Mutational Heterogeneity in Cancer

Akash Kumar

A dissertation

Submitted in partial fulfillment of

requirements for the degree of

Doctor of Philosophy

University of Washington

2014 June 5

Reading Committee:

Jay Shendure

Pete Nelson

Mary Claire King

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

ABSTRACT

Mutational Heterogeneity in Cancer

Akash Kumar

Chair of the Supervisory Committee:

Associate Professor Jay Shendure

Department of Genome Sciences

Somatic mutation plays a key role in the formation and progression of cancer. Differences in mutation patterns likely explain much of the heterogeneity seen in prognosis and treatment response among patients.

Recent advances in massively parallel sequencing have greatly expanded our capability to investigate somatic mutation. Genomic profiling of tumor biopsies could guide the administration of targeted therapeutics on the basis of the tumor's collection of mutations. Central to the success of this approach is the general applicability of targeted therapies to a patient's entire tumor burden. This requires a better understanding of the genomic heterogeneity present both within individual tumors (intratumoral) and amongst tumors from the same patient (intrapatient).

My dissertation is broadly organized around investigating mutational heterogeneity in cancer. Three projects are discussed in detail: analysis of (1) interpatient and (2) intrapatient heterogeneity in men with disseminated prostate cancer, and (3) investigation of regional intratumoral heterogeneity in gliomas. I conclude with a summary of my research and a discussion of future directions.

Table of Contents

Acknowledgements

First, I would like to thank my mentor, Jay Shendure, who has been a fantastic role model, source of inspiration and support. He has taught me how to think like a scientist, communicate and take risks with my work. His brilliance, energy and enthusiasm have made the past four years a joy to work in his lab and I continue to be amazed by his work ethic and work-life balance.

One of Jay's many talents has been to draw a group of exceptionally bright, motivated and collaborative people together. I have been so happy to be able to spend time with this group of people who have taught me how to think bigger and more creatively about scientific problems.

I owe special thanks to several members of the lab. Charlie Lee works tirelessly at the sequencer and provided company during many late nights in the lab. Sarah Ng and Emily Turner launched me on my first projects with encouragement and patience. Joe Hiatt and Stephen Salipante offered critical advice and insight at key moments on the way to become a physician scientist. Jacob Kitzman and Aaron McKenna were entirely selfless with their experience in computational and experimental approaches. Matthew Snyder and Andrew Adey shared their knack for statistics and beautifully illustrated figures, respectively. Alex Lewis, Riza Daza and Ruolan Qui provided countless hours of methodical and precise help with experimental procedures. I had the chance to work with two excellent undergraduates during my PhD: Evan Boyle who assisted with experimental methods for the hereditary prostate cancer work en route to what I am sure will be a prolific scientific career and subsequently, Jennifer Milbank for her patient help in validating results from sequencing. Finally, Rupali Patwardhan, Jerrod Schwartz, Martin Kircher and the rest of the Shendure Lab have been invaluable sources of advice friendship and camaraderie.

I was blessed to also collaborate with and learn from a number of excellent scientists and clinicians in the past few years including Peter Nelson, Janet Stanford, Matthew Rabinowitz, Robert Rostomily, Jim Olson, and Marshall Horwitz. Peter Nelson served as a co-advisor for nearly all of my work with prostate cancer and his dedication to improving the health of men with cancer shaped my approach to becoming a physician scientist. Liesel Fitzgerald, Ilsa Coleman, Tom White and many other members of the Nelson, Stanford and Olson labs were essential parts of the projects described here. James Maher, Yiannis Kaznessis, Jagesh Shah and Scott Nyberg played important roles in my early development as a scientist.

My thesis committee members Peter Nelson, Mary-Claire King and Jim Olson have provided helpful encouragement and guidance throughout these projects as well as the benefit of their wisdom and experience as I shape my own career.

I am also thankful to the MSTP administration, especially our program director Marshall Horwitz for his positive energy, careful advice and brilliant ideas. Marcie Buckner, Brian Giebel and Maureen Holstad ensured I didn't wander too far away while training. The Vicki and Gary Glant ARCS fellowship provided important financial support. I am particularly grateful to my E-08 MSTP classmates who have been supportive, fun companions over the past 6 years and to my medical school and graduate school colleagues as well.

I would like to thank my family: my parents Renu and Vipin and my siblings Vipasha and Avi. My parents spared no effort to encourage creativity and scholarship during my formative years and my family's love, laughter and encouragement have kept me smiling and inspired.

Finally I would like to thank my partner Neha for pushing me to follow my dreams even if it meant living thousands of miles away from her. Thank you for your love, patience and understanding throughout this journey.

Chapter 1 – Introduction

Somatic mutations in cancer

Cancer remains a major burden to public health in the United States. In 2010, more than 1.45 million people received a diagnosis and nearly 575,000 people died of cancer[1]. In the US, cost of care alone totaled nearly 125 billion dollars in 2010[2]. An important challenge facing oncologists is the fact that tumors of the same histopathological type can often have very different prognoses and responses to treatment.

One approach to understanding the molecular mechanisms underlying cancer has been to investigate somatic mutations that accumulate in cancer genomes. These mutations vary in type and include single nucleotide substitutions, small insertions and deletions ("indels"), larger copy number alterations and chromosomal translocations[3]. Some of these somatic mutations promote cancer progression by increasing mitotic activity, promoting metastasis or enabling therapeutic resistance. These mutations are commonly referred to as "driver" mutations to distinguish them from other "passenger" mutations that simply accumulate in the cancer genome.

The identification of driver mutations and the genes they alter has been a goal of cancer research for many decades. The field has historically used a variety of approaches to identify cancer genes that were largely dependent on the methods and technologies that were available at the time. Low resolution genomic surveys (e.g. cytogenetics), transformation assays using genetic material from cancers, and targeted investigation of candidate genes previously identified in biologic assays or hereditary cancer syndromes are some of the approaches used in the past[4-7]. These efforts have proven to be very successful and have yielded identification of hundreds of cancer genes for potential use in guiding diagnosis and therapy[7]. However, until ten years ago

unbiased high-resolution surveys of somatic mutations in cancer remained impossible due to the cost and difficulties associated with whole genome sequencing[7-9].

Improvements in DNA sequencing technologies enable profiling of somatic mutations

The advent of massively parallel sequencing has dropped the cost of whole genome sequencing (WGS) by five orders of magnitude, enabling WGS of tumors[10,11]. In common embodiments of this technology[10,11] complex shotgun libraries of DNA are constructed, amplified and arrayed on a solid surface such that millions of fragments of DNA are contained in an area no larger than a microscope slide. Individually labeled nucleotides are added with successful incorporation typically monitored using fluorescence. This process currently generates hundreds of millions of sequencing "reads" in each reaction with read lengths that currently range from 100-300 nucleotides. The massively parallel nature of this approach is a defining advantage over earlier methods that only investigated single DNA fragments in each reaction volume[11,12]. Analysis involves alignment of these reads to a (human) reference sequence, followed by base calling at each position covered by sequence reads. In cancer sequencing, somatic mutations are determined by direct comparison of tumor sequence with that from corresponding normal tissue. Various tools have been developed to perform each of these components of analysis[13-15].

Although WGS can be performed on tumor sequences, we commonly struggle to interpret driver events (point mutations) outside of coding regions. One strategy is to restrict sequencing and analysis to the "exome" or the set of exons across all protein coding genes[16]. The exome typically represents between 1-2% of the genome sequence and variation in these regions is comparatively easier to interpret as sequence changes typically alter proteins. Exome sequencing is currently cheaper than whole genome sequencing because of its focused approach and it has been widely deployed in

the study of human genetics and cancer[3,16-18]. However, an important limitation of exome sequencing is that it is unable to detect most chromosomal rearrangements and, by definition, noncoding mutations[17,19].

Several methods are used to selectively capture and sequence regions of the genome. The two methods I will be discussing in this thesis are Molecular Inversion Probe (MIP) sequencing and hybrid capture. The MIP method starts by synthesizing many single-stranded oligonucleotide "probes" that are each composed of a common sequence flanked by two sequences homologous to a specific target region of interest. During capture both ends of the MIP probe hybridize to the region of interest. Following a series of enzymatic steps, this region is copied into a covalently linked circular molecule that can be subsequently amplified in parallel to create sequencing libraries[18,20]. Hybrid capture involves hybridization of a sequencing library to complementary RNA or DNA probes designed against specified regions of the genome. Library molecules that specifically hybridize to the targeting probes are isolated and purified using magnetic beads and a series of wash steps. Both methods have been used in a variety of applications targeting different subsets of the genome[16,21-24]. In this thesis, I use hybrid capture to perform exome sequencing and MIPs to investigate the coding regions from a smaller subset of cancer-specific genes.

Results from tumor genome/exome studies have shed light on driver genes

Whole genome and whole exome sequencing have enabled many studies investigating the genomic landscape of cancer[25-28]. These investigations have identified new driver genes including *SPOP* in prostate cancer, *MLL2* and *MLL3* in medulloblastoma (and subsequently prostate cancer) and *ARID1A* in ovarian and gastric cancers[29-31]. Many of these driver genes are mutated in only a subset of patients. More generally, tumors from different individuals share only a small subset of mutations. While many of these differences are likely passenger mutations, some may drive cancer

progression and account for the variable clinical courses and responses to therapy observed by oncologists (interpatient heterogeneity).

Precision medicine

A growing compendium of cancer genome alterations combined with improvements in sequencing technology suggests an approach of "precision medicine" with a cancer patient's treatments tailored against the constellation of mutations in his/her tumor. Several previous successes motivate this approach. The development of the targeted tyrosine kinase inhibitor imatinib specifically inhibits growth of leukemic cells in patients with the *bcr-abl* translocation[32]. The *ERBB2* (*HER2*) amplification seen in 15-30% of breast cancers is now successfully targeted with trastuzumab with long term survival rates reaching >90% in some cases[33]. While inactivating mutations have been historically more difficult to target, pathway-directed therapies, such as the use of PARP inhibitors in *BRCA1/2* dysregulated cancers offers a compelling approach[34,35]. These decisions are currently made based using limited genetic/histologic investigation of tumor biopsies in conjunction with a patient's clinical history. Performing whole genome or targeted sequencing of tumor biopsies would greatly expand the scope of mutations assayed, and multiple studies are investigating its use to direct treatments with promising results[36,37].

Challenges posed by intrapatient/intratumoral heterogeneity

Mutation and clonal expansion are defining features of cancer. Mutations arising at every cell division introduce new differences in the genomes of neighboring cells. If the mutation results in even a slight growth advantage, cells with this mutation can clonally expand within a tumor[7,30]. Repeated cycles of mutation and clonal expansion can result in tumors with spatially segregated differences in mutation which I will hereafter refer to as intratumoral heterogeneity[38]. While evidence of intratumoral heterogeneity in individual genes or chromosomes was previously known, genome

sequencing has provided a more detailed view[27,28,39]. Gerlinger *et al* investigated multiple spatially distinct regions of primary and metastastic renal cell carcinoma and discovered regional differences in a number of driver genes including *PTEN*, *SETD2* and *KDM5C*[28]. The extent of this heterogeneity and the fact that it included known driver genes raised questions about the efficacy of single-tumor biopsies in informing targeted treatments. Additional work was needed to determine whether this was a general principle across cancer types or whether findings from renal cell carcinoma represented isolated examples of extreme cases of heterogeneity.

A related challenge is seen in patients with disseminated metastatic disease, in whom cells from one or more primary tumors have metastasized to different regions of the body. Founder cells of each metastasis might have originated from different regions within a primary tumor and thus differ at seeding. Resulting metastases could also independently acquire mutations that promote growth in the context of their unique environmental and therapeutic stressors. How well a single biopsy from a primary or metastasis captures the mutations across all tumors within a patient with disseminated disease depends on the degree of intrapatient heterogeneity across his/her tumors. This question is especially relevant in prostate cancer because most metastases occur in areas of the body that are difficult to access (e.g. bone and soft tissue) and each biopsy can be painful and carries risk of complication.

Thesis outline:

This thesis describes efforts to characterize the mutational heterogeneity of cancer using massively parallel sequencing. In Chapter 2, I describe a survey of aggressive prostate cancers that revealed a subset of tumors exhibiting hypermutation. This finding was subsequently confirmed and linked to defects in the mismatch repair pathway. In Chapter 3, I describe a follow-up study within a larger cohort of patients and tumors to both identify candidate genes and explore intrapatient heterogeneity in men with disseminated disease. In Chapter 4, I investigate regional intratumoral heterogeneity in gliomas. Appendices A-C contain supplemental information for chapters 2-4, respectively.

Over the course of my studies, I was also involved with a number of projects that did not fit within the scope of mutational heterogeneity in cancer. The remaining sections (Appendices D-F) contain the results of these studies. Appendix D discusses a study associating an inherited variant in *BTNL2* with aggressive prostate cancer. Appendix E describes a study involving whole genome sequencing of a patient with idiopathic pulmonary fibrosis.

Chapter 2- Molecular Landscape of Advanced Prostate Cancers

Note: This work was published as:

Akash Kumar, Thomas White, Alexandra P. MacKenzie, Nigel Clegg, Choli Lee, Ruth F. Dumpit, Ilsa Coleman, Sarah B. Ng, Stephen J. Salipante, Mark J. Rieder, Deborah A. Nickerson, Eva Corey, Paul H. Lange, Colm Morrissey, Robert L. Vessella, Peter S. Nelson, Jay Shendure. Exome Sequencing Identifies a Spectrum of Mutation Frequencies in Advanced and Lethal Prostate Cancers. Proceedings of the National Academy of Sciences USA. 2011. 108(41):17087-17092. PMC3193229.

Jay Shendure and Peter Nelson conceived the initial approach. Jay Shendure, Peter Nelson and I designed the research plan. Samples were provided by Colm Morrissey, Robert Vessella and Paul Lange. Ilsa Coleman and I performed DNA extractions. Alexandra MacKenzie and I performed library preparation and exome capture. Choli Lee performed Illumina sequencing. I performed all sequence and mutation analysis. Jay Shendure, Peter Nelson and I wrote the manuscript with input from all authors.

2.1- ABSTRACT

To systematically catalog protein-altering mutations that may drive the development of prostate cancers and their progression to metastatic disease, we performed whole exome sequencing of 23 prostate cancers derived from 16 different lethal metastatic tumors and 3 high-grade primary carcinomas. All tumors were propagated in mice as xenografts, designated the LuCaP series, in order to model phenotypic variation such as responses to cancer-directed therapeutics. Although corresponding normal tissue was not available for most tumors, we were able to take advantage of increasingly deep catalogs of human genetic variation to remove most germline variants. On average, each tumor genome contained ~200 novel, nonsynonymous variants, of which the vast majority was unique to individual carcinomas. A subset of genes was recurrently altered across tumors derived from different individuals, including *TP53*, *DLK2*, *GPC6*, and *SDF4*. Unexpectedly, 3 prostate cancer genomes exhibited substantially higher mutation frequencies, with 2,000-4,000 novel coding variants per exome. A comparison of castration-resistant and castration-sensitive pairs of tumor lines derived from the same prostate cancer highlights mutations in the Wnt pathway as potentially contributing to the development of castration resistance. Collectively, our results indicate that point mutations arising in coding regions of advanced prostate cancers are common, but with notable exceptions, very few genes are mutated in a substantial fraction of tumors. We also report a new subtype of prostate cancers exhibiting "hypermutated" genomes, with potential implications for resistance to cancer therapeutics. Our results also suggest that increasingly deep catalogs of human germline variation may challenge the necessity of sequencing matched tumor-normal pairs.

2.2- INTRODUCTION

Prostate carcinoma is a disease that commonly affects men, with incidence rates dramatically rising with advancing age[40]. Though the vast majority of these malignancies behave in an indolent fashion, a subset is highly aggressive and resistant to conventional cancer therapeutics. Though recent studies have detailed the landscape of genomic alterations in localized prostate cancers, including a report describing the whole genome sequencing of seven primary tumors[40-43], the genetic composition of lethal and advanced disease is poorly defined. Previous work demonstrates the importance of chromosomal rearrangements that include the TMPRSS2-ERG gene fusion as a frequent attribute of prostate cancer genomes with clear implications for tumor biology[44-46]. However, considerably less is known about the contribution of somatic point mutations to the pathogenesis of prostate cancer[42,43,47], including those specific somatic variants that may drive metastatic progression or the development of resistance to specific therapeutics such as those targeting the androgen receptor program[41-43]. In this study, we describe the application of whole exome sequencing[16] to determine the mutational landscape of 23 prostate cancers representing aggressive and lethal disease, including both metastases and primary carcinomas. All tumors were propagated in immunocompromised mice as tumor xenografts[48] in order to model the heterogeneity in tumor growth, response to treatment and lethality that exists in prostate cancer. Furthermore, these tumor xenografts have the advantage of little-to-no human stromal contamination, and provide the means to functionally test the consequences of mutations. Although corresponding normal tissue was not sequenced for most samples, we find that comparisons to increasingly deep catalogs of segregating germline variants based on unrelated individuals provides an effective filter, challenging the necessity of sequencing matched tumor-normal pairs. We identify a number of genes in which nonsynonymous alterations (somatic mutations or very rare germline mutations) are

recurrently observed, including variants in *TP53*, *DLK2*, *GPC6*, and *SDF4*. Surprisingly, we also identify 3 aggressive prostate cancers that exhibit a "hypermutated" phenotype, i.e. a gross excess of point mutations relative to the other tumors sequenced here as well as those prostate cancers that have been evaluated to date. Finally, a comparison of castration-resistant and castration-sensitive matched tumor pairs derived from the same site of origin highlights mutations in the Wnt pathway as potentially contributing to the development of resistance to therapeutic targeting of androgen receptor signaling.

