International PhD Program in Biomolecular Sciences

Centre for Integrative Biology

29th Cycle

Quantitative analyses to study

tumor clones dynamics and tumor heterogeneity

Tutor

Prof. Francesca DEMICHELIS

CIBIO, University of Trento

Co-tutor

Dr. Gerhardt ATTARD

The Institute of Cancer Research and Royal Marsden

Advisor

Dr. Alessandro ROMANEL

CIBIO, University of Trento

Ph.D. Thesis of

Nicola Andrea CASIRAGHI

CIBIO, University of Trento

Academic Year 2015-2016

INDEX

Introduction

Results – Section a

Methods – Section a

Results – Section b

Methods – Section b

ABSTRACT

Prostate cancer is a highly heterogeneous disease and its manifestations can vary from indolent localized tumor to widespread metastases. This heterogeneity is also observed at the molecular level both inter- and intra-patient. Intra-patient heterogeneity in the clinical setting of men with castration resistant prostate cancer (CRPC) might be informative in terms of treatment decision. Here I present analytical work on two approaches relevant to the characterization of intra-patient heterogeneity and applied to unpublished CRPC patients sequencing data. The first is based on the genome wide interrogation of multiple metastatic and primary tissue biopsies from single patients. I present genomic analyses to decipher the content of multiple tumor biopsies from CRPC patients and provide comparisons to highlight similarities and differences and to identify alternative patterns of aberrations. The second approach, alternative to tissue biopsies that might under-represent the genomic landscape of the patient's disease, relies on liquid biopsies, a minimally invasive test that is also amenable to serial sampling. Liquid biopsies contain circulating cell free DNA (cfDNA) released from widespread tumor cells, potentially uncovering the full tumor landscape. By using next generation sequencing on cfDNA obtained from plasma, I developed strategies aimed at systematically tracking the reiterative process of genetic diversification leading to disease evolution and to detect genomic aberrations. I specifically focused on an *ad hoc* computational procedure (ABEMUS) to detect somatic point mutations that could emerge under treatment pressure and as drug resistance mechanism. The work I present is relevant to the context of precision oncology that exploits detailed patient-specific molecular information to diagnose and follow cancer progression with the ultimate goal of promptly guiding treatment decisions to improve clinical outcome with transdisciplinary strategies. The analytical work I developed can be applied to the study of any tumor type.

INTRODUCTION

I first introduce tumor evolution models and the importance of tumor heterogeneity in the context of patient treatment and then focus on prostate cancer genomics and on the molecular pathways involved in its oncogenesis, progression and treatment resistance mechanisms. Advantages and drawbacks of both well-established and innovative clinical strategies to study tumor genomics through diverse next generation sequencing assays are also described.

1. Clonal evolution in cancer

In 1976 Peter Nowell published a landmark perspective on cancer as an evolutionary process that is driven by stepwise, somatic-cell mutations with sequential, subclonal selection [1]. Thus, tumor evolution is a Darwinian evolutionary system for the selection of the fittest heritable genetic variant. In this setting, the population of cancer cells, influenced by endogenous and exogenous mutational processes, provides the fuel for selection to act. Modern cancer biology together with the large amount of genomic data obtained from tissue sections, small biopsies and more recently single-cell analysis and genomics have validated cancer as complex, Darwinian, adaptive system.

Cancers evolve by an iterative process of clonal expansion, genetic diversification and clonal selection within the adaptive landscapes of tissue ecosystems [2]. Its evolution is conceptually similar to asexual microorganisms and should be governed by the dynamic interplay of the same three basic processes: a) The generation of heritable variation; b) The influence of random birth and death events on the fate of new genotypes, referred to as genetic drift; and c) Darwinian selection, which changes the frequency genotypes in the population based on their relative fitness advantage. The acquisition of heritable alterations (heritable somatic variation encompasses genetic alterations such as point mutations, insertions, deletions, and chromosomal aberrations, as well as epigenetic changes that are heritable over cell generations) and genetic drift are both random processes, while Darwinian selection is a deterministic process. During clonal selection, a new mutation that increases the ability of the cell to survive and reproduce under particular environmental conditions will gradually increase in its abundance within the population. Multiple intratumoral subclones harbouring different driver mutations, displaying distinct phenotypes, and evolving with branched phylogenies were identified in many cancer types. The presence of multiple subclones within a tumor can lead to clonal competition and the fitness of an individual subclone is then defined in relation to the fitness of other competing clones.

1.1 Gradualism and punctuated evolution

Gradual mutation accumulation occurs to some degree in all cancers, representing perpetual adaptation to the tumor environment but may be punctuated by highly disruptive episodes. Such a dichotomy has been framed in the context of micro- versus macro-evolution, with gradual accumulation of point mutations (micro-evolution) presented in opposition to "saltationist" view, which emphasizes the importance of large-scale chromosomal alterations and bursts of mutations (macro-evolution) [3]. Examples of these catastrophic events are mutational phenomena termed *kataegis*, a localized hypermutation that often colocalizes with somatic rearrangements and *chromotripsis*, a single event that causes genome shattering and reassembly, resulting in a characteristic pattern of oscillating copy number and up to several hundred genomic rearrangements localized to one or a few chromosomes. Additionally, WGS analysis of prostate tumors by Baca et *al.* [4] described large chains of rearrangements that coordinately affect multiple chromosomes in prostate cancer, a phenomenon defined as *chromoplexy*.

Thus, cancer genome evolution may not always be a gradual stepwise and the observation that tumors with an extreme level of chromosomal instability appear associated with improved prognosis [5], compared to intermediate levels, supports the hypothesis that there may be a delicate balance between too much and too little instability and that there may be potent selection pressures in cancer evolution for a "just-right" level of cell-to-cell variation.

1.2 Clonal heterogeneity

Tumoral clones evolve through the interaction of selectively advantageous "driver" lesions, selectively neutral "passenger" lesions and deleterious lesions. In addition, "mutator" lesions increase the rate of other genetic changes [6,7]. In this scenario, selective pressures allow some mutant subclones to expand while others become extinct or remain dormant. The mutational profile of a tumor represents a historical record of alterations that have accumulated during its evolutionary history. These data together with heterogeneity among cancer cells can be used to understand the temporal order of mutational events. Alterations identified in every sequenced cancer cell can be considered to form the trunk of a cancer's somatic evolutionary tree, while subclonal mutations, present in only a subset of cancer cells, make up the branches. Bioinformatics tools have been developed to help decipher the temporal order of mutations and determine which are clonal or subclonal [8]. Somatic mutations within cancer genes that confer a clonal advantage are causally implicated in oncogenesis and are positively selected during cancer evolution. Accumulating evidence suggests that certain driver alterations may be more likely to be subclonal than others [3,9–14] and other driver mutations exhibit a tendency to be clonal in certain cancer type, but not others [9,14,15]. Such differences may reflect the importance of epistasis in cancer evolution and are in agreement with findings that co-occurrence and mutual exclusivity relationships between cancer driver alterations can vary extensively in different cancer types [16].

2. Tumor heterogeneity

As a result of evolutionary forces of variation and selection, extensive genetic and phenotypic variations exist not only between tumours (inter-tumor heterogeneity) but also within individual tumours (intra-tumor heterogeneity).

Tumours that originate from different tissues and cell types vary in terms of their genomic landscapes, prognosis and their response to treatments. Mutational frequencies of oncogenes and tumor suppressors vary between tumours of different tissues, probably reflecting the importance of distinct tissue dependent signalling pathways. Additionally, recent advances in Next Generation Sequencing (NGS) have revealed that very few mutations were observed in more than 5-10% of tumours of a particular tissue type [17]. Nevertheless, studies described that histone-modifying genes are recurrently mutated in a range of tumors [17–19] and genomic instability, occurring through various distinct routes [20–22], is a unifying feature of many genetically diverse malignancies.

Within tumors, genetically distinct subclonal populations of cells arise through inter-cellular genetic variation, followed by selective outgrowth of clones that have a phenotypic advantage within a given tumor environmental context [1,23]. If a new clone takes over the entire population by replacing ancestral ones, this will result in a homogenous cell population. Otherwise, if during linear evolution a new clone fails to outcompete its predecessors, a degree of heterogeneity will be observed [24] and if distinct subclones evolve in parallel (branched tumor evolution) this will result in extensive subclonal diversity [2].

Analysis of large cancer databases support evidences of the genetic heterogeneity between cancers and even within individual cancer types. For instance, when the Cancer Genome Atlas (TCGA) Project analysed 489 high-grade serous ovarian cancers [25] only 10 among the thousands of identified somatic mutations identified were recurrently mutated cancer genes, and all but TP53 mutations were present in less than 10% of cases. The genomic analysis of 77 estrogen-receptor-positive breast cancers also identified that most recurrent mutations occur infrequently, but they do cluster within a limited number of cellular pathways that are central to tumour biology [26]. In addition to the heterogeneity of cancer genes, there is considerable diversity in the nature, number and distribution of mutations within and across different cancer histologies [27]. These studies have revealed that the degree of intra-tumor heterogeneity can be highly variable, with between zero and thousands coding mutations found to be heterogeneous within primary tumors or between primary and metastatic or recurrence sites [28]. Genomic copy number heterogeneity can also be extensive within tumors. Large scale chromosomal alterations may have profound impact upon the genome, disrupting hundreds of genes, and can be considered macro-evolutionary events, which may contribute to tumor progression [15,29,30].

Sequential analysis of tumors has also revealed evidence that intra-tumor heterogeneity temporally evolves during the disease course and can have important implications for predictive or prognostic biomarker strategies.

2.1 Clinical implications of intra-tumor heterogeneity

Recognition of tumor heterogeneity led to the concept of personalized cancer medicine: deciphering individual cancer genomic profiles should provide precise insights into disease biology and allow the targeting of genetically encoded susceptibilities for therapeutic benefit. At the same time, observation of intra-tumor heterogeneity pose a challenge to targeted therapies and raise important questions regarding future drug-development strategies. Indeed, intra-tumor genetic heterogeneity results in phenotypic diversity affecting clinically relevant parameters such as gene expression signatures that reflect prognosis and response to therapeutic agents. It is also important to note that phenotypic heterogeneity is not only mediated through genetic diversity; genetically homogenous subclones can behave in functionally distinct ways after exposure to chemotherapy [31]. Therapeutic intervention may destroy subclones and alter their favourable microenvironment, but it can also provide a potent selective pressure for the expansion of resistant variants. Increasing evidence suggests that efforts to forecast outcome of an individual cancer require the identification of low-frequency genetically and functionally distinct subclones at diagnosis [32]. Indeed, tumor deep-sequencing analyses attempt to stratify therapeutics based on identification of “actionable mutations” where a clinician matches a tumor aberration to a cancer drug.

3. Primary and metastatic tumors

Tumor metastasis is frequently cited to be responsible of about 90% of all cancer-related deaths [33]. The process has been linked to a speciation event with macro-evolutionary leaps required to endow a tumor cell with metastatic potential [34]. Next-generation DNA sequencing has made it apparent that most primary tumours do not consist of a single population of genetically identical cells. Instead they are a collection of subpopulations of genetically identical cells that can be distinguished from other subclones by the mutations they harbour. The evolutionary paths from primary tumors to metastasis that are taken by tumour cells are many and represent a challenging research issue. It is often assumed that one disseminated tumour cell initiates metastatic outgrowth (monoclonal seeding) and there is debate about whether metastases derive from multiple branched spreading events involving disseminated cells from the primary tumour as well as metastases (polyclonal seeding). Many studies have shown that primary prostate cancers are multifocal and separate foci are not only spatially and pathologically distinct, but are composed of multiple distinct cancer cell clones [35]. An open question is if and how these independent multiclonal tumor foci that are present in the primary tumor give rise to multiclonal or monoclonal metastases [36]. The monoclonal origin model indicates that all metastatic lesions are derived from a common cancer cell ancestor traceable back to one distinct focus that is present in the primary prostate tumor [37,38]. In the polyclonal origin model, multiple genomically distinct foci in the primary tumor, without sharing a common cancer cell ancestor, can independently progress and metastases can harbour multiple distinct clonal aberrations originating from the primary tumor [37,39]. Additionally, in both models the acquisition of subsequent mutations can also occur during disease progression and/or metastasis-to-metastasis cross-seeding (in which subclones within a metastasis originated from another metastatic site, rather than from the primary tumour) [37,39], leading to substantial genomic diversity.

4. Prostate cancer

Prostate cancer is a significant public health burden and a major cause of morbidity and mortality among men. Worldwide, prostate cancer is the second most frequently diagnosed cancer and the fifth leading cause of cancer death among men. Its greater prevalence in the west and migration population implicates lifestyle and environmental as risk factors [40]. Established risk factors for prostate cancer are limited to advancing age, ethnicity, a family history of this malignancy (the risk for first-degree relatives of men with prostate cancer is about twice that for men in the general population [41]) and certain genetic polymorphisms [42].

Genetic predisposition can result from penetrant mutations, from genetic variants associated with risk, or from a combination of these two. Most men are diagnosed at an early stage and are either followed with active surveillance or treated with curative intent where the choice of therapy depends upon clinical features such as tumor stage, Gleason grade and serum prostate specific antigen (PSA) levels. Despite best attempts at risk stratification the overtreatment of patients with indolent disease and the potential undertreatment of patients with aggressive disease remain a concern.

Prostate cancer is commonly multifocal often harbouring pathologically and genomically distinct foci [35]. Comparison of the genomic landscape in both inter-related and geographically distinct regions within prostates has revealed independent tumour origins in different studies [35,43,44]. Additionally, whole genome sequencing (WGS) data of multiple metastatic sites from 10 tumours has revealed a common clonal origin containing 40-90% of total mutations and the majority of driver mutations suggesting that metastases originate from only one tumor foci [37]. In another recent study, WGS data of multiple tumor foci in patients with clinically localized prostate cancer indicates quiet point mutation profiles but extensive structural heterogeneity between foci. Overall, the observation of very few copy number alterations shared between tumour foci supports the independent origin of distinct foci [45].

4.1 Androgen dependent prostate cancer

The androgen receptor (*AR*) signalling axis plays a critical role in the development, function and homeostasis of the prostate. The classical action of *AR* is to regulate gene transcriptional processes via AR nuclear translocation, binding to androgen response elements on target genes and recruitment of, or crosstalk with, transcription factors. Prostate cancer initiation and progression is also uniquely dependent on *AR* [46].

Next generation sequencing has allowed characterisation of the clonal hierarchy of genomic lesions in prostate tumours, providing information about carcinogenesis and identification of genomic rearrangements that result in androgen-driven *ETS* gene expression. These rearrangements are clonal, suggesting that they occur early and might result from activated androgen receptors generating DNA damage through transcription at *AR* binding sites [47]. However, *ETS* gene fusions alone are not sufficient to result in cancer and other genomic events, such as activation of the PI3K/AKT by *PTEN* loss, are needed [48,49]. Several studies exploiting whole exome sequencing (WES) data indicate that prostate cancer genome is characterised by relatively few focal chromosomal gains or losses and overall low mutation rate (roughly one per megabase) [4,50–52]. *SPOP*, *TP53* and *PTEN* are among the most frequently mutated genes across several studies of localised prostate cancers [50,53]. About 50-60% of PSA-screened prostate cancers have recurrent gene fusions, typically fusing the 5' untranslated region of an androgen-regulated gene (i.e. *TMPRSS2*) to nearly the entire coding sequence of an ETS transcription factor family member (i.e. *ERG*) [54]. Genomic, epigenetic, and expression profiling studies support the premise that tumors with ETS fusions (ETS-positive) are distinct from those without (ETS-negative); driving changes in several genes have been identified that occur exclusively in ETS-negative prostate cancers [51,55]. Mutations in *SPOP,* which cluster in the encoded protein's substrate binding cleft, occur in about 5-10% of prostate cancers, and *SPOP* mutated cancers are exclusively ETS-negative [47,50–52]. Loss or mutation of the tumor-suppressor genes *PTEN* and *TP53* are among the most frequent events in prostate cancer, and occur in both ETS-positive and ETS-negative cancers.

4.1.1 The androgen receptor pathway

The human *AR* gene is a nuclear transcription factor and a member of the steroid hormone receptor superfamily of genes. It is located on the X chromosome (q11-12) and consists of 8 exons. It codes for a protein of 919 amino acids with a mass of 110 kDa. The *AR* consists of four structurally and functionally distinct domains; a poorly conserved N-terminal domain, a highly conserved DNA-binding domain and a moderately conserved ligand-binding domain. A short amino acid sequence separates the ligand-binding domain from the DNA-binding domain and also contains part of a bipartite ligand-dependent nuclear localization signal for *AR* nuclear transport.

The AR ligand-binding domain (LBD, amino acids 669-919) facilitates binding of the AR ligands testosterone and dihydrotestosterone (DHT) which represents the primary control mechanism of the androgen-signalling pathway. The cytochrome P450 enzyme converts testosterone to DHT and both can bind to and activate AR under physiological conditions, with DHT having a significantly greater affinity for AR. In the absence of ligands, the AR is located primarily in the cytoplasm where it associates with heat shock proteins (HSP)-90, -70, -56, cytoskeletal proteins and other chaperones. Binding of ligand to the AR ligand-binding pocket induces a conformational change in AR within the LBD, forming the principal protein-protein interaction surface that facilitates intramolecular and intermolecular interaction resulting in the dimerization of AR. Several AR-associated coactivators facilitate the nuclear targeting of AR and once inside the nucleus, AR binds to specific recognition sequences known as androgen response elements (AREs) in the promoter and enhancer regions of target genes. The AR transcriptional complex is completed by recruitment of coregulators, which ultimately results in modulation of target genes expression.

4.2 Primary prostate cancer

Men with localised prostate cancer can have very different prognoses and face a wide array of treatment options. Men are advised on treatment based on risk assessments that often combine patient age, clinical tumor stage, serum PSA, Gleason score, number of positive prostate biopsies and amount of malignant tissue per core to select patients for treatment ranging from active surveillance alone through multimodality treatment. Active surveillance protocols that are amenable to a subset of low-grade localised prostate cancer patient avoid unnecessary treatments and typically monitor patients over time with serum PSA measurements, repeated prostate biopsies and MRI. For prostate cancer patients who do not favour or are not eligible for active surveillance protocols, radical prostatectomy, external-beam radiotherapy and brachytherapy are standard local treatments.

Multiple studies have identified recurrent somatic mutations, copy number alterations, and oncogenic structural DNA rearrangements in primary prostate cancer [4,50,51,54]. These include point mutations in *SPOP*, *FOXA1*, and *TP53*; copy number alterations involving *MYC*, *RB1*, *PTEN*, and *CHD1*; and ETS fusions, among other biologically relevant genes. While certain primary prostate cancer alterations or signatures have prognostic clinical significance, the therapeutic impact of primary prostate cancer genomic events has not yet been realized.

4.3 Castration resistant prostate cancer

In many cases, local therapy is not effective and rising of PSA levels indicates disease recurrence. First-line therapeutic intervention for metastatic prostate cancer is hormone deprivation therapy, which is designed to ablate *AR* activity. Although initially effective, hormone therapy resistant tumors arise, representative of the transition to incurable castration resistant prostate cancer (CRPC). Frequent reactivation of *AR* signalling has been reported in studies of CRPC patients as well as frequent disruption of chromatin and histone modellers and tumor suppressors. Indeed, frequent copy number gains of 8q as well as copy number losses of 8p, 13q, 16q, and 18q were observed and the landscape of copy number alterations is characterized by recurrent amplification peaks (frequent *AR*, 8q gain) and deletion peaks (*CHD1*, *PTEN*, *RB1*, *TP53*). To provide a systematic analysis of the genomic landscape of CRPC and its potential relevance for patient care, the Stand Up To

Cancer (SU2C)-Prostate Cancer Foundation (PCF) International Dream Team pursued whole-exome and transcriptome sequencing of 150 biopsies from metastatic CRPC (mCRPC) [56]. Results indicate presence of established biological “driver” aberrations in a cancer-related gene (i.e., known oncogenes or tumor suppressors) in nearly all the cases. While 99% of the CPRC cases harbours a potential driver single nucleotide variant (SNV) or indel, other classes of driver aberrations were also highly prevalent. These include driver gene fusions in 60%, driver homozygous deletions in 50% and driver amplifications in 54%. While informative mutations were present in virtually all CRPC cases, 63% harbored aberrations in *AR*.

4.4 DNA repair and cell cycle defects in prostate cancer

Prostate carcinogenesis is mediated, as other cancers, by the accumulation of genetic and epigenetic aberrations; these molecular changes can be inherited or be the result of altered *AR* transcriptional activity, changes in chromatin architecture, oncogenic replication, error-prone DNA repair, or defective cell division. Deficient DNA repair response and defective apoptotic checkpoint control can then lead to permanent incorporation of these genome abnormalities.

4.4.1 Overview of the DNA damage response pathway

DNA damage continuously occurs in human cells. If repair mechanisms are impaired, genome stability is compromised, therefore contributing to tumorigenesis. Damage can occur endogenously (due to spontaneous hydrolysis of bases or reaction of DNA with naturally occurring reactive oxygen species or alkylating agents) or can be induced by exogenous agents (*i.e.* radiation and toxins). In order to protect their genome, cells have evolved several biological pathways with complementary and partially overlapping functions for recognizing and accurately repairing damages. Different forms of DNA damage trigger a response from different branches of this complex system. The main workflow is as follows; when genomic insults are detected, cell-cycle checkpoints are activated to halt the cell cycle and allow the cellular machinery to repair the DNA damage. If the repair is successful, the cell continues its normal cycle; otherwise, programmed cell death or senescence programs are triggered. If the DNA repair mechanisms are dysfunctional, genomic instability, which is one of the hallmarks of carcinogenesis, occurs. When damage is limited to one of the DNA strands (single-strand breaks or base modifications), different repair mechanisms can be deployed. These include base-excision repair (BER), single-strand break repair (SSBR), nucleotide-excision repair (NER), and mismatched repair (MMR). Each of these pathways uses the complementary undamaged strand as a template to ensure fidelity of repair. The primary mechanisms involved in DNA double-strand break (DSB) repair comprise the homologous recombination (HR) system and the non-homologous end joining (NHEJ). HR requires a sister chromatid as template and is therefore restricted to the S/G2 phases of the cell cycle. It restores the original DNA code error-free. Key mediators of this pathway include BRCA1, BRCA2, PALB2, ATM, ATR, RAD51, MRE11, CHEK2, and XRCC2/3. In contrast, NHEJ functions by ligating broken DNA ends without the use of a template and is therefore functional throughout the cell cycle. The error-prone mode of NHEJ action leads to errors that are permanent and can drive genomic instability.

4.4.2 The role of DNA repair defects in prostate cancer

Interestingly, prostate cancer is often characterized by high numbers of genomic rearrangements. Many of these tumors have oncogenic mutations in the *SPOP* gene that stabilize proteins including *AR* and its transcriptional regulators [57]. Mechanistically, *SPOP* mutant tumors rely predominantly on NHEJ-based DSB repair (while reducing error-free HR-mediated DSB repair activity). The pattern of genomic aberrations may partly depend on deficiencies in specific DNA repair pathway branches. It has been shown that loss of MMR function induces a hyper-mutated microsatellite unstable genotype [56]. Somatic complex rearrangements in *MSH2* and *MSH6*, as well as somatic and germline truncating mutations in these two genes, have been described as the most common mechanism for MMR-deficient prostate tumors [58]. BRCA2-deficient prostate cancers also present specific mutation signatures enriched in deletions and with higher mutational burden than BRCA2 wild type tumors [59]. Moreover, hereditary germline mutations in DNA repair genes are associated with a higher risk of prostate cancers. This results in one gene allele being dysfunctional in every cell, with the second allele commonly lost by a second hit (mutation, deletion, epigenetic silencing) [60]. While the proportion of patients carrying a germline BRCA1/2 mutation is low (1–2%) among the general population of primary prostate cancer patients, a multicenter study lead by the SU2C PCF consortium in metastatic CRPC patients estimated the prevalence of germline BRCA2 mutations as 5.3% in the setting of advanced disease; when a panel of 20 DNA repair genes was considered, 82/692 (11.8%) of patients with metastatic disease carried an underlying germline mutation [61].

4.5 The tumor suppressor *RB1* gene in prostate cancer

The retinoblastoma gene (*RB1*) is implicated in many cellular processes such as regulation of the cell cycle, DNA-damage response, DNA repair, DNA replication, protection against apoptosis, and differentiation, all of which contribute to its function as a tumor suppressor (the first one to be cloned). The retinoblastoma protein (RB, encoded by *RB1*) has a key role in repressing the transcriptional activity of the activator class of E2F transcription factors [62,63]. Briefly, E2Fs regulate the expression of several genes and particularly they control the transcriptional regulation of genes required for cell cycle, nucleotide synthesis, and checkpoint control. During early cell cycle phase G1, the RB and the RB-related p107 and p130 (altogether known as the “pocket protein” family of cell cycle regulators) bind to the E2F transcription factors. Specifically, RB binds and represses activator E2F transcription factors (E2F1–E2F3), while p107 and p130 bind E2F4 and E2F5 to form complexes that repress transcription of G1 to S promoting factors. Upon the decision to progress past the G1 checkpoint, cyclin D forms a complex with cyclin dependent kinases (CDK) CDK4 and CDK6, which in turn phosphorylate the pocket proteins. The phosphorylation causes the release of their bound targets, thereby relieving the repression of the E2F1-3 activators and translocating repressor E2F4-5 from the nucleus to cytoplasm. This results in the transcriptional activation of downstream targets which promote the G1 to S transition (mediated by cyclin E in complex with CDK2).

Although RB plays an important role in the response to hormone therapy in vitro [64], the frequency and impact of RB deregulation during prostate cancer development and progression is not well defined. Based on the observation that in prostate cancer *RB1* copy number loss is overrepresented in metastatic and CRPC [56,65,66], data suggested that RB deficiency may be specifically associated with the transition to castration resistance rather than with tumor initiation [66]. RB deficiency alone did not confer a significant tumor growth advantage in vivo, however castration of host animals with RB-depleted unmasked a growth advantage specific to RB-deficient tumors. These data suggest that RB depletion is sufficient to induce castration-resistant tumor growth as monitored by tumor growth kinetics and serum PSA level, the latter one indicative of enhanced AR signalling. Since CRPC phenotype is associated with alterations of the AR pathway, RB loss may act in concert with or impinge upon the AR axis. Under conditions mimicking therapeutic intervention (androgen ablation, *AR* antagonist bicalutamide and a combination of both) RB-depleted cells showed significantly higher AR target expression suggesting that aberrations in RB function result in enhanced AR signalling [66].

By exploiting double knockout and triple-knockout mice (with *RB1* deletion and *RB1* plus *TP53* deletion, respectively) using a *PTEN-null* mouse model of prostate, authors showed in a recent study that loss of *RB1* drives increased epigenetic deregulation, metastatic progression, and lineage plasticity. Moreover, additional loss of *TP53* converts the disease into a fully antiandrogen-resistant, neuroendocrine variant and causes upregulation of stem cell reprogramming factors. These data indicate that *RB1 and TP53* cooperate to suppress lineage plasticity, metastasis, and resistance to antiandrogens in prostate cancer [67] and therefore nominate *RB1* as main player in advanced prostate cancer.

