. *et al.* Genetic heterogeneity and clonal evolution during

metastasis in breast cancer patient-derived tumor xenograft models. *Computational and Structural Biotechnology Journal* **18**, 323–331 (2020).

39. Reiter, J. G. *et al.* Evolutionary bottleneck than distant metastases. *Nature Genetics* **52**, (2020).

40. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**, 256–259 (2014).

41. Simmons, C. *et al.* Does confirmatory tumor biopsy alter the management of breast cancer patients with distant metastases? *Annals of Oncology* **20**, 1499–1504 (2009).

42. Thompson, A. M. *et al.* Prospective comparison of switches in biomarker status between primary and recurrent breast cancer: The Breast Recurrence In Tissues Study (BRITS). *Breast Cancer Research* **12**, (2010).

43. Curigliano, G. *et al.* Should liver metastases of breast cancer be biopsied to improve treatment choice? *Annals of Oncology* **22**, 2227–2233 (2011).

44. Chang, H. J. *et al.* Discordant human epidermal growth factor receptor 2 and hormone receptor status in primary and metastatic breast cancer and response to trastuzumab. *Japanese Journal of Clinical Oncology* **41**, 593–599 (2011).

45. Barbier, A. *et al.* Coexpression of biological key modulators in primary colorectal carcinomas and related metastatic sites: Implications for treatment with cetuximab. *Bulletin du Cancer* **97**, (2010).

46. Italiano, A. *et al.* Epidermal growth factor receptor (EGFR) status in primary colorectal tumors correlates with EGFR expression in related metastatic sites: Biological and clinical implications. *Annals of Oncology* **16**, 1503–1507 (2005).

47. Di Renzo, M. F. & Corso, S. Patient-derived cancer models. vol. 12 1–3 (2020).

48. Byrne, A. T. *et al.* Interrogating open issues in cancer precision medicine with patient-derived xenografts. *Nature Reviews Cancer* **17**, 254–268 (2017).

49. Wang, K. *et al.* Patient-derived xenotransplants can recapitulate the genetic driver landscape of acute leukemias. *Leukemia* **31**, 151–158 (2017).

50. Johnson, J. I. *et al.* Relationships between drug activity in NCI preclinical *in vitro* and *in vivo* models and early clinical trials. *British Journal of Cancer* **84**, 1424–1431 (2001).

51. Gao, H. *et al.* High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature Medicine* **21**, 1318–1325 (2015).

52. Bertotti, A. *et al.* The genomic landscape of response to EGFR blockade in colorectal cancer. *Nature* **526**, 263–267 (2015).

53. Corso, S. *et al.* A comprehensive PDX gastric cancer collection captures cancer cell–intrinsic transcriptional MSI traits. *Cancer Research* **79**, 5884–5896 (2019).

54. Savaikar, M. A. *et al.* Preclinical PERCIST and 25% of SUVmax Threshold: Precision Imaging of Response to Therapy in Co-clinical 18F-FDG PET Imaging of Triple-Negative Breast Cancer Patient-Derived Tumor Xenografts. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* **61**, 842–849 (2020).

55. Inoue, A. *et al.* Current and future horizons of patient-derived xenograft models in colorectal cancer translational research. vol. 11 (2019).

56. Pauli, C. *et al.* Personalized *in vitro* and *in vivo* cancer models to guide precision medicine. *Cancer Discovery* **7**, 462–477 (2017).

57. Lazzari, L. *et al.* Patient-derived xenografts and matched cell lines identify pharmacogenomic vulnerabilities in colorectal cancer. *Clinical Cancer Research* **25**, 6243–6259 (2019).

58. Yan, H. H. N. *et al.* A Comprehensive Human Gastric Cancer Organoid Biobank Captures Tumor Subtype Heterogeneity and Enables Therapeutic Screening. *Cell Stem Cell* **23**, 882–897.e11 (2018).

59. Shi, J., Li, Y., Jia, R. & Fan, X. The fidelity of cancer cells in PDX models: Characteristics, mechanism and clinical significance. vol. 146 2078–2088 (2020).

60. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. vol. 144 646–674 (2011).

61. Mantovani, A., Allavena, P., Sica, A. & Balkwill, F. Cancer-related

inflammation. vol. 454 436–444 (2008).

62. Dougan, M. & Dranoff, G. Immune therapy for cancer. vol. 27 83–117 (2009).

63. Zitvogel, L., Apetoh, L., Ghiringhelli, F. & Kroemer, G. Immunological aspects of cancer chemotherapy. vol. 8 59–73 (2008).

64. Lin, W. W. & Karin, M. A cytokine-mediated link between innate immunity, inflammation, and cancer. vol. 117 1175–1183 (2007).

65. Smyth, M. J., Dunn, G. P. & Schreiber, R. D. Cancer Immunosurveillance and Immunoediting: The Roles of Immunity in Suppressing Tumor Development and Shaping Tumor Immunogenicity. vol. 90 1–50 (2006).