2.3- RESULTS

Landscape of prostate cancer mutations. We performed whole exome sequencing of 23 prostate cancers derived from 16 different lethal metastatic tumors and 3 high grade primary carcinomas using solution-based hybrid capture (Nimblegen) followed by massively parallel sequencing (Illumina). Samples were designated as LuCaP 23.1 through LuCaP 147 in the order in which they were initially established as xenografts in mice (**Table 2.1**). Three tumors representing castration-resistant variants of the original cancers (LuCaP 23.1AI, LuCaP 35V and LuCaP 96AI) were also analyzed. Eight samples were captured against regions defined by the NCBI Consensus Coding Sequence Database (CCDS, 26.6 Mb), while the remaining 15 samples were captured using a more inclusive definition of the exome (RefSeq, 36.6 Mb) (**Table A.1**).

In order to filter contamination by mouse genomic DNA, sequence reads were independently mapped to both the mouse (mm9) and human (hg18) genome sequences, and only sequences that mapped exclusively to the latter were considered further. In each xenograft, 4 to 19% of total reads were discarded due to mapping to the mouse genome. After also removing duplicates, we achieved an average of ~100 fold coverage of the 26.6 Mb target in samples captured using the CCDS target definition, and an average of ~140 fold coverage of the 36.6 Mb target in samples captured using the RefSeq definition. Samples had between 90 to 95% of their respective target

definitions covered to sufficient depth to enable high quality base calling (see **Supplementary Table A.2** and **Supplementary Figures A.1**, **A.2**, and **A.3**). Across 23 tumors, we identified a non-redundant set of ~80,000 single nucleotide variants occurring within coding regions.

Most tumor sequencing analyses use matched tumor-normal pairs to distinguish somatic mutations present in the tumor from variants present in the germline of a given individual, with few exceptions[49]. However, the fact that the overwhelming majority of germline variation in an individual human genome is "common", coupled with the availability of increasingly deep catalogs of germline variation segregating in the human population, challenge the assumption that this is essential. As corresponding normal tissue was not available for many of these tumor samples, we used the approach of sequencing tumor tissue only, removing from consideration all variants that were also observed in the pilot dataset of the 1000 Genomes Project[50,51], as well as variants present in any of ~2,000 additional exomes sequenced at the University of Washington. After filtering, three tumors (LuCaP 58, LuCaP 73 and LuCaP 147) were observed to contain a very large number of single nucleotide variants relative to all other tumors: 4,067, 2,972 and 2,714 respectively (**Figure 2.1**). We refer to these xenografts as "hypermutated" and discuss their features below ("**Prostate cancers with "hypermutated" genomes**"). Excepting these three tumors, the applied filters reduced the number of coding variants under consideration from ~13,500 to ~350 per tumor (**Figure 2.1** and **Tables 2.1** and **A.3**). Of the 14,705 novel variants observed across the 23 tumors, 13,827 variants were called as heterozygous and 878 were called as homozygous, and 8,617 variants are predicted to cause amino acid changes (non-synonymous) including 8,176 missense, 346 nonsense, and 95 splice site variants (**Supplementary Table A.4**). These likely comprise a mixture of: (a) somatic mutations that were present in the original tumor, (b) somatic mutations occurring after tumor

propagation and evolution in the mouse hosts, (c) germline variants that were present in the individual-of-origin but are very rare in the population (i.e. "private" germline variation), and (d) false positive variant calls.

We next sought to assess the efficiency of filtering against databases of germline variation in enriching for somatic variants. For 3 tumors, LuCaP 92, LuCaP145.2 and LuCaP 147, benign tissue and tumor tissue was also collected directly from patients prior to propagation as xenografts. For two xenografts, LuCaP145.2 and LuCaP147, the fresh tumors were neighboring metastases from the same patient, whereas the fresh tumor for LuCaP 92 is the exact precursor lesion from which the xenograft was derived. However, based on the observations of Liu et. al, these metastases are likely to be closely related[52]. We sequenced the exomes of these tissues and determined somatic mutations by identifying positions that were called as variant in the xenograft, but not within normal tissue. We required that each base be covered by at least 24x in xenograft, tumor and normal tissue and used less stringent requirements to call a variant within the germline tissue to reduce the number of false positive somatic calls. In two of these three tumors (LuCaP 92 and LuCaP145.2), filtering against germline databases reduced the number of variants under consideration from ~21,000 to ~400 (**Table 2.1**), such 0.2% of all SNVs but ~30% of novSNVs (**Table 2.2**) were somatic mutations, i.e. a ~150-fold enrichment. Of note, ~11% of apparently true somatic mutations were removed by filtering against our databases of germline variation. These could either represent false negative variant calls within normal tissue, or else true recurrence of a somatic mutation in the same position as found in the germline database. The third tumor, LuCaP147, retained a high number of variants and represents a tumor class we term "hypermutated" (discussed below).

Recurrent nonsynonymous genomic sequence alterations in prostate cancers. We examined the set of novel, nonsynonymous single nucleotide variants (nov-nsSNVs) to

identify those genes that may be recurrently affected by protein-altering point mutations across different tumors. In order to reduce spurious findings due to inconsequential passenger mutations, we excluded the three "hypermutated" tumors from this analysis. We also manually examined read pileups for variants in genes with potential recurrence attributable to basecalling artifacts due to either insertions/deletions or poorly mapping reads. Across 16 tumors from unrelated individuals, 131 genes had nov-nsSNVs in two or more exomes, and 23 genes had nov-nsSNVs in three or more exomes (**Supplementary Table A.5**).

A subset of the novel variants are likely due to instances where very rare germline variants (i.e. not seen in several thousand other chromosomes) occur in the same gene, as we cannot distinguish these from somatic mutations. We therefore excluded from consideration the 1% of genes with the highest rate of very rare germline variants, i.e. singletons, based on an analysis of control exomes (as some genes are much more likely to contain very rare germline variants than other genes) [53,54]. This reduced the number of candidates to 104 genes with nov-nsSNVs in two or more exomes, and 12 genes with nov-nsSNVs in three or more exomes. To further segregate candidate genes with the goal of identifying those with recurrent somatic mutations, we estimated the probability of recurrently observing germline nov-nsSNVs in each candidate gene by iterative sampling from 1,865 other exomes sequenced at the University of Washington. We excluded from consideration genes for which the probability of observing the genes recurrently mutated due to germline variation was greater than 0.001. This reduced the number of candidates to 20 genes with nov-nsSNVs in two or more exomes, and 10 genes with nov-nsSNVs in three or more exomes (**Table 2.3**). Notably, whereas we began with 4 genes with nov-nsSNVs in four or more exomes (*MUC16*, *SYNE1*, *UBR4*, and *TP53*), only one of these (*TP53*) remained in our final candidate list, where it is the most significant (**Table 2.3**).

To estimate the "background" rate for calling genes as recurrently mutated via this approach, we analyzed 16 germline exomes from normal individuals that were captured using equivalent methods and applied the same filters. With the caveat that the overall number of coding alterations was lower in this set (an average of ~250 instead of ~350 novel variants per individual tumor), we identified 58 genes with nov-nsSNVs in two or more exomes with no p-value cutoff. Using the same threshold criteria (i.e. removing the top 1% of genes with the highest rate of germline variants and a p-value threshold of 0.001) reduced the number of genes with nov-nsSNVs in two or more exomes to 4 genes.

To further segregate candidate genes, we annotated positions with their conservation as measured with the Genomic Evolutionary Rate Profiling (GERP) score; variants at highly conserved positions would be predicted to be functionally significant [55]. This allowed us to identify a subset of "best candidates" that includes several previously determined to be mutated in advanced prostate cancer (e.g. *TP53*), and others with described roles in tumorigenesis, but not previously implicated in prostate cancer, including *DLK2* and *SDF4* (**Table 2.3; Discussion**). Determining which of these genes may be true driver mutations in prostate cancer will require the interrogation of larger cohorts, as well as functional characterization.

Mutations associated with castration resistant (CR) prostate cancer. Castration, or androgen deprivation therapy (ADT), is a commonly used treatment for advanced, disseminated prostate cancer. While effective initially, resistance inevitably develops, leading to a disease state termed castration resistant prostate cancer (CRPC) with high rates of cancer-specific mortality[41,51]. Our study included three tumors with castration sensitive (CS) and castration resistant (CR) derivatives: LuCaP 96/LuCaP 96AI, LuCaP 23.1/LuCaP 23.1AI and LuCaP 35/LuCaP 35V [51] (**Figure 2.2**). A comparison of exomes from each CR xenograft with that of its CS counterpart identified approximately 12-50

genes with nonsynonymous mutations that were present uniquely in the CR xenografts (**Supplementary Table A.6**). There were no genes recurrently mutated exclusively in CR tumors. To look for enrichment of mutations in genes encoding proteins comprising specific biochemical pathways in CRPC, we examined 880 gene sets using the MSigDB pathways database (http://www.broadinstitute.org/gsea/msigdb/). We found a significant enrichment for genes participating in Wnt signaling in castration resistant tumors: of 86 mutations unique to CRPCs, each tumor had at least one mutation in a member of the Wnt pathway ($q< 0.01$)[56]. These included *FZD6* (in LuCaP 23.1AI), *GSK3B* (in LuCaP 96AI), and *WNT6* (in LuCaP 35V) (**Supplementary Table A.6**).

Prostate cancers with "hypermutated" genomes. The genomes of three prostate cancers, LuCaP 58, LuCaP 73 and LuCaP 147 possessed a strikingly high number of nov-nsSNVs, nearly tenfold more than other tumors ($p=0.0097$) (**Figure 2.1**). There were no distinctive features to suggest why these tumors should have more variants. Each tumor originated as a high grade Gleason 9 cancer, all were from individuals of Caucasian ancestry, one represented a primary neoplasm, one a lymph node metastasis, and one a metastasis to the liver. The "hypermutated" phenotype also does not appear to be solely determined by the length of time a tumor was passaged in animals, as LuCaP 147 was started nearly ten years after most other xenografts in this panel. Further, tumors with hypermutated genomes did not exhibit substantially different patterns of structural changes compared to non-hypermutated tumors. As ascertained by array CGH, LuCaP 58, LuCaP 73 and LuCaP 147 had 1,582, 1,577, and 1,295 copy number variation (CNV) calls, respectively, compared to 1,470, 1,769, and 2,129 CNVs in non-hypermutated LuCaP70, LuCaP92, and LuCaP145.2 tumors (**Supplementary Table A.7**).

We hypothesized that the large number of nov-SNVs observed in three prostate cancers may be due to a ‘mutator phenotype’ that either developed during the initial

stages of tumorigenesis as a consequence of therapeutic pressures and subsequent clonal selection, or evolved while being passaged in the mouse hosts. To determine if these results reflect truly elevated numbers of somatic mutations within human tumors and not as a result of passage within mice, we sequenced the exomes of paired normal and directly resected non-xenografted, tumor samples corresponding to one hypermutated xenograft (LuCaP 147), and two non-hypermutated xenograft lines (LuCaP 92 and LuCaP 145.2) (**Supplementary Tables A.7 and A.8**). Of 2,122 novSNVs in LuCaP147 able to be called across all three samples (xenograft, derivative tumor and normal tissue) 1,464 were somatic and present in metastasis tissue (**Table 2.4**). In contrast, the other two non-xenografted tumors (corresponding to LuCaP 92 and LuCaP 145.2) had 31 and 57 somatic mutations respectively. Furthermore, because we sequenced a neighboring metastasis, rather than the exact metastasis from which LuCaP147 was derived, this result indicates that at least these ~1,400 somatic mutations were shared between these metastases. The vast majority of the ~600 somatic mutations observed in the LuCaP147 xenograft but not observed in the metastasis likely occurred during passage within mice, or else were mutations specific to the metastasis from which LuCaP147 was derived. Whole genome shotgun sequencing of this metastasis and its corresponding normal was also performed to 10x depth. Transition mutations account for ~90% of somatic mutations within the metastasis (**Supplementary Figure A.4**).

2.4- DISCUSSION

In this study, we performed a genome-wide analysis of protein-coding variation to identify sequence alterations in highly aggressive, lethal prostate cancers. Despite having only limited access to matched normal tissue for comparisons, we were able to exploit increasingly deep catalogs of segregating germline variation to highlight genes that may be recurrently mutated in prostate cancer. This strategy may be highly relevant

for the genomic analysis of carcinomas or tumor-derived cell-lines for which corresponding benign tissue is not available.

Overall, we identified 131 genes that had nov-nsSNVs in two or more tumors. Additional analysis based on the likelihood of observing rare germline variation highlighted 20 genes as candidates for recurrent somatic alteration, with the known cancer gene TP53 emerging as the top candidate from the analysis. We acknowledge that the genetic alterations observed in xenograft lines may not reflect changes originally present in a tumor, or may be a result of previously unseen germline variation, and it will be important to validate these candidates by establishing their prevalence in larger numbers of tumors for which matched normal tissue is available. However, these data provide an intriguing set of candidates for follow-up analysis. Several of these are discussed in further detail below.

We identified nov-nsSNVs in p53 in five out of the 16 independent tumors used to evaluate recurrence as well as in one of the "hypermutated" tumors. These variants included two positions that were called as homozygous (likely due to loss of heterozygosity) and are predicted to cause premature termination of the protein (**Table 2.3**). Hypermutated LuCaP 73 possessed two nov-nsSNVs in *TP53* after filtering including one in a mutational hotspot (175 ARG -> CYS). LuCaP 77 possessed a homozygous nov-nsSNV (278 PRO -> SER) that is also present in dbSNP 131. This SNV was previously described in a case of familial cancer syndrome (Li-Fraumeni Syndrome), and would have been lost if we had filtered against positions within dbSNP [57]. Three tumors harbored nov-nsSNVs within the gene encoding *DLK2*, a protein that shares similarity with the Delta transcription factor and has recently been shown to be involved in *NOTCH1* signaling during development [58]. Two *DLK2* nov-nsSNVs are in close proximity (at positions 361 and 371) in what is predicted to be a cytoplasmic domain and are in residues that are highly conserved evolutionarily (GERP score above

4.5). Three tumor genomes encoded variants in stromal derived factor (*SDF4*), a 363 amino acid calcium-binding protein whose function is poorly understood [59]. Two of the residues affected by nov-nsSNVs are highly conserved evolutionarily, with a GERP score above 4. Recent work has correlated low levels of *SDF4* expression with a poor prognosis in metastatic breast cancer[60].

Recently, whole genome sequencing of localized primary prostate cancers identified 165 genes that harbored somatic nonsynonymous mutations[40]. Of these, *PCDH15*, *LAMC1* and *GPC6* also had nov-nsSNVs in two or more advanced prostate cancers characterized in the present study. Both *PCDH15* and *LAMC1* are large (>1500 AA) and complex extracellular proteins that have a higher prior probability for somatic mutation or rare germline variants. *GPC6* encodes a smaller protein (~350 AA), and contains nov-nsSNVs at positions that are highly conserved (GERP score above 5) in two out of 16 non-hypermutated tumors as well as in one hypermutated tumor. *GPC6* encodes a glypican, a class of cell surface coreceptors for proteases implicated in cell growth and division[61-63].