5. Circulating tumor DNA and liquid biopsy

The presence of fragments of cell-free nucleic acids in human blood was first described in 1948 by Mandel and Metais. In healthy individuals, cell free DNA (cfDNA) concentrations tend to range between 1 and 10 ng ml^{-1} in plasma [68,69]. Although it is unclear what are the release and clearance mechanisms, raised levels of cfDNA were observed in the serum of patients with cancers [70] and is thought to be released from tumoral cells. The modal size of cfDNA was first determined by gel electrophoresis as ~180bp and later by sequencing-based approaches refining this measure as 166bp [71,72]. These measures indicate that cfDNA is likely to be associated with nucleosomes. In 1994, Sorenson *et al.* [73] discovered that the *KRAS* mutated sequence found in the plasma was identical to the patient's tumor, thereby confirming that the mutant DNA fragments in the plasma were of tumor origin [73]. Since mutations in cfDNA are highly specific markers for cancer, this observation gave rise to the term circulating tumor DNA (ctDNA).

The development of NGS-based technologies has facilitated the interrogation of the genome at a broader scale than previously possible. Studies in the last decades explored the ctDNA as a prognostic and predictive biomarker supporting its potential in the clinical setting [74,75]. Although the levels of ctDNA in different clinical contexts were not yet accurately defined, the concentration of ctDNA in plasma has been shown to correlate with tumor size and stage and variability in inter-individual ctDNA concentration is partially explained by differences in the extent of disease burden [72,76]. Analysis of ctDNA ranges in scale from single mutations (custom assay to achieve high sensitivity) to whole-genome assays. Since conventional sampling methods such as needle biopsies are subject to experimental complications (difficulty in obtaining sufficient material of adequate quality for genomic profiling) and biological limitations (sampling bias from genetic heterogeneity), the analysis of tumoral material obtained in a minimally invasive manner through blood sampling (liquid biopsy) could represent an extreme valid alternative to produce clinical benefit in multiple areas of oncology such as cancer diagnosis and prognosis, treatment selection and monitoring of treatment response and disease burden. Even if it is unclear whether all tumor subclones contribute proportionately to the total ctDNA pool or whether their representation in the bloodstream is biased by other biological factors, liquid biopsies sample ctDNA released from multiple tumor regions and thereby reflect both intra-tumor heterogeneity and spatially separated disease foci. Although individual tumor biopsies from different tumour regions may differ in their mutation profiles owing to intra-tumor heterogeneity, ctDNA analysis has detected mutations that have been missed in corresponding tissue samples [77–79]. Appropriately designed assays allow interrogating ctDNA also for the detection of copy number alteration. Indeed, in a cohort of 80 patients with prostate cancer, *AR* copy number gain before anti-androgen (abiraterone) therapy predicted a worse overall survival, thus identifying patients with primary resistance [80]. The ease and reduced risk of repeating liquid biopsies enables to use them for real-time monitoring of cancer burden in response to therapy.

A better understanding of the origin and biology of cfDNA and ctDNA would aid implementation of liquid biopsies. Indeed, the limited understanding of the release and clearance mechanisms of cfDNA should be explored to better define the relative contributions of apoptosis, necrosis and active release and moreover to improve the interpretation of current research studies. So far, proof-of-concept studies have provided an excellent starting point for larger prospective studies of the clinical utility of cfDNA and have demonstrated that ctDNA may be useful research tool for the study of intra-tumor heterogeneity and clonal evolution. In June 2016, the FDA approved the first companion diagnostic test based on ctDNA. The test is designed for the detection of exon 19 deletions or exon 21 (L858R) substitution mutations in the epidermal growth factor receptor (EGFR) gene to identify pateints with metastatic non-small cell lung cancer (NSLC) eligible for treatment with erlotinib. However, randomized trials comparing ctDNA-guided decision-making against standard of care would be definitive to demonstrate the potential utility across a range of applications for patients benefit.

6. High-throughput DNA sequencing

Since the completion of the human genome project in 2003, extraordinary progress has been made in genome sequencing technologies, which has led to a decreased cost per megabase (US$1,000 is currently the average cost of sequencing of a human genome) [81].

These NGS strategies, that in a short period of time transitioned from novelty to almost routine approaches in biomedical research, are providing researchers and clinicians with a variety of experimental assays to profile genomes in greater depth and to translate genomic information into clinically actionable results. Some approaches maximise the number of bases sequenced in the least amount of time, generating a wealth of data that can be used to understand increasingly complex phenotypes. Alternatively, other approaches now aim to sequence longer contiguous pieces of DNA, which are essential for resolving structurally complex regions. WGS is becoming one of the most widely used applications in NGS. Through this technology, the most comprehensive view of genomic information can be obtained revealing the complete DNA make-up of an organism. Whole-exome and targeted sequencing are also providing invaluable advantages to sequencing research. By reducing the size of targeted DNA sequence and by limiting the amount of the genomic material used, more individual samples can be sequenced within a single sequencing run, which can increase both the breadth and the depth of a genomic study. Indeed, a WES experiment targets the set of exons representing nearly the 2% or 55 Mb, of the euchromatic human genome (2.85 Gb). Targeted sequencing assays further reduce the size of DNA to be read by focusing on a limited number of genes (*i.e.* cancer related genes) or other genomic regions of interest.

Although exciting, these advancements are not without limitations. As new technologies emerge, existing problems are exacerbated or new problems arise. NGS platforms provide vast quantities of data but the associated error rates (~0.1-15%) are higher and the read (the sequence of bases from a single molecule of DNA) lengths generally shorter (35-700bp) than those of traditional Sanger sequencing platforms [82], requiring careful examination of the results, particularly for variants discovery.

RATIONALE

Analyses of large cancer datasets support evidence of the genetic heterogeneity among cancers types and even within individual cancer types. Prostate cancer is a highly heterogeneous disease where multiple and coexisting aberrant molecular mechanisms make challenging the selection of effective clinical treatments. The improvements in genomic studies and their association to clinical properties have been driven by the last sequencing methods and proper development of specific algorithms and bioinformatics methodologies have proved useful for defining the mutations, genes and molecular networks that drive diverse cancer phenotypes and that determine clonal architectures in tumor samples. This scenario raises the purpose to quantify the level of inter- and intra-tumor heterogeneity and to explore alternative molecular mechanisms causing the inactivation of key player cancer genes in prostate cancer by setting an automated analysis of high-throughput sequencing data (WGS, WES and targeted) with the goal of being quantifiable, analysable, scalable and reproducible. Tumor DNA from tissue biopsies represent a reliable source to examine intra-tumor heterogeneity, nonetheless this strategy can be unpractical due to complications and pain. Cell free DNA released in blood stream from widespread metastatic cells can be exploited as alternative. Particularly, detection of hot-spot mutations in serum and plasma is challenging since this biological scenario is characterized by little DNA material and high admixture. Allele-specific PCR methods and some assays are available as kits that were approved for clinical use [83,84], but have limited analytical sensitivity. Since these assays rely on differential binding affinities of mutant and wild-type alleles, they require primers or probes that are specific to each genomic locus of interest; this issue limits the multiplexing capacity and reduces the number of mutations that can be investigated concurrently. Conversely, targeted sequencing using PCR amplicons or hybrid capture are used to interrogate a larger number of loci simultaneously. However, sequencing platforms are limited by errors which makes single-nucleotide variant (SNV) detection challenging. The limit of detection represents the threshold below which mutations cannot be confidently discriminated from background noise; for sequencing-based approaches, this is often determined by technical artefacts such as PCR and/or sequencing errors. To improve SNVs detection using this technology, there is a clear need to quantify these limitations and exploit them as useful information. Indeed, my research aimed to set up computational methodology to discriminate between biological and artefactual signals by using locus-specific and data-driven thresholds instead of general and *a priori* selected ones. I also investigated the hypothesis that some genomic loci are more prone than others to be sequenced wrongly and I checked whether this drawback depends on diverse library preparation kits and sequencing chemistries.

RESULTS

In line with the nature of this work, the *Results* section includes both clinically and biologically relevant results (sections 3, 6, 8 and 9), and technical advances and analysis approaches I designed and developed to proper address the biological questions (sections 2, 4 and 5). The *Results* section also includes description of bioinformatics strategies aimed to assemble an efficient computational workflow (section 1) and to achieve methodological improvements (section 7). For clarity, I organized the result and method sections in two parts, **a. Multi-sampling tissue biopsies** and **b. Cell free DNA in plasma samples**. Each results part is followed by relevant methods.

a. Multi-sampling tissue biopsies

The future of cancer treatment is represented by personalized clinical strategies designed to specifically target tumorigenic cell populations present in one individual. Although very promising, this scenario is often complicated due to mutagenesis that plays a key role in shaping the cell subpopulations that characterize patient tumors and this heterogeneity may show dramatic differences in drug response, with the limit condition that some tumor cells are very sensitive and some other show high resistance to the same drug. As a consequence, the treatment may promote the emergence of resistant and more aggressive tumors[85]. Even if a single biopsy produces only limited information on the site from which it is taken, multiples biopsies, when possible, are a reliable strategy for examining different foci within a primary tumor or disseminated metastatic sites. At present, most precision medicine programs rely on high-throughput sequencing (or more commonly NGS) to examine tumor DNA from patient samples and ad-hoc computational methods can be applied to gain proper assessment of the major oncogenic drivers in an individual.

1. Computational toolbox for studying cancer genomes

Here I describe the workflow (**Figure 1**) I assembled to exploit diverse computational tools in order to retrieve an exhaustive genomic characterization of a tumor sample and its matched germline both profiled via NGS techniques. Each element of this pipeline is described in detail in *Methods* sections 1a-6a.

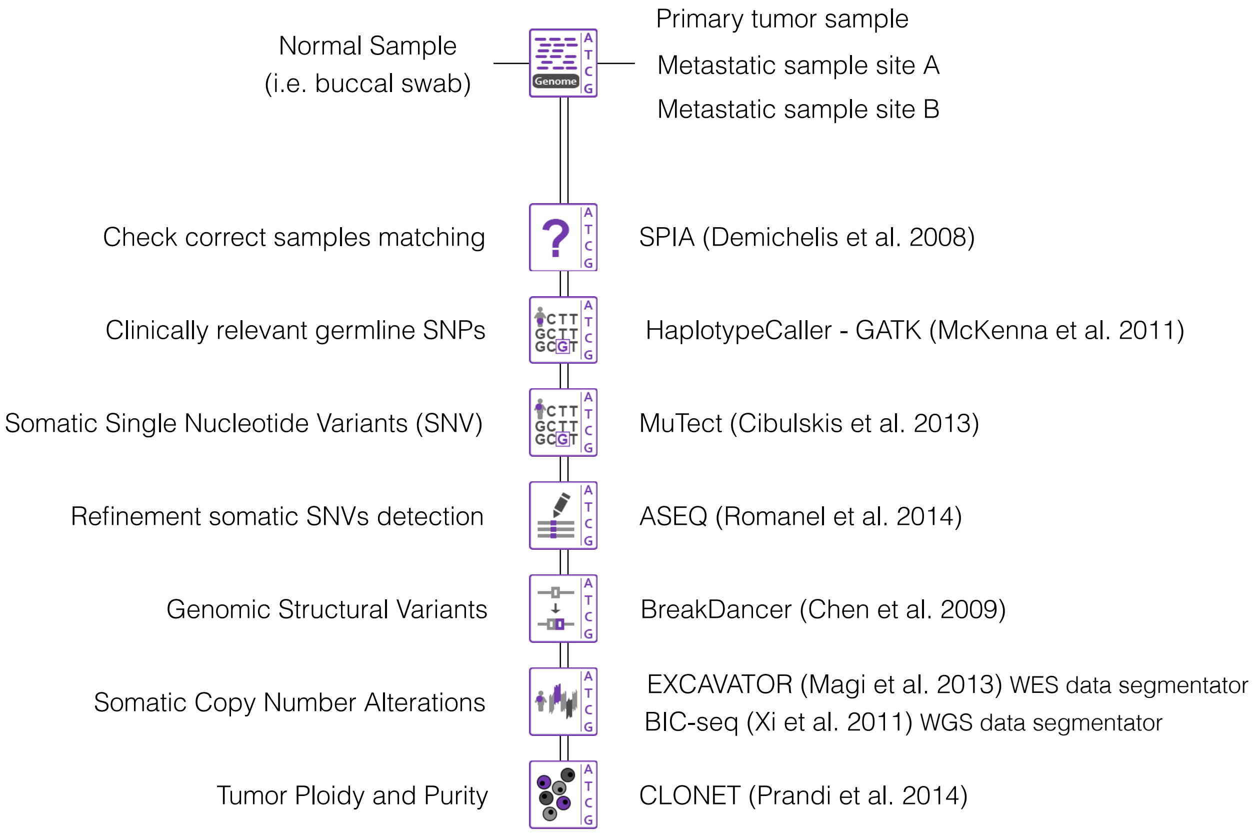

Normal Sample
(i.e. buccal swab)
Genome
Primary tumor sample
Metastatic sample site A
Metastatic sample site B
Check correct samples matching
SPIA (Demichelis et al. 2008)
Clinically relevant germline SNPs
HaplotypeCaller - GATK (McKenna et al. 2011)
Somatic Single Nucleotide Variants (SNV)
MuTect (Cibulskis et al. 2013)
Refinement somatic SNVs detection
ASEQ (Romanel et al. 2014)
Genomic Structural Variants
BreakDancer (Chen et al. 2009)
Somatic Copy Number Alterations
EXCAVATOR (Magi et al. 2013) WES data segmentator
BIC-seq (Xi et al. 2011) WGS data segmentator
Tumor Ploidy and Purity
CLONET (Prandi et al. 2014)

Figure 1. Computational toolbox for studying cancer genomes.

2. Comparative analysis of computational tools for WES data segmentation

Somatic copy number alterations (SCNAs), defined as duplication and deletion genomic events that occur in somatic cells, are common in cancer genomes and recurrent alterations of gatekeeper genes have been associated with specific cancer types[86]. SCNAs can result both in the amplification of oncogenes and in the deletion of tumor suppressors, significantly contributing to cancer genesis and progression. The advent of massively parallel sequencing methods has revolutionized structural variation studies by exploiting the single-base resolution provided by deep sequencing data to precisely predict boundaries of altered genomic regions. In particular, the identification of somatic genetic alterations from WES data is an active research field because it is a cost-effective and powerful technology that represents a valid alternative to whole genome sequencing. Moreover, large-scale collaborative efforts such as TCGA[87] are generating thousands of WES experiments for multiple tumor types amenable for SCNAs analysis. However, WES data are generally affected by a non-uniform read-depth among scattered genomic target regions, making the analysis of structural variants particularly challenging. In this context, although several algorithms have been recently developed, key differences in sensitivity and specificity make none of the computational tools eligible as gold standard method[88].

I therefore performed an analytical comparative analysis to assess the power of three recent computational methods and an in-house developed method tailored on SCNAs detection from WES data. Selected computational tools EXCAVATOR[89], ADTEx[90] and Control-FREEC[91] implement different algorithms to split chromosomes into segments with the same putative DNA copy number. The comparison is made by considering a prostate cancer data set including 16 tumors and matched normal tissues[4,92]. For these samples, regions of SCNAs were previously detected by SNP array (reference data), FISH assay (TMPRSS2:ERG fusion) and segmented regions generated through the processing of WGS data.

Concordance among SCNA callers and reference data has been tested looking at regions spanning a set (N=822) of selected cancer related genes (**Figure 2A**, **3A**) and a set of equally spaced (1Mb) positions across the genome (**Figure 2B**, **3B**). First, I estimated mean correlations among methods considering the overall dataset, both in the gene-based and in the genomic-sampling context (**Figure 2**). Next, to avoid samples heterogeneity issues, a one-by-one sample analysis was conducted estimating correlations between each method and reference data (**Figure 3**).

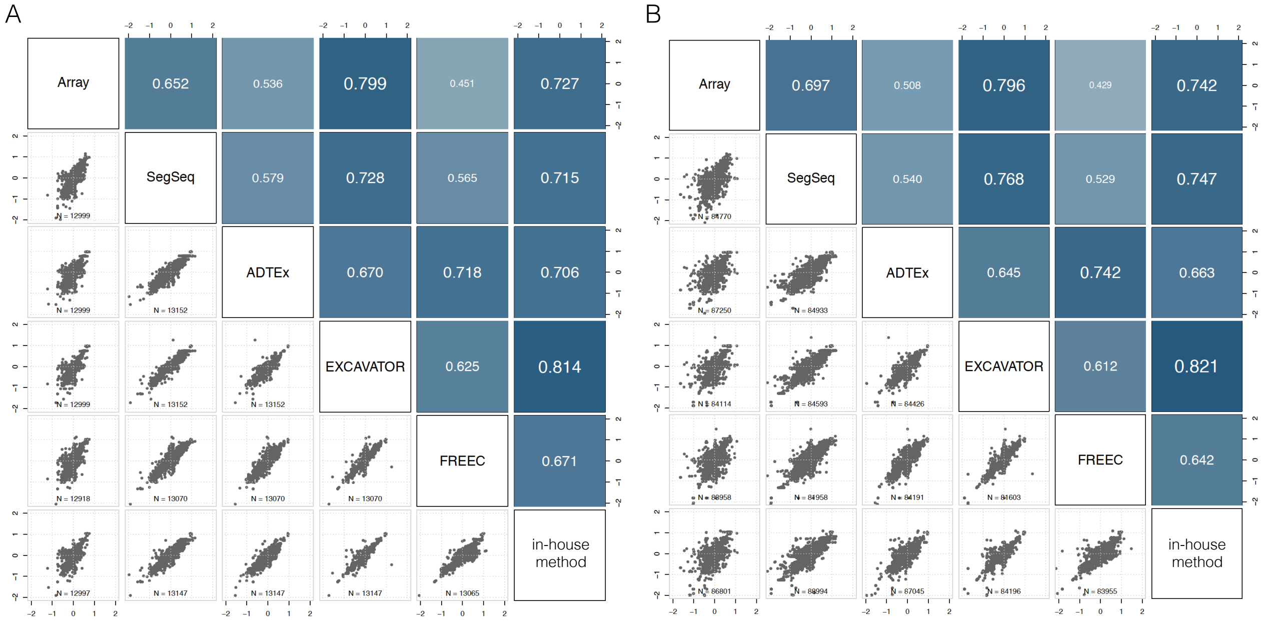
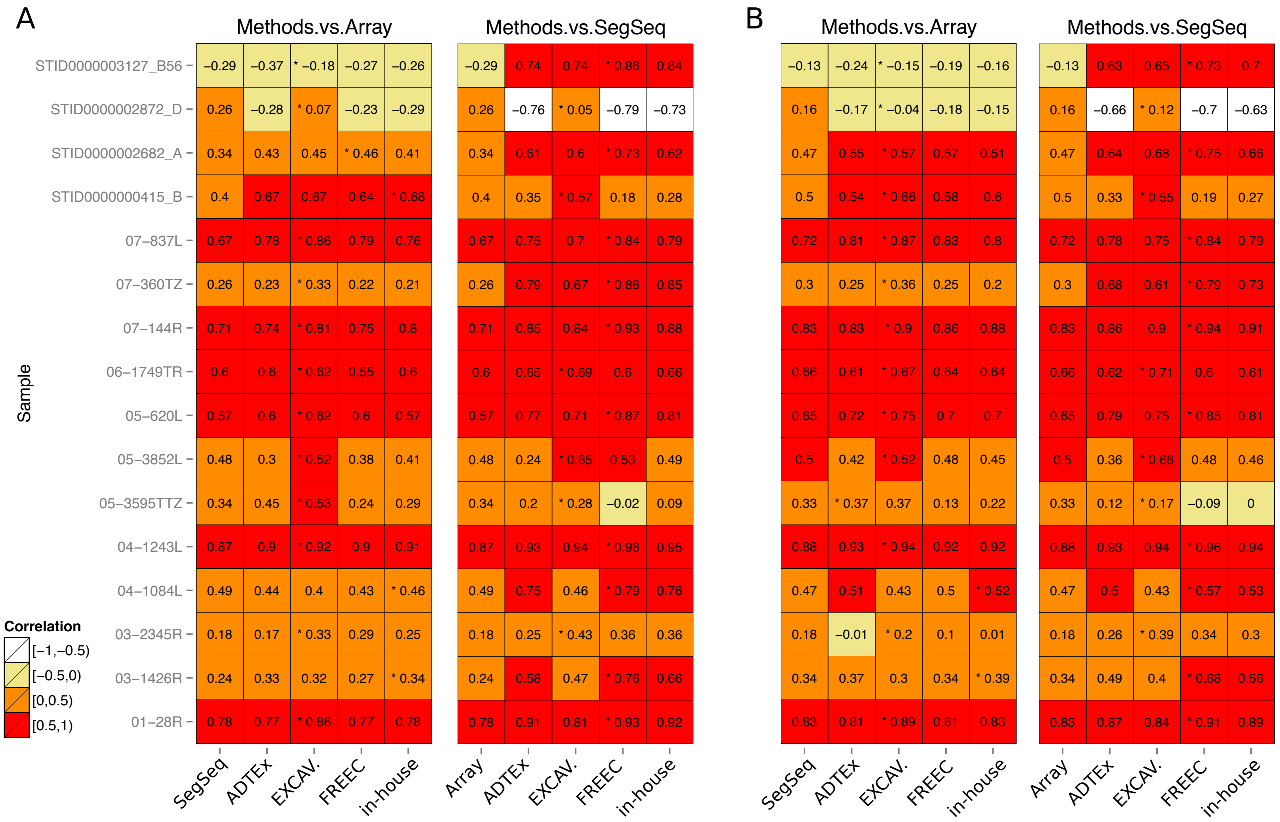


Figure 2. Concordance among SCNA callers and reference data tested (**A**) looking at regions spanning a set of selected cancer related genes (N=822) and (**B**) a set of equally spaced (1Mb) positions across the genome.



A

Sample	Methods.vs.Array					Methods.vs.SegSeq				
	SegSeq	ADTEx	EXCAV.	FREEC	in-house	Array	ADTEx	EXCAV.	FREEC	in-house
STID0000003127_B56	−0.29	−0.37	* −0.18	−0.27	−0.26	−0.29	0.74	0.74	* 0.86	0.84
STID0000002872_D	0.26	−0.28	* 0.07	−0.23	−0.29	0.26	−0.76	* 0.05	−0.79	−0.73
STID0000002682_A	0.34	0.43	0.45	* 0.46	0.41	0.34	0.61	0.6	* 0.73	0.62
STID0000000415_B	0.4	0.67	0.67	0.64	* 0.68	0.4	0.35	* 0.57	0.18	0.28
07–837L	0.67	0.78	* 0.86	0.79	0.76	0.67	0.75	0.7	* 0.84	0.79
07–360TZ	0.26	0.23	* 0.33	0.22	0.21	0.26	0.79	0.67	* 0.86	0.85
07–144R	0.71	0.74	* 0.81	0.75	0.8	0.71	0.85	0.84	* 0.93	0.88
06–1749TR	0.6	0.6	* 0.62	0.55	0.6	0.6	0.65	* 0.69	0.6	0.66
05–620L	0.57	0.6	* 0.62	0.6	0.57	0.57	0.77	0.71	* 0.87	0.81
05–3852L	0.48	0.3	* 0.52	0.38	0.41	0.48	0.24	* 0.65	0.53	0.49
05–3595TTZ	0.34	0.45	* 0.53	0.24	0.29	0.34	0.2	* 0.28	−0.02	0.09
04–1243L	0.87	0.9	* 0.92	0.9	0.91	0.87	0.93	0.94	* 0.96	0.95
04–1084L	0.49	0.44	0.4	0.43	* 0.46	0.49	0.75	0.46	* 0.79	0.76
03–2345R	0.18	0.17	* 0.33	0.29	0.25	0.18	0.25	* 0.43	0.36	0.36
03–1426R	0.24	0.33	0.32	0.27	* 0.34	0.24	0.58	0.47	* 0.76	0.66
01–28R	0.78	0.77	* 0.86	0.77	0.78	0.78	0.91	0.81	* 0.93	0.92

B

Sample	Methods.vs.Array					Methods.vs.SegSeq				
	SegSeq	ADTEx	EXCAV.	FREEC	in-house	Array	ADTEx	EXCAV.	FREEC	in-house
STID0000003127_B56	−0.13	−0.24	* −0.15	−0.19	−0.16	−0.13	0.63	0.65	* 0.73	0.7
STID0000002872_D	0.16	−0.17	* −0.04	−0.18	−0.15	0.16	−0.66	* 0.12	−0.7	−0.63
STID0000002682_A	0.47	0.55	* 0.57	0.57	0.51	0.47	0.64	0.68	* 0.75	0.66
STID0000000415_B	0.5	0.54	* 0.66	0.58	0.6	0.5	0.33	* 0.55	0.19	0.27
07–837L	0.72	0.81	* 0.87	0.83	0.8	0.72	0.78	0.75	* 0.84	0.79
07–360TZ	0.3	0.25	* 0.36	0.25	0.2	0.3	0.68	0.61	* 0.79	0.73
07–144R	0.83	0.83	* 0.9	0.86	0.88	0.83	0.86	0.9	* 0.94	0.91
06–1749TR	0.66	0.61	* 0.67	0.64	0.64	0.66	0.62	* 0.71	0.6	0.61
05–620L	0.65	0.72	* 0.75	0.7	0.7	0.65	0.79	0.75	* 0.85	0.81
05–3852L	0.5	0.42	* 0.52	0.48	0.45	0.5	0.36	* 0.66	0.48	0.46
05–3595TTZ	0.33	* 0.37	0.37	0.13	0.22	0.33	0.12	* 0.17	−0.09	0
04–1243L	0.88	0.93	* 0.94	0.92	0.92	0.88	0.93	0.94	* 0.96	0.94
04–1084L	0.47	0.51	0.43	0.5	* 0.52	0.47	0.5	0.43	* 0.57	0.53
03–2345R	0.18	−0.01	* 0.2	0.1	0.01	0.18	0.26	* 0.39	0.34	0.3
03–1426R	0.34	0.37	0.3	0.34	* 0.39	0.34	0.49	0.4	* 0.68	0.56
01–28R	0.83	0.81	* 0.89	0.81	0.83	0.83	0.87	0.84	* 0.91	0.89

Figure 3. One-by-one sample analysis estimating correlations between each method and reference data looking at regions spanning a set (N=822) of selected cancer related genes (**A**) and a set of equally spaced (1Mb) positions across the genome (**B**).

Figure 4 shows the distributions of log2 ratio values computed by each method for segments spanning the ETS2 gene (located within the 3Mb interstitial deletion between ERG and TMPRSS2 genes) in each sample. Blue and red boxes represent distributions of ETS2 log2 ratio values of samples annotated as TMPRSS2:ERG deletion positive or not by in-situ FISH.