66. Bankert, R. B., Egilmez, N. K. & Hess, S. D. Human-SCID mouse chimeric models for the evaluation of anti-cancer therapies. vol. 22 386–393 (2001).

67. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: Integrating immunity's roles in cancer suppression and promotion. vol. 331 1565–1570 (2011).

68. Hylander, B. L. *et al.* Origin of the vasculature supporting growth of primary patient tumor xenografts. *Journal of Translational Medicine* **11**, 110 (2013).

69. Wegner, C. S. *et al.* Increasing aggressiveness of patient-derived xenograft models of cervix carcinoma during serial transplantation. *Oncotarget* **9**, 21036–21051 (2018).

70. Coleman, O. *et al.* A comparative quantitative LC-MS/MS profiling analysis of human pancreatic adenocarcinoma, adjacent-normal tissue, and patient-derived tumour xenografts. *Proteomes* **6**, 45 (2018).

71. Kersten, K., Visser, K. E., Miltenburg, M. H. & Jonkers, J. Genetically engineered mouse models in oncology research and cancer medicine. *EMBO Molecular Medicine* **9**, 137–153 (2017).

72. Frese, K. K. & Tuveson, D. A. Maximizing mouse cancer models. vol. 7 645–658 (2007).

73. Walrath, J. C., Hawes, J. J., Van Dyke, T. & Reilly, K. M. Genetically Engineered Mouse Models in Cancer Research. in *Advances in cancer*

research vol. 106 113–164 (Academic Press Inc., 2010).

74. Becher, O. J. & Holland, E. C. Genetically engineered models have advantages over xenografts for preclinical studies. vol. 66 3355–3358 (2006).

75. Kemper, K. *et al.* Intra- and inter-tumor heterogeneity in a vemurafenib-resistant melanoma patient and derived xenografts. *EMBO Molecular Medicine* **7**, 1104–1118 (2015).

76. Kemper, K. *et al.* BRAFV600E Kinase Domain Duplication Identified in Therapy-Refractory Melanoma Patient-Derived Xenografts. *Cell Reports* **16**, 263–277 (2016).

77. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2015).

78. Bruna, A. *et al.* A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds. *Cell* **167**, 260–274.e22 (2016).

79. Marangoni, E. *et al.* A new model of patient tumor-derived breast cancer xenografts for preclinical assays. *Clinical Cancer Research* **13**, 3989–3998 (2007).

80. Li, S. *et al.* Endocrine-Therapy-Resistant ESR1 Variants Revealed by Genomic Characterization of Breast-Cancer-Derived Xenografts. *Cell Reports* **4**, 1116–1130 (2013).

81. Cassidy, J. W., Caldas, C. & Bruna, A. Maintaining tumor heterogeneity in patient-derived tumor xenografts. vol. 75 2963–2968 (2015).

82. Kim, T. M. *et al.* Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. *Clinical Cancer Research* **21**, 4461–4472 (2015).

83. Ben-David, U. *et al.* Patient-derived xenografts undergo mouse-specific tumor evolution. *Nature genetics* **49**, 1567–1575 (2017).

84. Ben-David, U., Beroukhim, R. & Golub, T. R. Genomic evolution of cancer models: perils and opportunities. (2019) doi:10.1038/s41568-018-0095-3.

85. Hoge, A. C. H. *et al.* DNA-based copy number analysis confirms genomic evolution of PDX models. *bioRxiv* 2021.01.15.426865 (2021) doi:10.1101/2021.01.15.426865.

86. Ben-David, U., Mayshar, Y. & Benvenisty, N. Virtual karyotyping of pluripotent stem cells on the basis of their global gene expression profiles. *Nature protocols* **8**, 989–997 (2013).

87. Bertotti, A. *et al.* A molecularly annotated platform of patient- derived xenografts ("xenopatients") identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer. *Cancer Discovery* **1**, 508–523 (2011).

88. Galimi, F. *et al.* Genetic and expression analysis of MET, MACC1, and HGF in metastatic colorectal cancer: Response to Met inhibition in patient xenografts and pathologic correlations. *Clinical Cancer Research* **17**, 3146–3156 (2011).

89. Hu, B. *et al.* KPNA3 confers sorafenib resistance to advanced hepatocellular carcinoma via TWIST regulated epithelial-mesenchymal transition. *Journal of Cancer* **10**, 3914–3925 (2019).

90. Roth, R. B. *et al.* Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics* **7**, 67–80 (2006).

91. Huang, Y. *et al.* Identification of a two-layer regulatory network of proliferation-related microRNAs in hepatoma cells. *Nucleic Acids Research* **40**, 10478–10493 (2012).

92. Malouf, G. G. *et al.* Transcriptional profiling of pure fibrolamellar hepatocellular carcinoma reveals an endocrine signature. *Hepatology* **59**, 2228–2237 (2014).

93. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research* **17**, 1665–1674 (2007).

94. Belmont, J. W. *et al.* The international HapMap project. *Nature* **426**, 789–796 (2003).

95. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 16910–16915 (2010).

96. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

97. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

98. Scheinin, I. *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Research* **24**, 2022–2032 (2014).

99. Desmedt, C. *et al.* Uncovering the genomic heterogeneity of multifocal breast cancer. *Journal of Pathology* **236**, 457–466 (2015).

100. Kluin, R. J. C. C. *et al.* XenofilteR: Computational deconvolution of mouse and human reads in tumor xenograft sequence data. vol. 19 (2018).

101. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

102. Conway, T. *et al.* Xenome-a tool for classifying reads from xenograft samples. *Bioinformatics* **28**, 172–178 (2012).

103. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).

104. Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–501 (2011).

105. Favero, F. *et al.* Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology* **26**, 64–70 (2015).

106. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, (2011).

107. Weissbein, U., Schachter, M., Egli, D. & Benvenisty, N. Analysis of chromosomal aberrations and recombination by allelic bias in RNA-Seq. *Nature communications* **7**, 12144 (2016).

108. Ben-David, U. *et al.* The landscape of chromosomal aberrations in breast cancer mouse models reveals driver-specific routes to tumorigenesis. *Nature communications* **7**, 12160 (2016).

109. Lingjærde, O. C., Baumbusch, L. O., Liestøl, K., Glad, I. K. & Børresen-Dale, A. L. CGH-Explorer: A program for analysis of array-CGH data. vol. 21 821–822 (2005).

110. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).

111. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178–192 (2013).

112. Skidmore, Z. L. *et al.* GenVisR: Genomic Visualizations in R. *Bioinformatics* **32**, 3012–3014 (2016).

113. Coussy, F. *et al.* A large collection of integrated genomically characterized patient-derived xenografts highlighting the heterogeneity of triple-negative breast cancer. *International Journal of Cancer* **145**, ijc.32266 (2019).

114. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337.e10 (2018).

115. Patterson, S. E., Statz, C. M., Yin, T. & Mockus, S. M. Utility of the JAX Clinical Knowledgebase in capture and assessment of complex genomic cancer data. *npj Precision Oncology* **3**, (2019).

116. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. vol. 18 696–705 (2018).

117. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

118. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

119. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature* **511**, 543–550 (2014).

120. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).

121. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human

cancers. *Genome Biology* **12**, R41 (2011).

122. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–15550 (2005).

123. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

124. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems* **1**, 417–425 (2015).

125. Derose, Y. S. *et al.* Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nature Medicine* **17**, 1514–1520 (2011).

126. He, S. *et al.* PDXliver: A database of liver cancer patient derived xenograft mouse models. *BMC Cancer* **18**, 550 (2018).

127. Zare, F., Hosny, A. & Nabavi, S. Noise cancellation using total variation for copy number variation detection. *BMC Bioinformatics* **19**, 361 (2018).

128. Wineinger, N. E. & Tiwari, H. K. The Impact of Errors in Copy Number Variation Detection Algorithms on Association Results. *PLoS ONE* **7**, e32396 (2012).

129. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature Genetics* **45**, 1134–1140 (2013).

130. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676–689.e3 (2018).

131. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Computational Biology* **12**, 1–18 (2016).

132. Bass, A. J. *et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).

133. Schumacher, S. E. *et al.* Somatic copy number alterations in gastric adenocarcinomas among Asian and Western patients. *PLOS ONE* **12**, e0176045 (2017).

134. Ally, A. *et al.* Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327–1341.e23 (2017).

135. Iacopetta, B. TP53 mutation in colorectal cancer. vol. 21 271–276 (2003).

136. Li, X. L., Zhou, J., Chen, Z. R. & Chng, W. J. P53 mutations in colorectal cancer- Molecular pathogenesis and pharmacological reactivation. *World Journal of Gastroenterology* **21**, 84–93 (2015).

137. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: Different roles in a common pathway of genome protection. vol. 12 68–78 (2012).

138. Rebbeck, T. R. *et al.* Association of type and location of BRCA1 and BRCA2 mutations with risk of breast and ovarian cancer. *JAMA - Journal of the American Medical Association* **313**, 1347–1361 (2015).

139. Kim, H. *et al.* High-resolution deconstruction of evolution induced by chemotherapy treatments in breast cancer xenografts. *Scientific Reports* **8**, 1–16 (2018).

140. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine* **376**, 2109–2121 (2017).

141. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20007–20012 (2007).

142. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology* **12**, R41 (2011).

143. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–15550 (2005).

144. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267–273 (2003).

145. Woo, X. Y. *et al.* Conservation of copy number profiles during engraft-

ment and passaging of patient-derived cancer xenografts. *Nature Genetics* **53**, 86–99 (2021).