Unexpectedly, we identified three tumors (representing 15% of those analyzed) with very high numbers of nov-nsSNVs. We confirmed that this hypermutated phenotype arose before passage in mice for at least one of these tumors (LuCaP 147), for which a non-xenografted tumor was available for comparison. These mutation frequencies far exceed those found in primary prostate cancers, as well as in most neoplasms arising in the breast, pancreas and brain, where comprehensive exome or genome sequencing studies have been performed[25,26,64]. However, cancers in the colon with mismatch repair gene defects[65], and those that arise in the lung and skin, where environmental genotoxins such as tobacco or UV sun exposure are implicated in disease etiology, have numbers of mutations that approach those present in these hypermutated prostate cancers[66,67]. The pattern of mutation observed in whole genome data argues against

tobacco exposure within the metastasis corresponding to LuCaP 147 as the characteristic predominance of G->T transversion mutations caused by polycyclic aromatic hydrocarbons was lacking[66]. Several nov-nsSNVs in the hypermutated tumors affect genes previously implicated in prostate cancer. For example, nov-nsSNVs in Androgen Receptor (*AR*) are observed in two of the hypermutated tumors, LuCaP147 and LuCaP73, including one well characterized gain of function mutation (877 THR ->ALA)[68]. However, the very large number of nov-nsSNVs within these tumors renders it difficult to distinguish disease-relevant mutations from likely passenger events.

One potential explanation for the large number of mutations seen in these samples is acquisition of a "mutator phenotype", in which alterations in DNA polymerase or DNA repair genes result in an accelerated rate of mutations[69,70]. In support of this, LuCaP58 possessed three candidate mutations in *MSH6*, a gene known to promote mismatch repair and microsatellite stability, including a particular substitution, 1284 THR-> MET, observed in individuals with Lynch syndrome[71]. This gene was previously seen to be mutated in prostate cancer where it associated with an increase in overall mutation rate, although with a more limited assessment of genomic sequence (1.3 Mb)[43]. Tumors with microsatellite instability are known to possess more mutations than other cancers; a recent analysis of colorectal cancer genomes detected approximately eight-fold more nonsynonymous variation in a tumor that displayed microsatellite instability, consistent with the number of mutations seen here[65]. We did not find nov-nsSNVs within DNA mismatch repair genes within the other two hypermutated prostate tumors (LuCaP 73 and 147), and thus a plausible explanation for the elevated mutation frequencies in these cancers remains to be established.

One limitation of this study is the use of tumor xenografts that may not precisely reflect the status of the tumor genome sampled directly from the patient. For those

xenografts where a corresponding non-xenograft tumor was available, the xenograft harbored ~2-fold more mutations (**Table 2.4**). This finding likely reflects continued tumor evolution and genotoxic stress over numerous population doublings, or further selective pressure to adapt to a murine host. However these xenografts are able to recapitulate many aspects of prostate cancer *in vivo*[72,73]. Thus, defining the genetic landscapes of these tumors allows one to use the xenografts as a means to both functionally test the consequences of mutation and evaluate therapeutics directed against pathways that are disrupted by specific genetic lesions.

In summary, by sequencing the exomes of 23 tumors representing a spectrum of aggressive advanced prostate cancers, we identified a large number of previously unrecognized gene coding variants with the potential to influence tumor behavior. Our results also indicate that with notable exceptions, very few genes are mutated in a substantial fraction of tumors. Furthermore, while the overall mutation frequencies approximate those found in other cancers of epithelial origin, we also identified a distinct subset of tumors that exhibit a hypermutated genome. It will be important to determine the mechanism(s) responsible for the enhanced point mutation rates in these malignancies, particularly if further studies demonstrate enhanced resistance to cancer therapeutics.

2.5- METHODS

Xenograft Tissues. The LuCaP series of prostate cancer xenografts used in this study were obtained from the University of Washington Prostate Cancer Biorepository and developed by co-author RLV within the Department of Urology[74]. DNA was isolated from frozen tissue blocks using the QIAGEN DNeasy Blood and Tissue kit.

Exome capture and massively parallel sequencing. The Nimblegen EZ SeqCap kit (Roche) was used as previously described in order to capture exons[75]. Shotgun libraries were constructed by shearing gDNA and ligating sequencing adaptors. Libraries were hybridized to either the EZSeqCap V1 or V2 solution-based probes, amplified and sequenced on either the Illumina GAIIx or HiSeq platforms (**Supplementary Table A.1**). V1 probes (used in eight samples) targeted 26.6 Mb corresponding to the CCDS definitions of exons, while V2 probes (used in 15 samples) targeted 36.6 Mb corresponding to the RefSeq gene database.

Read mapping and base calling. We dealt with the possibility of mouse gDNA contamination by mapping sequence reads to both the human (UCSC hg18) and mouse (mm9) genome sequences using BWA[14]. Reads that mapped to the mouse genome were excluded from further analysis. See **Supplementary Figures A.1, A.2,** and **A.3** for mapping statistics and calculations of mapping complexity. Sequence variant calls were performed by *samtools*[76] after removing potential PCR duplicates, and were filtered to consider only positions with more than 8x coverage and a Phred-like consensus quality of at least 30[16].

Identification of genes with sequence variation. To eliminate common germline polymorphisms from consideration, variants that had the same position as variants present in pilot data from the 1000 Genomes Project or in ~2,000 exomes corresponding to normal (non-tumor, non-xenografted) tissues sequenced at the University of Washington were removed from consideration (**Figure 2.1**). Genotypes were annotated

using the *SeattleSeq* server (http://gvs.gs.washington.edu/SeattleSeqAnnotation/) and only nonsynonymous variants (missense/nonsense/splice-site mutations) were considered in identifying genes with recurrent mutations. The subset of genes that were recurrently mutated was validated manually using IGV, the Integrated Genomics Viewer, to identify and remove false positive calls due to the presence of an insertion/deletion or incorrectly mapping read[50]. In order to estimate the significance of the three "hypermutated xenografts", the number of nov-SNVs present in LuCaPs 58, 73, and 147 were compared against other xenografts using a one-sided t-test assuming unequal variance.

Estimation of significance in genes with recurrent nov-nsSNVs. To distinguish genes that are observed to be recurrently mutated as a result of sampling germline variation between individuals from genes that are recurrently mutated as a result of somatic mutation in the tumor, we used exome sequence data from 1,865 individuals sequenced at the University of Washington.

To identify genes with the highest rate of very rare germline variants, i.e. singletons, we tabulated the genes that were affected by rare variants (nov-nsSNVs, defined as protein-altering mutations seen uniquely in this individual relative to all other exomes in the set of 1,865) for each individual. We estimated the likelihood of seeing a rare protein altering mutation in an individual by dividing the number of individuals with nov-nsSNVs in a given gene by the total number of individuals sampled.

We also used exome data on these 1,865 individuals to identify the likelihood of observing recurrence in a gene as a result of germline polymorphism. 16 individuals were randomly selected in each iteration, and for each of these 16 exomes we identified genes that were affected by nov-nsSNVs. We then looked for genes that recurrently contained nov-nsSNVs within the set of 16 individuals, and repeated this process 20,000

times to generate an estimate for the probability that a given gene would be observed to contain recurrent nov-nsSNVs due to previously unobserved germline polymorphism.

Assessments of filtering approaches. To test the effectiveness of our method of filtering germline variants, we sequenced normal and tumor tissue corresponding to each of three xenografts. Sequence data was processed through the same mapping pipeline (mapping to the mouse and human (hg18) reference, variant calling using *Samtools*) as was used for xenograft exome data. Positions called as a high quality variant (position has 24x coverage and a Phred-like consensus probability of at least 50) in the xenograft line were queried within both non-xenografted metastasis and normal tissues. To increase accuracy, only those positions with at least 24x coverage in both the metastasis and normal tissue were considered for each of three xenografts. A position was considered a somatic mutation that arose before xenografting if it was called as a variant within the xenograft tumor and metastasis, but not within the normal tissue. To account for the possibility of low coverage resulting in a miscall within normal tumor tissue, we used less stringent criteria to determine if a position was variant within normal tissue (At least 10% or 10 reads covering this position support this call). A position was considered a somatic mutation that arose after xenografting if it was called as variant in the xenograft and invariant within metastasis and normal tissue. If a position was variant within a xenograft as well as its corresponding metastasis and normal tissue, it was considered to be a germline polymorphism. This process was repeated only considering thos positions previously determined to be nov-SNVs in order to estimate the sensitivity of the germline filtering approach.

Genome Copy Number Analysis. Copy number variation (CNV) analysis was carried out using Illumina Infinium 660W-Quad Beadchips following manufacturer's standard protocols. Genotyping calls were generated for six samples (three "hypermutated", three randomly chosen other xenografts) using the Illumina BeadStudio software with Illumina

Human660W-Quad_v1_A.egt HapMap genotype cluster definitions. Data analysis was performed with Biodiscovery Nexus Copy Number 6.0 software. The SNP-FASST2 segmentation algorithm and default Illumina settings for significance, number of probes per segment, and gain and loss thresholds were used to identify regions of CNV for each sample. Statistical analysis was done using a two-sided t-test assuming unequal variance.

Estimating contribution of rare germline variation. To estimate the number of nov-nsSNVs that are germline in origin, we used exome sequence data from 16 normal individuals that had been both captured and sequenced at the University of Washington in a similar manner to tumors in this study, although they had been sequenced to a moderately lower depth. Sequence data was processed through the same variant calling and filtering pipeline (mapping to the mouse and human (hg18) reference, variant calling using *Samtools* and manual validation using IGV) as was used for xenograft exome data.

Identification of Castration Resistant (CR) specific mutations. To identify genes potentially involved in the development of Castration Resistance (CR), we compared the sequences of Castration Resistant lines with their corresponding Castration Sensitive (CS) lines (**Figure 2.2**). Variants were called as mentioned in "Read mapping and base calling", except positions were only considered if both CR and CS sequences had an 8x coverage and base quality of 30 as determined by *samtools*. Resulting genotypes were annotated using *SeattleSeq* server (http://gvs.gs.washington.edu/SeattleSeqAnnotation/) and only nonsynonymous variants (missense/nonsense/splice-site mutations) were considered. This subset of genes was then validated manually using IGV to ensure that variant alleles were not present in CS xenografts. We entered these genes into the MSigDB website (http://www.broadinstitute.org/gsea/msigdb/) using the "Investigate

Gene Sets" option. We looked for overlap with "KEGG gene sets", and report the q-value from the website [56].

2.6-FIGURES

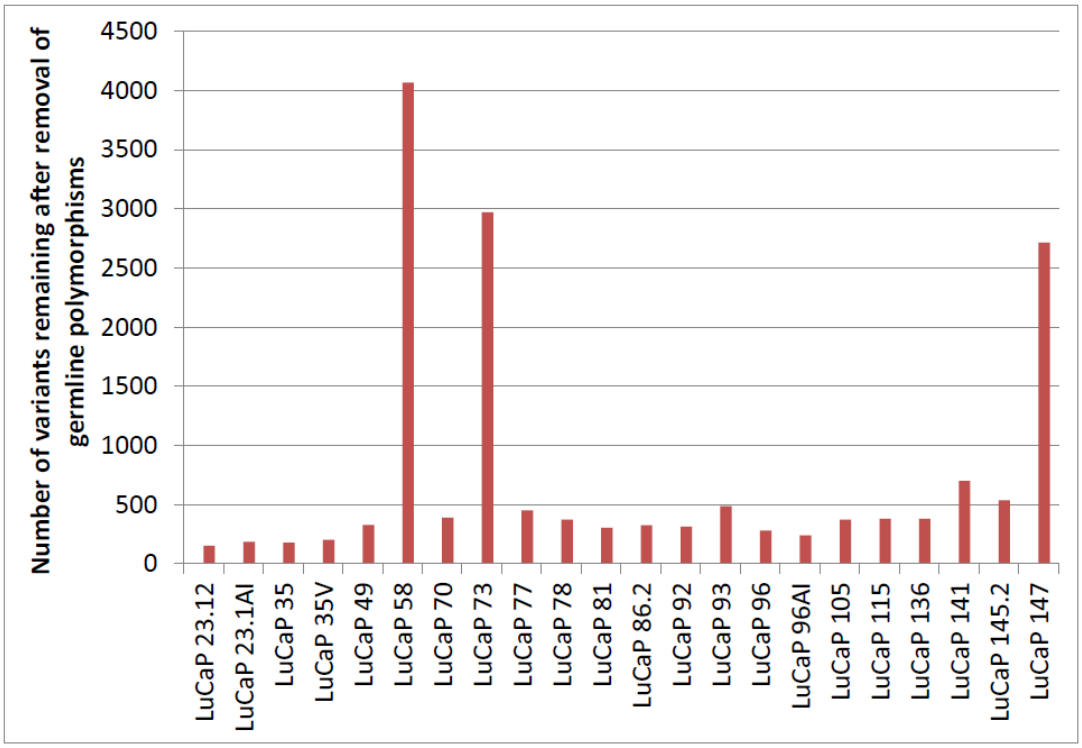


Fig. 2.1. A subset of xenografts exhibit a high number of mutations
After filtering to remove common germline polymorphisms, three xenografts (LuCaP 73, LuCaP 147 and LuCaP 58) exhibit a "hypermutated" phenotype, with several thousand novel SNVs each. This contrasts with the other 20 xenografts, which have 362 +/- 147 coding alterations remaining after filtering.

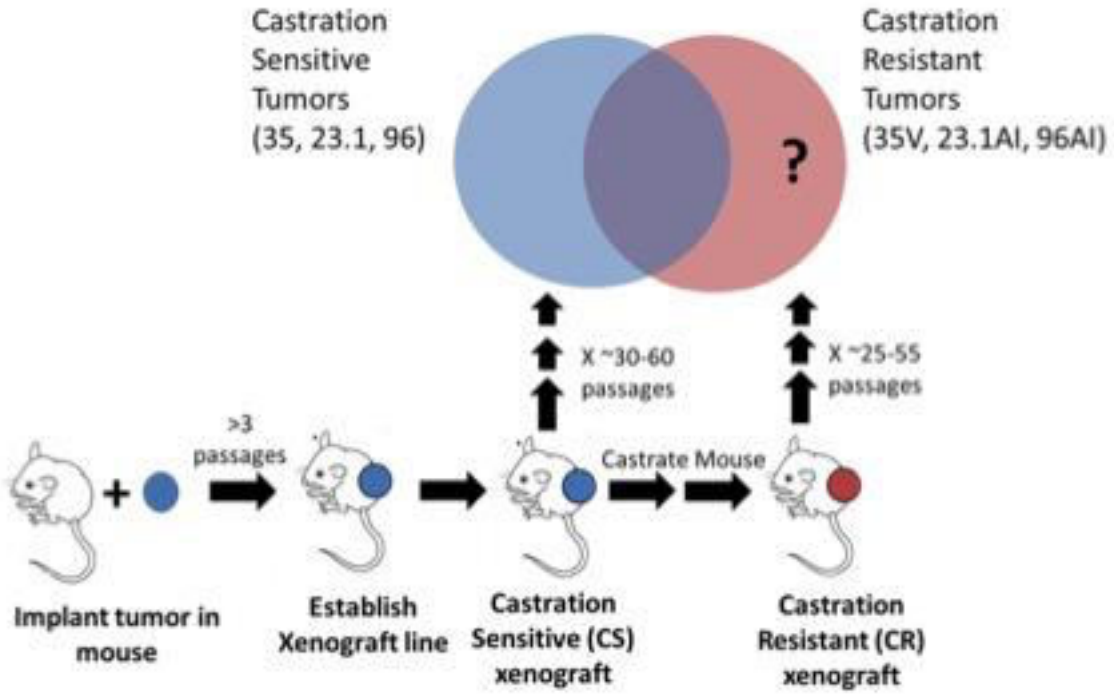


Fig. 2.2. Approach to identify genes involved in castration resistance
Immunocompromised mice are seeded with human prostate cancer tumors. Tumors grow and are serially passaged in mice, and are defined as a xenograft line after 3 sequential passages in mice. A subset of these mice is castrated, reducing androgen levels in the mouse and therefore causing the (Castration Sensitive) tumor to regress. After a period of time, an androgen insensitive tumor grows, after more than three passages in a castrated mouse the xenograft is called Castration Resistant (CR). We sequenced the exomes of both CR xenografts as well as their Castration Sensitive counterparts to identify genes with nonsynonymous mutations unique to CR tumors.