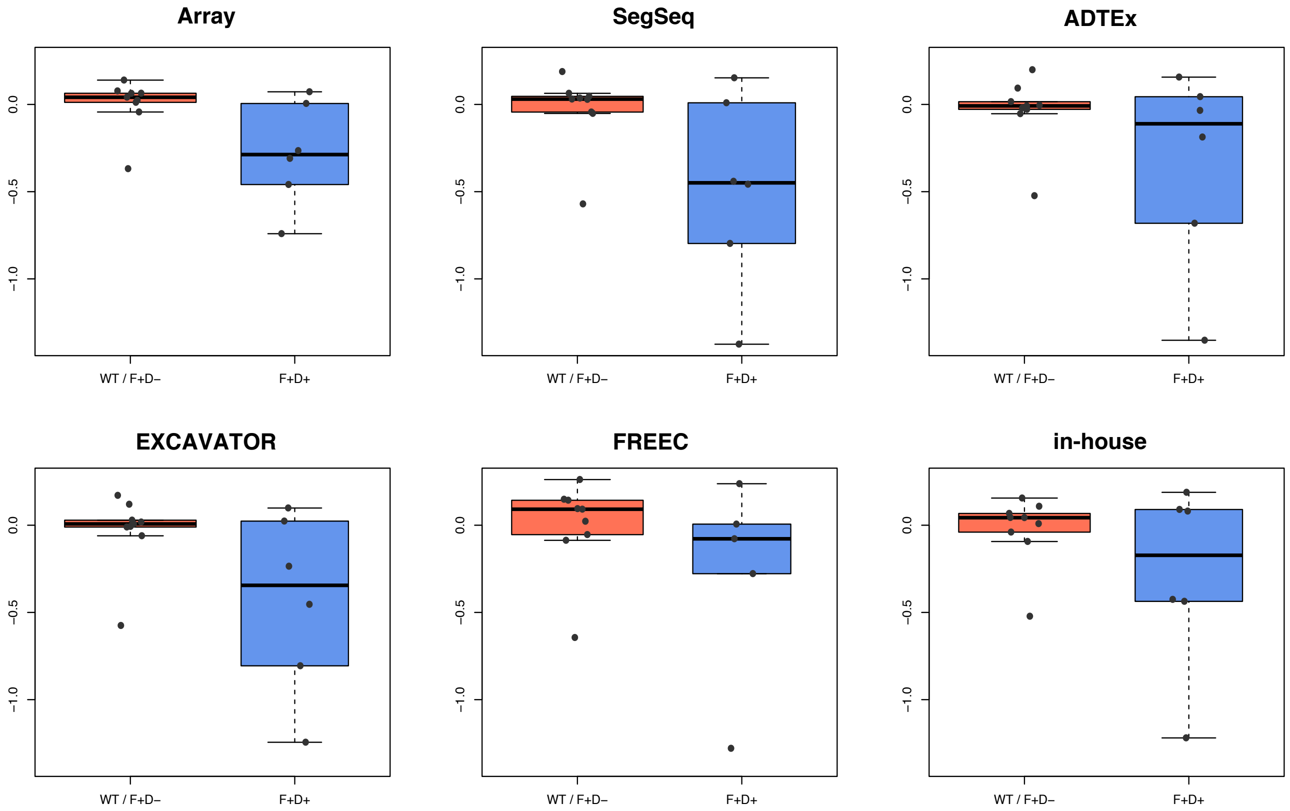


Figure 4. Log2 ratio distributions computed by each method for segments spanning the ETS2 gene in each sample.

These formal analyses suggest that, on average, EXCAVATOR achieves higher concordance than ADTEx and Control-FREEC both with SNP array data (Spearman's correlation: 0.799, 0.536, 0.451, respectively) and WGS segmented regions (Spearman's correlation: 0.728, 0.579, 0.565, respectively).

3. Tumor heterogeneity in castration resistant prostate cancer patients

I studied a cohort composed by 10 advanced prostate cancer patients looking for evidence of intra- and inter-patient tumor heterogeneity (Table 1 and supplementary Table 1). This work was performed in collaboration with Prof. Johann De Bono and members of his team (ICR London, UK). The manuscript is in preparation (Nava Rodrigues D*, Casiraghi N*, *et al*).

Table 1. List of patients and samples included in the study cohort. All tissue samples are incisional biopsies. WGS, whole genome sequencing; WES, whole exome sequencing; TURP, Transurethral resection of the prostate; LN, lymph node.

Patient ID	Tumor ID	Sequencing data	Sites	Site classification
904	LP2000338	WGS	Left Axillary LN	Metastatic tumor
904	LP2000339	WGS	Inguinal LN	Metastatic tumor
V4002	LP2000360	WGS	Anterior left supraclavicular LN	Metastatic tumor
V4002	LP2000361	WGS	Posterior left supraclavicular LN	Metastatic tumor
V4038	LP2000301	WGS	Right Supraclavicular LN	Metastatic tumor
V4038	LP2000302	WGS	Right Retropectoral LN	Metastatic tumor
V4074	LP2000115	WGS	Dorsal Glans	Metastatic tumor
V4074	LP2000116	WGS	Ventral Glans	Metastatic tumor
V4074	LP2000117	WGS	Right Coronal Sulcus	Metastatic tumor
V4074	Sample_A34_0001	WES	TURP	Primary tumor
V5033	LP2000119	WGS	Liver	Metastatic tumor
V5033	LP2000120	WGS	Thigh Muscle	Metastatic tumor
V5128	LP2000184	WGS	Right Inguinal LN	Metastatic tumor
V5128	LP2000185	WGS	Left Supraclavicular LN	Metastatic tumor
V5149	LP2000304	WGS	Right Supraclavicular LN	Metastatic tumor
V5149	LP2000305	WGS	Left Supraclavicular LN	Metastatic tumor
V5149	Sample_B16_0001	WES	Prostatectomy microdissected area	Primary tumor
V5149	Sample_B16_0002	WES	Prostatectomy microdissected area	Primary tumor
V5162	LP2000341	WGS	Left Inguinal LN	Metastatic tumor
V5162	LP2000342	WGS	Right Inguinal LN	Metastatic tumor
V5162	Sample_B16_0004	WES	Prostatectomy microdissected area	Primary tumor
V5162	Sample_B16_0005	WES	Prostatectomy microdissected area	Primary tumor
V5164	LP2000313	WGS	Left Supraclavicular node	Metastatic tumor
V5164	LP2000314	WGS	Liver	Metastatic tumor
V5191	LP2000363	WGS	Left axillary LN	Metastatic tumor
V5191	LP2000364	WGS	Right Axillary LN	Metastatic tumor

The genomic profiles of primary (N=4) and metastatic (N=21) samples (**Figure 5**) have been generated by applying the computational workflow described in section 1. Primaries tissue samples were formalin-fixed paraffin-embedded (FFPE) and WES was performed, where metastatic tissue samples were fresh-frozen (FF) tissue samples and WGS was performed. Overall, the total number of non-synonymous SNVs and the burden of genomic copy number alterations are concordant among tumors within the same patient. These genomic features and additional analysis did not show evidence of tumor heterogeneity in most cancer genes but interesting results emerged when I focused the analysis on a subset of key player prostate cancer genes. Indeed, in two patients I observed intra-patient heterogeneity in the genomic status of *RB1*, the gene encoding for the retinoblastoma protein RB. In one patient, only one (out of two) metastatic site has sequencing reads supporting the presence of a single base substitution introducing a premature stop codon in exon 7 and then causing a truncated product. In the second patient, the allele specific copy number status of RB1 is different between two metastatic sites: the first shows a heterozygous deletion, the latter a neutral loss of heterozygosity. Additionally, the WGS data allowed extending the analysis to structural variants detection. While focusing on *RB1*, I found putative structural genomic rearrangements whose breakpoints lie within the coding region of *RB1* in 3 of the 10 patients. Sequencing-based genomic analysis results were integrated and validated by experimental in situ assays (supplementary Table 2) to estimate *RB1* copy number (fluoresce in situ hybridization) and RB quantification (immunohistochemistry). In summary, the genomic status of *RB1* at different metastatic sites is altered by heterogeneous aberrations such as point mutations, deletion events and structural variants that together can cause RB functionality impairment.

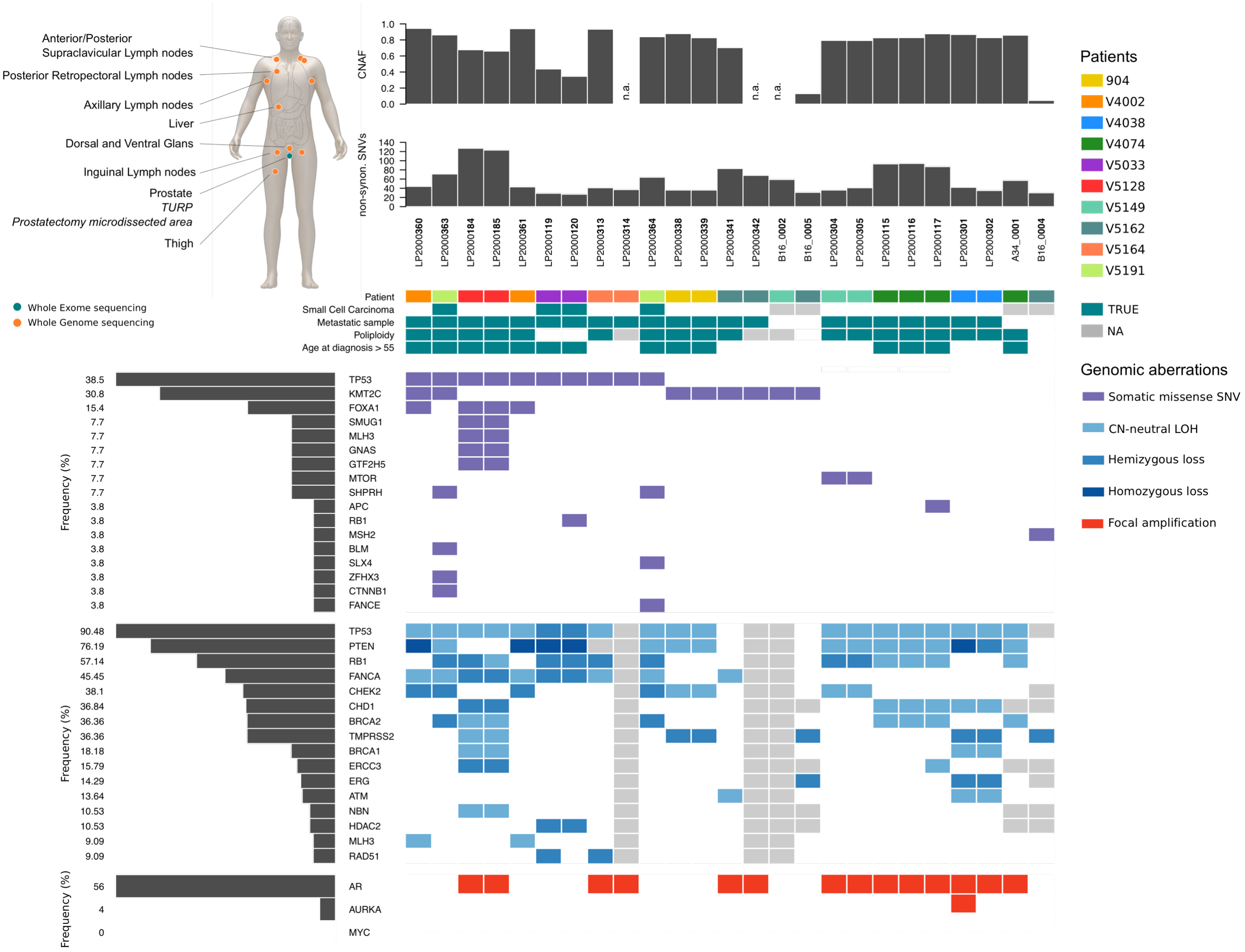


Figure 5. Clinical and genomic profiles of the study cohort. Top left**,** schematic illustrating biopsies at primary (green) and metastatic (orange) sites. Primary sites and metastatic sites were profiled throughout whole-exome and whole-genome sequencing, respectively. Main figure, focus on study

cohort most aberrant cancer related genes. Each row represents a gene and each column a tumour sample. Samples have been sorted accordingly with the genomic status of a subset of cancer related genes. Specifically, the genomic status (wild type or mutated) is here established based on the presence or absence of at least one missense somatic point mutation. Grey bars at top correspond to the fraction of genome altered by a SCNA event (Copy Number Altered Fraction, CNAF) and to the total number of somatic non-synonymous SNVs. Grey bars on the left indicate, for each gene, the fraction of samples affected by somatic missense SNV (violet), copy number loss (blue) or focal amplification (red). In this study, small cell carcinoma status was determined based on histology and lack of AR protein expression. Polyploidy here refers to samples with more than 2 paired sets of chromosomes. The quantification of number of paired sets of chromosomes is based on computational genomic analysis. Overall, the total number of non-synonymous SNVs and CNAFs indicate low intra-patient tumor heterogeneity. Analysis focused on a reduced set of known cancer-associated genes revealed heterogeneous genomic status of *RB1* in patients V5128 and V5033.

3.1 Limited intra-tumor heterogeneity

As first genomic characterization and genomic instability measure, I counted the total number of SNVs detected in coding regions for each tumoral sample analysed. A median number of 170 SNVs are found considering all tumor samples (min = 70, max = 494, mean = 220). The set of primary samples and metastatic sites have a median number of 103 (mean = 112, SD = 50) and 173 (mean = 241, SD = 109) SNVs, respectively. When the analysis is stratified by using information provided by functional annotation, the median number of non-synonymous SNVs observed across primary and metastatic samples is 44 (mean = 44, SD = 15) and 43 (mean = 60, SD = 30), respectively.

However, genomic instability is mainly caused by extended rearrangements, such as duplications or losses of DNA portions, that alter the normal architecture of the genome. Thus, I estimated for each sample the fraction of the genome that is affected by a somatic copy number alteration, defined as Copy Number Altered Fraction (CNAF). The median CNAF across all tumors is 0.83 (mean = 0.72, SD = 0.25). CNAF computed separately in primary and metastases is 0.13 (mean = 0.35, SD = 0.45) and 0.82 (mean = 0.78, SD = 0.16), respectively. Overall, these results confirm that primary samples are less aberrant than metastatic ones and genomic profiles of tumor samples within the same patient are fully comparable indicating a low level intra-tumor heterogeneity.

The number of non-synonymous SNVs and the CNAFs are reported as barplots at the top of Figure 5.

3.2 Allelic Fraction (AFs) comparison and evolutionary trees

For each patient, available tumor samples have been compared pairwise looking at SNVs detected in coding regions. The fraction of SNVs that is shared between tumors or private to only one sample varies across patients. Generally, AFs of missense somatic point mutations that are shared between tumor samples are higher than AF of private ones. Moreover, to better represent and highlight the evolutionary process of tumorigenesis, SNVs and SCNAs affecting a set of cancer genes were used to build phylogenetic trees in each patient. In Figure 6 are reported pairwise comparison between metastatic samples of patient V4074 (**Figure 6A**) and V4002 (**Figure 6B**).

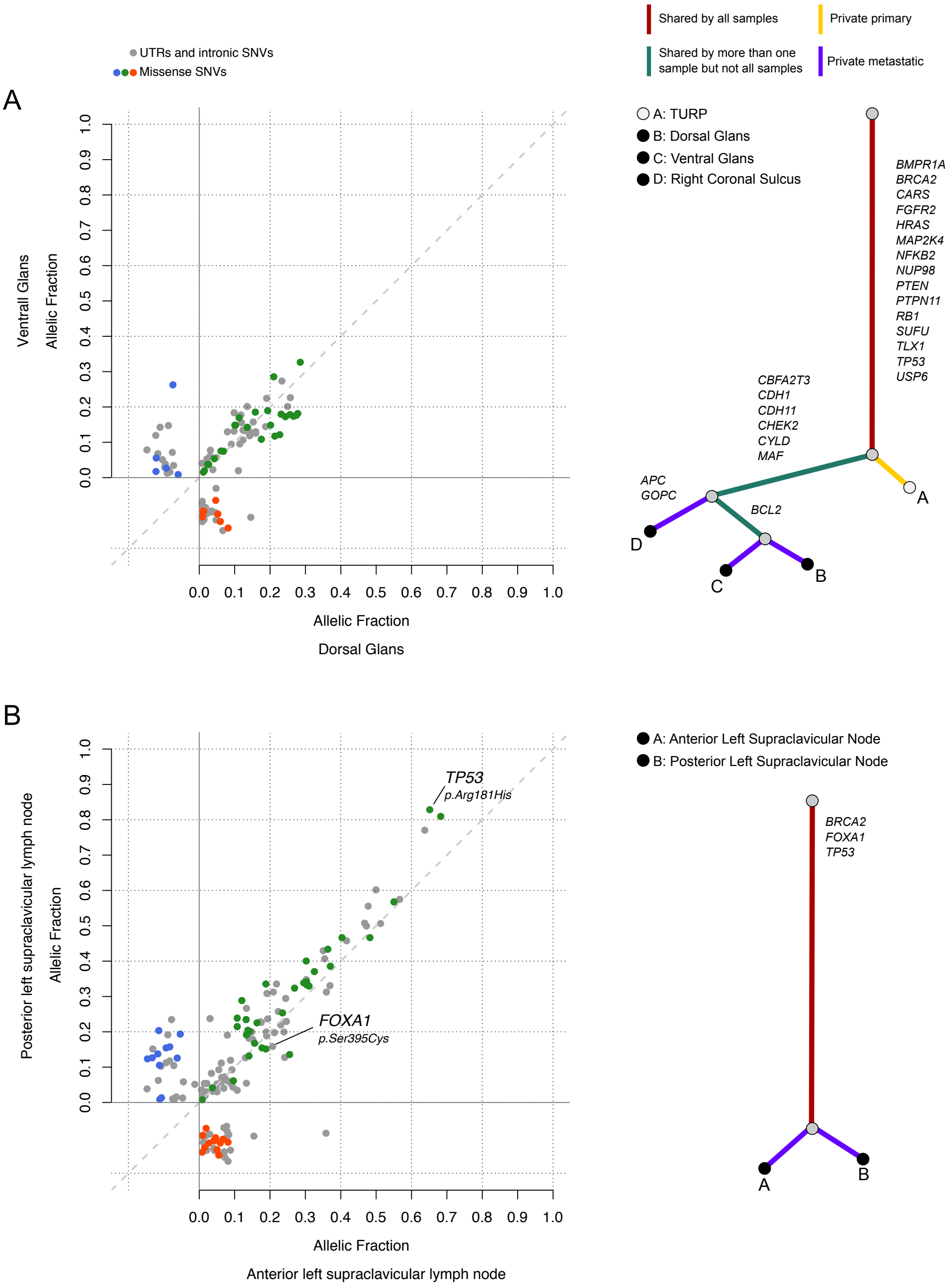


Figure 6. Genomic comparison between metastatic sites from patient V4074 **(A)** and V4002 **(B)**. Left panels show pairwise comparison of variant allelic fractions of SNVs detected in two metastatic sites. Each dot is a somatic SNV; red and blue dots indicate missense SNVs private to the metastatic site on x-axis and y-axis, respectively. Green dots indicate missense SNVs shared between the two

samples. Grey dots indicate private/shared SNVs located in UTR or intronic regions. On the right, phylogenetic trees built from allele-specific copy-number and single-nucleotide variant calls are shown. The length of each branch is proportional to the number of aberrant cancer related genes that are shared by all samples (red), by more than one but not all samples (green), private to primary or to metastatic sites (yellow and violet). Clonality of aberrations is not considered. Color of tree leaves distinguishes between primary and metastatic sites.

3.3 Different molecular mechanism for *RB1* inactivation

I focused the study of tumor heterogeneity by considering genomic aberrations affecting a subset of known cancer-associated genes (N=231) (**Figure 5** central heatmap). Evidence of tumor heterogeneity can be observed in two patients: metastatic sites of patient V5033 differ in *RB1* genomic status by somatic single nucleotide variant and metastatic sites of patient V5128 differ in *RB1* genomic status by somatic copy number status.

3.3.1 Diverse *RB1* genomic status by somatic single nucleotide variant

In patient V5033, heterogeneity in the genomic status of *RB1* is represented by the presence of a single nucleotide substitution only in one metastatic site. The Guanine (G) to Thymine (T) substitution in position chr13:48934155 (hg19) causes a premature stop codon with consequent protein truncation (p.Glu204*). Sequencing reads mapped on 40bp region spanning upstream region of *RB1* exon 7 and supporting the presence of the alternative allele T can be observed exclusively in the thigh metastasis (**Figure 7**). Immunohistochemistry analysis shows absence of RB protein expression at both sites and FISH data indicates loss of one allele, in agreement with NGS data (**Figure 5**).

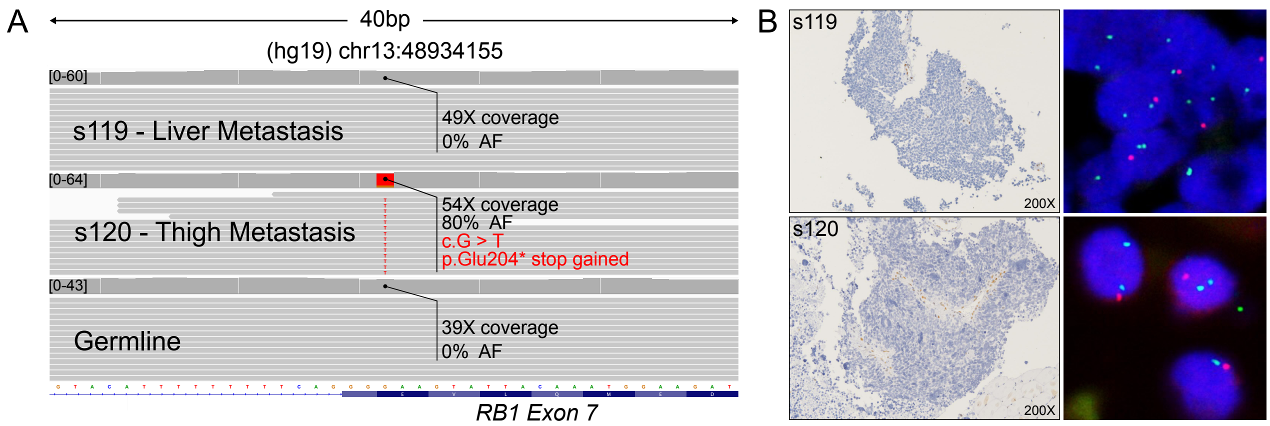


Figure 7. Metastatic sites of patient V5033 differ in *RB1* genomic status by somatic single nucleotide variant. (**A**) Each track shows sequencing reads (horizontal grey bars) mapped on a 40bp region spanning the upstream region of *RB1* exon 7 from two metastatic samples (top first and second tracks) and matched germline (bottom track). Red highlights the nucleotide base Thymine (T) in mapped reads crossing position chr13:48934155 (hg19) (reference allele is Guanine (G)). This

single nucleotide substitution, observed only in one metastatic site, causes a premature stop codon with consequent protein truncation (p.Glu204*). Variant allelic fraction is indicated with AF. (**B**) Left, immunohistochemistry of metastatic samples for RB. Right, FISH data for *RB1* (target probes in red and reference probes in green).

3.3.2 Diverse *RB1* genomic status by copy number status

In patient V5128, data show that the copy number status of *RB1* is compatible with a heterozygous loss event (loss of one allele) in the right inguinal lymph node and, differently, with a copy-number neutral loss of heterozygosity event (two copies of the same allele and none of the other one) in the left supraclavicular lymph node (**Figure 8**).

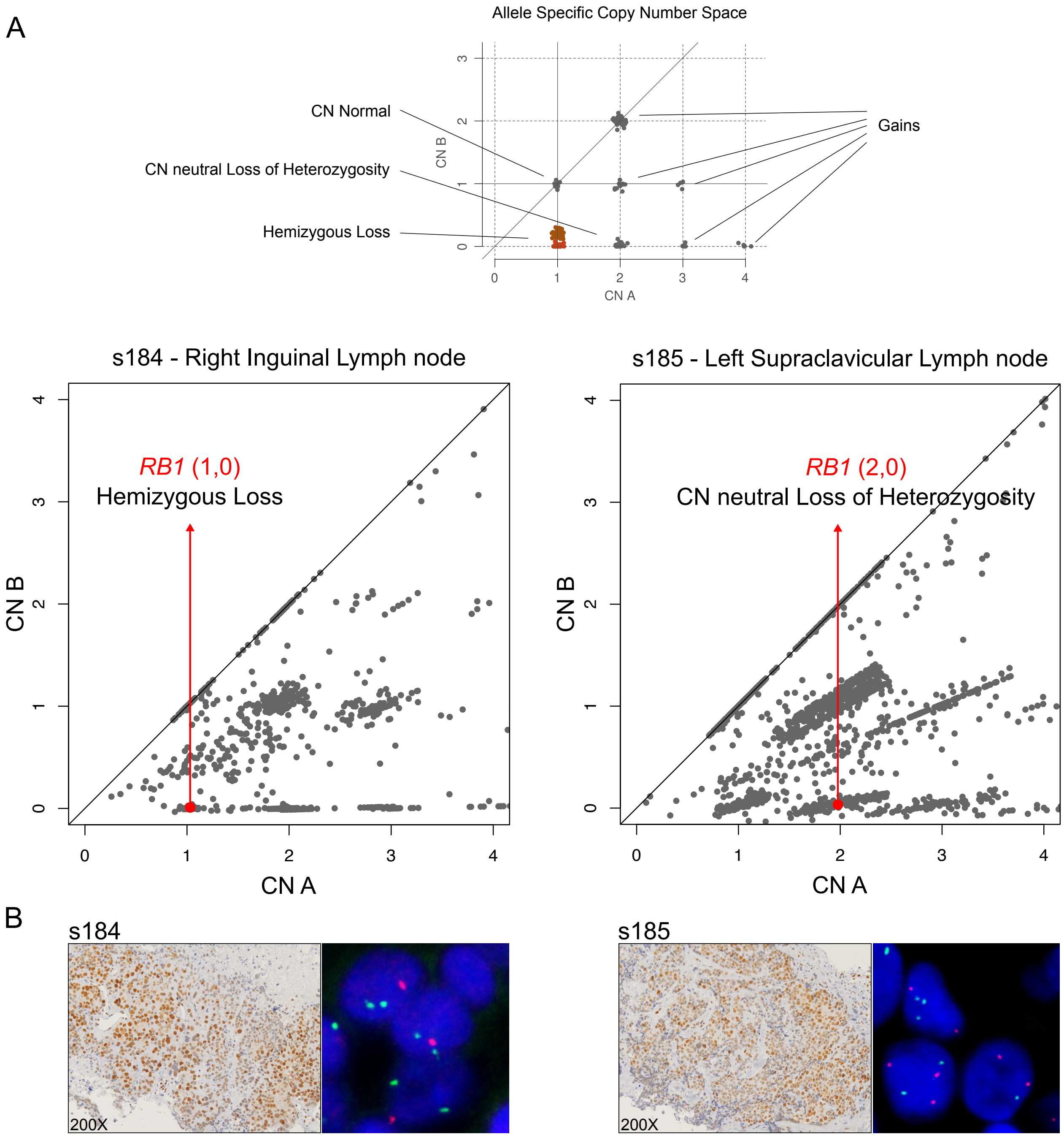


Figure 8. Metastatic sites of patient V5128 differ in *RB1* genomic status by somatic copy number. (**A**) Schematic representation of the allele specific copy number space is shown on top. Each dot is a genomic segment; position in the x-y space reflect allele A and B copy number estimates by CLONET (values are discretized copy numbers where CN A represents the major allele, while CN B the minor allele). Data show that *RB1* copy number status (red dot) of patient V5128 is compatible with monoallelic loss (loss of one allele) in the right inguinal lymph node (bottom left) and with a copy-number neutral loss of heterozygosity (two copies of the same allele and none of the other one) in the left supraclavicular lymph node (bottom right). (**B**) RB immunohistochemistry and *RB1* FISH data (target probes in red, reference probes in green) in metastatic samples s184 (left) and s185 (right). IHC data shows comparable expression levels of RB (Hscores = 100) in both sites. *RB1* hemizygous deletion is supported by FISH data in s184; non homogenous signal is observed for sample s185.

3.4 Alternative molecular mechanism causing *RB1* inactivation in patient V4074

The paired-end whole genome sequencing data available for metastatic samples allowed to extend the analysis to structural variants detection. Data inspection at BreakDancer identified break points reveals that all the 3 metastatic sites of patient V4074 share structural variant break points within *RB1* (**Figure 9A**). Analysis of orientation and insert size of paired reads spanning the break points indicates that some pairs of reads have an anomalous orientation that is compatible with a tandem duplication event involving exons 7 to 17 altering the normal *RB1* genomic architecture (**Figure 9C**). RNA sequencing data were also queried for anomalous pair orientations and reads supporting the structural variant were found around expected junctions (**Figure 9B**). Moreover, available WES data for the V4074 patient allowed to verify that coverage increases in exons involved in the tandem duplication as carefully described in next section 3.4.1.

3.4.1 Tandem duplication analysis in patient V4074

Data inspection at BreakDancer identified break points in the 3 metastatic sites of patient V4074 revealed paired reads with anomalous orientations suggesting that a genomic segment of DNA spanning *RB1* 7 to 17 exons is duplicated and inserted adjacent to the original sequence. I checked if the tandem duplication event is also supported by increased coverage in the genomic region comprised between the two putative break points with respect to both upstream (exons 1 to 6) and downstream (exons 18 to 27) *RB1* regions.

To confirm this yet unknown mechanism of potential RB1 impairment, I queried WES and RNA sequencing data available for the same patient from the PCF SU2C[56] cohort. I computed the following steps on WES data to verify the presence of the tandem duplication event:

1. Compute the mean coverage of each *RB1* exon in tumor sample;
2. Normalize the mean coverage of each *RB1* exons on the total coverage of all *RB1* exons in tumor sample;
3. Repeat step 1 and 2 on matched normal sample;
4. For each *RB1* exon, compute the ratio between tumor and matched normal normalized mean coverages;
5. Compute the median ratio grouping exons 1 to 6 ($R_{upstream}$), 7 to 17 ($R_{tandem\_duplication}$) and 8 to 27 ($R_{downstream}$).

Among WES data of the PCF SU2C cohort (N=149), tumor sample of patient V4074 shows the highest ratios $R_{tandem\_duplication}$ / $R_{upstream}$ and $R_{tandem\_duplication}$ / $R_{downstream}$ (specifically 1.75 and 2.18) indicating a significant enrichment in the coverage of exons involved in the tandem duplication (**Figure 9E**).

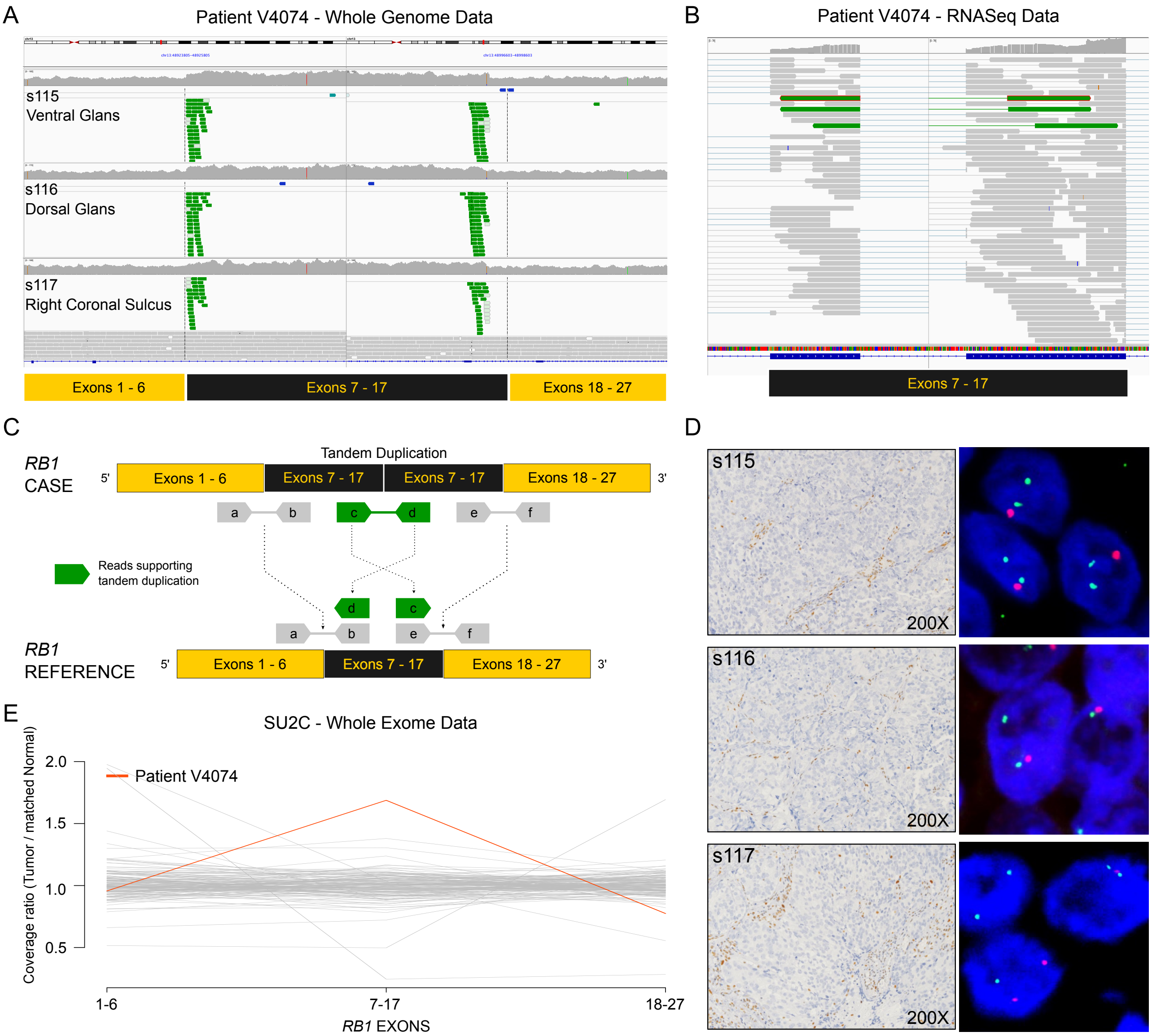


Figure 9. Alternative molecular mechanism causing RB1 inactivation in patient V4074. **(A)** Paired-end whole genome sequencing data mapped on *RB1* locus for 3 metastatic sites of patient V4074. Histograms at the top of each alignment track show coverage profiles. Grey sequencing reads indicate expected (correct orientation and insert size of paired reads) mapping. Vertical black lines indicate break points detected by BreakDancer algorithm. Green highlights pairs of reads with anomalous orientation compatible with a tandem duplication event. **(B)** Paired-end RNA sequencing reads from V4074 supporting (green) the tandem duplication event involving exons 7 to 17. **(C)** Schematic representation of detected tandem duplication. On the top it is shown the potential *RB1* genomic architecture when tandem duplication involving exons 7 to 17 occurs (black boxes). Read pairs with coordinates or insert size spanning break points (conjunctions between yellow and black boxes) will correctly map (grey reads) on the reference genome (bottom), while read pairs with insert size spanning the region between the repeated region (conjunctions between consecutive black boxes) will map with an anomalous orientation (green reads). **(D)** Left, immunohistochemistry of metastatic samples for RB show absence of expression. Right, FISH data for *RB1* (target probes in red and reference probes in green) support concordantly hemizygous loss across all sites. **(E)** Analysis extended to PCF SU2C cohort (N=149) confirms that the event is detectable in WES data by ad hoc coverage bases computation. In red, the median tumor/normal normalized coverage log-ratio profile (y-axis) computed across *RB1* exons (grouped as indicated on x-axis) for a metastatic

site of patient V4074 affected by the tandem duplication. Grey lines represent profiles of all other PCF SU2C patients.

3.5 Genomic data explains RB protein levels

In order to have an exhaustive characterization of RB1 in tumor samples, genomic and experimental data have been integrated and analysed together. Results show that immunohistochemistry (IHC) Retinoblastoma protein levels correlate with the number of genomic aberrations at *RB1* genomic locus with potential damaging functional impact. Indeed, data show that the higher the number of genomic aberrations at *RB1* the lower the protein level measured. Specifically, tumors carrying none (N=7), one (N=8) or two (N=6) genomic aberrations in *RB1* have a median IHC score of 120 (mean = 120, SD = 35.11), 65 (mean = 61.25, SD = 57.92) and 0 (mean = 0, SD = 0), respectively (**Figure 10**).

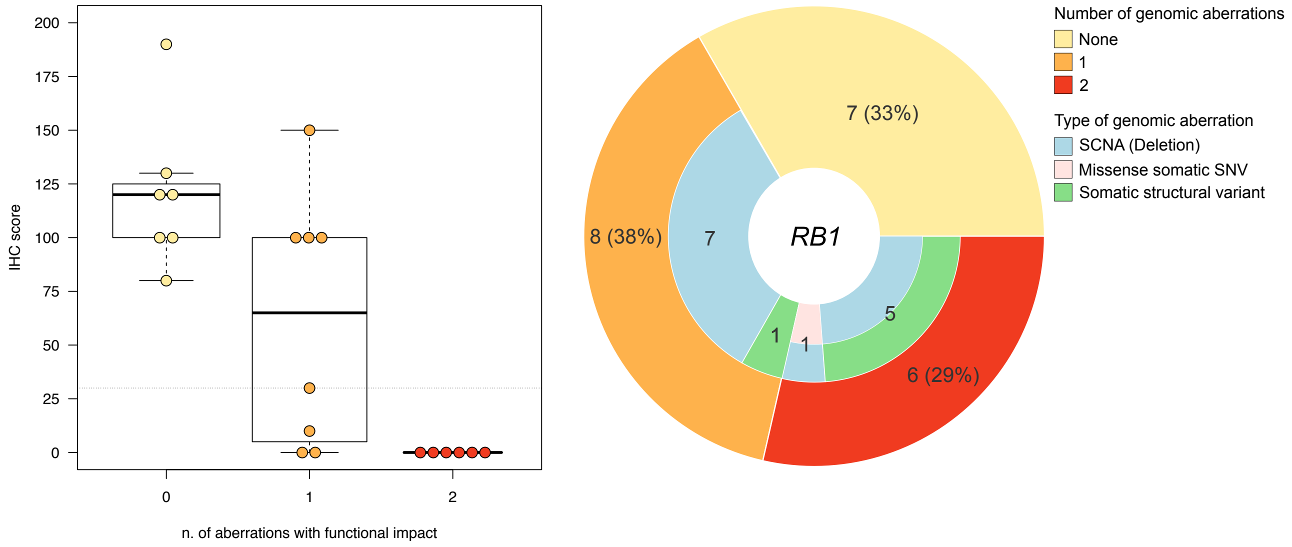


Figure 10. Genomic data explains Rb protein levels*.* (Left) Immunohistochemistry Retinoblastoma protein levels against the number of genomic aberrations at *RB1* genomic locus with potential damaging functional impact*.* (Right) Pie chart summarizing genomic data: the external level indicates the number of samples carrying none, one or two (yellow, orange and red sections, respectively) putative deleterious genomic hits in *RB1*. Inner level specifies the type of *RB1* aberrations (light blue, copy number loss; pink, missense SNV; green, structural variant) affecting samples in each previous section.

METHODS (Section a)

1a. Characterization of germline single nucleotide polymorphisms

Germline single nucleotide polymorphisms (SNPs) are called using *HaplotypeCaller* tool from GATK package. The *HaplotypeCaller* is capable of calling SNPs via local de-novo assembly of mapped reads in a region demonstrating signal of variation from existing mapping information. This allows the tool to be more accurate at complex sites, for instance containing different types of variants close to each other. The web-based ANNOVAR tool (wANNOVAR)[93] was used to annotate functional consequences of detected germline genetic variants. Effects of coding non-synonymous variants on protein function and carrier's phenotype were estimated by considering the functional importance scores predicted by the SIFT algorithm (included in wANNOVAR output)[94]. The SIFT prediction score ranges from 0 to 1 and corresponds to the scaled probability of an amino-acid substitutions being tolerated. Amino-acid substitutions with scores below 0.05 are predicted to affect the protein function.

2a. Check normal-tumor pairs consistency

In order to surmise whether or not the matched samples originate from the same patient, genotypes of 334 high MAF SNPs, selected to be well represented also on most WES platforms, are computed using and the genotypic distance of these SNPs are calculated for each sample using the SNP panel identification assay (SPIA)[95]. These 334 SNPs are chosen such that the genotypes of the 334 SNPs should be very similar in paired tumor/control samples that originate from the same patient versus paired samples that originate from different patient. For a negative control the genotype distance between the paired tumor/control samples with a random sample that originates from a different patient is also computed.

3a. Detection of somatic single nucleotide variants in exons

To identify and characterize somatic single-nucleotide variants (SNVs) in exons, MuTect[96] from the Broad Institute Genome Analysis Toolkit is applied; MuTect uses Bayesian statistical analysis to nominate putative SNVs upon coverage, allelic fraction, and base-qualities information. In order to reduce false positives, calls are refined using a pileup approach using ASEQ[97]. Stringent filtering quality criteria were applied and for each tumor sample are retained only those SNVs for which no reads supporting any alternative allele at the genomic locus in the matched germline sample were present. Using the same ad-hoc pileup approach, a SNV is considered private of a tumor sample if and only if no more than two reads supporting an alternative allele are observed at the same genomic position in other tumor samples of the same patient. Finally, each SNV is annotated with genomic features and effect predictions using SnpEff software[98].

4a. Genomic structural variations

For each tumor and matched normal sample WGS data, the BreakDancer algorithm has been run with default parameters to detect genomic structural variations[99].

The raw output of BreakDancer was filtered using the following criteria:

- Quality score ≥ 95;
- Minimum number of reads supporting the structural variant ≥ 30;
- At least one of the two break points of the structural variant located within a coding region of a gene of interest;
- Retain structural variants identified as inter-chromosomal (CTX) and intra-chromosomal translocations (ITX).

All results are visually inspected using the Integrative Genome Viewer (IGV). Specifically, sequencing reads are coloured "*by pair orientation" or "by insert size"* to flag anomalous pair orientations or insert sizes, respectively.

5a. Estimates of somatic copy number alterations

For WGS data, the reference genome is partitioned by using the BICseq algorithm[100] using default parameters except for *lambda* (set to 50) and *bin size (set to* 1000) when to accommodate for high mean coverage of the study data. Somatic copy number alterations (SCNA) in WES samples are instead identified from the read count based EXCAVATOR software[89] with default parameters and somatic mode which performs a pairwise tumor-normal comparison for each case by minimizing systematic biases, such as guanine-cytosine (GC) content, mappability, and exon length (EXCAVATOR software was selected based on results from tools comparison analysis described in section 2). Thus, each segment data generated using BICseq and EXCAVATOR is represented by the log2 of the ratio between values proportional to the tumor and normal local coverage within the genomic segment.

6a. Ploidy and purity

Segmented data generated using BICSeq and EXCAVATOR were used by CLONET[8] to a) estimate ploidy and purity for each tumor sample; b) adjust log2 ratios for tumor ploidy and purity; c) to determine the copy-number landscape through allele-specific copy-number analysis as previously described[65].

7a. Sequencing data

Study patient's tumor biopsies from metastatic sites (N=21) and matched germline samples were profiled with whole genome sequencing (WGS) protocol. Available specimens from primary tumor sites (N=4) from 3 patients were profiled with whole exome sequencing (WES) protocol using Agilent Sure Select Human All Exon V4 kit. Paired-end sequencing reads generated from both protocols were aligned using Illumina Isaac aligner[101] to the human reference genome (GRCh37/hg19). The Genome Analysis Toolkit (GATK)[102] best practices for variant calling that include marking of duplicate reads, recalibration of base quality scores and local realignment were adopted for all aligned samples.

8a. Immunohistochemistry

A mouse monoclonal anti-RB1 antibody (Clone G3-245, BD Biosciences, San Jose – CA, USA) was selected for this study. For validation of specificity, protein lysates were isolated from 22RV1 and MDA-MB-468 cell lines and run on western blot for positive and negative controls, respectively. Immunohistochemistry (IHC) staining was performed using conventional diaminobenzidine method using cell pellets of the aforementioned cell lines as positive and negative controls. Briefly, FFPE samples were cut at 4um thick sections onto superfrost glass slides and heat based antigen retrieval was performed by boiling slides in a pressure cooker at 125ºC for 2 minutes then 90°C for 1 minute in a pH 6 citrate buffer solution. Endogenous peroxide was blocked using a 3% H2O2 solution. Non-specific staining was blocked using Dako protein block serum-free X0909. RB1 staining was semi-quantitatively assessed by means of an H-Score determined by the formula: (% of weak positivity)x1 + (% of moderate positivity)x2 + (% of strong positivity)x3, yielding a result between 0 and 300 [103]. Immunohistochemistry was performed and evaluated at ICR by Dr. Daniel Nava Rodrigues.

9a. Fluorescent in situ Hybridisation

Fluorescent in situ Hybridisation (FISH) was performed to determine copy number status of *RB1*. Dual color FISH assay was optimized; commercially available FISH probes for 13q14, ~202 Kb locus spanning *RB1*, and 13q34, ~612 Kb locus in the subtelomeric region of 13q were used (Abbot Laboratories, Lake Bluff, IL-USA). Up to 50 intact non-overlapping nuclei were counted per sample and the number of cells with >2, 2, 1, or 0 signals was recorded for both probes. FISH was performed at evaluated at ICR by Dr. Daniel Nava Rodrigues.

b. Cell free DNA in plasma samples

To overcome limitations due to intra-patient tumor heterogeneity in the clinical setting, here I report studies aimed to characterize the genomics of metastatic prostate cancer through “liquid biopsy”, i.e. plasma form patient’s blood samples. Cell free DNA (cfDNA) released in blood stream from widespread metastatic cells is exploited as alternative to multiple metastatic tissue biopsies that are unpractical due to complications and pain[104–108]. This approach could provide the full landscape of all tumor lesions present at a certain time point; that is knowledge is key information to evaluate treatments response/resistance and thus the most effective clinical therapy. Moreover, by exploiting serial sampling and looking for specific somatic aberrations, liquid biopsies can provide the opportunity to survey genetic material potentially representative of multiple metastases and efficiently track tumoral clones evolution. My work in the context of cancer patient cell free DNA quantification and interpretation for the understanding of patient’s heterogeneity ranged from advanced quality control procedure that highlighted generalizable features highly relevant for optimized design of targeted sequencing assay, detection of alternative mechanisms of AR enhancement during abiraterone treatment, to the development of a *per-base error measure* (*pbem*) for local sequencing error to accurately detect single nucleotide variants in highly challenging samples.

4. Quality control assessment of sequencing targeted panel

Recently published works from my laboratory showed the efficacy of the NGS targeted assay to systematically track tumor evolution and evaluate response to treatments by detecting somatic aberrations in circulating-free DNA from advanced castration resistance prostate cancer patients’ plasma[80,109]. Nevertheless, in some cases the biological interpretation of experimental data was extremely challenging due to a non-homogeneous distribution of the sequencing signal across amplicons capturing different genomic regions within the same sample. In particular, the estimation of copy number variations (gain or loss of genomic regions) required the implementation of sophisticated computational methods able to deal with biological signal fluctuations which combined tumoral and control sample’s data. In addition, quantitative analysis showed that the efficacy (that is proportional to the number of sequencing reads aligned on the corresponding genomic region) of a specific amplicon varies across samples making, in some cases, comparisons unconfident and hence requiring preprocessing computational steps able to identify and filter out amplicons demonstrating excessive instability.

To investigate the potential source of both intra-sample and inter-samples amplicon instability, I evaluated a deep sequencing targeted panel used in Carreira *et al.*[109] and in Romanel *et al.*[80] covering 8 genomic regions for a total of about 40 kb using 367 amplicons optimized for IonTorrent Personal Genome Machine (PGM). Specifically, data suggests that this targeted deep sequencing panel was affected, in specific target regions, by sub-optimal depth of coverage both at gene and amplicon level. The aim of my analysis was to understand whether particular intrinsic genomic features or other technical aspects could explain these problems. Thus, I estimated the mappability of target regions, I studied the relationships between GC content and both amplicons length and depth of coverage and I characterized possible undesired drawbacks caused by inaccurate primers design.

4.1 Sequencing coverage and GC content bias

Amplicons included in the targeted panel were stratified based on type of genomic region covered (87 coding, 114 intronic, 166 intergenic) and lengths (7 bins of 10bp range, minimum length is 68, maximum length is 141bp). The median GC content for the entire set of amplicons is 44.1%. Specifically, the median GC content estimated for coding, intronic and intergenic amplicons is 54.3%, 40.7% and 43.6%, respectively. As shown in **Figure 11**, GC content is higher in coding than non-coding (intronic and intergenic) amplicons (p-value = 1.537e-14, Welch Two Sample t-test) and this is in line with previous observations that chromosomal regions of high GC exhibit higher genes density[110,111] and human housekeeping genes contain relatively high GC content and were found to include short introns[112].

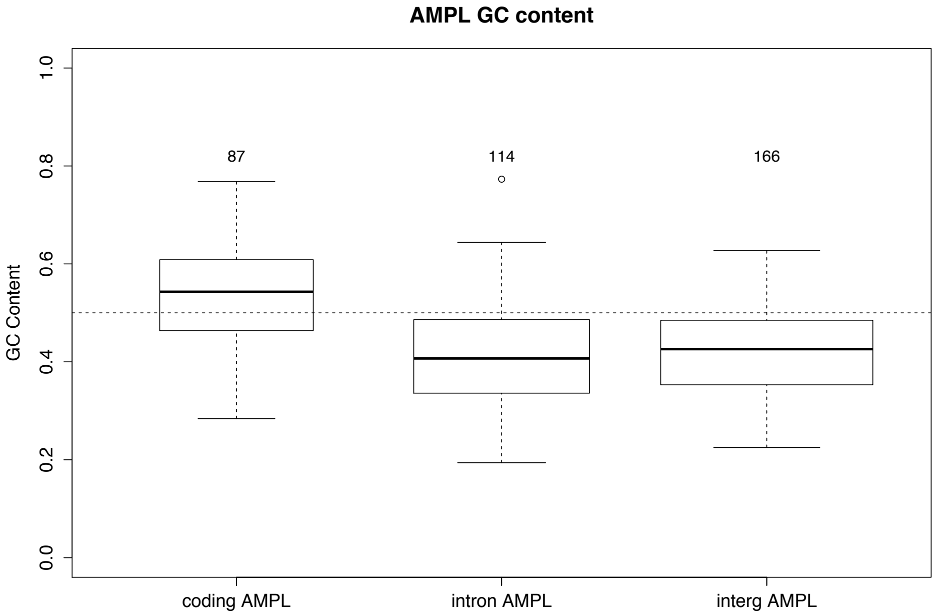


Figure 11. GC content comparison among amplicons covering coding, intronic and intergenic genomic regions. Data show that GC content is higher in coding than non-coding (intronic and intergenic) amplicons (p-value = 1.537e-14, Welch Two Sample t-test).

Next, I investigated the relationship between GC content and amplicon size in each class. Amplicons shorter than 80bp and covering coding regions have a median GC content significantly lower than larger amplicons of the same class and this characteristic is observed also for intronic amplicons. Instead, the size of amplicons covering intergenic regions does not associate with the median level of GC content that is stable around 40% across length bins (**Figure 12**).

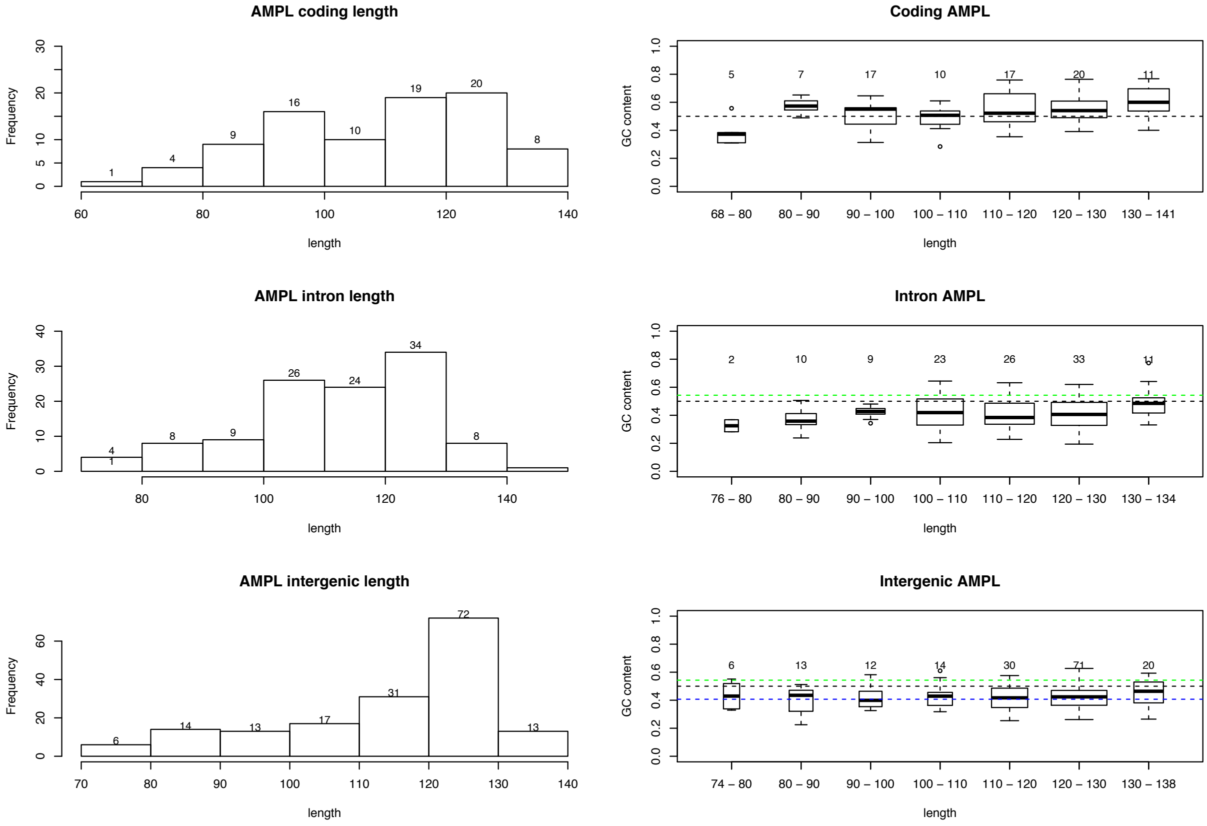


Figure 12. GC content and amplicon size in coding (**A**), intronic (**B**) and intergenic (**C**) amplicons. Barplots on the left show amplicon length distributions for each amplicon class. On the right, boxplots show distribution of GC content stratified by bins of size. Dashed horizontal green and blue lines indicates the median GC content computed across all coding and intronic amplicons, respectively.