2.7-TABLES

	Source	Tissue Origin	Responsive to Castration	Histology	Ethnicity	# of variants within targeted region	# of variants within targeted region after filtering
LuCaP 23.1	Autopsy	LN	CS	adenocarcinoma	CAU	22165	643
LuCaP 23.12	Autopsy	Liver	CS	adenocarcinoma	CAU	14504	153
LuCaP 23.1AI	EXP	23.1	CR	adenocarcinoma	CAU	14230	186
LuCaP 35	OR	Inguinal LN	CS	adenocarcinoma	CAU	13847	179
LuCaP 35V	EXP	35	CR	adenocarcinoma	CAU	13858	203
LuCaP 49	OR	Omental Fat Met	CR	small cell carcinoma	CAU	16909	328
LuCaP 58	OR	LN	CR	adenocarcinoma	CAU	27339	4067
LuCaP 70	Autopsy	Liver Met	CS	adenocarcinoma	CAU	21910	389
LuCaP 73	OR	Prostate	CS	adenocarcinoma	CAU	26687	2972
LuCaP 77	Autopsy	Femur Met	CS	adenocarcinoma	CAU	23288	451
LuCaP 78	Autopsy	Peritoneal Met	ND	adenocarcinoma	CAU	24044	372
LuCaP 81	Autopsy	R. Pelvic Met	CR	adenocarcinoma	CAU	22992	304
LuCaP 86.2	OR	Bladder Met	CR	adenocarcinoma	UNK	16891	326
LuCaP 92	Autopsy	Peritoneal Met	CS	adenocarcinoma	CAU	21065	313
LuCaP 93	OR	TURP	CR	small cell carcinoma	CAU	23011	487
LuCaP 96	OR	TURP	CS	adenocarcinoma	CAU	14780	282
LuCaP 96AI	EXP	96	CR	adenocarcinoma	CAU	14518	240
LuCaP 105	Autopsy	Rib Mets R5,L5	CS	adenocarcinoma	CAU	21795	374
LuCaP 115	OR	LN	ND	adenocarcinoma	CAU	22898	380
LuCaP 136	OR	Ascites Fluid (cells)	CS	adenocarcinoma	CAU	22056	382
LuCaP 141	OR	Prostate TURP	CS	adenocarcinoma	ASIAN	23133	701

LuCaP 145.2	Autopsy	LN	CR	small cell carcinoma	CAU	23037	538
LuCaP 147	EXP	Liver Met	CR	adenocarcinoma	CAU	26741	2714

Table 2.1. The 23 prostate cancer xenografts analyzed for coding mutations
Our sample set consists of 23 LuCaP xenografts derived from 16 lethal metastatic tumors and three primary carcinomas from 19 unrelated patients with prostate cancer. Three pairs of xenografts (LuCaP 23.1AI and LuCaP 23.1, LuCaP 35 and LuCaP 35V, LuCaP 96 and LuCaP 96AI) were derived from the same individual. After removing from consideration all variants that were observed in the pilot dataset of the 1000 Genomes Project [50,51] as well as any variants present in any of ~2,000 additional exomes sequenced at the University of Washington, 20/23 samples had ~350 variants remaining. TURP refers to a sample derived from a transurethral resection of prostate procedure. Androgen status refers to whether or not a xenograft is able to grow when deprived of external androgens. Met, metastasis; LN, lymph node metastasis; CAU, Caucasian descent; EXP, Experimental; OR, operating room; ND, Not determined

Sample ID	# of coding variants	# of xenograft novSNVs	# of true somatic mutations	# of true somatic mutations observed within set of xenograft novSNVs
LuCaP 92	17092	193	56	51
LuCaP 145.2*	18455	281	122	106
LuCaP 147*	22458	2122	2045	1823

Table 2.2. **Efficiency of germline filtering in identifying somatic mutation**
We sequenced the exomes of normal and metastatic cancer tissue corresponding to three xenografts (LuCaP 92, 145.2 and 147), and, for this analysis, considered only those positions called at high confidence across all three tissues. The first two columns represent the number of coding variants and novSNVs (variants observed in xenograft exome that remained after filtering) occurring at coordinates that could be confidently base-called in all three samples. The next two columns describe the number of true somatic mutations (defined by comparison of the exomes of normal and metastatic cancer tissue) within the set of all variants and the set of novSNVs. For example, filtering reduced the number of variants in LuCaP 92 from 17,092 to 193 while preserving 51 of 56 somatic mutations (sensitivity of 91%). *Original tumor sample could not be identified, so a neighboring metastasis was used.

# of samples seen out of 16	Gene ID	Gene Name	Estimated P-value of being germline	Individual mutations seen
5	TP53	tumor protein p53 (Li-Fraumeni syndrome)	< 0.00005	LuCaP73(ARG306GLN), LuCaP136(ARG280stop), LuCaP96AI(CYS238TYR), †LuCaP92(GLU198stop), LuCaP73(ARG175CYS), LuCaP70(TYR163HIS), LuCaP77(PRO278SER)
3	SDF4	stromal cell derived factor 4	< 0.00005	LuCaP108(ASP276ASN), LuCaP78(GLY76SER), LuCaP115(ALA9SER)
3	PDZRN3	PDZ domain containing RING finger 3	< 0.00005	LuCaP96AI(ARG727CYS), LuCaP108(GLY570SER), LuCaP73(ARG463CYS), LuCaP92(ILE331LEU)
3	DLK2	delta-like 2 homolog	0.00005	LuCaP70(ARG371HIS), *LuCaP145.2(SER361ARG), LuCaP23.1AI(HIS280GLN)
3	FSIP2	fibrous sheath interacting protein 2	0.00005	LuCaP81(LYS22ASN), †LuCaP92(THR698ILE), LuCaP136(GLN1526HIS)
3	NRCAM	neuronal cell adhesion molecule	0.00015	LuCaP115(MET1094ILE), LuCaP86.2(LYS645GLU), †LuCaP145.2(SER329CYS)
3	MGAT4B	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme B	0.0002	LuCaP108(ALA504THR), LuCaP96AI(ARG168CYS), LuCaP136(VAL150MET)
3	PCDH11X	protocadherin 11 X-linked	0.0003	*LuCaP145.2(VAL38PHE), LuCaP58(MET867VAL), LuCaP108(VAL1007ILE), LuCaP49(THR1296ASN)
3	GLI1	glioma-associated oncogene homolog 1 (zinc finger protein)	0.0003	LuCaP86.2(ARG20TRP), LuCaP78(ARG81GLN), LuCaP23.1AI(PRO210THR)
3	KDM4B	Lysine-specific demethylase 4B	0.00035	LuCaP73(ALA265VAL), LuCaP108(ARG534TRP), LuCaP35V(ALA555VAL), LuCaP73(ALA827VAL), LuCaP86.2(SER1036CYS)
2	DKK1	dickkopf homolog 1 (Xenopus laevis)	< 0.00005	†LuCaP92(GLU151GLN), LuCaP93(SER244TYR)
2	RAB32	RAB32, member RAS oncogene family	0.00005	LuCaP93(VAL66ILE), LuCaP141(SER109stop)
2	PLA2G16	phospholipase A2, group XVI	0.00015	LuCaP115(SER85LEU), LuCaP35V(PRO19HIS)

2	TFG	TRK-fused gene	0.00015	LuCaP96AI(ASN134HIS), LuCaP141(GLN318stop), LuCaP147(TYR319stop)
2	TBX20	T-box 20	0.0002	LuCaP77(ARG437HIS), LuCaP23.1AI(ALA52SER)
2	ZNF473	zinc finger protein 473	0.00025	LuCaP108(VAL465ILE), LuCaP115(GLY652ARG)
2	SF3A1	splicing factor 3a, subunit 1, 120kDa	0.0006	LuCaP70(PRO558LEU), LuCaP23.1AI(VAL479ILE)
2	NMI	N-myc (and STAT) interactor	0.00075	LuCaP141(ILE302ARG), LuCaP86.2(GLN101ARG)
2	IKZF4	IKAROS family zinc finger 4 (Eos)	0.0008	LuCaP93(ASP106ASN), LuCaP81(ASP498ASN)
2	BDH1	3-hydroxybutyrate dehydrogenase, type 1	0.00095	LuCaP73(VAL190ILE), LuCaP96AI(THR176MET), LuCaP147(VAL142ILE), LuCaP115(HIS74TYR), LuCaP147(ALA50VAL)

Table 2.3. **Genes with recurrent novel, nonsynonymous alterations**
This analysis excludes LuCaP 73, LuCaP 147 and LuCaP 58 as well as the castration resistant lines LuCaP 35V, LuCaP 96AI and LuCaP 23.1AI. P-values were estimated by randomly sampling from 1,865 other exomes sequenced at the University of Washington to estimate the probability of recurrently observing nov-nsSNVs in a given candidate gene. These are the 20 genes with the best p-values; a full list of 131 candidates is provided in **Supplementary Table A.5**. *This nov-nsSNV was determined to be a rare germline mutation within this xenograft. †This nov-nsSNV was determined to be a somatic mutation within this xenograft.

Sample ID	Number of somatic mutations in unique to metastasis	Number of somatic mutations shared by metastasis and xenograft	Number of somatic mutations unique to xenografts
LuCaP 92	8	31	25
LuCaP 145.2*	35	57	65
LuCaP 147*	91	1464	581

Table 2.4. Hypermutation phenotype arose before xenografting within LuCaP 147
After sequencing metastases and normal tissue corresponding to three xenografts, we calculated the number of somatic mutations shared by xenografts and a corresponding metastasis. In this table, somatic mutations are classified according to their presence in the metastasis and xenograft (in metastasis only, both metastasis and xenograft, and xenograft only). 1464 of ~2045 novSNVs within LuCaP 147 were also present within a different lung metastasis from the same individual. However, in all xenografts a substantial number of mutations (25 within LuCaP 92 and 65 within LuCaP 145.2) appear to have developed after xenografting. *Original tumor sample could not be identified, so a neighboring metastasis was used. These numbers therefore represent the minimal overlap between xenograft and the metastasis from which it was derived.

Chapter 3- Intrapatient Heterogeneity in Advanced Prostate Cancer

3.1- ABSTRACT

Genome surveys of mutations occurring in tumors from populations of patients with prostate cancer have provided many insights into the molecular landscape of this common malignancy. However, considerably less is known about the mutational heterogeneity that exists within patients with disseminated disease. Here we sequenced the exomes of aggressive prostate cancer from 55 patients including 16 primary tumors and 115 metastases. We identify a number of mutations in novel genes potentially involved in disease pathogenesis including *DGKB* and *FOXA2* and confirm our earlier findings of a hypermutated subtype. We also characterize the extent of intrapatient heterogeneity across potentially clinically relevant genes including *AR*, *PTEN* and *RB*. We find that while all metastases from a patient share a clonal origin, there is substantial mutational heterogeneity in clinically-relevant genes, especially in those genes involved in therapeutic resistance.

3.2- INTRODUCTION

Prostate cancer (PrCa) remains a major public health problem primarily because a subset of cases progress to systemic disease, with metastases that are generally resistant to therapy[77-79]. A number of whole exome and whole genome investigations of localized PrCa have expanded our understanding of molecular drivers to include genes like *SPOP* and *CHD1* [78,80-82]. While fewer studies have examined mutations in aggressive prostate cancer, these have revealed molecular alterations in genes involved with androgen signaling in addition to members of the *PI3K* and *AKT* pathways[41,52,78,83].

As improvements in technology drive down the cost of sequencing, and pathway-specific therapies continue to emerge, oncology is moving towards a paradigm of "precision medicine"[37,78]. In this model, tumor genomes would be profiled upon patient presentation and treatments tailored towards the specific pathways disrupted by mutation. This approach relies

on an ability to sample and detect clinically relevant mutations that are representative of a patient's tumor burden. Patients with prostate cancer often present to physicians with disseminated disease, with multiple metastases in lymph nodes, soft tissue and bone. This raises a number of practical questions regarding the implementation of precision medicine. If a man presents with metastases and has an intact prostate (i.e. has not undergone radical prostatectomy), do the mutations in the primary tumor adequately represent the mutations present within the metastases? In addition, how well do metastases that are more accessible (e.g. in lymph node) represent metastases to bone or soft tissue? While previous work investigating copy number alterations in metastatic prostate cancer suggested that all metastases in a patient share a clonal origin[52], the extent of intrapatient mutational heterogeneity is presently unclear, especially for point mutations.

To address these questions, we investigated mutations within a total of 131 tumors from 55 patients who died of aggressive prostate cancer by whole exome sequencing. Our findings confirm that a subset of prostate cancer tumors are hypermutated, with 5-10 fold more mutations that the typical prostate cancer[83]. We also discover recurrent mutations involving several novel genes with plausible roles in driving tumor development or progression. We also generat estimates of the extent intrapatient genomic heterogeneity in patients with disseminated prostate cancer, and use this information to evaluate the accuracy of sampling a single tumor for targeted treatments in patients with multiple metastases.

3.3- RESULTS AND DISCUSSION

Tumors investigated

The tumors used in this study were obtained at rapid autopsy from 55 patients and include a total of 115 metastases and 16 treated matched primary prostate cancers. Metastases were obtained from a variety of sites including the lymph nodes (n=58), liver (n=20), lung (n=12) and bone (n=9) in addition to the adrenal gland, kidney, peritoneum, scrotum, skin and spleen. For 38 patients, we investigated more than one tumor per patient (range of 2 to 9). All patients were previously treated with standard androgen-deprivation and/or cytotoxic therapies. Further clinical information can be found in **Supplementary Table 1**.

Somatic Mutations

We performed exome sequencing (Nimblegen capture and Illumina sequencing) on all tumors as well as corresponding normal tissue to identify single nucleotide variants (SNVs) and small insertions and deletions. We also ran array CGH (Agilent) on 125 tumors to determine copy number alterations. We previously observed a subset of tumors with hypermutated genomes in a panel of aggressive prostate cancer xenografts[83,84]. Within our set of 55 patients, five patients (9%) had high numbers of mutations, slightly lower than our previously estimated frequency of 15-20%. Further investigation revealed that hypermutated samples displayed defective mismatch repair activity with structural alterations in MSH2 and MSH6 (C. Pritchard *et al.,* manuscript submitted). We sequenced additional metastases from any patients displaying a hypermutation phenotype. In each of these cases, all tumors were hypermutated, consistent with the scenario of a defect in mismatch repair arising prior to metastasis (**Supplementary Figure B.1**).

We next looked for frequently mutated genes within the combined set of castration resistant tumors including results from Grasso *et al.* using MutSigCV[85]. After correcting for multiple hypothesis testing, *TP53* and *AR* were the only significantly mutated genes. However, the top ranked genes included other genes previously found to be significantly mutated in

prostate cancer such as *SPOP* and *ZFHX3* (**Figure 3.1**). We also identified stereotyped mutations that are known to be driver mutations in other cancers, including the T41A mutation in *CTNNB1* and K601E mutation in *BRAF* (**Supplementary Figures 2 and 3**).

Metastatic tumor-specific candidate genes

We next sought to identify the subset of genes that may be specifically enriched in advanced/aggressive disease. For this analysis we used several publically available data sources including 48 patients with Castration Resistant Prostate Cancer (CRPC)[86], as well as exome data from three cohorts of primary tumors: Barbieri et al.[80], Lindbergh *et al.*[87] and TCGA (unpublished) which had 112, 55 and 300 tumors respectively. We performed a two sample t-test comparing the frequency of mutation in each gene in the two groups (primary or metastasis) and combined the p-value from this analysis with that obtained previously from MutSigCV to prioritize candidate prostate-cancer specific driver genes.

This approach yielded multiple candidate genes (**Table 3.1**) with TP53 and AR at the top of the list. Interestingly, *AR* was mutated at 12.7% and 10.4% within two studies of metastatic castration-resistant prostate cancer (CRPC) but *AR* mutations were not seen in any of the three studies containing more than 350 untreated primary tumors. This is likely due to the fact that *AR* mutation occurs after androgen deprivation therapy (ADT) and that primaries are almost always treatment-naïve. Several other genes followed a similar pattern of mutation, although with lower frequency in CRPCs. One of these genes was *DGKB*, which encodes the beta subunit of diacylglycerol kinase. Mutations in *DGKB* were seen in three patients within our study and in no primary tumors across our meta-analysis. Interestingly, DGKB mutations clustered in the catalytic domain of the kinase (with recurrence seen after including other studies), suggesting a possible functional impact of mutations. *DGKB* plays a role in regulating the concentration of diacylglycerol and plays a role in multiple pathways including mTOR signaling.

Another gene, *FOXA2* (closely related to *FOXA1*) was mutated in three metastases. While this gene did not score highly in the previous analysis, we were intrigued to find that

mutations (H410Y, N429K and A435T) clustered within the C-terminus of the protein, mirroring the pattern of mutation seen in the known androgen co-activator *FOXA1* (**Figure 3.2a**). *FOXA2* is closely related to *FOXA1*, and has been associated with the neuroendocrine subtype of prostate cancer and is also currently thought to function as a transcriptional activator for liver-specific genes[88]. Expression of mutant *FOXA2* within HEK293 cells revealed increased AR signaling when compared with wild type *FOXA2*, suggesting that these mutations may play a role in castration resistance (**Figure 3.2b**).