To understand if this behavior is peculiar only for the considered amplicons or it is a general genomic sequence property, I performed the same GC content versus size analysis on random genomic regions. Specifically, the GC content distribution observed for each amplicon bin size is compared with the GC content distribution computed in 25K randomly sampled regions with comparable size. This exercise was performed in the three amplicon classes (coding, intronic, intergenic) separately and by performing random sampling of genomic regions belonging to the considered amplicon class (i.e. the GC content computed in *n* amplicons of length in range 90-100bp and targeting coding regions will be compared with the GC content computed in 25K coding regions of 90-100bp randomly sampled genome wide). As a result, designed coding amplicons have generally a median GC content that is comparable with randomly sampled positions whose median is stable around 50% across bins of size considered; unique exception is represented by the bin with the shortest sizes that have a median GC content significantly lower than random one. Supporting the previous observation, both sets of random intronic and random intergenic amplicons have a median GC content that is lower with respect to that observed in coding ones. Intronic amplicons with size in range 76-90bp have a lower GC content than that computed in the corresponding random sets; for larger sizes the two distributions are fully comparable. Finally, the median GC content computed in intergenic amplicons is comparable with that computed in all corresponding random sets across the whole size range (**Figure 13**).

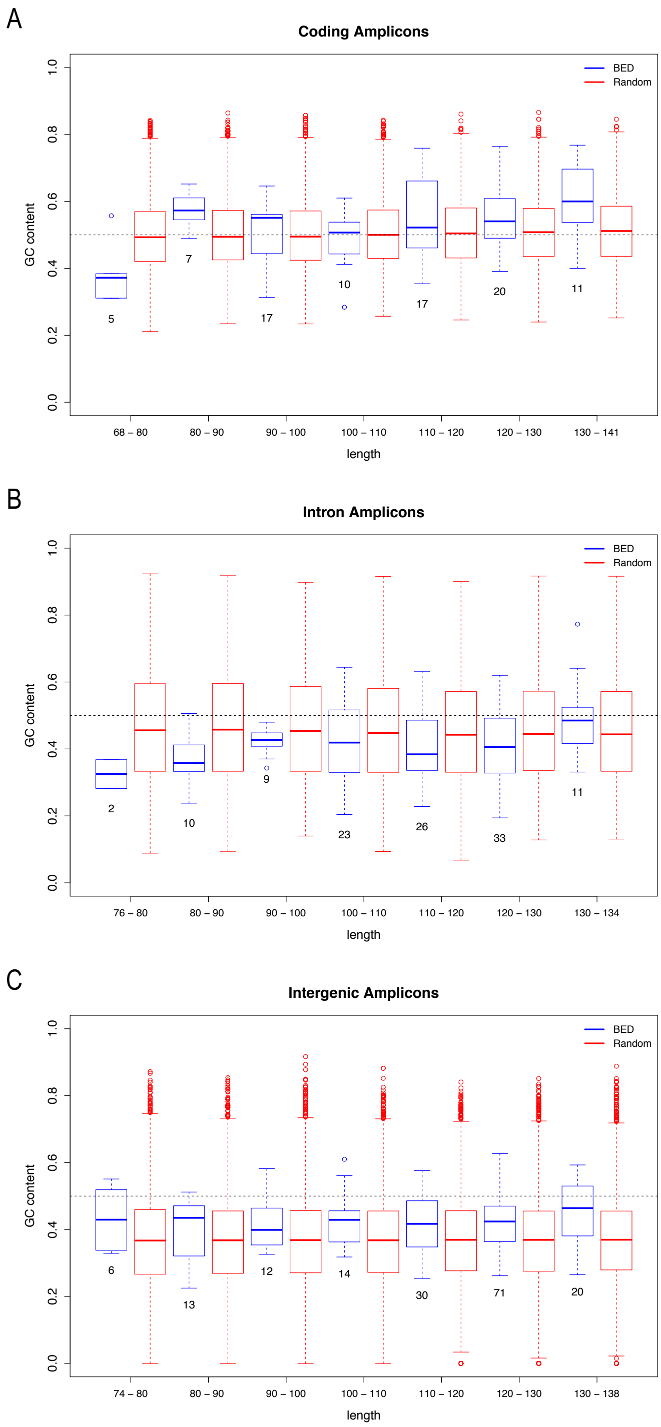


Figure 13. Comparison, stratified by lengths, between GC content computed in designed and random amplicons covering coding (**A**), intronic (**B**) and intergenic (**C**) genomic regions. For each panel, on y-axis is reported the GC content and on the x-axis the bins of lengths by which amplicons are grouped. Blue and red boxplots show the GC content distribution computed in designed amplicons and in a set of 25K randomly sampled amplicons with comparable size and type of genomic region covered.

Effects of extreme GC content are evaluated in terms of sequencing coverage level (**Figure 14**). Indeed, amplicons having a GC content in range 40-60% (N=199) have a median coverage of 1205. Amplicons with GC content lower than 40% (N=132) or higher than 60% (N=36) suffer decrease in their median coverage: 1132 and 631, respectively (p-value = 2.2e-16, Kruskal-Wallis rank sum test). In addition, as shown in **Figure 15**, high GC content levels associate with decreased amplicon coverage stability, hence partially explaining this phenomena for a subset of amplicons.

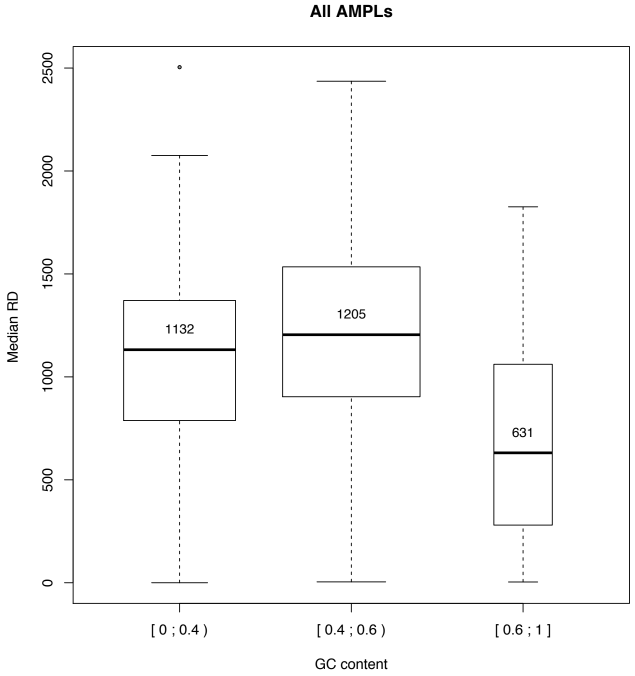


Figure 14. Effects of extreme GC content on median coverage. Amplicons with GC content lower than 40% (N=132) or higher than 60% (N=36) suffer decrease in their median coverage (p-value = 2.2e-16, Kruskal-Wallis rank sum test).

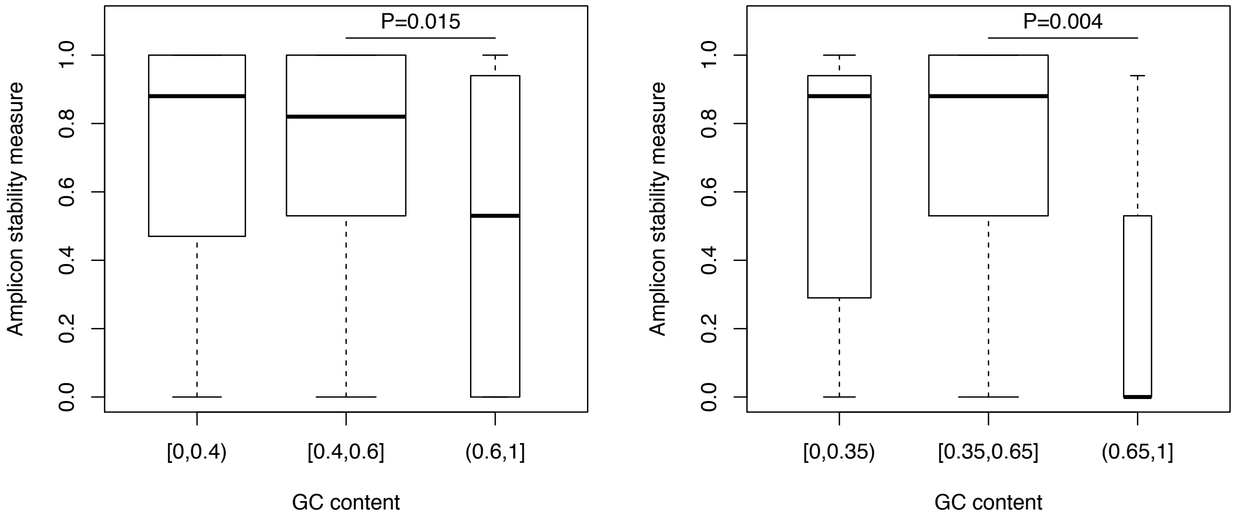


Figure 15. Amplicon sequencing coverage stability and GC content stratified in three main intervals. Two different GC content discretization intervals (left and right panels) have been considered producing the same result.

4.2 Genomic alignment and mappability

Mean mappability of each amplicon sequence was evaluated in terms of alignability and uniqueness measures. The mean alignability is optimal (equal to 1) for most of the designed amplicons except for a relevant fraction of those covering *PTEN* (21 out of 73, 29%) and *FOXA1* (5 out of 14, 36%) genes. These amplicons are also characterized by suboptimal mean values of uniqueness, a more stringent measure of mappability (**Figure 16**). Amplicons covering genes *FOXA1* and particularly *PTEN* are enriched for sub-optimal mappability measures (**Figure 17**). More generally, amplicons characterized by sub-optimal mappability measures result more problematic in terms of sequencing coverage stability (**Figure 18**).

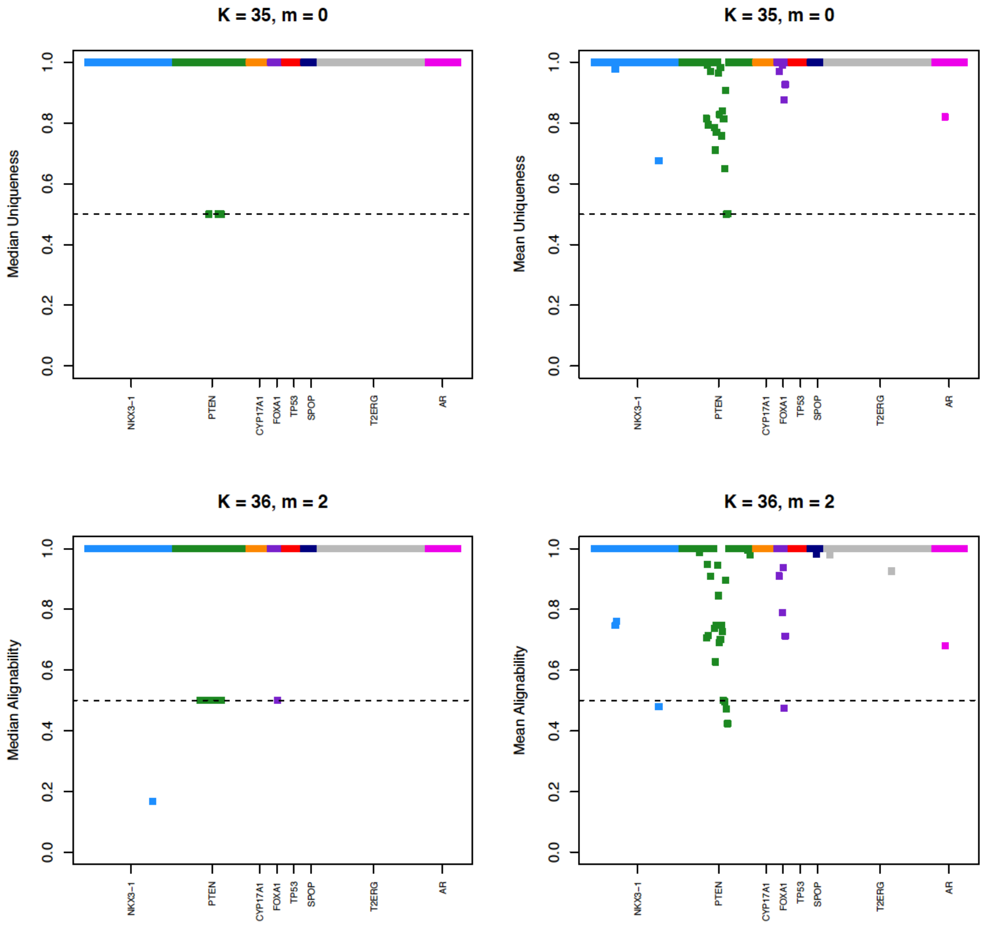


Figure 16. Measures of mappability. Median and mean uniqueness (top) and alignability (bottom) computed in amplicons (each coloured square) covering genes (depicted in different colors) targeted in the assay. Amplicons covering *PTEN* (green) and *FOXA1* (violet) show sub-optimal (lower than 1) levels mappability both in terms of uniqueness and alignability. At top of each panel, parameter *k* indicates, the size of the sliding window used to compute alignability and uniqueness, *m* indicates the maximum number of mismatches allowed during sequence search genome wide.

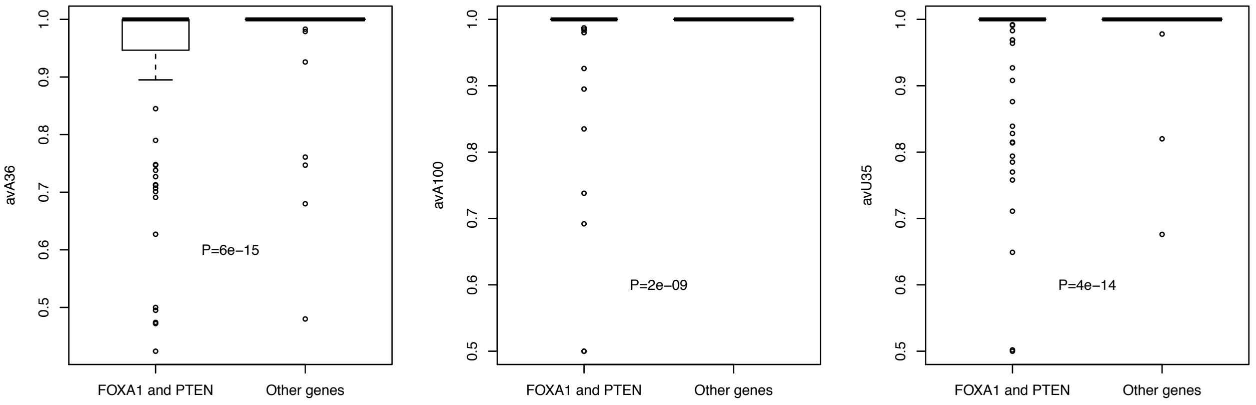


Figure 17. Comparison of mappability in amplicons covering FOXA1 and PTEN versus amplicons covering all other panel genes. From left to right, boxplots report on y-axis mean alignability (computed with parameter k=36), mean alignability (k=100) and mean uniqueness (k=35). In each panel, left distribution is computed in amplicons covering *FOXA1* and *PTEN*, right distribution show mappability distribution computed in all other amplicons.

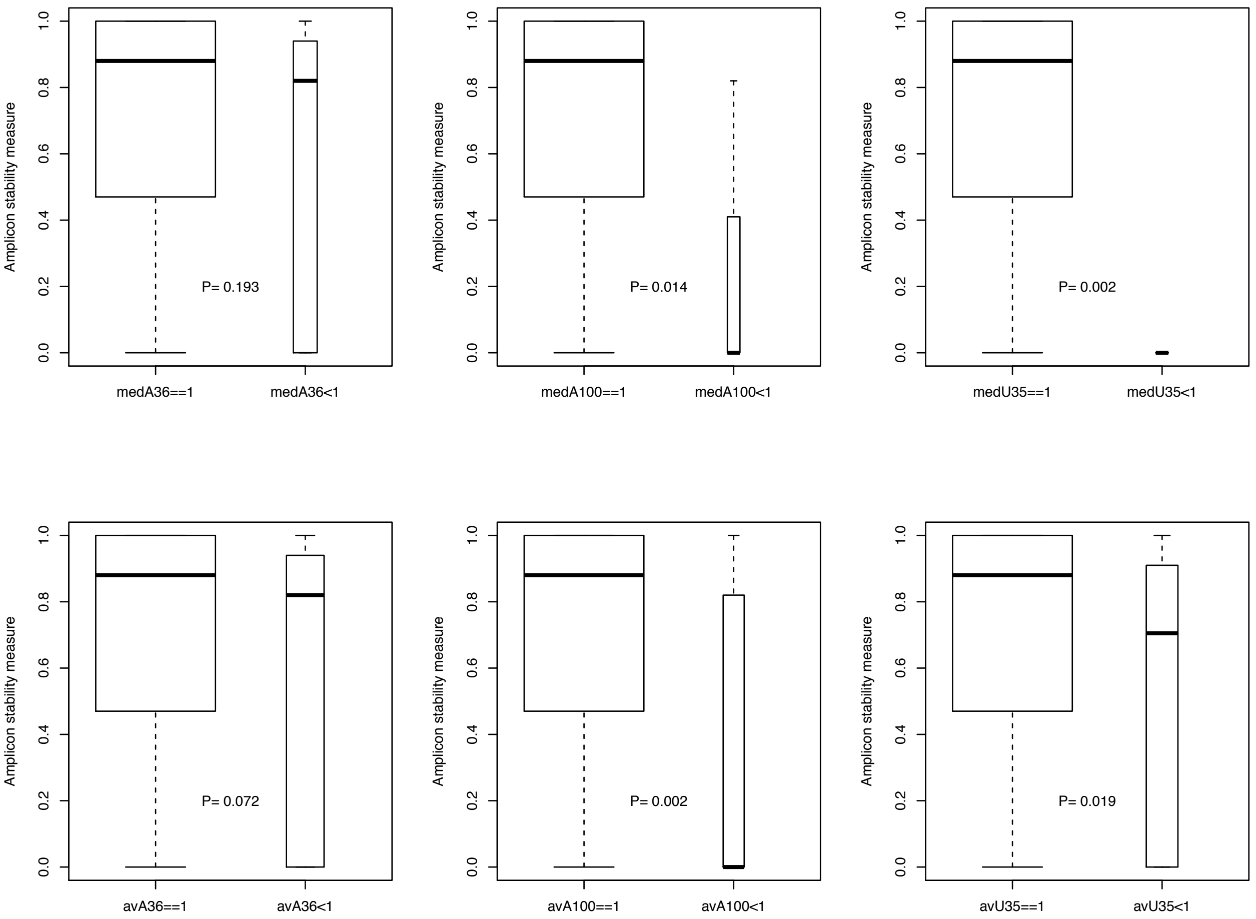


Figure 18. Amplicon stability versus amplicon mappability measures. Distributions of amplicon stability measure (y-axis) compared between optimal (left boxplot in each panel) and sub-optimal (right boxplot in each panel) values of median (top boxplots) and mean (bottom boxplots) mappability measures. From left to right, mappability measures considered are alignability (computed with parameter k=36), alignability (k=100) and uniqueness (k=35).

4.3 Characterization of multiplex PCR primers to avoid or be aware of possible undesired amplification products

I designed a computational strategy to verify if drops in coverage observed in specific target regions can be caused by poor primers design that can compromise their activity during the DNA amplification step by multiplex PCR during library preparation. First, I quantified how much a designed primer is specific to the targeted genomic region by looking for all possible hits (allowing none, one or more than one mismatches) of the studied primer sequence across the genome. Second, since multiplex PCR uses multiple pairs of primers to amplify different genomic regions in a single reaction, I checked if two primers, given their expected or/unexpected hits on the DNA sequence, can sit on the genome at reasonable distance and right orientation to properly work and give rise to undesired PCR amplification products. Indeed, amplicons amplified by primers that have multiple hits along the genome (not specific for a single genomic region) are less represented than expected in the sequencing library and thus their sequencing coverage is negatively influenced. Moreover, undesired PCR products can subtract sequencing reads from targeted regions and decrease their median coverage. Results show that amplicons having both primers that align only to the sequence of interest (the targeted one) have the higher median coverage. Instead, the set of amplicons having one or both primers that can align to sequences different from the targeted one, have lower median coverage.

5. Design of a new custom sequencing targeted panel

Results and computational strategies derived from the in-silico characterization of the targeted sequencing panel used in Carreira *et al.*[109] and in Romanel *et al.*[80] was exploited to design a new amplicon-based assay for targeted resequencing. The novel targeted panel was designed in collaboration with *Illumina* company and based on the TruSeq Custom Amplicon (TSCA) assay. Most of the TSCA's amplicons were designed to cover genomic regions found to be frequently aberrant in both primary and advanced prostate cancers (N=20 genes). Additionally, a set of amplicons is dedicated to cover non-aberrant regions (N=3 genes): this will assure the presence of reference controls regions within the assay that can be used to improve copy number evaluation. The designed custom panel including includes 1161 amplicons designed to cover approximately 105 kb in genomic regions of interest (on-targets) and also 671 potential amplicons that may amplify non-targeted regions (off-targets). All amplicons are summarized in a file provided by Illumina called *TSCA Manifest*.

Taking advantage of strategies developed for estimating tumor purity and lesion hierarchy from whole-genome sequencing[8], the targeted panel aimed to exploit the genetic information of single individuals at heterozygous SNPs, informative SNPs, to computationally determine the fraction of total DNA in circulation that contained common monoallelic deletions. Thus, SNPs at high MAF were covered by amplicons within commonly aberrant genomic regions

(21q, *NKX3-1, PTEN, MYCN, CHD1, MYC, ATM, BRCA2, RB1, MAP2K5, CYLD, FANCA, TP53, BRCA1, AURKA*) and control genes (*HP1BP3, FGFBP2, UGT2B17*). Additionally, all *AR* exons were covered.

Moreover, I designed amplicons to cover missense somatic point mutations in *BRAF, PTEN, ATM, BRCA2, RB1, FOXA1TP53, BRCA1, SPOP, AR* genes previously reported in extended cohorts[56,87]. Amplicons covering regions dense of somatic point mutations of interest, were designed taking into account that the number of point mutations that lie underneath one of the probes is limited. Specifically, probe (both forward and reverse) design can tolerate up to 3 point mutations and they are still functional. The position of the SNVs within the probe does matter: forward probes cannot include SNVs in the last third of the probe and reverse probes cannot include SNVs within the centre of the probes.

As quality control, I estimated the GC content for forward probes, reverse probes and target regions comprised between them. The median GC content values are 43.48%, 45.45%, 41.07%, respectively. The median mappability values are 77.32%, 75.86%, 74.73%, respectively. As expected, I observed that GC content is slightly higher in target regions located in exonic regions respect to intronic and intergenic ones. Similarly, amplicons covering exonic regions are characterized by better mappability values respect to other amplicons. There are 24 target regions (2.5%) that show a GC content higher than 70% and 74 (8%) with a mappability value lower that 30%. There are 17 forward probes (1.8%) that show a GC content higher than 70% and 91 (9.7%) with a mappability value lower that 30%. There are 10 reverse probes (1%) that show a GC content higher than 70% and 95 (10%) with a mappability value lower that 30%. Differently from PGM technology, probes in this design can tolerate very wide range of GC content because of diverse amplification process. Indeed, hybridization and extension-ligation processes are not strict (in terms of annealing temperature and time) as in a typical PCR reaction that allows higher hybridization specificity.

5.1 Comparative analysis of sequencing data generated by two sequencing targeted assays

Explorative analyses were performed using DNA extracted from different sources: plasma and serum of both prostate cancer patients and healthy individuals, tumor biopsies and matched germline controls from prostate cancer patients, four prostate cancer cell lines (DU145, LNCAP, PC3, RWPE) and four HAPMAP samples. All samples were sequenced using Illumina MiSeq. Preliminary results suggest that TSCA-MiSeq output data characteristics are comparable with data generated using the first custom panel sequenced on the IonTorrent PGM. In particular, TSCA technology does not provide evident improvements in terms of amplicon signal stability.

5.2 Qualitative and quantitative assessment of plasma and serum samples using TSCA panel

Overall statistics on sequencing data were performed for qualitative and quantitative assessment of plasma and serum samples from low-grade prostate cancer patient.

The library prepared with cfDNA extracted from plasma sample was sequenced in 3 different sessions. The total number of reads generated per plasma sample in each session was 1205858, 1085458 and 1642594. Plasma samples showed an average fraction of mapped reads and properly-paired reads of 0.78 and 0.74 indicating good quality reads and successful alignment procedure. Plasma samples mean coverages in the three sessions were 754.13, 651.49 and 977.14. In each session, plasma samples have about 5% of the reads mapped in off-targets and this value is in line with expected fraction.

DNA extracted from serum sample was used included in 2 libraries. The first library was sequenced twice (first and third sequencing session) and the second library once (second sequencing session). Both library preparations provided comparable sequencing outputs: mean coverages of library one were 1050.32 and 1155.61, 942.52 for library two; the fractions of mapped reads (and properly-paired reads) were 0.82 (0.78) and 0.79 (0.76) for library one and 0.80 (0.75) for library two. The two libraries slightly differ in the fraction of reads mapped in off-targets of Manifest: 5% and 9% for library one and two, respectively.

In order to obtain a unique BAM file for plasma and serum samples, reads generated in the three sequencing sessions were merged together. Merged plasma and serum samples reached a mean coverage of 2382.76 and 3148.46. With respect to the total number of reads, more than 77% are successfully mapped and more than 73% are mapped as properly-paired in both sample types. Reads in off-targets account for approximately 6% of the total filtered reads mapped to regions in Manifest file.

6. Plasma *AR* and abiraterone-resistant prostate cancer

Androgen receptor (*AR*) gene aberrations are rare in prostate cancer before primary hormone treatment but emerge with castration resistance. To determine *AR* gene status using a minimally invasive assay that could have broad clinical utility, in collaboration with Dr. Gerhardt Attard at the Royal Mardsen, we developed a targeted next-generation sequencing approach amenable to plasma DNA, covering all *AR* coding bases and genomic regions that are highly informative in prostate cancer. Here I present part of results from the analysis of 274 sequenced plasma samples from 97 castration-resistant prostate cancer patients treated with abiraterone at two institutions[80]. cfDNA in patients' circulation was analyzed and for 217 samples (80 patients) there was sufficiently high tumor DNA fraction to quantify *AR* copy number state (tumor DNA fraction above 7.5%) (**Figure 19**). Through computational analysis the genomic status of *AR* in terms of somatic point mutations and somatic copy number was determined for all considered tumoral patient's samples.

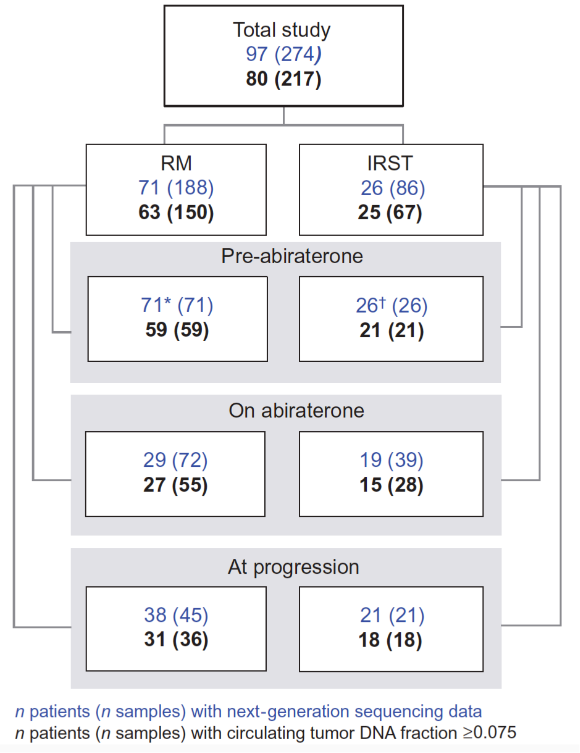


Figure 19. Study profile showing the number of patients and samples with next-generation sequencing data and with a circulating tumor DNA fraction ≥ 0.075. Twenty-six patients (*) and one patient (†) had pre-abiraterone samples only.

6.1 Mutant *AR* alleles do not acquire copy number gain

We detected (*Methods* section 4b) somatic *AR* non-synonymous point mutations described recently in sequencing studies of CRPC tissue[56] in 41 plasma samples (15%) from 16 patients. W742C and W742L *AR* mutations were observed in the same sample collected prior to initiation of abiraterone in a patient who had progressed on and discontinued bicalutamide 36 days previously. L702H was only observed in patients (five) receiving prednisolone. The L702H, H875Y and T878A mutations were validated using digital droplet PCR (**Figure 20**).

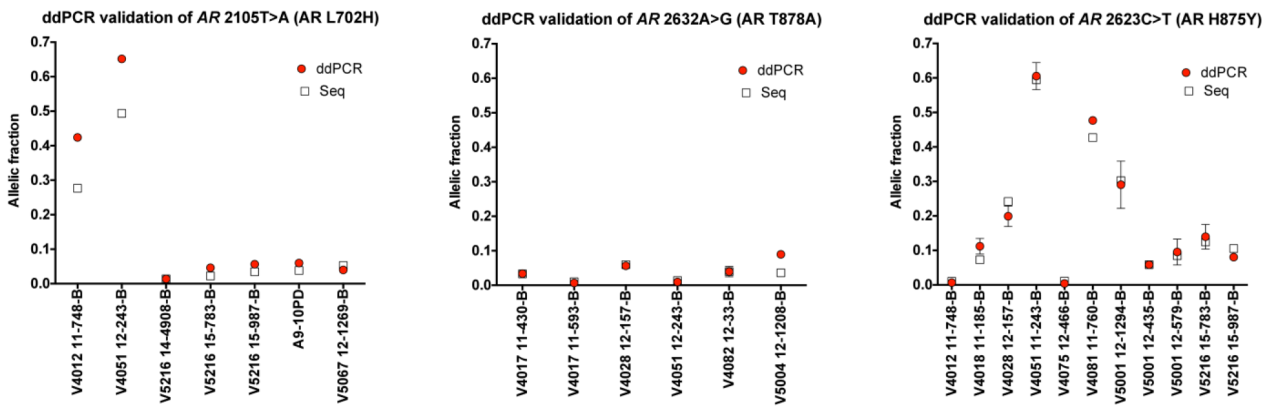
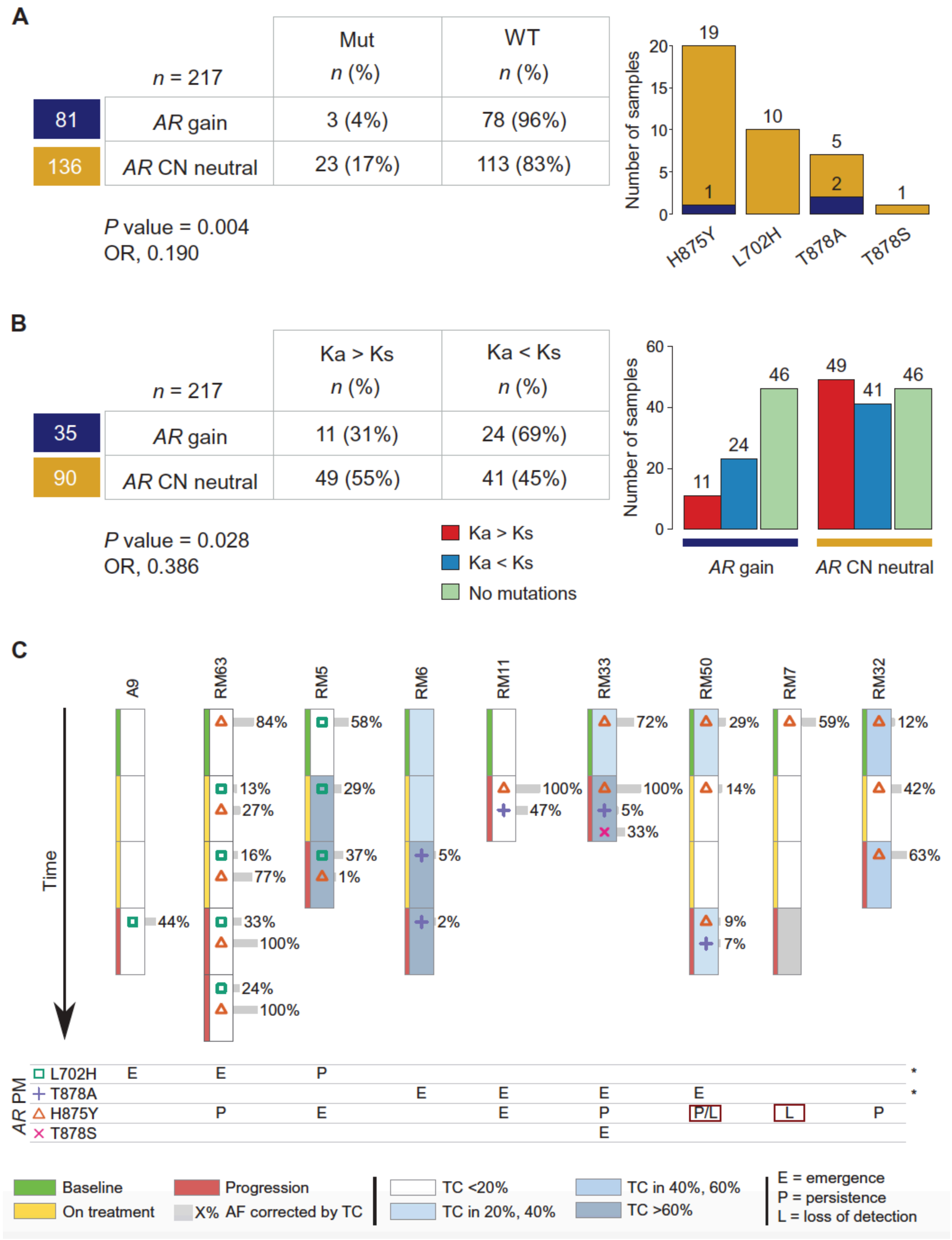


Figure 20. The allelic frequencies of the AR mutations 2105T>A (L702H), 2632A>G (T878A) and 2623C>T (H875Y) were determined with ddPCR as a validation of the sequencing estimation. For samples with sufficient input DNA (6ng) PCR reactions were setup in duplicate with primer/probe mixes detecting either the wild type and mutant allele for L702H (top panel), T878A (middle panel) and H875Y (bottom panel). Based on the number of droplets positive for the mutant or wild type allele the allelic frequency was calculated (red circles) and compared to the sequencing estimation (white squares).

Amongst samples with detectable DNA fraction, we observed a significant inverse correlation between detection of *AR* copy number gain and *AR* point mutation (**Figure 21A**) and no instances where the fraction of reads suggested gain of a mutant *AR* allele. As we had sequence data on all the bases in coding regions of the *AR*, we proceeded to identify a significantly higher rate of non-synonymous with respect to synonymous *AR* point mutations in the samples with no *AR* gain compared to those with gain, supporting selection of non-synonymous mutations in the absence of gain (**Figure 21B**). To identify *AR* point mutations that specifically associate with resistance to abiraterone, we selected lesions that were consistently detected and showed an increase in circulating abundance with disease progression. We included 59 patients with both baseline and progression samples. *AR-L702H* (three patients) and *AR-T878A* (four patients) were the only two mutations that met these criteria (**Figure 21C**). Both mutations are activated by non-androgenic ligands present at increased levels in patients treated with abiraterone[8,109]. Overall, we observed emergence of T878A or L702H AR amino acid changes in 13% of tumors at progression on abiraterone.



A

	n = 217	Mut n (%)	WT n (%)
81	AR gain	3 (4%)	78 (96%)
136	AR CN neutral	23 (17%)	113 (83%)

P value = 0.004
OR, 0.190

B

	n = 217	Ka > Ks n (%)	Ka < Ks n (%)
35	AR gain	11 (31%)	24 (69%)
90	AR CN neutral	49 (55%)	41 (45%)

P value = 0.028
OR, 0.386

C

Figure 21. *AR* gain in non-mutant *AR* alleles. **(A)** Distribution of AR point mutations in all samples, stratified by AR copy number (CN) status. OR, odds ratio. **(B)** The prevalence of nonsynonymous (Ka) and synonymous (Ks) substitutions in *AR* gain and AR CN neutral samples. Fisher's exact test

was applied to test differences between the number of mutated (Mut) versus wild-type (WT) samples across AR gain and AR CN neutral (A) and nonsynonymous versus synonymous substitutions in AR gain versus AR copy number neutral samples (B). Fisher's exact test was applied to test differences between the number of mutated (Mut) versus wild-type (WT) samples across AR gain and AR CN neutral (A) and nonsynonymous versus synonymous substitutions in AR gain versus AR copy number neutral samples (B). **(C)** Presence of AR point mutations (PM) in serial plasma samples from study patients. For every patient, the temporal pattern of mutation detection is shown, distinguishing baseline (green), on-treatment (yellow), and progression (red) samples, along with fractions of circulating tumor DNA (TC). Mutations are marked with different colors and symbols, and the corresponding allelic fractions (AF) corrected for tumor DNA fraction are reported. Temporal patterns observed for each specific patient/mutation combination are annotated as emergence (E), persistence (P), or loss of detection (L and marked with a red box). Stars are used to mark AR point mutations that are consistently detected with disease progression. Corresponding to original Figure 2 in manuscript Romanel *et al.* [80]*.*

7. Improved detection of somatic point mutations in circulating free DNA

Building on data features I observed during the assessment of mutant *AR* alleles via sequencing targeted assay, I recognized the need for a dedicated computational method that combines genetic knowledge and empirical signal to readily detect and quantify somatic point mutations in cfDNA by fully exploiting single base resolution information from targeted sequencing data using patient's plasma (case) and matched germline sample (control). I will here describe the main steps of the approach ABEMUS (Adaptive per Base Error Model for Ultra-deep Sequencing data) and then present the results I obtained on WES data of 36 CRPC patients samples from the *Caryl and Israel Englander Institute for Precision Medicine* (New York Presbyterian Hospital-Weill Cornell Medicine).

First, each targeted base is genotyped in controls to build the allelic fraction distribution necessary to determine the cut-off to call a somatic point mutation with a desired specificity in plasma samples. Analyses showed that the estimation of this threshold is more stable by increasing the number of controls and furthermore its value significantly varies according to coverage levels (**Figure 22**).

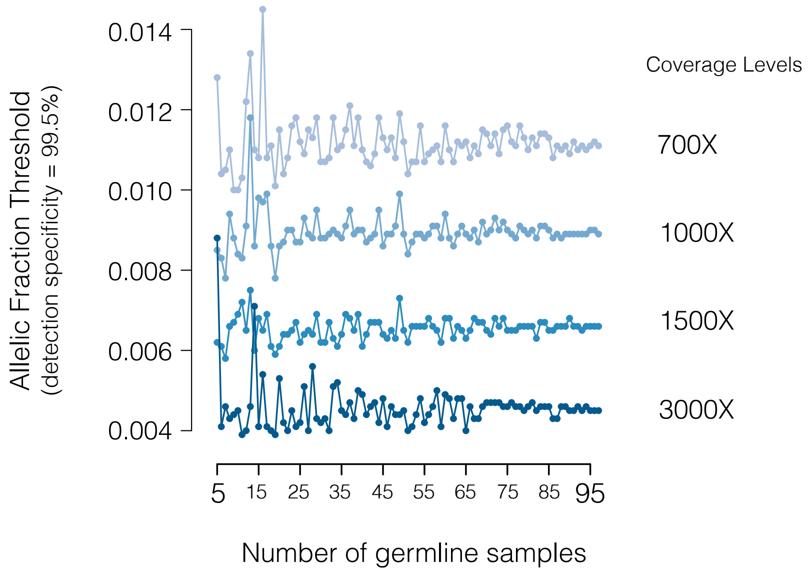


Figure 22. AF thresholds (computed at 99.5% detection specificity) vary across coverage levels (blue lines) and are getting more accurate by increasing the number of germline samples used (x-axis) for the computation.

Second, control samples are exploited to build a genomic locus-specific error model to estimate the probability that observed case allelic fraction is indeed evidence of a somatic event. Specifically, this error model is computed for each locus as the ratio between the number of reads supporting alleles different from the reference one (alternative allele) and the total coverage across all control samples.

Third, plasma samples are analysed and only targeted positions passing stringent custom filtering criteria are retained for next analysis. Indeed, by exploiting results computed in the first step, the combination of desired detection specificity and locus coverage gives the most appropriate allelic fraction cut-off to apply to the interrogated position. Loci passing also the ad hoc allelic fraction filtering constitute the set of putative somatic point mutations.

Fourth, additional filtering analyses based on the computed base-error model and aimed to mitigate effects of potential experimental biases (i.e. strand bias effect) are used to end up with a final set of putative somatic point mutations.

Next paragraphs are dedicated to a detailed description of the method.

7.1 From BAM file to per-base pileup

NGS sequencing files (BAM files) are read in order to extract information available for genomic positions covered by a sequencing assay. Currently, the here presented strategy is optimized for the analysis of sequencing information ranging from targeted panels (in the order of kb) to whole exome sequencing assay (in the order of Mb). The pileup format (PILEUP) describes the base-pair information at each genomic locus interrogated. Proper input for this computational workflow requires that the PILEUP must have the following information:

a. Chromosome, genomic position and Reference allele (In this study, the assembly GRCh37/hg19 of the Human Genome was used for all the analyses);
b. Flag if the locus is a known single nucleotide polymorphism as reported in the dbSNP catalogue (In this study, the release 144 was used for all the analyses);
c. Coverage, that is the total number of reads spanning the genomic locus;
d. Number of reads supporting each of the 4 possible bases at that genomic locus;
e. The variant Allelic Fraction (AF), in case of reads supporting an allele different from the reference one. AF is computed as the ratio between the number of reads supporting the most represented alternative base and the total number of reads covering the position. Whether two or more alternative bases have the same coverage, no AF is computed.

During the pileup process, quality filtering criteria were applied to ensure reliability of downstream analyses. Specifically, for each position only reads showing both read quality and base quality values greater than 20 were considered. Moreover, all genomic positions showing at least one read supporting a base different from the reference one were saved also in a separate file, indicated as SNVS format. In addition to information saved in the PILEUP format, the SNVS format reports following information:

a. The putative alternative allele, that is the alternative base with the highest coverage;
b. For each base, the number of reads covering the forward and reverse strand. This information is used to compute strand bias.

To efficiently scale the computational process, both PILEUP and SNVS formats are generated split by chromosomes. Although ad-hoc in-house tools were used to provide

PILEUP and SNVS files from BAM files, any other NGS pipeline providing outputs in the required formats can be used.

7.2 Annotations of targeted genomic regions

The second required input is the list of genomic regions, covered by the sequencing assay, provided as BED format (each entry indicates a genomic range). Since the aim of this strategy is to characterize single loci, the BED input is processed to obtain a file where each entry indicates a single genomic locus. Whether an entry is a known single nucleotide polymorphism this will be flagged with the corresponding identifier reported in the dbSNP catalogue. Each entry is annotated with values (min = 0, max = 1) of mean uniqueness, mean mappability and GC content computed for the target region (default setting) in which the entry is located.

7.3 Retrieve information at each genomic locus across germline samples

In this step of the computation, the workflow considers data from normal-germline samples only.

For each targeted genomic position, corresponding PILUEP information are extracted from the set of germline sample and collected together. For the sake of clarity, a temporary output table is generated for each position interrogated in the sequencing assay; each entry of this table reports PILEUP information of that entry computed in each of the *N* germline samples considered. These data are exploited to build the overall distribution of variant allelic fractions observed in the set of germline samples (global sequencing error estimation) and to compute a locus-specific measure that indicates the probability of observing in that position a sequencing read supporting an allele different form the reference one (local sequencing error estimation).

Operatively, the collection process is performed in parallel by distributing on *k* threads the step of extracting M loci of interest across germline sample PILEUPs.

7.3.1 Collect AFs stratified by bins of coverage for global sequencing error estimation

As described in PILEUP format, each locus is characterized by a value of coverage and a value of allelic fraction (AF is zero if there are not reads supporting alternative alleles). To increase true positives (here intended as real sequencing errors), loci that were flagged as previously reported germline SNPs or having an AF higher than 10% will not be considered as element of the overall variant AF distribution. Indeed, this two information suggest that the position in the normal sample is a germline variant and not an artefactual error. AFs measured in all retained positions are saved accordingly with their coverage level in pre-defined bins. Reasonable bins of coverage can be specified by the user depending on the median coverage of germline BAMs.

7.3.2 Per-base error measure (pbem) for local sequencing error estimation

For each locus, the *pbem* is computed as the ratio between the total number of sequencing reads supporting an allele different from the reference one in that locus and the total coverage of that locus across germline samples (**Figure 23**). The *pbem* is computed also in positions that were flagged as germline SNPs and the AF cut-off to retain or not the considered position in a germline sample can be modulated.

A more allele-specific version of the *pbem*, indicated as *pbem_allele* is computed for each locus as the ratio between the total number of sequencing reads supporting a specific alternative base (i.e. the base A) and the total coverage of that locus across germline samples.

Besides the *pbem* and *pbem_allele,* additional information computed using the germline set is saved for each position:

a. Locus total coverage;
b. Total coverage of each base;
c. Number of germline sample in which the considered position has enough coverage (default min coverage = 10) to be included in the computation;
d. Number of germline samples in which the considered position has an allelic fraction higher than a specified AF cut-off (default 10%).

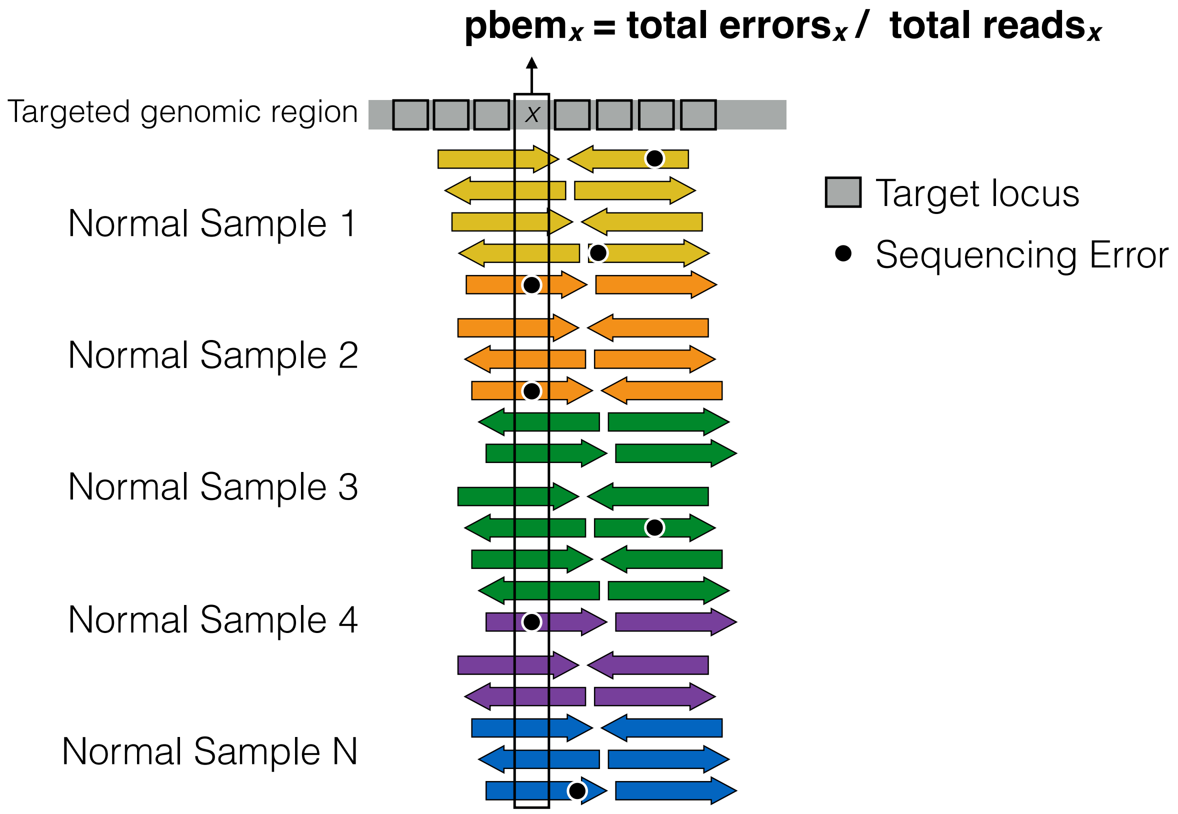
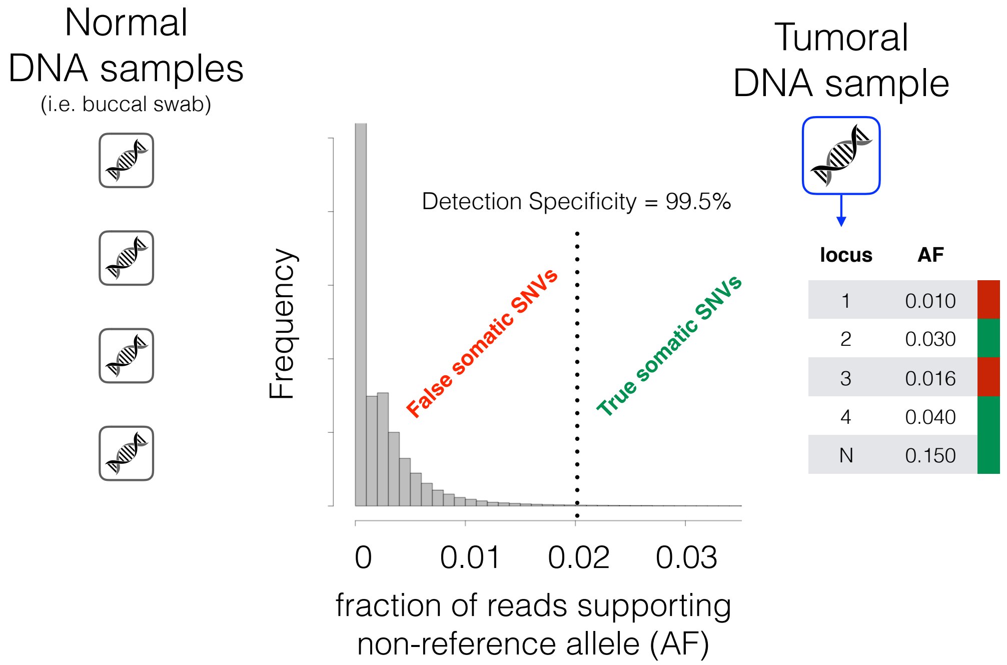


Figure 23. Sketch showing the computation of the per-base error measure (pbem) by exploiting mapped sequencing reads (coloured arrows) collected from a set of germline samples.

7.4 Assessment of coverage-dependent and independent AF thresholds

The overall AF distribution, built as described in section 7.3.1, is used to estimate the most suitable threshold to discriminate between sequencing errors (false positives) and somatic point mutations (true positives). Once a desired level of specificity is decided, the AF threshold represents the corresponding quantile within the overall AF distribution (coverage independent AF threshold). Additionally, since AF were stratified also by coverage levels, the workflow provides AF thresholds by considering AF distributions at each coverage bin (coverage independent AF thresholds) (**Figure 24**).



locus	AF	
1	0.010	🟥
2	0.030	🟩
3	0.016	🟥
4	0.040	🟩
N	0.150	🟩

Figure 24. Germline data-driven threshold to accurately call somatic SNV.

7.5 Calling somatic point mutations in tumour samples

Here the workflow analyses tumour samples with matched normal-germlines.

For each tumor samples, its SNVS file is read and the list of positions found is extracted from the PILEUP of the matched germline sample. This operation allows to fully compare information available for a considered position both in tumor and matched germline samples. Two consequential filtering steps are then performed.

First, a somatic putative SNV is filtered out if the AF of the corresponding position in the matched germline sample has an AF lower than a settled threshold (default is 0) or there are less than *x* sequencing reads (default is 0) supporting an alternative allele.

Second, filtering criteria is applied to the AF of the somatic putative SNVs. The method allows to use alternatively a user-defined threshold, the coverage-independent or -dependent AF thresholds (both computed as described in section 7.4). In the first two cases (user-defined and coverage-independent thresholds) the same AF cut-off is applied indistinctly to all putative somatic SNVs. Otherwise, the coverage in the tumor samples of the considered somatic SNV is retrieved and the AF threshold computed using germline positions in the corresponding coverage bin is applied. As a result, putative somatic SNVs with different coverage will be examined using different AF thresholds.

Outputs of this step are two tables listing putative somatic SNVs passing only the first or both the filtering criteria.

7.6 Additional refinement of the set of putative somatic point mutations

Both lists of putative somatic SNVs generated at step 5 are annotated with information computed as described in section 7.3.2. Specifically, AF in matched germline, *pbem* and *pbem_allele* allow to define 5 classes of putative somatic SNVs (Table 2).

Table 2. Classes of putative somatic SNVs based on AF observed in matched germline, pbem and pbem_allele values.

	Computed on matched germline	Computed on a set of germline samples	
CLASS	Allelic Fraction in germline	Per-base error Locus specific (*pbem*)	Per-base error Allele specific (*pbem_allele*)
1	0	0	0
2	0	> 0	0
3	0	> 0	> 0
4	> 0	> 0	0
5	> 0	> 0	> 0

Putative SNVs for which none alternative alleles are observed across all germline samples (matched normal included) belong to class 1. This class is the most reliable. Class 2 and 3 indicate that there are no evidences of alternative alleles in the matched sample but these are observed in the whole set of germline sample. More precisely, class 2 indicates that the alternative allele of the putative SNV observed in tumor is never found in one or more non-matched germline samples, conversely class 3 specifies that the somatic alternative allele is found in the set of non-matched germline samples.

Class 4 indicates, since the pbem_allele is 0, that the alternative allele observed in tumor and in matched germline is different. Class 5 indicates that the same alternative allele can be observed both in plasma and in matched germline or other germline samples show that alternative allele. Summarizing, the higher the class the higher the probability of observing a systematic error or a germline SNPs, and not a somatic SNV, in tumor sample. Finally, the computational workflow outputs a final table containing the refined set of putative somatic SNVs and functionally annotated using the software tool Annovar [93].