Intrapatient Heterogeneity: Metastasis vs. Metastasis

To determine how well a single metastasis would represent the tumor burden within a patient with disseminated disease, we investigated the extent of heterogeneity in possible actionable/driver genes across tumors within a patient. We started by curating a list of genes that are either clinically actionable or else putative drivers of prostate cancer. We identified a total of 150 genes, composed of 100 genes thought to be clinically actionable across any cancer type and 50 genes with a putative driver role in prostate cancer (**Figure 3.3**). We next investigated tumors in each patient for mutations in these genes. A gene was considered to have a positive hit in a tumor if it either possessed a nonsynonymous point mutation (missense, splice, nonsense) or was subject to either homozygous copy number loss or high-level amplification (**Supplementary Figure B.4**). All patients shared somatic mutations, consistent with previous findings of a clonal origin of metastases[52]. This was also true for mutations thought to be early driver events; for example in four patients with *SPOP* mutation all eight metastases shared the same mutations. However, tumors displayed heterogeneity in genes that affect resistance such as *AR*, *PTEN* and *RB1*. Out of 18 patients with *AR* mutation or amplification, six had evidence of intrapatient heterogeneity (**Figure 3.4**, **Supplementary Figure B.4**). These findings suggest that tumors had already metastasized prior to treatment and assuming that all tumors were proliferating prior to autopsy, metastases from the same

patient may have acquired resistance to treatment through different mutations in the same gene or different genes in the same pathway.

Intrapatient Heterogeneity: Primary vs. Metastasis

To address the question of how a treated primary tumor compares with its metastasis, we investigated 13 patients for which we sequenced a treated primary and metastasis. To perform comparisons we restricted our analysis to those positions that were covered to greater than 30x depth and we required an 20% of reads to contain a mutation for it to be considered. We found that all primaries we investigated shared more than half of mutations with the metastasis, indicating a shared clonal origin of tumors (**Figure 3.5**). One exception was the primary and metastasis from patient 04-149, which exhibited substantial divergence. In this case, the primary had high frequency mutations in a number of genes including a missense mutation within SETD7 that were absent from the metastasis.

Our exome study has shed more light on the landscape of aggressive prostate cancer, providing greater insight into the mechanisms of metastasis and resistance to therapy. We identified possible driver mutations in genes within known pathways, including *FOXA2*. Our findings suggest that while metastases share a clonal origin, they may exhibit heterogeneity in several genes including those involved in drug resistance. Our findings are especially relevant for cases of men with disseminated prostate cancer who relapse after treatment. Moving forward, it will be important to explore alternative, less invasive methods such as DNA sequencing of circulating tumor cells (CTCs)[89] or circulating cell-free tumor DNA in individuals to capture this heterogeneity.

3.4- METHODS

Tissue and gDNA extraction

All tumor samples were obtained from the Rapid Autopsy Program at the University of Washington. All samples were collected with the informed consent of patients and Institutional Review Board approval. gDNA was isolated from frozen tissue blocks using the QIAGEN DNEasy Blood and Tissue kit.

Copy number analysis

Copy number analysis was carried out using Agilent Sureprint G3 2X 400K custom aCGH array following manufacturer's standard protocols and using Promega Male gDNA as a control. Data analysis was performed with Biodiscovery Nexus Copy Number 6.0 software. The SNP-FASST2 segmentation algorithm was used with default settings to identify regions of copy number variation (CNV) for each sample.

Exome sequencing

Exome sequencing was performed using the Nimblegen V2 or V3 platforms as previously described with the following modification: in a subset of tumors, individually barcoded libraries were pooled in pairs prior to capture (**Supplementary Table 1**). Sequencing was performed using the Illumina Hiseq 2000 with either 50 bp paired reads or 100 bp paired end sequences. Reads were mapped to the human reference genome sequence (hg19) with bwa v0.7.1[14]. After removal of PCR duplicate pairs we performed local realignment around indels using the Genome Analysis Toolkit (GATK)[13]. We subsequently called mutations using the Mutect software package with the following parameters: "--minimum_normal_allele_fraction 0.02 --max_alt_alleles_in_normal_count 12 --intervals poscont.list --fraction_contamination 0.02". To remove common polymorphisms and enrich for likely somatic mutations, we imposed a number of additional requirements, including requiring variants to be observed in at least 10% of reads at a position and removing variants present within a modified database of SNPs (dbSNP v137) that had first been stripped of all COSMIC variants. We investigated mutations for significance

with MutSigCV using standard parameters, and inspected mutations in the top 50 significant genes manually using the Integrated Genomics Viewer (IGV) to remove sequence artifacts.

Barcode Crosstalk

In the course of our analysis, we found that samples paired with each other prior to exome capture suffered from a low level of barcode cross talk. This crosstalk results in reduced sensitivity of detection of somatic mutations and complicates comparisons between one tumor to another. To attempt to correct for this phenomenon, we compared the frequency of mutation in each sample with that of its paired sample and flagged those sites where mutation in one sample could be explained by barcode crosstalk. These cases are referred to as "ambiguous mutations" in **Supplementary Figure B.4**.

3.5- FIGURES

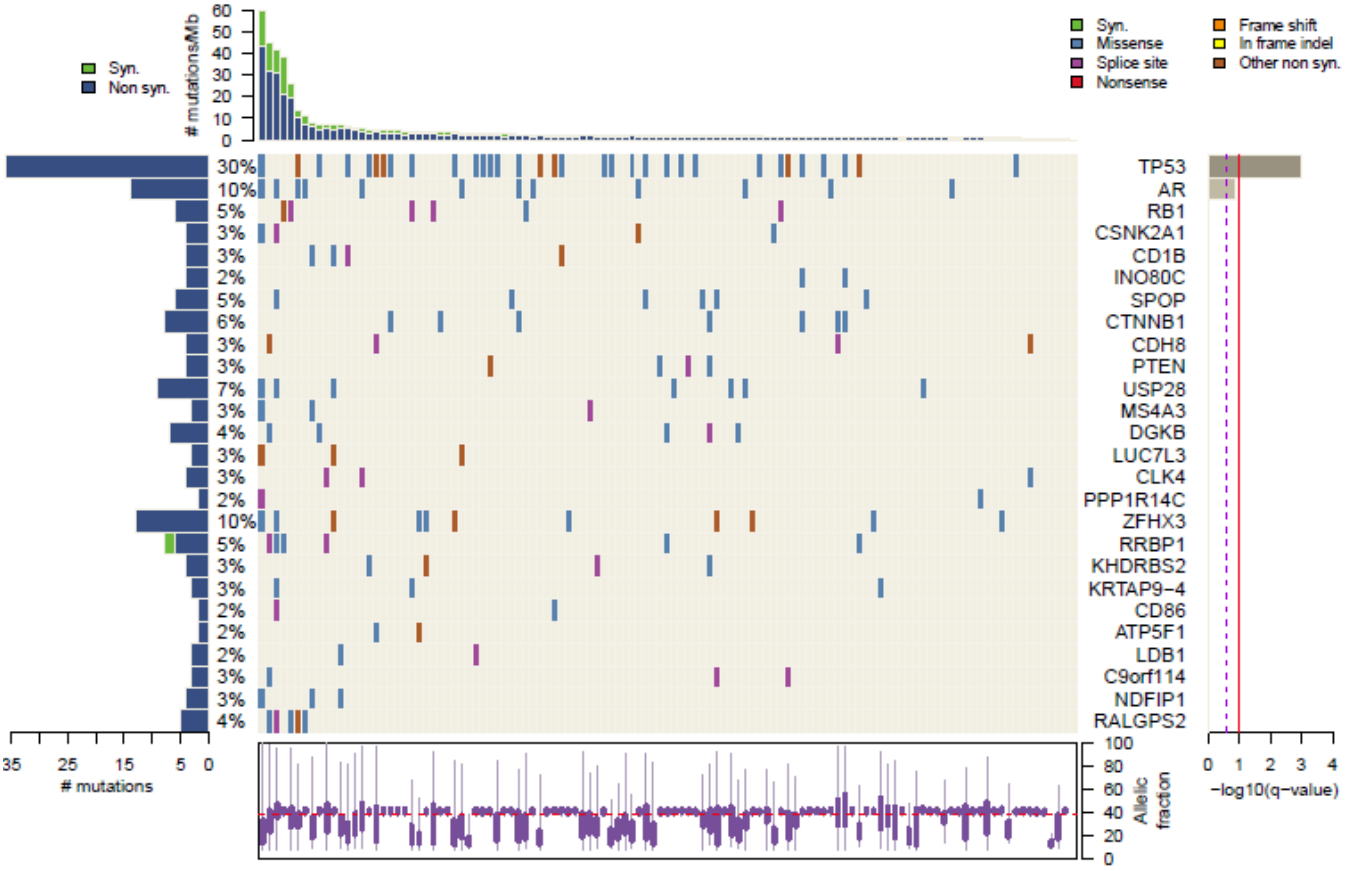


Figure 3.1: Summary of candidate genes involved with aggressive PrCa. Plots show (clockwise from top): the mutations per megabase in each individual, significance of each gene (as determined by MutSigCV), a boxplot of the allele fraction of mutations across each tumor, a barplot of the frequency of mutation across each gene and finally a heat map of the distribution of mutations across individuals (mutation type coded by color). This plot incorporates data from Grasso et al., yielding data from 103 patients with castration resistant prostate cancers (55 from this study).

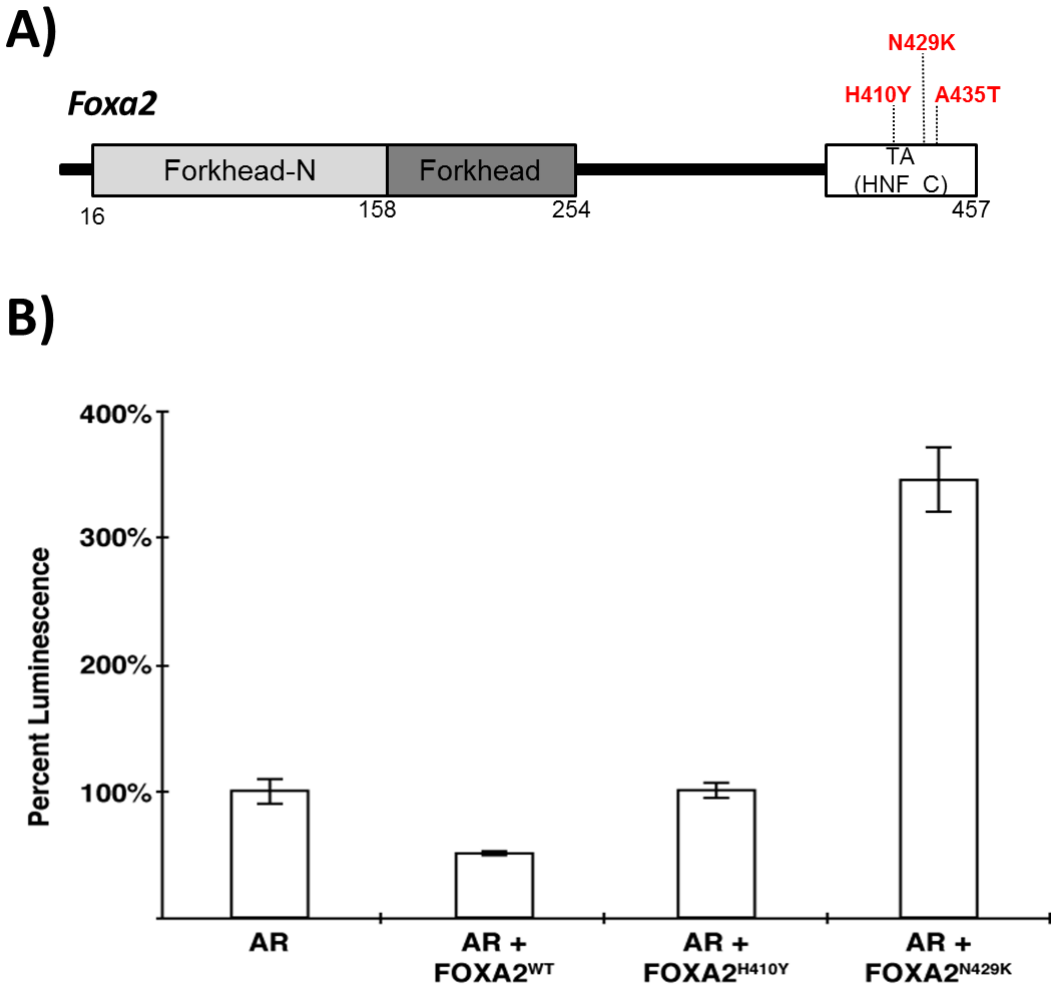


Figure 3.2: Mutations in the gene *FOXA2* promote AR signaling. A) Exome sequencing of tumors identified 3 individuals with *FOXA2* mutations. The locations of each of these mutations are indicated on the domain structure of *FOXA2*. B) *FOXA2* mutants increase AR signaling. WT *FOXA2* and *FOXA2* mutants observed in clinical samples were stably expressed in HEK293 cells and AR signaling was assayed using a luciferase reporter assay. Mutant N249K activated reporter activity higher than AR alone.

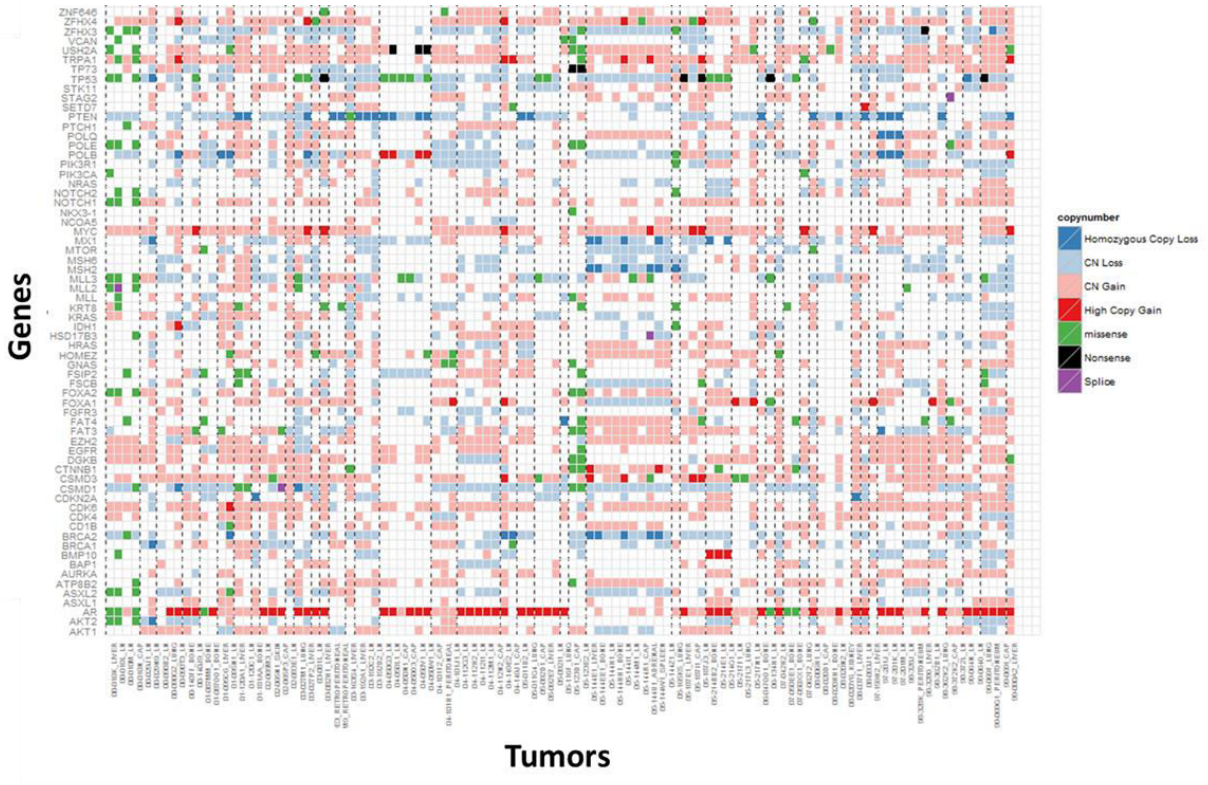


Figure 3.3: Extent of mutational heterogeneity in clinically relevant genes across 55 patients. For the purpose of clarity, only mutations in a subset of putative prostate cancer driver genes are shown. Tumors are grouped by patient (dotted lines). While all metastases from each patient appear to share a clonal origin, many contain unique mutations in genes that may be important. See **Supplementary Figure B.4** for further information on each patient.