7.7 DNA damage as cause of sequencing errors

Costello *et al.* [113] observed that as a consequence of oxidative stress generated by acoustic sharing of DNA during sequencing library preparation step, DNA lesion 7,8-dihydro-8-oxoguanine (8-oxo-dG) can emerge. As main consequence, the 8-oxo-dG is often missed by a polymerase as a Thymine (T) instead of a Guanine (G) and analysis of both tumor and normal samples evidenced an enrichment of Cytosine>Adenine and Guanine>Thymine transversion genomic variants occurring at low allelic fraction in targeted capture data. Moreover, a recent study [114] described the impact of this bias in confounding variant identification in extend public genomic datasets. Based on these observations, I decided to include in the workflow a step in which this source of artefactual bias can be quantified.

7.7.1 Assessment of alternative alleles occurrences

For each patient, genomic loci in germline samples having at least one read supporting an allele different from the reference one are studied. In these positions, the occurrence of each alternative allele observed and the occurrence of each reference-to-alternative change are quantified. The same count is performed in the matched plasma samples. Together, these data allow for the comparison of the probability to observe a specific base as an alternative and a certain transition/transversion event. This assessment can be used to check for DNA damage status looking, for example, at levels of C>A and G>T transversions. Additionally, putative somatic point mutations can be evaluated in terms of how frequent is to observe that allele and that reference-to-alternative change across all the positions showing one alternative allele. For example, in a plasma sample a putative somatic point mutation is observed as transition C>T. Based on the hypothesis that all transitions/transversion events occurs at the same frequency, if the frequency of the C>T event is significantly overrepresented with respect to all other transition/transversion events in plasma and/or germline sample, the considered variant is more likely to be a false positive. All identified putative somatic point mutations are hence annotated indicating which of them support transitions/transversions that are overrepresented in the corresponding sample.

8. The *pbem* is a sequencing platform dependent feature

I formally tested the hypothesis that sequencing errors, quantified using the *pbem,* depend on the experimental platform used (platform is here intended to include both the library preparation kit and the machine/chemistry adopted to sequence a DNA sample). To test this hypothesis, I collected normal samples that have been profiled using different platforms as reported in **Table 3**. Thus, I first exploited a set of 113 germline samples sequenced on IonTorrent PGM using a custom targeted panel covering a total of 40 kbp. Samples were randomly divided into two random subsets including 56 and 57 samples respectively. I computed the *pbem* for all targeted genomic loci independently using the two subsets of normal samples. As a result, each targeted locus is characterized by two values of pbem that were then tested for correlation. Distributions of pbem and coverage in the two subsets are fully comparable and the correlation (Pearson's product-moment correlation) between *pbems* in S1 and S2 is 0.72 (**Figure 25A**, S1 and S2). Then I considered 20 normal samples equally profiled using a WES assay (Agilent HaloPlex Exome) covering 36 Mbp. As in the previous experiment, I split the set into two subsets each including 10 normal samples to compute independently *pbems*. The correlation between *pbems* at corresponding targeted loci is 0.13 (**Figure 25A**, S3 and S4). Finally, I made two subsets of same cardinality using 20 normal samples sequenced with a targeted panel (Roche NimbleGen N250 targeted panel) covering approximately 3 Mbp across 250 genes of relevant clinical interest. Again, the distribution of *pbem* and the coverage are fully comparable and the correlation between *pbems* at corresponding targeted loci is 0.47 (**Figure 25A**, S5 and S6).

Next I compared normal samples profiled through two partially overlapping targeted panels (Ion AmpliSeq Targeted Custom Amplicon Panel and Illumina True Seq Custom Amplicon) and different sequencing machines: IonTorrent PGM and Illumina MiSeq. Specifically, for this experiment I used data derived from 3 DNA samples and each of them was profiled on both considered platforms. The *pbems* were computed and compared on the 7201 bp shared between the two targeted designs. Results show lack of correlation (*r* = -0.02) between *pbems* at corresponding loci in samples sequenced using different platforms (**Figure 25B**, S7 and S8). Similarly, in the next experiment I selected 40 normal samples sequenced using both NimbleGen (Roche NimbleGen SeqCap Exome v3) and HaloPlex (Agilent HaloPlex Exome) WES kits. As in previous experiment, no correlation (*r* = 0.03) results when comparing *pbems* at corresponding loci shared (N=31 Mbp) between the two assays (**Figure 25B**, S9 and S3+S4). The experiments clearly indicate the majority of *pbems* of sets of normal samples sequenced using the same platform are concordant, either equal or greater than zero. Conversely, in experiments considering samples profiled with different platforms, genomic loci are annotated with discordant *pbems.* Indeed, approximately 50% of targeted positions show evidence of errors (*pbem* > 0*)* only when data are derived from one platform (**Figure 25C**). Altogether these results suggest that some targeted loci are recurrently subject to sequencing errors and this propensity strictly depends on the sequencing platform used.

Table 3. Information on normal samples collected to study the relationship between sequencing platform and *pbems*. Column "Target" reports the extended name of the sequencing assay; the corresponding abbreviation used in Figure 25 legend in brackets.

Target	Sequencing	Target size	Number of Normal samples	Institution
Ion AmpliSeq Targeted Custom Amplicon panel (AmpliSeq)	IonTorrent PGM	40 kbp	113	The Institute of Cancer Research (London)
Illumina True Seq Custom Amplicon (TSCA)	Illumina MiSeq	110 kbp	3	Computational Oncology Laboratory (Trento)
Roche NimbleGen SeqCap Exome v3 (NimbleGen)	Illumina HiSeq 2000	64 Mbp	40	Weill Cornell Medicine Englander Institute for Precision Medicine (New York)
Roche NimbleGen N250 targeted panel (NimbleGen_N250)	Illumina HiSeq 2000	3.2 Mbp	20	Weill Cornell Medicine Englander Institute for Precision Medicine (New York)
Agilent HaloPlex Exome (HaloPlex)	Illumina HiSeq 2000	36 Mbp	40	Weill Cornell Medicine Englander Institute for Precision Medicine (New York)

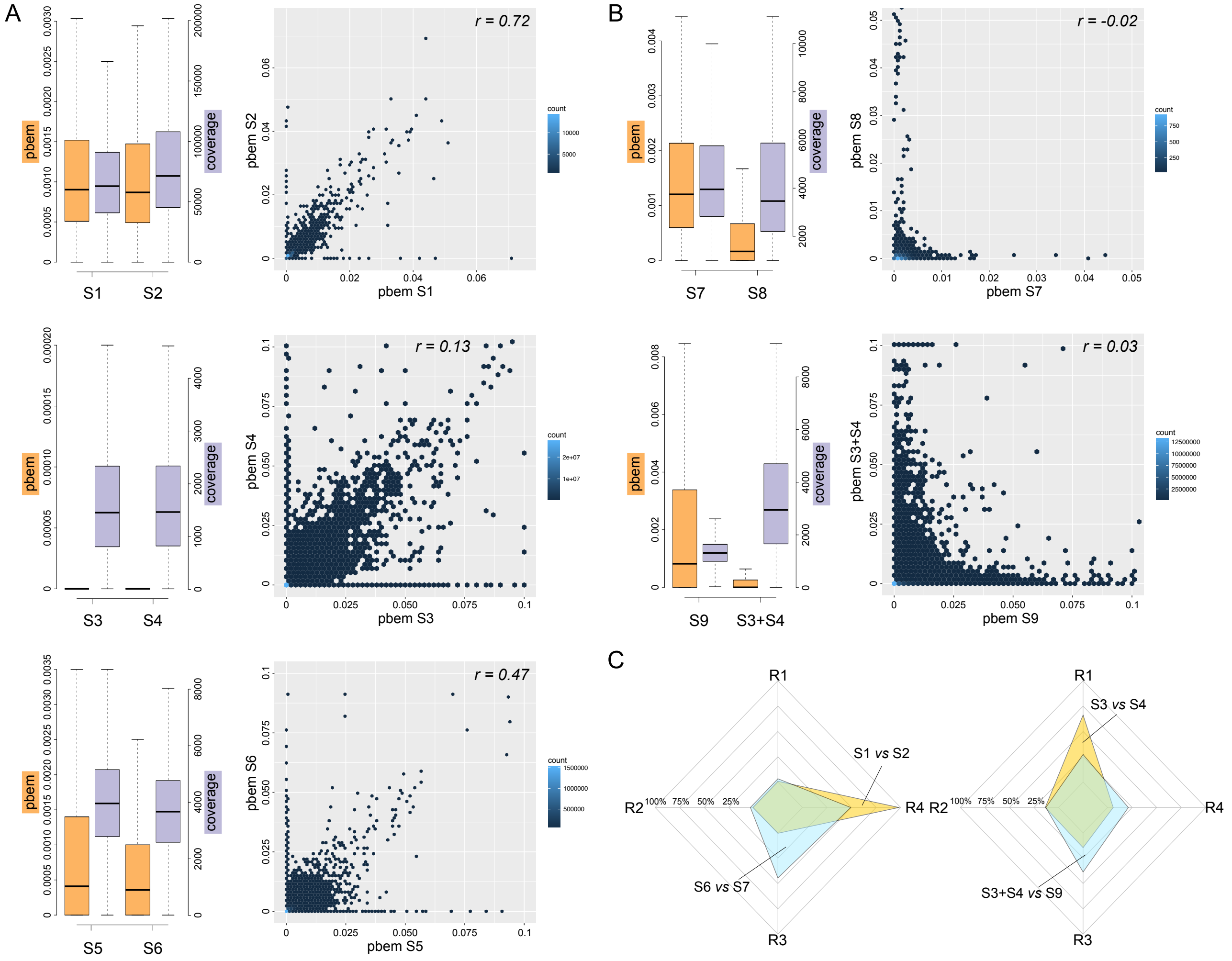


Figure 25. (**A**) Good correlations among *pbems* when computed using sets of normal samples sequenced on the same platform. S1 (N=56) and S2 (N=57) are normal samples sequenced using AmpliSeq (40kbp; IonTorrent PGM); S3 (N=20) and S4 (N=20) are normal samples sequenced using HaloPlex (36Mbp; Illumina HiSeq2000); S5 (N=10) and S6 (N=10) are normal samples sequenced using NimbleGen_N250 (3.2Mbp; Illumina HiSeq 2000). (**B**) Low correlation among *pbems* when computed using sets of normal samples sequenced on different platforms. S7 (N=3) and S8 (N=3) loci shared (7 kbp) between targeted custom AmpliSeq TSCA; S9 (N=40) and S3+S4 (N=40) loci shared (26Mbp) between NimbleGen (64Mbp; Illumina HiSeq 2000) and HaloPlex. (**C**) Proportion of concordant and discordant *pbems* when comparing samples profiled using the same platform (yellow polygons) or different ones (light blue polygons). R1 and R4 axes indicate the proportion of loci characterized by two concordant *pbems* since they are both equal or greater than zero, respectively. R2 and R3 axes indicate the proportion of genomic loci with discordant *pbems:* a genomic locus showing the first *pbem* equal to zero and the second one greater than zero, or contrariwise.

9. Somatic point mutations in plasma and matched tissue biopsies

The method described in section 7 has been partially exploited to detect somatic point mutations in 36 plasma WES samples. Matched control samples (buffy-coat) were available for all 36 patients; moreover, matched tissues biopsies (one or more) with matched germline (blood) were also available for 34 (94%) patients considered. The aim of the study was to quantitatively compare somatic point mutations detected in plasma and matched tissue biopsies. The set of 36 control samples was used to estimate overall and allele specific per-base error measures as described in sections 7.3.2 and ad-hoc stringent filtering criteria were applied as described in *Methods* section 8b.

455 putative SNVs were detected across the 34 plasma samples by applying an AF thresholds of 0.05. Among these, 42% (N=191) were observed exclusively in plasma samples and 53% (N=239) were found in plasma and in all other biopsies available. Lastly, 25 SNVs (5%) were detected in one or more, when available, biopsies. Fractions change by considering only the 10 plasma samples for which 2 or more biopsies are available. Indeed, the total number of SNVs detected is 249 (AF ≥ 0.05%); 19%, 70% and 10% are the fraction of SNVs found only in plasma, shared among plasma and all biopsies and found in at least one biopsy, respectively. Decreasing the AF threshold down to 0.03 and 0.01 the total number of detected SNVs and specifically the fraction of somatic point mutations detected exclusively in plasma increases (**Figure 26**).

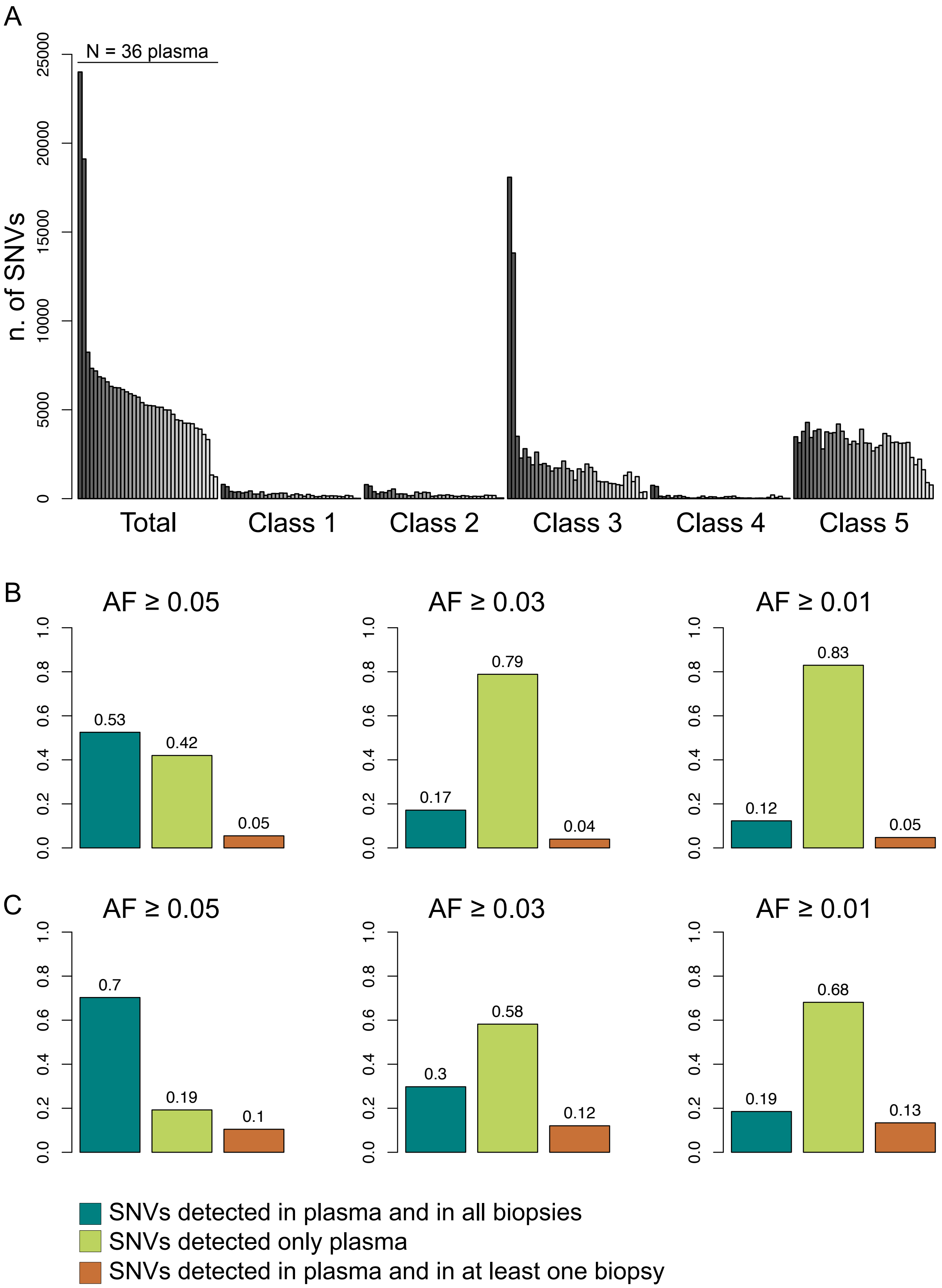


Figure 26. Top barplot, number of putative somatic SNVs called in 36 plasma samples and stratified by classes as described in section 7.3.2. Bottom, coloured barplots represents the fraction of SNVs

detected in plasma and biopsies using different AF thresholds and considering samples for which at least one (N=34, panel **B**) and at least two (N=10, panel **C**) matched tissue biopsies are available.

The median AF of SNVs detected (AF ≥ 0.05%) in plasma and all biopsies is 0.15 (mean=0.17, SD=0.09), that is higher than the median of SNVs found only in plasma (median=0.07, mean=0.08, SD=0.05) or in at least one biopsy (median=0.08, mean=0.10, SD=0.05). This suggests that mutations are those more clonal in metastatic lesions and more represented in ctDNA. Outliers in the AF distributions of SNVs detected only in plasma or in only some tissue biopsy will be object of further investigations. An example of this AF outlier is observed in patient PM189 (**Figure 27**). Indeed, a missense somatic SNV (previously reported in SU2C dataset) is detected with an AF of 0.39 but not evidences of this mutation are found in matched tissue biopsy. Additionally, other 7 SNVs are found only in plasma but with AF ranging between 0.02 and 0.07. 9 SNVs are detected in both plasma and biopsy. Interestingly, plasma-biopsy shared SNVs detected at low AF in plasma (0.07, 0.06 and 0.03) belong to class 1 providing robustness to the classification method based on the per-base error measures.

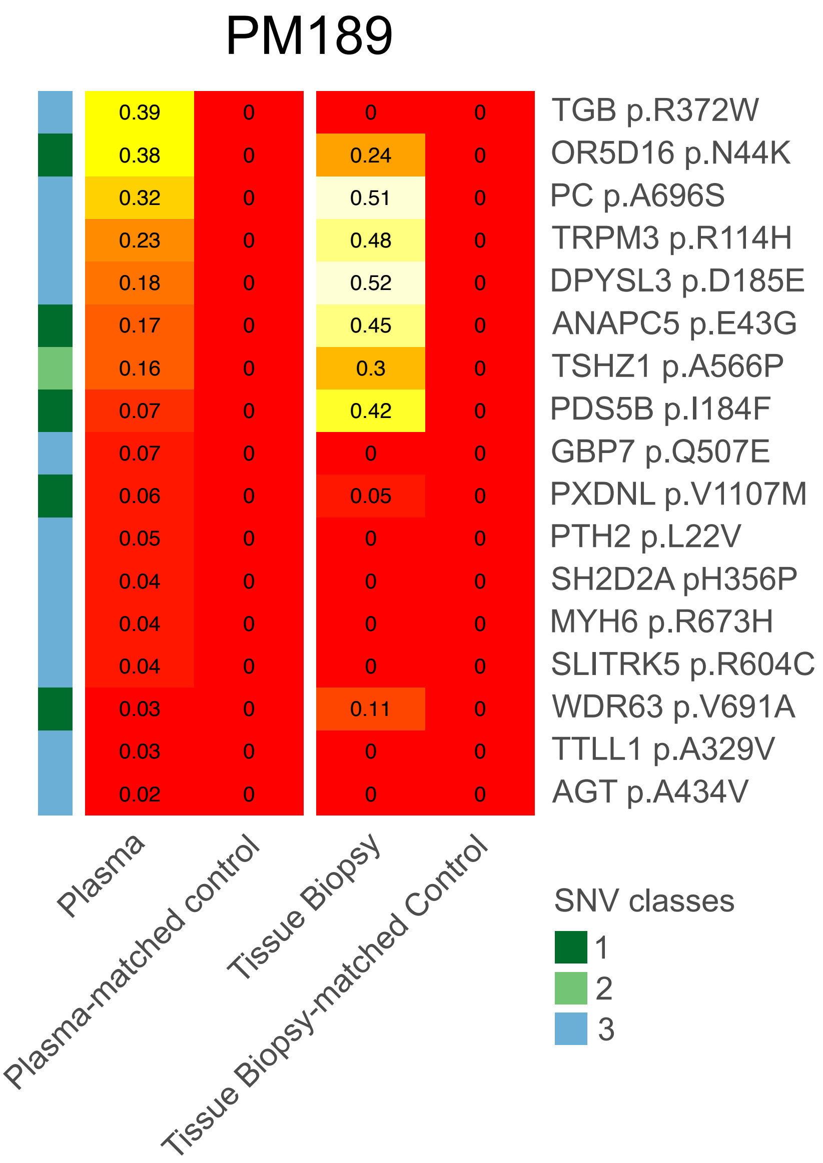


Figure 27. AF comparison among SNVs detected in plasma and tissues biopsies in patient PM189.

Among others, PM90 (**Figure 28A**) is an example of homogeneous presence of SNVs in plasma and tissue biopsies. SNVs are observed in plasma at different AF levels and in most of the cases are comparable with AFs measured in biopsies. To notice the detection in plasma (AF=0.2, class 1) of *SPOP* missense mutation p.F133I (previously reported in both SU2C-PCF and TCGA-PRAD datasets), detected also in all 6 available biopsies accordingly is an evidence that aberrations in this gene represent early and driver events in prostate tumorigenesis. Similar scenario is observed among samples of patient PM161 (**Figure 28B**) where clonal mutation p.I195T affecting *TP53* is detected (AF=0.41, class 1) in plasma and all 3 tissue biopsies. Finally, 9 SNVs (all previously reported in SU2C-PCF dataset) are detected in plasma (0.13 ≤ AF ≤ 0.05, class 2 and 3) and in only 2 out of 3 biopsies suggesting a probable polyclonal metastatic seeding.

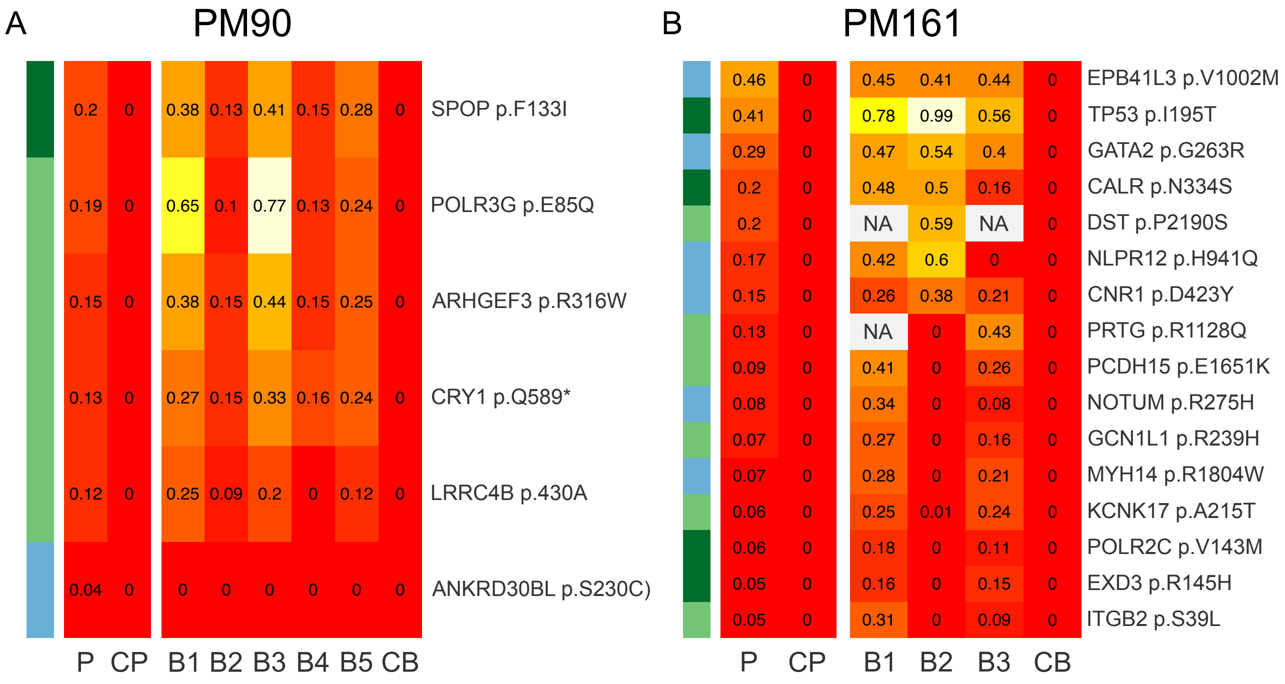


Figure 28. AF comparison among SNVs detected in plasma and tissues biopsies in patients PM90 (**A**) and PM161 (**B**). SNVs are sorted by decreasing AF as measured in plasma sample. On the left of each row is depicted the class of the corresponding SNV: Dark green class 1, light green class 2, light blue class 3. P plasma, CP plasma-matched control, B*N* tissue biopsies, CB biopsy-matched control.

METHODS (Section b)

1b. Measures of genomic mappability

ENCODE mappability tracks provide a measure of how often the sequence found at the particular location will align within the whole genome[115–117]. Depending on the number of mismatched tolerated during search alignment across the genome, mappability can be expressed as uniqueness and alignability. Unlike measures of uniqueness (none mismatches tolerated), alignability will tolerate up to 2 mismatches. These tracks are in the form of signals ranging from 0 to 1. Indeed, mappability scores are calculated as 1/the number of places a sequence of length *k* maps (0 mismatches for uniqueness, up to 2 mismatches for alignability) to the genome. Mappability of a single position is then computed as the average mappability scores from the *k* sliding sequences of length *k* spanning that position.

2b. Amplicon stability measure

To minimize non-informative signal, we applied amplicon selection for autosomal and non-autosomal regions, separately, based on the germline sample set. For each germline sample, we calculated the amplicon mean coverage distribution. Mean and standard deviation (SD) of that distribution were then computed. Amplicon stability is measured as the fraction of samples for which the mean coverage of the considered amplicon is in range [mean-SD, mean+SD].

3b. Analysis of sequencing data from plasma and serum samples

DNA extracted from plasma and serum samples were sequence in 3 sessions through paired-end sequencing protocol on Illumina MiSeq set to generate reads of 150bp length. Reads (FASTQ files) were mapped with Isaac Genome Alignment Software (human genome reference sequence hg19/GRC37) to targeted regions reported in the *TSCA Manifest* file provided by *Illumina*. First, the total number of reads generated for each sequenced sample was estimated by *flagstat* utility included in samtools software[118]. I stratified this analysis making distinction among total, mapped and properly-paired reads (mates of a read pair map to the same chromosome, oriented towards each other, and with a sensible insert size). Second, the mean sequencing coverage across designed on-targets was computed with the computational tool *DepthOfCoverage* included in the Genome Analysis Tool Kit (GATK) software package for analysis of high-throughput sequencing data[102]. For this analysis, I considered only reads with mapping quality and base quality greater than 20. Third, an *ad-hoc* coverage estimation is performed in Manifest on- and off-targets by retaining only properly-paired reads characterized by a mapping quality greater than 20.

4b. Detection of candidate somatic point mutations

Strict criteria were applied for the detection of somatic point mutations (PM) to contribute to the assessment of tumor content. Similar to previous work[109], the detection procedure includes the following filters:

a. the local total coverage is >=100;
b. the alternative base is supported by at least 5 reads;
c. the allelic fraction (AF) >=1%;
d. if any, the same alternative base is detected in additional samples from the same individual;
e. exclusion of all genomic positions close to amplicon edges (<=4 bases from internal edges);
f. exclusion of all positions not satisfying strand bias criteria;
g. the allelic fraction of the position for the patient normal sample is <1%.

Filters from a-f apply to all patients' plasma samples. The strand bias filter (f) combines the Fisher Exact Test and an ad-hoc test that computes the strand bias distribution from all candidate PMs across all tumour samples and retains only first quartile values. To decrease the impact of false positives, we retain PMs present in single samples only if the AF>=2%.