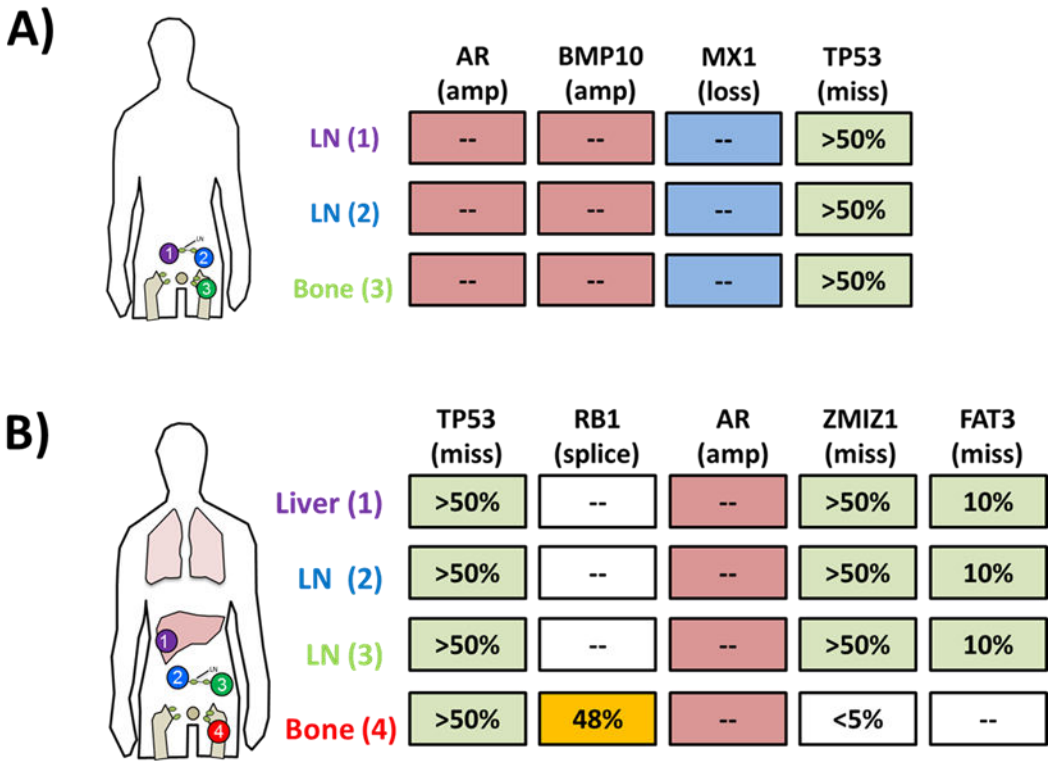


Figure 3.4: Intrapatient comparisons of mutations in clinically relevant genes. A) Tumors from this patient share all clinically relevant genes. B) Tumors from another patient likely share clonal origin (evidenced by sharing of TP53 mutation), but differ with respect to a splice mutation in RB1.

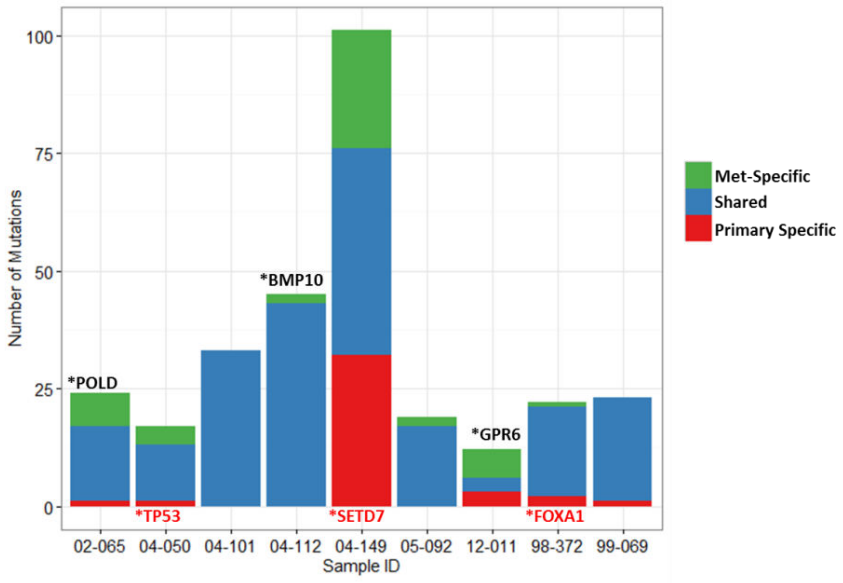


Figure 3.5: Comparison of treated primary vs. metastasis across nine patients. Color indicates whether or not mutations were shared or specific to either metastasis or primary.

3.6- TABLES

Gene	Kumar et al. (55, CRPC)	Grasso et al. (48, CRPC)	Barbieri et al. (112, Prim)	Lindberg et al. (55, Prim)	TCGA (175, Prim)	p from MutSigCV	p value (2 sample t)
TP53	27.3%	33.3%	5.4%	12.7%	6.9%	1.0E-09	4.7E-03
AR	12.7%	10.4%	0.0%	0.0%	0.0%	1.4E-05	4.5E-04
ZFHX3	10.9%	8.3%	0.9%	0.0%	1.1%	6.2E-03	1.7E-03
USP28	9.1%	4.2%	0.9%	1.8%	0.6%	3.0E-03	3.1E-02
CTNNB1	7.3%	6.3%	0.0%	1.8%	2.9%	2.5E-03	9.9E-03
ATP8B2	7.3%	4.2%	0.0%	0.0%	0.6%	5.5E-02	9.3E-03
RRBP1	7.3%	4.2%	1.8%	1.8%	0.6%	6.4E-03	2.1E-02
RALGPS2	7.3%	2.1%	0.0%	0.0%	0.0%	2.7E-03	4.7E-02
RB1	7.3%	2.1%	0.0%	0.0%	0.6%	6.6E-04	5.2E-02
DGKB	5.5%	4.2%	0.0%	0.0%	0.0%	3.8E-03	1.1E-03
ROBO2	5.5%	4.2%	0.9%	0.0%	1.7%	7.1E-02	8.0E-03
CSNK2A1	5.5%	2.1%	0.9%	0.0%	0.0%	9.2E-04	3.9E-02
ASXL2	3.6%	8.3%	2.7%	0.0%	0.0%	5.1E-03	4.7E-02
CD1B	3.6%	4.2%	0.9%	0.0%	0.6%	1.4E-03	1.6E-03
KHDRBS2	3.6%	4.2%	0.9%	1.8%	0.0%	8.7E-03	1.2E-02
ABCG2	3.6%	2.1%	0.0%	0.0%	0.0%	8.5E-02	7.9E-03
ADSSL1	3.6%	2.1%	0.0%	0.0%	0.0%	9.8E-02	7.9E-03
ANXA10	3.6%	2.1%	0.0%	0.0%	0.0%	5.6E-02	7.9E-03
C19orf55	3.6%	2.1%	0.0%	0.0%	0.0%	2.3E-02	7.9E-03
CLK4	3.6%	2.1%	0.0%	0.0%	0.0%	4.9E-03	7.9E-03

Table 3.1: Candidate CRPC-specific genes obtained via meta-analysis. We compared mutation frequency of CRPCs in our 55 patients as well as an additional 48 patients with CRPCs [86], with that of primary tumors from three studies: Barbieri et al.[80], Lindbergh *et al.*[87] and TCGA (112, 55 and 300 tumors respectively). P values refer to the result of either a two-sample t-test comparing mutation frequencies within metastases and primary tumor groups and estimates of significance obtained via MutSigCV.

Chapter 4- Intratumoral Heterogeneity in Glioblastoma

Note: This chapter is based on a recently submitted manuscript:

Akash Kumar, Evan A. Boyle, Mari Tokita, Andrei M. Mikheev, Michelle C. Sanger, Emily Girard, John R. Silber, Luis F. Gonzalez-Cuyar , Joseph B. Hiatt, Andrew Adey, Choli Lee, Jacob O. Kitzman, Donald E. Born, Daniel L. Silbergeld, James M. Olson, Robert C. Rostomily, Jay Shendure. Deep sequencing of multiple regions of glial tumors reveals spatial heterogeneity for mutations in clinically relevant genes.

This project grew out of discussions between Jay Shendure, James Olson, Robert Rostomily and I. Daniel Silbergeld, Michelle Sanger, Emily Girard, Andrei Mikheev and I obtained samples. I performed DNA extractions with help from Evan Boyle. Mari Tokita performed FISH and IHC with help from Donald Born and Luis Gonzalez-Cuyar. Evan Boyle and I performed TaqMan validations. I prepared MIP sequencing libraries and performed all analyses of sequence data. Jay Shendure, Robert Rostomily and I wrote the manuscript with input from all other authors.

4.1- ABSTRACT

The extent of intratumoral mutational heterogeneity remains unclear in gliomas, the most common primary brain tumors, especially with respect to point mutation. To address this, we applied single molecule Molecular Inversion Probes (smMIPs) targeting 33 cancer genes to assay both point mutations and gene amplifications within spatially distinct regions of 14 glial tumors. We found evidence of regional mutational heterogeneity in multiple tumors, including mutations in *TP53* and *RB1* in an anaplastic oligodendroglioma and amplifications in *PDGFRA* and *KIT* in two glioblastomas (GBMs). IHC confirmed heterogeneity of *TP53* mutation and *PDGFRA* amplification. In all, 3 out of 14 of glial tumors surveyed had evidence for heterogeneity for clinically relevant mutations. Our results underscore the need to sample multiple regions in GBM and other glial tumors when devising personalized treatments based on genomic information, and furthermore demonstrate the importance of measuring both point mutation and copy number alteration while investigating genetic heterogeneity within cancer samples.

4.2- INTRODUCTION

Regional heterogeneity of mutations has been observed in a variety of tumor types[28,90]. This intratumoral heterogeneity has broad implications for the clinical management of cancer patients, especially in the current paradigm of personalized medicine based on genomic analysis of a single cancer biopsy. Within the context of primary brain tumors, several groups have previously identified heterogeneity of gene amplifications in genes *EGFR* and *PDGFRA* in GBM using FISH and Array-CGH on multiple regions within primary tumors[91,92]. However, despite the dropping cost of DNA sequencing, the extent of point mutational heterogeneity in brain tumors remains limited to a single case of GBM[93]. This is in part because the investigation of intratumoral heterogeneity requires both sampling and deep sequencing of multiple regions in a tumor.

We recently developed a method to identify low frequency mutations across known cancer genes[23] using the single molecule Molecular Inversion Probe (smMIP) assay, which combines multiplex target capture with single molecule tagging[18,23]. Here, we extend this technique to detect gene amplifications and examine intratumoral heterogeneity by targeting 33 cancer genes across 62 spatial sections of 14 glial tumors, including ten grade IV gliomas (all GBMs), three grade III gliomas (one each of ependymoma, astrocytoma, and anaplastic oligodendroglioma) and one grade II astrocytoma. We detected intratumoral heterogeneity in both point mutations and amplifications of genes implicated as glioma tumor drivers and therapeutic targets.

4.4- RESULTS

Study Design:

To assess heterogeneity within gliomas, we dissected each of 14 tumors into 3 to 5 regions per tumor (**Figure 4.1a, Supplementary Table C.1**). We used the single molecule Molecular Inversion Probe (smMIP) assay on gDNA isolated from each region to identify single nucleotide variants and high level copy amplifications (**Figure 4.1b, Supplementary Figure C.1**). smMIP probes capture target sequence into covalently linked circular molecules after polymerase extension and ligation. Following barcoding-PCR, sample pooling, sequencing, deduplication and alignment, we identified high level amplifications and point mutations (**Figure 4.1b,c, Figure C.1**).

Across the 14 tumors and 33 genes considered in this analysis, we identified a total of 33 putative protein-altering mutations (**Supplementary Table C.1**, **Supplementary Table C.2**). Tumors had between zero and 16 putative protein-altering mutations, with a median of two. *TP53* was the most commonly mutated gene, with mutations found in 8/14 tumors (**Supplementary Table C.3; Figure 4.2a**). One tumor, BI12, had many more candidate somatic mutations than other tumors (n=16 vs. median

n=2 in other tumors). Mutations in this GBM were predominantly G -> T (or C -> A) transversions (8 of 16 total), possibly representing mutation from unrepaired 8-oxo-guanine damage. Most mutations were observed across all tumor regions of BI12, consistent with a defect in DNA repair arising early in the development of the tumor.

To identify high level gene amplifications in tumors, we compared read depth of smMIP-targeted regions in each tumor against that of a control tissue. As smMIP sequencing suggested that a subset of control tissues were contaminated with tumor cells, we performed analyses using either patient-matched controls (**Supplementary Figure C.2**) or a "universal" control (**Figure 4.2b**). For the latter, we selected control tissue from tumor BI12, as it appeared to have the least contamination based on its frequency of known pathogenic point mutations, and restricted copy number analyses to targets with >30X coverage in control tissue from BI12 as well as targets whose GC percentage ranged from 30-60% (n=885 capture probes). A careful review of discrepant calls when using patient-matched vs. a universal control indicated that use of the universal control was more sensitive in identifying bona fide amplification events, secondary to the contamination of a subset of control tissues with tumor cells. After applying our filters (see **Methods**), a total of 21 genes could be assayed in a total of 62 regions across 14 tumors (**Figure 4.2b**).

The ratio of coverage of each probe was calculated relative to the control tissue (from BI12). We used DNACopy[94] to segment genes and obtain R, the mean ratio of coverage relative to control for each gene. We estimated the copy number for each gene by dividing R for each gene by the median value of R across all genes for each tissue. Genes with ratios above 3 were called as amplified. Genes with ratios above 6 were called as highly amplified. We did not measure deletion of genes using this method.

This process identified five tumors with gene amplifications with three having one or more regions with a highly amplified gene (**Figure 4.2b**). Three tumors had amplification of both *PDGFRA* and *KIT*, and three tumors had amplification of *EGFR*. We validated copy number estimates for a subset of calls using a variety of different methods including Taqman qPCR (across all tumors for *EGFR* and tumors BI05, BI06 and BI15 for *PDGFRA*), as well as whole genome sequencing (in tumor BI15 for *EGFR*). MIP copy number estimates of *EGFR* were highly correlated (R^2=0.90) with delta Ct obtained by Taqman qPCR when compared across all 62 regions sequenced (**Supplementary Figure C.3**). Additionally for five tumor regions of BI15 that were subjected to light-whole genome sequencing (WGS), *EGFR* copy number estimates were consistent between WGS and smMIP techniques (**Supplementary Methods, Supplementary Figure C.4**).

Tumors in which only a subset of regions possess an amplification or point mutation with no other mutation shared across regions can either be the result of mutational heterogeneity within a tumor, or the result of varying levels of tumor content between different tumor regions. As an example, tumor BI15 was called as amplified for *EGFR* in 2 out of 5 regions with no other somatic mutations/point mutations detected across the tumor (**Figures C.5 and C.6**). Upon close inspection of histologic slides prepared from adjacent tissue, the observed difference in amplification was most likely due to lower tumor cellularity within other regions of this tumor rather than intratumoral genetic heterogeneity. This was also seen in tumor BI04, where one region without detectable *PDGFRA* amplification also had lower frequencies of a *TP53* mutation seen across all regions. For this reason, we chose to restrict our interpretation of intratumoral heterogeneity to tumors in which all regions also shared a high frequency point mutation or gene amplification. Three tumors met these criteria and are described below.

Spatial heterogeneity of *TP53* and *RB1* point mutations:

One tumor exhibited clear spatial heterogeneity with respect to point mutations within the 33 genes investigated (**Figure 4.3**). BI09, an *IDH1*-mutant anaplastic oligodendroglioma, had a high frequency (>30% reads supporting mutation) inactivating mutation (R248H) in *TP53* in only two regions of the tumor (A and B). This tumor had high frequency mutations in *RB1* exclusively in two other regions (D and E) within the same tumor. Both *TP53* and *RB1* mutations were present at trace levels (<1%) within region C. As clinical workup indicated that BI09 had an *IDH1* mutation, we investigated all regions of this tumor by Sanger sequencing and found that regions A-E shared the *IDH1* R132H mutation. Sanger sequencing also validated the *TP53* mutation in regions A and B as well as the *RB1* mutation in regions D and E (**Supplementary Figure C.7**). Immunohistochemistry of p53 and IDH1-R132H expression on tissue adjacent to regions A-E provided additional confirmatory evidence (**Supplementary Figure C.8**). These findings are consistent with an *IDH1*-mutant tumor subsequently diverging to form subclones with mutations in *RB1* and *TP53*[95,96]. A pathologist scored these tissues blinded to the mutation type. Interestingly, the grade of each region correlated with the mutation type in each individual, with *TP53* mutation correlating with the histology of a grade III anaplastic oligodendroglioma. The clinical significance is unknown but this serves as a potential example of how genomic heterogeneity may affect histology of a tumor.

Spatial heterogeneity of *PDGFRA* and *KIT* amplifications

Our smMIP technique detected amplification of *PDGFRA*, *KIT* and *EGFR* within tumor BI05, an *IDH1*-wild type glioblastoma. In this tumor *EGFR* amplification was seen across all tumor regions, while amplification of both *PDGFRA* and *KIT* was detected in

two of five regions (**Figure 4.4a**). As *KIT* is located near *PDGFRA* on chromosome 4, shared amplification of these genes is expected[97]. Taqman real-time PCR assays performed in quadruplicate confirmed both the amplification in *EGFR* and the amplification in *PDGFRA* across all assayed regions (**Figure 4.4b**). Immunohistochemistry of *PDGFRA* and *EGFR* on tissue adjacent to regions A-E provided additional confirmatory evidence (**Supplementary Figure C.9**).

Similarly, we detected heterogeneity of *PDGFRA* amplification within BI06, an *IDH1*-mutant glioblastoma. This tumor had amplification of *PDGFRA* and *KIT* in region A not detected within other regions (**Figure 4.5a**). Taqman qPCR confirmed amplification of region A, mild amplification in region B and no amplification in region C, D and E (**Figure 4.5b**). All other regions of this tumor had high frequency somatic mutations in *PTEN*, such that reduced tumor cellularity is an unlikely explanation for our observations.

Additional cases of heterogeneity are potential passenger mutations

A missense mutation in *KRAS* was observed at low read depth (10% of reads) in region D of the glioblastoma BI12 and was not detected in at least one other region (**Supplementary Table C.3**). As this mutation does not occur within known mutation hotspots and is in a tumor with signs of hypermutation (BI12), the clinical significance of this heterogeneity remains unclear. Other somatic point mutations are heterogeneous across an individual tumor but occur within genes that have another, ubiquitously distributed mutation. BI12 has missense mutations in *PTEN* that are observed in regions A, B and C and not in region D (**Supplementary Table C.3**). This tumor also has another high frequency mutation in this gene that is present across all regions of this tumor. A similar scenario is seen in the astrocytoma BI08. Regions D and E of this tumor have low frequency mutations in *TP53*, but all regions share another high frequency mutation in *TP53*.

4.4- DISCUSSION

These results demonstrate that intratumoral spatial heterogeneity with respect to clinically relevant genes occurs among multiple types of brain tumors, and spans the mutational spectrum from copy number to point mutations. Across a set of recurrently mutated cancer genes (33 genes examined for point mutations, 21 genes for amplifications), we observed heterogeneity for clinically relevant mutations in 3 of 14 (21%) glial tumors. These include point mutations in *TP53* and *RB1* as well as amplifications in *PDGFRA*/*KIT*. All cases of mutational heterogeneity that we detected in a tumor occur in adjacent regions, consistent with the hypothesis that spatially distinct regions represent divergent subclones of a single tumor.

Historically, in anaplastic oligodendroglioma with intact 1p, mutations in *TP53* were found to stratify outcomes, with median survival of 71 versus 16 months in patients with mutant versus wild-type *TP53*, respectively[98]. While not specifically applying to our patient in whom 1p/19del is deleted, our data demonstrating discrete differences in *TP53* status from different regions within an individual tumor nevertheless shows the potential of genetic heterogeneity to confound the assignment of prognostication based on the detection of specific molecular markers.

In addition, decision-making regarding the use of receptor tyrosine kinase inhibitors could be influenced by the status of amplifications/mutations in *PDGFRA*. Our finding of regional heterogeneity of *PDGFRA* within tumors BI05 and BI06, confirms recent work by Sottoriva et al. and others and suggests that a single biopsy may not be sufficient to allow for informed application of targeted therapies against these presumed oncogenic drivers [91,92,99]. Clinical decision-making at recurrence will also likely be impacted by regional heterogeneity. Nickel et al. compared mutations within a group of 10 genes from 2 regionally distinct samples of a single GBM at initial resection and 2 subsequent recurrences. No heterogeneity was detected at initial resection but

heterogeneity of *PIK3CA* and *PTEN* mutation was detected at the first recurrence, and heterogeneity of *PIK3CA*, *TP53* and *EGFR* mutation was detected at the second recurrence [93].

In this study we were able to identify amplification of only a subset of genes of interest, as some genes had too few probes to accurately determine copy number. Our study also did not detect genomic rearrangements and deletions such as the *EGFR* VIII deletion commonly found in GBM. However, one can imagine expanding this assay to consider amplifications and deletions with smMIPs by tiling probes at higher density and incorporating known SNP positions to aid in identifying cases of loss of heterozygosity (LOH). One could also capture additional glioma-relevant genes like *IDH1* and *IDH2* by adding probes targeting these genomic regions.

Our investigation focused on regional heterogeneity within a tumor, instead of the microscopic heterogeneity that is likely present within a given tumor biopsy. As we performed the smMIP assay on DNA extracted from tissue pieces that likely contained millions of cells, we would likely miss cases of heterogeneity where only a small population of cells within a biopsy contained a mutation (such as an amplification). Use of techniques such as IHC, FISH and more recently, single cell sequencing, remain necessary to characterize the extent of microscopic heterogeneity in tumors.

These results validate the single-molecule molecular inversion probe (smMIP) approach as a scalable and cost-effective platform for deep sequencing of cancer genomes to examine subclonal variation. Despite deeply sequencing multiple sections of 14 tumors, our survey required only one lane of sequencing on the Illumina HiSeq because we focused on well-known gene targets of mutation in cancer. In contrast to the technique used by a similar investigation[93], our method is also easily scaled and amenable automation with samples processed in 96-well formats. This advantage in scalability enables one to easily assay many more regions (10s to 100s) per tumor to

obtain much finer scale picture of intratumoral heterogeneity, as we are likely underestimating its extent even here by sampling of only a few regions. While our study represents an improvement over previous studies, analysis of greater number of genes in a greater number of tumors will be necessary to determine rates of regional heterogeneity in different driver mutations across GBMs.

4.5- CONCLUSIONS

We find multiple instances of regional heterogeneity in clinically relevant cancer genes within malignant gliomas at the time of diagnosis. We also demonstrate a scalable technique that can be used to efficiently characterize regional genetic heterogeneity for both point mutations and copy number alterations in tumors. Future challenges will include how best to interpret cases of intratumoral heterogeneity and test its impact in the context of clinical trials using targeted therapy approaches.

4.6- METHODS

Samples

Freshly resected brain tumor specimens from adult patients were obtained with informed consent as part of the Genomics Big Idea pilot program (UW/FHCRC). Tissue, patient demographics and final diagnosis were obtained in accordance with protocols approved by the IRB at the University of Washington. Tumors were divided into three to five regions, depending on size. Tissue from each region was then sub-divided into four pieces for use in next generation sequencing (NGS), histology, cell culture and xenotransplantation (**Figure 5.1**). In ten cases, brain grossly uninvolved by tumor was resected to provide adequate surgical access and was utilized as a source of germline or "control" DNA to identify somatic mutations (**Supplementary Table C.1**). For all samples, DNA was isolated from snap-frozen tissue pieces using the QIAGEN DNEasy Blood and Tissue kit.

Targeted capture and sequencing

The single molecule Molecular Inversion Probe (smMIP) assay was used to genotype candidate genes. Probes were previously designed by Hiatt et al.[23] against 33 genes that are commonly mutated in cancer (**Supplementary Table C.2**). Targeted capture and PCR amplification was performed as previously described, except that 200 ng of genomic DNA was used for each sample instead of 500 ng[23]. After smMIP capture, amplified products were pooled and sequenced on a single lane of the Illumina HiSeq 2000 platform with paired 100-nt reads and an 8-nt index read.

Primary analysis and variant calling

Initial analysis steps through to read mapping were performed as previously described [23], except that instead of constructing a consensus read from tagged smMIP molecules, we chose one read per unique molecular tag event at random for subsequent analysis.

Variants were called using SAMtools, and were filtered for positions with *phred* base quality ≥30, ≥30X coverage and the absence of a neighboring homopolymer run of four bases or more (**Supplementary Table C.2**). To remove common polymorphisms and enrich for likely somatic mutations, we imposed a number of additional requirements, including requiring variants to be observed with an allele balance of at least 5% within a sample, removing variants present within a modified database of the Exome Sequencing Project (ESP)[100] and 1000 Genomes[101] pilot project that had first been stripped of all COSMIC variants, removing variants that were present at an allele balance of at least 5% in two or more control samples.

Copy number analysis

We compared read depth of smMIP-targeted regions in each tumor against that of the control tissue BI12 to identify high level gene amplifications in tumors. We restricted the copy number analysis to targets with greater than 30X coverage in control

tissue and a GC content ranging from 30-60%. To reduce the number of potential artifacts remaining, we removed from consideration (for the purposes of copy number analysis only) twelve genes (*AKT1, AKT2, CDK4, CDKN2A, FGFR3, HRAS, KRAS, MYC, NRAS, SRC, STK11,* and *VHL)* that had fewer than 15 probes with sufficient coverage in the control tissue (BI12).

After calculating the ratio of coverage for each probe relative to control tissue from BI12, we used DNACopy[94] to segment genes into discrete levels of coverage and obtain R, the mean ratio of coverage relative to control for each gene. We estimated the copy number for each gene by dividing R for each gene by the median value of R across all genes for each tissue. Genes with ratios above 3 were called as amplified and genes with ratios above 6 were called as highly amplified.

Sanger validation

DNA from five regions of tumor BI09 were subjected to Sanger sequencing (Genewiz) against positions within *IDH1*, *TP53* and *RB1*.

Copy number validation

Tumors with regional heterogeneity in *EGFR* and *PDGFRA* detected using smMIP sequencing were confirmed using Taqman qPCR analysis. DNA from each region were analyzed in quadruplicate using commercially available probes against *PDGFRA* (assay ID: Hs02749151_cn; Life Technologies Biosystems) and *EGFR* (assay ID: Hs07526740_cn; Life Technologies). Reference primers amplified a fragment from *TERT* (no. 4403316; Life Technologies). Finally, to compare sensitivity of the smMIP approach, all regions from all tumors were assayed in duplicate for *EGFR* copy number.

Immunohistochemistry and FISH

Immunohistochemistry for IDH1 and p53 was performed on 4 micron paraffin sections using mouse anti-human p53 clone (1:2000 dilution, DAKO) and mouse anti-human IDH1 R132H (1:200 dilution, Dianova). All tumors were investigated for IDH1

mutation by neuropathology, while only a subset of tumors was investigated for p53 expression by IHC (Supplementary Table C.1). Immunohistochemistry for EGFR and PDGFRA was performed on 5-6 micron paraffin sections using mouse anti-human EGFR, clone 2-18C9 (pharmDx kit, DAKO) and rabbit anti-human PDGFRα, clone D1E1E (1:500 dilution, Cell Signaling). Dual-color *EGFR* FISH was performed using commercially available probes (LSI *EGFR* SpectrumOrange/CEP 7 SpectrumGreen, no 32-191053; Abbott Molecular) with DAPI counterstain using standard methods. Slides were imaged using an Olympus DP72 digital camera mounted on a Nikon E400 microscope. 50 nuclei were scored for each region. *EGFR* amplification was called if more than 10% of nuclei either contained many *EGFR* signals or exhibited a *EGFR:CEP7* ratio greater than 2. 1p19q deletion FISH was performed using commercially available probes (no. 04N60-020 Abbot Molecular) using standard methods.

4.7- FIGURES

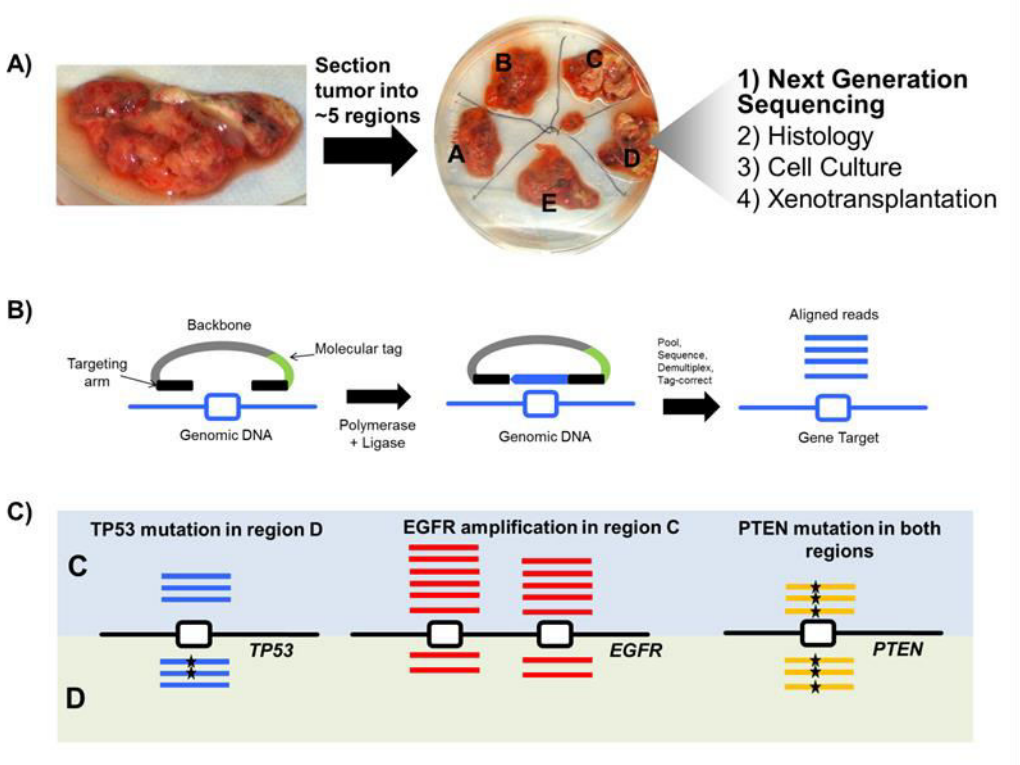


Figure 4.1: Experimental approach. a) Each tumor was divided into 3-5 regions to assay intratumoral heterogeneity. Each individual region was subdivided into 4 pieces for use in Next Generation Sequencing (NGS), histology, cell culture and xenotransplantation. Results from NGS and histology are described here **b)** Molecular inversion probe method. Oligonucleotide probes were previously designed against 33 cancer genes [23]. MIPs have a common backbone sequence, molecular tag sequence as well as targeting arms homologous to regions flanking targets of interest. After polymerase extension and ligation, targeted sequence is captured within a circular molecule. Captured sequences are amplified in a barcoding-PCR reaction and multiple samples are pooled and sequenced on the same lane. After tag-correction (not shown), reads corresponding to each tumor region are mapped to the human reference sequence to be used to identify copy number amplifications and point mutations specific to one region or another. Additional details are provided in **Supplementary Figure C.1**. **c)** Example of comparisons: MIP captures of regions C and D can detect both *TP53* point mutation heterogeneity and *EGFR* amplification heterogeneity within a tumor. Tumors with heterogeneity were required to share either a point mutation or copy number alteration (in this case mutation of *PTEN*) across all regions to ensure that differences in observed mutation was not due to varying levels of tumor cellularity.