5b. Specificity and false positives of point mutation detection

Specificity of point mutation detection was computed from the distribution of AFs at all positions except for germline SNPs across all germline and HV samples; the proportion of positions with local coverage >=100 and an alternative base supported by at least 5 reads was computed. The specificity resulted in 99.63% with a median AF for false positives of 2% (with standard deviation of 3%).

6b. Synonymous and non-synonymous substitutions rates *(Ks and Ka)*

Within a genomic region, *Ks* and *Ka* rates are defined as the number of synonymous substitutions per synonymous sites and the number of nonsynonymous substitutions per nonsynonymous sites, respectively[119]. To quantify *Ks* and *Ka* for each study sample, I first considered genomic positions within *AR* coding regions spanning 9 exons and covered by targeted amplicons (in this targeted design, all exons are fully covered except for exon 1 where amplicons cover 86% of the entire coding region). SNPs loci were excluded. Given the reference base (allele; GRCh37/hg19) at each locus, I simulated all possible transition and transversion events. Then, I annotated as synonymous or not the amino acid changes caused by each combination of reference and simulated alternative alleles. The total number of synonymous and nonsynonymous sites, denominators of *Ks* and *Ka* rates respectively, are computed by summing the proportions of synonymous or nonsynonymous simulated events across considered loci. Next, based on the annotation of synonymous and nonsynonymous sites, *Ks* and *Ka* rates were computed for each sample based on the number of synonymous and nonsynonymous point mutations detected in the *AR* region of interest. In this study, among the 2469 loci selected in *AR* coding regions the analysis annotated 570 and 1899 as synonymous and nonsynonymous sites.

7b. Plasma samples from CRPC patients

Whole exome sequencing is performed for matched circulating tumor DNA, germline DNA, and metastatic biopsies from 34 patients with CRPC using minimum 50ng DNA, Roche NimbleGen SeqCap EZ Human Exome Kit v3.0 library prep, Illumina platform (mean coverage in ctDNA >300X).

8b. Detection of somatic point mutations in plasma and matched biopsies

For plasma samples, 3 starting sets of putative somatic point mutations were generated requiring an AF ≥ 0.01, 0.03 and 0.05, respectively. For each position, the corresponding locus in all matched controls (plasma-matched and biopsy-matched) samples were checked and required to have an AF = 0. Using classification criteria described in section 7.3.2, only variants belonging to classes 1, 2 and 3 were considered. Passing loci were functionally annotated using Oncontator [120] software and further filtered by keeping only those variants labelled as *missense* and *nonsense*. Finally, to increase true positives, I retained only variants affecting a cancer related gene (N_{genes}=1391) or previously reported in TCGA-PRAD ($N_{variants}$=24844) or SU2C-PCF ($N_{variants}$=21406) datasets.

DISCUSSION

Recent work highlighted the spread of tumor heterogeneity in advanced prostate cancer patients as evidenced by tissue based and circulating material studies [37,39,109]. The extent to which this is relevant in the context of patients' treatment is still poorly understood, partially due to sub-optimal characterization approaches. To help address this clinical question, I focused on two strategies, a tissue based and a cell free DNA based one. Upon the setup of a computational toolbox for studying cancer genomes by fully exploiting high resolution data provided by next generation sequencing experiments, I generated exhaustive genomic characterization of a cohort of 25 samples from 10 advanced CRPC patients for which multiple biopsies from primary and metastatic lesions were sampled and genomically profiled via whole exome and whole genome sequencing. Based on global CNAF assessment and the genomic status of a comprehensive cancer genes list, the results showed overall modest intra-patient tumor heterogeneity. However, key genes as *RB1* demonstrated variable genomic status across metastases; for instance, in liver and thigh metastases of patient V5033 and in inguinal and supraclavicular lymph nodes of patient V5128. Metastases of patient V5033 have lost one allele of *RB1* (hemizygous deletion revealed by genomic data and validated by FISH) and do not express RB protein (IHC data); the impairment of the last *RB1* allele can be explained by genomic data only for the liver metastasis where a disruptive somatic point mutation is detected. Since none *RB1* aberrations are found, RB inactivation in the thigh metastasis is caused by a different molecular mechanism. In patient V5128, intra-tumor heterogeneity is disclosed as different *RB1* copy number status. Specifically, hemizygous loss and neutral loss of heterozygosity (two copies of one allele, none for the other one) is revealed by genomic data and FISH assays. RB protein level has been measured in both lymph nodes and found to be expressed at comparable levels. Whole genome data allowed to discover a never reported alternative molecular mechanism for *RB1* inactivation (V4074). Indeed, a tandem duplication event damaging the canonical *RB1* architecture is detected in all the 3 metastatic sites studied in this patient. This event, concomitant with hemizygous loss of one *RB1* allele as detected by genomic data and validated by FISH, explains full lack of RB expression in all the 3 lesions. Although the specific 7-17 tandem duplication is not frequent in CRPC, as demonstrated by the focused analysis of 149 additional patients from the SU2C-PCF cohort [56], this result suggest that *RB1* inactivation might exist by multiple deleterious molecular mechanisms in addition to point mutations and genomic loss, leading to retinoblastoma protein functionality impairment (Nava Rodrigues D*, Casiraghi N*, et al, *in preparation*). Understanding these mechanisms in prostate cancer is important also in the context of recent studies showing that inactivation of *RB1* when combined with *TP53* loss promotes lineage plasticity, metastasis and antiandrogen resistance [67,121] and for ongoing clinical trials of CDK4/6 inhibitors that imply careful patient selection. In the presented study, a comprehensive assessment of tumor heterogeneity was limited by the few number of primary tissue samples available and lack for these samples of both IHC and FISH data.

The second part of this thesis work focused on the genomic analysis of plasma samples. Using a targeted assay covering all the *AR* coding regions we sequenced plasma samples from CRPC patients immediately before starting abiraterone, on treatment, and after progression, concurrently evaluating both copy number and somatic point mutations. Amongst the samples with detectable DNA fraction, we observed a significant inverse correlation between detection of *AR* copy number gain and *AR* point mutation. Moreover, we identified a significantly higher rate of non-synonymous with respect to synonymous *AR* point mutations in the samples with no *AR* gain compared with gain, supporting selection of non-synonymous mutations in the absence of gain. Finally, abiraterone resistance in up to 30% of patients with no detectable *AR* gain at progression was associated with an *AR* somatic point mutation, which is often observed several months before confirmed clinical progression and putatively activated by nonandrogenic ligands. This suggests that analysis of plasma *AR,* whose genomic status may be predictive for abiraterone resistance, could complement other modalities for evaluating CRPC patients and allow early treatment change before overt radiological progression. This work was published in 2015 in Science Translational Medicine [80] and has been since highly cited.

Challenges posed by this study, such as a biological scenario characterized by little DNA material and high admixture, represented the rationale to develop a computational strategy to readily detect and quantify somatic point mutations in sequencing data from patient's plasma and by fully exploiting matched germline DNA. The method I implemented provides data-driven AF thresholds, extensive estimations of per-base sequencing errors and indicators for experimental biases to limit false positives within the final set of putative somatic point mutations. First, I used the per-base sequencing error measure (*pbem*) to show that some targeted loci are recurrently subject to sequencing errors. This observation indicates that, in addition to randomly distributed noise, sequencing assays can be affected by error hot-spots. Moreover, by comparing different experimental conditions (number of samples, size of the sequencing assay, DNA library preparation, sequencing machine) I demonstrated that these hot-spots are consistent only among DNA samples profiled using the same library preparation kit and sequencing machine.

Second, I used the developed computational strategy to detect somatic point mutations in ctDNA profiled via whole exome sequencing and sampled from 34 CRPC patients. This is the first study to show that WES of ctDNA is feasible in CRPC and can help elucidate intra-patient heterogeneity. Analysis showed that, depending on the AF, the fraction of mutations detected in both ctDNA and in available biopsies vary. Overall, 53% of highly trustable somatic SNVs detected (AF ≥ 0.05) in plasmas are also observed in biopsies, supporting the advantage of liquid biopsy to outline lesions mutational landscape. Mutations found only in plasma samples showing high AF (20% with AF ≥ 0.05) could help to better outline the landscape of circulating tumoral clones possibly released by not collected tissue biopsies. Data show that fractions of SNVs found only in plasma increases by lowering the AF threshold. Indeed, ability to detect mutations occurring only in plasma at low frequencies (AF ≤ 0.03) is helpful to accurately shape the profile of clonal/subclonal aberrations but still present challenging discrimination between true and false positives. The ABEMUS methodology (Casiraghi N, et al, manuscript in preparation) can be used for the detection of SNV in any challenging tumor/normal sample pair. To improve ABEMUS methodology, ongoing analyses are dedicated to better delineate the *pbem* underlying features with the final goal to enhance its power as filtering criteria. Moreover, further analyses are now focused on the comparison between the ABEMUS and existing computational tools expressly developed for somatic SNVs detection. Indeed, in-silico normal and tumoral samples will be generated through an ad-hoc computational procedure to obtain fully customizable datasets where teste methods will be evaluated based on precision and recall measures. Then, tools comparison exercises will be performed also considering an extended set of ctDNA profiled via WES and sampled from CRPC patients.

SUPPLEMENTARY INFORMATION

Supplementary Table 1. Additional information on tumor samples included in the study cohort.

Patient ID	Tumor ID	Sites	Size	Tumor purity based on Histology (%)
904	LP2000338	Left Axillary LN	6.2x0.82mm	60
904	LP2000339	Inguinal LN	4.1x0.86mm	90
V4002	LP2000360	Anterior left supraclavicular LN	2.73x2mm	70
V4002	LP2000361	Posterior left supraclavicular LN	3.5x1.3mm	80
V4038	LP2000301	Right Supraclavicular LN	7.9x1.4mm	90
V4038	LP2000302	Right Retropectoral LN	4x0.67mm	40
V4074	LP2000115	Dorsal Glans	3.34x2.5mm	80
V4074	LP2000116	Ventral Glans	6.7x3.8mm	80
V4074	LP2000117	Right Coronal Sulcus	7.9x4.0mm	80
V4074	Sample_A34_0001	TURP	NA	70
V5033	LP2000119	Liver	3.8x0.83mm	10
V5033	LP2000120	Thigh Muscle	3.8x1.57mm	60
V5128	LP2000184	Right Inguinal LN	1.5x0.61mm	60
V5128	LP2000185	Left Supraclavicular LN	1.9x1.37mm	80
V5149	LP2000304	Right Supraclavicular LN	7.38x0.62mm	60
V5149	LP2000305	Left Supraclavicular LN	5.27x0.8mm	60
V5149	Sample_B16_0001	Prostatectomy microdissected area	NA	NA
V5149	Sample_B16_0002	Prostatectomy microdissected area	NA	NA
V5162	LP2000341	Left Inguinal LN	7.12x0.88mm	80
V5162	LP2000342	Right Inguinal LN	11x0.88mm	90
V5162	Sample_B16_0004	Prostatectomy microdissected area	NA	NA
V5162	Sample_B16_0005	Prostatectomy microdissected area	NA	NA
V5164	LP2000313	Left Supraclavicular node	7x1.2mm	60
V5164	LP2000314	Liver	2.6x0.6mm	25
V5191	LP2000363	Left axillary LN	3.68x1.18mm	80
V5191	LP2000364	Right Axillary LN	8.7x1.3mm	60

Supplementary Table 2. *RB1* copy number (average ratios target over control probes) and RB quantification estimated by FISH and IHC, respectively.

Patient ID	Tumor ID	Sites	FISH	IHC
904	LP2000338	Left Axillary LN	1.031	100
904	LP2000339	Inguinal LN	1.022	100
V4002	LP2000360	Anterior left supraclavicular LN	1.020	120
V4002	LP2000361	Posterior left supraclavicular LN	1.040	120
V4038	LP2000301	Right Supraclavicular LN	1.614	150
V4038	LP2000302	Right Retropectoral LN	1.220	190
V4074	LP2000115	Dorsal Glans	0.977	0
V4074	LP2000116	Ventral Glans	1.048	0
V4074	LP2000117	Right Coronal Sulcus	Na	0
V5033	LP2000119	Liver	0.494	0
V5033	LP2000120	Thigh Muscle	0.540	0
V5128	LP2000184	Right Inguinal LN	0.885	100
V5128	LP2000185	Left Supraclavicular LN	1.072	100
V5149	LP2000304	Right Supraclavicular LN	0.685	30
V5149	LP2000305	Left Supraclavicular LN	0.663	100
V5162	LP2000341	Left Inguinal LN	1.000	130
V5162	LP2000342	Right Inguinal LN	1.011	80
V5164	LP2000313	Left Supraclavicular node	0.615	0
V5164	LP2000314	Liver	0.841	10
V5191	LP2000363	Left axillary LN	0.638	0
V5191	LP2000364	Right Axillary LN	0.831	0

REFERENCES

1. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–8 (1976).
2. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012).
3. Gerlinger, M. *et al.* Cancer: Evolution Within a Lifetime. *Annu. Rev. Genet.* **48,** 215–236 (2014).
4. Baca, S. C. *et al.* Punctuated Evolution of Prostate Cancer Genomes. *Cell* **153,** 666–677 (2013).
5. Birkbak, N. J. *et al.* Paradoxical relationship between chromosomal instability and survival outcome in cancer. *Cancer Res.* **71,** 3447–3452 (2011).
6. Bardelli, A. *et al.* Carcinogen-specific induction of genetic instability. *Proc Natl Acad Sci U S A* **98,** 5770–5. (2001).
7. Cahill, D. P., Kinzler, K. W., Vogelstein, B. & Lengauer, C. Genetic instability and darwinian selection in tumours. *Trends Biochem. Sci.* **24,** 57–60 (1999).
8. Prandi, D. *et al.* Unraveling the clonal hierarchy of somatic genomic aberrations. *Genome Biol.* **15,** 439 (2014).
9. de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science (80-.).* **346,** 251–256 (2014).
10. Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21,** 751–759 (2015).
11. Uchi, R. *et al.* Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution. *PLoS Genet.* **12,** 1–24 (2016).
12. Harbst, K. *et al.* Multiregion whole-exome sequencing uncovers the genetic evolution and mutational heterogeneity of early-stage metastatic melanoma. *Cancer Res.* **76,** 4765–4774 (2016).
13. Hao, J.-J. *et al.* Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nat. Genet.* **48,** 1500–1507 (2016).
14. Bashashati, A. *et al.* Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J. Pathol.* **231,** 21–34 (2013).
15. Murugaesu, N. *et al.* Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discov.* **5,** 821–832 (2015).
16. Park, S. & Lehner, B. Cancer type-dependent genetic interactions between cancer

driver alterations indicate plasticity of epistasis across cell types. *Mol. Syst. Biol.* **11,** 824–824 (2015).

17. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science (80-.).* **339,** 1546–1558 (2013).
18. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152,** 714–726 (2013).
19. Dalgliesh, G. L. *et al.* Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463,** 360–363 (2010).
20. Halazonetis, T. D., Gorgoulis, V. G. & Bartek, J. An Oncogene-Induced DNA Damage Model for Cancer Development. *Science (80-.).* **319,** 1352–1355 (2008).
21. Bartkova, J. *et al.* DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature* **434,** 864–870 (2005).
22. Bartkova, J. *et al.* Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. *Nature* **444,** 633–637 (2006).
23. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255,** 197–200 (1975).
24. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150,** 264–278 (2012).
25. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
26. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486,** 353–360 (2012).
27. Nathanson, D. A. *et al.* Targeted Therapy Resistance Mediated by Dynamic Regulation of Extrachromosomal Mutant EGFR DNA. *Science (80-.).* **343,** 72–76 (2014).
28. Johnson, B. E. Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. *Science (80-.).* **189,** 189–194 (2014).
29. Notta, F. *et al.* A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538,** 378–382 (2016).
30. McPherson, A. *et al.* Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* **48,** 758–767 (2016).
31. Kreso, A. Variable Clonal Repopulation Dynamics Influence Chemotherapy Response in Colorectal Cancer. **19,** 1–7 (2012).
32. Keats, J. J. *et al.* Clonal competition with alternating dominance in multiple myeloma

Clonal competition with alternating dominance in multiple myeloma Running Title = Clonal Evolution in Multiple Myeloma. **120,** 1067–1077 (2012).

33. Gupta, G. P. & Massagué, J. Cancer Metastasis: Building a Framework. *Cell* **127,** 679–695 (2006).

34. Turajlic; Swanton. Metastasis as an evolutionary process. *Science (80-.).* (2016).

35. Cooper, C. S. *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat. Genet.* **47,** 367–372 (2015).

36. Beltran, H. & Demichelis, F. Prostate cancer: Intrapatient heterogeneity in prostate cancer. *Nat. Rev. Urol.* 1–2 (2015). doi:10.1038/nrurol.2015.182

37. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* (2015). doi:10.1038/nature14347

38. Liu, W. *et al.* Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med* **15,** 559–565 (2009).

39. Hong, M. K. H. *et al.* Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat. Commun.* **6,** 6605 (2015).

40. Attard, G. *et al.* Prostate cancer. *Lancet* **387,** 70–82 (2016).

41. Schaid, D. J. The complex genetic epidemiology of prostate cancer. *Hum. Mol. Genet.* **13,** 103R–121 (2004).

42. Gann, P. H. Risk factors for prostate cancer. *Rev. Urol.* **4 Suppl 5,** S3–S10 (2002).

43. Svensson, M. A. *et al.* Testing mutual exclusivity of ETS rearranged prostate cancer. *Lab. Investig.* **91,** 404–412 (2011).

44. Lindberg, J. *et al.* Exome sequencing of prostate cancer supports the hypothesis of independent tumour origins. *Eur. Urol.* **63,** 347–353 (2013).

45. Boutros, P. C. *et al.* Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47,** 736–745 (2015).

46. Lonergan, P. & Tindall, D. Androgen receptor signaling in prostate cancer development and progression. *J Carcinog* (2011). at <http://www.carcinogenesis.com/text.asp?2011/10/1/20/83937>

47. Lin, C. *et al.* Nuclear Receptor-Induced Chromosomal Proximity and DNA Breaks Underlie Specific Translocations in Cancer. *Cell* **139,** 1069–1083 (2009).

48. Carver, B. S. *et al.* Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate. *Nat. Genet.* **41,** 619–624 (2009).

49. King, J. C. *et al.* Cooperativity of TMPRSS2-ERG with PI3-kinase pathway activation

in prostate oncogenesis. *Nat. Genet.* **41,** 524–526 (2009).

50. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44,** 685–9 (2012).

51. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–220 (2011).

52. Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487,** 239–243 (2012).

53. Taylor, B. S. *et al.* Integrative Genomic Profiling of Human Prostate Cancer. *Cancer Cell* **18,** 11–22 (2010).

54. Tomlins, S. A. Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. *Science (80-.).* **310,** 644–648 (2005).

55. Rubin, M. A., Maher, C. A. & Chinnaiyan, A. M. Common gene rearrangements in prostate cancer. *J. Clin. Oncol.* **29,** 3659–3668 (2011).

56. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **161,** 1215–1228 (2015).

57. Boysen, G. *et al.* SPOP mutation leads to genomic instability in prostate cancer. *Elife* **4,** 1–18 (2015).

58. Pritchard, C. C. *et al.* Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer. *Nat. Commun.* **5,** 4988 (2014).

59. Castro, E. *et al.* High burden of copy number alterations and c-MYC amplification in prostate cancer from BRCA2 germline mutation carriers. *Ann. Oncol.* **26,** 2293–2300 (2015).

60. Carter, B. S., Beaty, T. H., Steinberg, G. D., Childs, B. & Walsh, P. C. Mendelian inheritance of familial prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 3367–71 (1992).

61. Pritchard, C. C. *et al.* Inherited DNA Repair Gene Mutations in Men with Metastatic Prostate Cancer Short Title: Germline DNA repair gene mutations in prostate cancer. (2016). doi:10.1056/NEJMoa1603144

62. Chen, H.-Z., Tsai, S.-Y. & Leone, G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat. Rev. Cancer* **9,** 785–797 (2009).

63. Burkhart, D. L. & Sage, J. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nat. Rev. Cancer* **8,** 671–682 (2008).

64. Sharma, A. *et al.* Retinoblastoma tumor suppressor status is a critical determinant of therapeutic response in prostate cancer cells. *Cancer Res.* **67,** 6192–6203 (2007).

65. Beltran, H. *et al.* Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat. Med.* **22,** 298–305 (2016).

66. Sharma, A., Yeow, W. & Ertel, A. The retinoblastoma tumor suppressor controls androgen signaling and human prostate cancer progression. *J. Clin. Invest.* **120,** (2010).

67. Ku, S., Rosario, S. & Wnag, Y. Rb1 and Trp53 prostate cancer lineage plasticity, metastasis, and antiandrogen resistance. *Science (80-.).* (2017). doi:10.1126/science.aah4199

68. Mouliere, F. *et al.* High fragmentation characterizes tumour-derived circulating DNA. *PLoS One* **6,** (2011).

69. Mouliere, F., El Messaoudi, S., Pang, D., Dritschilo, A. & Thierry, A. R. Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer. *Mol Oncol* **8,** 927–941 (2014).

70. Leon, S. A. *et al.* Free DNA in the Serum of Cancer Patients and the Effect of Therapy Free DNA in the Serum of Cancer Patients and the Effect of Therapy. *Cancer Res* **37,** 646–650 (1977).

71. Lo, Y. M. D. *et al.* Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus. *Sci. Transl. Med.* **2,** 61ra91-61ra91 (2010).

72. Thierry, A. R. *et al.* Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. *Nucleic Acids Res.* **38,** 6159–6175 (2010).

73. Sorenson, G. D. *et al.* Soluble Normal and Mutated Dna-Sequences From Single-Copy Genes in Human Blood. *Cancer Epidemiol. Biomarkers Prev.* **3,** 67–71 (1994).

74. Kimura, H. *et al.* Detection of epidermal growth factor receptor mutations in serum as a predictor of the response to gefitinib in patients with non-small-cell lung cancer. *Clin. Cancer Res.* **12,** 3915–3921 (2006).

75. Sozzi, G., Musso, K., Ratcliffe, C., Goldstraw, P. & Pierotti, M. A. Detection of Microsatellite Alterations in Plasma DNA of Non-Small Cell Lung Cancer Patients : A Prospect for Early Diagnosis Advances in Brief Detection of Microsatellite Alterations in Plasma DNA of Non-Small Cell Lung Cancer Patients : A Prospect for E. **5,** 2689–2692 (1999).

76. Bettegowda, C. *et al.* Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Sci. Transl. Med.* **6,** 224ra24-224ra24 (2014).

77. De Mattos-Arruda, L. *et al.* Capturing intra-tumor genetic heterogeneity by de novo

mutation profiling of circulating cell-free tumor DNA: A proof-of-principle. *Ann. Oncol.* **25,** 1729–1735 (2014).

78. Murtaza, M. *et al.* Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. *Nat. Commun.* **6,** 8760 (2015).

79. Siravegna, G. *et al.* Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. *Nat. Med.* **21,** 795–801 (2015).

80. Romanel, A. *et al.* Plasma AR and abiraterone-resistant prostate cancer. *Sci. Transl. Med.* **7,** 312re10-312re10 (2015).

81. Check Hayden, E. Technology: The $1,000 genome. *Nature* **507,** 294–295 (2014).

82. Liu, L. *et al.* Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012,** (2012).

83. US Food and Drug Administration. Premarket approval P150044 — Cobas EGFR MUTATION TEST V2. FDA http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpma/pma.cfm?id=P150044.

84. QIAGEN. Therascreen EGFR Plasma RGQ PCR Kit - https://www.qiagen.com/gb/resources/resourcedetail?id=eb32e329-3422-4eda- b3d6-e44ed787002a&lang=en. (2014).

85. Fedele, C., Tothill, R. W. & McArthur, G. A. Navigating the challenge of tumor heterogeneity in cancer therapy. *Cancer Discov.* **4,** 146–148 (2014).

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

87. Abeshouse, A. *et al.* The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163,** 1011–1025 (2015).

88. Alkodsi, a., Louhimo, R. & Hautaniemi, S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Brief. Bioinform.* **16,** bbu004- (2014).

89. Magi, A. *et al.* EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol.* **14,** R120 (2013).

90. Amarasinghe, K. C. *et al.* Inferring copy number and genotype in tumour exome data. *BMC Genomics* **15,** 732 (2014).

91. Boeva, V. *et al.* Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28,** 423–425 (2012).

92. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44,** 685–9 (2012).

93. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

94. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).

95. Demichelis, F. *et al.* SNP panel identification assay (SPIA): A genetic-based assay for the identification of cell lines. *Nucleic Acids Res.* **36,** 2446–2456 (2008).

96. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

97. Romanel, A., Lago, S., Prandi, D., Sboner, A. & Demichelis, F. ASEQ: fast allele-specific studies from next-generation sequencing data. *BMC Med Genomics* **8,** 9 (2015).

98. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* **6,** 80–92 (2012).

99. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–81 (2009).

100. Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl. Acad. Sci.* **108,** E1128–E1136 (2011).

101. Raczy, C. *et al.* Isaac: Ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29,** 2041–2043 (2013).

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

103. Detre, S., Saclani Jotti, G. & Dowsett, M. A ‘quickscore’ method for immunohistochemical semiquantitation: validation for oestrogen receptor in breast carcinomas. *J. Clin. Pathol.* **48,** 876–8 (1995).

104. Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat. Rev. Clin. Oncol.* **10,** 472–484 (2013).

105. Diaz, L. A. J. & Bardelli, A. Liquid Biopsies: Genotyping Circulating Tumor DNA. **33,** 395–401 (2015).

106. Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4,** 136ra68-136ra68 (2012).

107. Siravegna, G. & Bardelli, A. Genotyping cell-free tumor DNA in the blood to detect residual disease and drug resistance. *Genome Biol.* 4–9 (2014).

108. Haber, D. a. & Velculescu, V. E. Blood-Based Analyses of Cancer: Circulating Tumor Cells and Circulating Tumor DNA. *Cancer Discov.* **4,** 650–661 (2014).

109. Carreira, S. *et al.* Tumor clone dynamics in lethal prostate cancer. *Sci. Transl. Med.* **125,** (2014).

110. Bernardi, G. Isochores and the evolutionary genomics of vertebrates. *Gene* **241,** 3–17 (2000).

111. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

112. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes are compact. *Trends Genet.* **19,** 362–365 (2003).

113. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41,** 1–12 (2013).

114. Chen, L., Liu, P., Evans, T. C. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science (80-.).* **355,** 752–756 (2017).

115. Feingold, E. *et al.* The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (80-.).* **306,** 636–40 (2004).

116. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci.* **111,** 6131–6138 (2014).

117. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7,** (2012).

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

119. Hurst, L. D. The Ka/Ks ratio: Diagnosing the form of sequence evolution. *Trends Genet.* **18,** 486–487 (2002).

120. Ramos, A. H. *et al.* Oncotator: Cancer variant annotation tool. *Hum. Mutat.* **36,** E2423–E2429 (2015).

121. Mu, P. *et al. SOX2* promotes lineage plasticity and antiandrogen resistance in *TP53* - and *RB1* -deficient prostate cancer. *Science (80-.).* **355,** 84–88 (2017).

Declaration

I Nicola Andrea Casiraghi confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Nicola Casiraghi