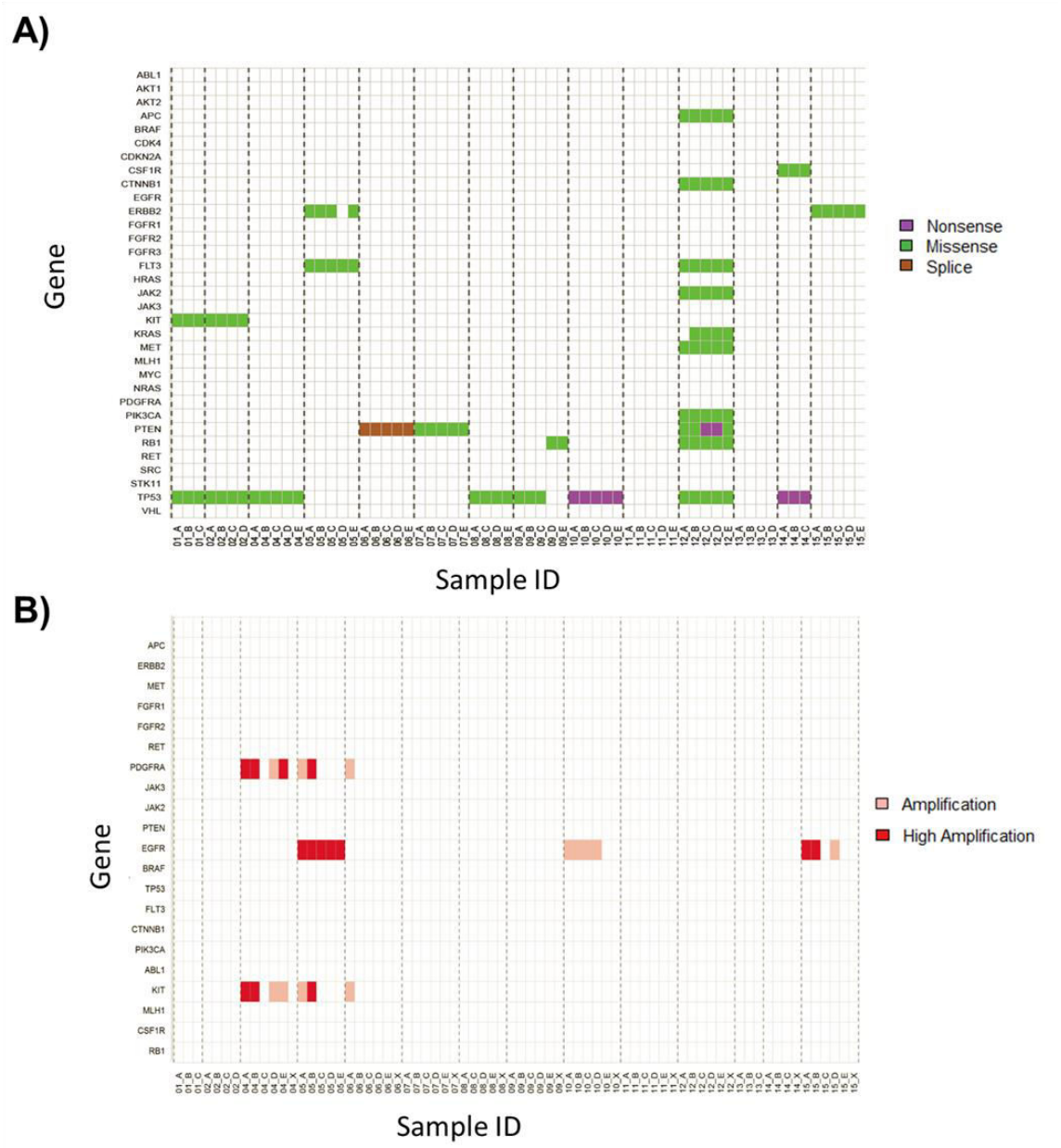


Figure 4.2: Summary of heterogeneity observed across all samples. a) Protein-altering mutations detected across all tumor regions. Genes mutated twice in the same tumor region are not identified here. A table of all mutations can be found in **Supplementary Table C.3**. **b)** High level gene amplifications detected by smMIP assay. Copy number was estimated by comparing all tumor samples against a universal control from BI12 (See **Supplementary Figure C.2** for analysis using patient matched controls). Amplification indicates genes with coverage three fold higher than median coverage across a sample. High Amplification indicates genes with coverage six-fold higher than median coverage across a sample. Our approach would miss any low-level gene amplifications that within these tumors.

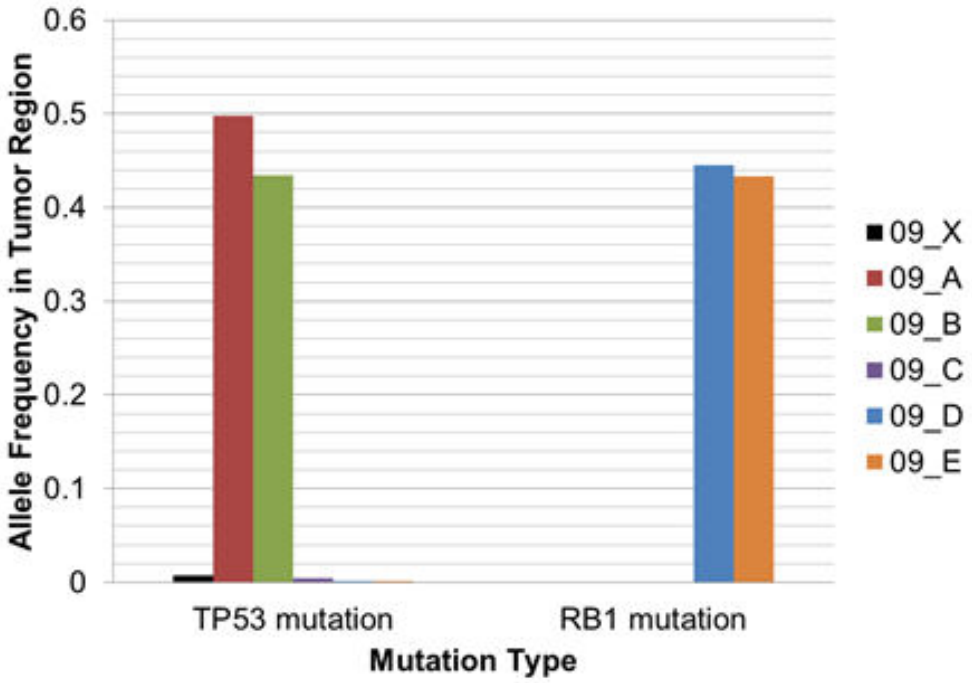


Figure 4.3: Intratumoral heterogeneity of *TP53* and *RB1* determined from smMIP sequencing. Tumor BI09 was sectioned into five regions A-E. Each region was assayed for mutations in 33 genes, including *TP53* and *RB1*. This plot shows the allele balance of *TP53* and *RB1* within each individual tumor region. Tumor regions A and B have a high frequency mutation in *TP53*, while regions D and E have a high frequency mutation in *RB1*. Sanger results validated *TP53* and *RB1* mutations in each region and also revealed that all regions shared a R132H mutation in *IDH1* (**Supplementary Figure C.7**).

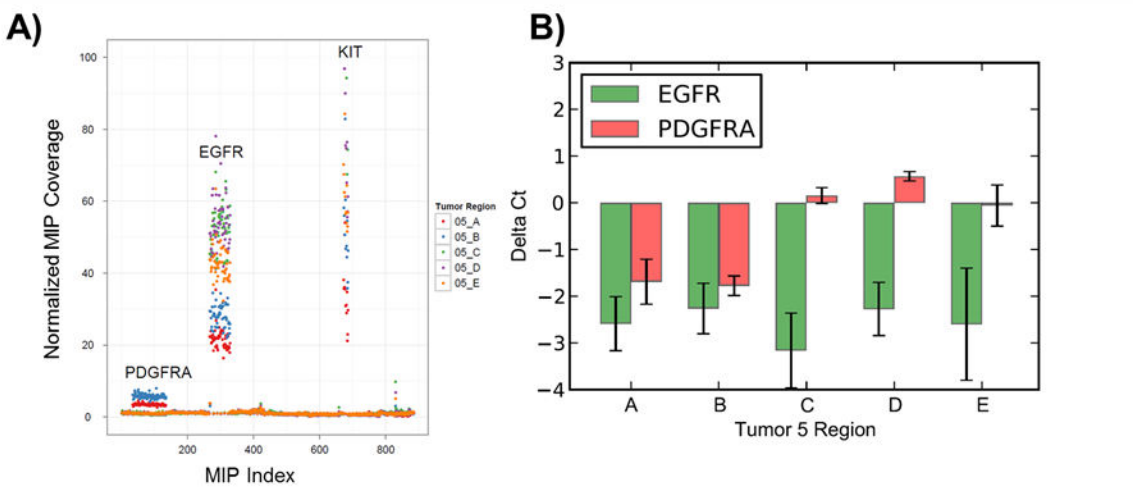


Figure 4.4: Heterogeneity of *PDGFRA* amplification in BI05. a) Copy Number estimates based on smMIP probe data. *PDGFRA* amplification (labeled) occurs in region A and B with no amplification in C, D or E. **b)** Results from Taqman qPCR performed against fragments of *PDGFRA* and *EGFR* in quadruplicate. *PDGFRA* amplification occurs in region A and B (between 4-8 fold amplification) with no significant amplification in regions C, D and E. *EGFR* amplification occurs in all regions of BI05, consistent with MIP sequencing results. Heterogeneity of *PDGFRA* amplification was also confirmed by IHC in region A and E (**Supplementary Figure C.9**).

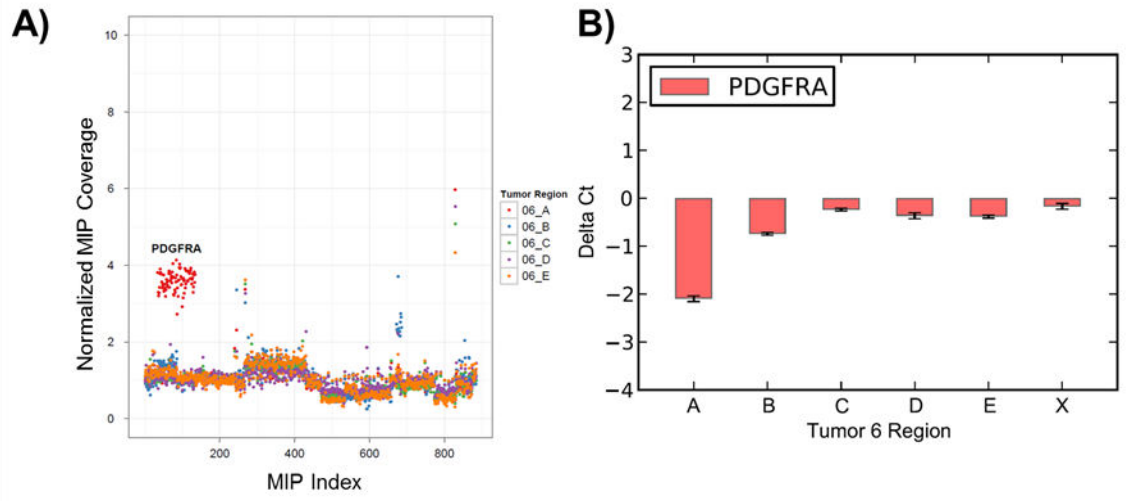


Figure 4.5: Heterogeneity of *PDGFRA* amplification in BI06. a) Copy Number estimates based on smMIP probe data. *PDGFRA* amplification (labeled) occurs in region A with only mild amplification in region B and no clear detectable amplification in regions C, D or E. **b)** Results from Taqman qPCR performed against fragments of *PDGFRA* performed in quadruplicate. *PDGFRA* amplification occurs in region A (~4 fold amplification) with only mild amplification in regions B, C, D and E.

Chapter 5- Conclusions and Future Directions

Research Summary:

My dissertation has centered on investigating genomic heterogeneity in cancer with the goals of identifying distinct subtypes and characterizing intrapatient and intratumoral heterogeneity. Chapters 2 and 3 summarized the results from a genomic survey of aggressive prostate cancer with the finding of a hypermutated subset of prostate cancer and a set of genes and mutations associated with aggressive disease. Chapter 3 also characterized intrapatient heterogeneity in men with disseminated disease. Chapter 4 discussed the application of a targeted sequencing technique to address intratumoral heterogeneity in gliomas.

Interpatient Heterogeneity: Expanding molecular subtypes

The work described in Chapters 2 and 3 is part of a growing body of work cataloging the genes and mutations that commonly occur in cancer. These studies have provided a detailed landscape of somatic mutations for most common cancers[3,37,102]. In most cancer types a few genes are mutated across a majority of tumors; these are likely driver genes and these have been much of the focus of previous investigations. However, each study (including ours) also identified a much larger set of genes that are mutated in only a few tumors. While many of these less frequently mutated genes are likely not essential to tumor growth and progression, some are clinically relevant. Interpreting and testing these mutations on the basis of sequence information alone is challenging, especially for missense mutations where the effect on protein function is often unclear. In these cases, experimental validation of these less frequently mutated genes will play an increasingly important role in rounding out molecular subtypes. Currently, mutations are introduced and tested one at a time. The current practice of testing the functional impact of mutations is poorly suited to the job of investigating the thousands of mutations identified for each cancer type. Recent improvements in genome engineering have been used to generate mutants directly into the genomes of cancer cell lines. A recent study by Choi *et al.* used the CRISPR-Cas system to generate and validate the functional effects of

specific rearrangements within tumors[103]. I imagine that this process can be scaled using principles from massively parallel reporter assays[104-107] to develop high-throughput methods of introducing and testing somatic mutations in cancer cells genome-wide.

Much of the work described in Chapters 2 and 3 involved cohorts of patients that had died from castration resistant prostate cancer. Mutation frequency in resistant tumors was sufficient to identify several known drivers of resistance (e.g. *AR*), but this approach was limited by passenger mutations in tumors and also by the types of treatments the used on the patients. An alternative approach would be to perform experimental evolution of resistance followed by sequencing of resulting tumors[108,109]. This approach would reduce the impact of passenger mutations and could be performed on drugs that are experimental. I described some preliminary findings of the approach (exome sequencing of 3 pairs of tumors), but I think repeating this analysis with additional tumors and experimental replicates could identify additional mechanisms of resistance.

Noninvasive measurement of tumor genomes

Recent discovery of tumor products in the bloodstream of cancer patients offers an attractive route to investigating heterogeneity. Cell-free circulating tumor DNA is free-floating DNA in plasma that is derived from necrotic tumor material shed into a patient's bloodstream[110,111]. In addition to providing another early biomarker for disease, this material can also be used to sample intrapatient heterogeneity as circulating DNA may be composed of material derived from multiple tumors. Circulating tumor cells (CTC's) are found in the blood of patients with multiple types of cancer and several groups have recently investigated genetic alterations in these cells[112,113]. Most recently, one group used exome sequencing to find that CTC's captured some of the mutational differences among metastases in a patient with prostate cancer[89]. Both circulating tumor DNA and CTC's can be obtained from peripheral blood facilitating sampling of the tumor population across many timepoints[89]. In addition to being used

to better track a patient's response to a series of therapies, these methods can provide a better understanding of the role of intrapatient heterogeneity in the development of resistance.

Dissecting intratumoral heterogeneity at the cellular level

In this thesis, I applied targeted sequencing to investigate regional heterogeneity within tumors. Similar approaches have now been applied to a variety of tumor types including pancreatic, breast and kidney cancers using methods that vary from low-resolution genome profiling of single cells to high resolution genome sequencing of bulk tissue samples[28,90,114-116]. As sequencing technologies improve both in terms of cost and requirements for input material, it will be possible to profile tumors at higher spatial and genomic resolution possibly leading to whole genome sequencing of all cells in a tumor[15,21-24]. These investigations will not only guide practical decisions on the number/scale of biopsies needed for precision medicine, but also will address basic questions of tumor biology and evolution by enabling the reconstruction of the full timeline and sequence of events leading to a clinically detected mass.

Final thoughts

This thesis discussed heterogeneity of somatic mutation as it manifests at many scales in cancer: across tumor types, individual patients and individual tumors. This heterogeneity simultaneously presents challenges and opportunities to develop new diagnostics and treatments for cancer patients. As technologies to both detect mutations and predict their functional impact continue to improve, I am hopeful that we will be able to address some of these challenges and provide future cancer patients with better care.

Appendix A- Supplementary Material for Chapter 2

Sample ID	Capture Method	Indexing	Sequencer	Run-type
LuCaP 23.1	V2	no	HiSeq	PE-100
LuCaP 23.12	V1	no	Illumina GAIIx	PE-76
LuCaP 23.1AI	V1	no	Illumina GAIIx	PE-76
LuCaP 35	V1	no	Illumina GAIIx	PE-76
LuCaP 35V	V1	no	Illumina GAIIx	PE-76
LuCaP 49	V1	no	HiSeq	PE-100
LuCaP 58	V2	no	HiSeq	PE-100
LuCaP 70	V2	no	HiSeq	PE-100
LuCaP 73	V2	yes	HiSeq	PE-100
LuCaP 77	V2	yes	HiSeq	PE-100
LuCaP 78	V2	no	HiSeq	PE-100
LuCaP 81	V2	no	HiSeq	PE-100
LuCaP 86.2	V1	no	HiSeq	PE-100
LuCaP 92	V2	no	HiSeq	PE-100
LuCaP 93	V2	no	HiSeq	PE-100
LuCaP 96	V1	no	Illumina GAIIx	PE-76
LuCaP 96AI	V1	no	Illumina GAIIx	PE-76
LuCaP 105	V2	no	HiSeq	PE-100
LuCaP 115	V2	no	HiSeq	PE-100
LuCaP 136	V2	no	HiSeq	PE-100
LuCaP 141	V2	no	HiSeq	PE-100
LuCaP 145.2	V2	yes	HiSeq	PE-100
LuCaP 147	V2	no	HiSeq	PE-100

Table A.1: Methods used to capture and sequence xenograft exomes. We used two versions of Nimblegen EZ SeqCap capture probes in this study. Eight samples were captured using V1 probes (targeting the 26.6 Mb Consensus Coding Sequence Database (CCDS), while the remainder of samples were captured using V2 probes (targeting the 36.6 Mb RefSeq database). Four samples were indexed with barcodes prior to capture and sequencing. V1, Nimblegen V1 solution capture probes targeting CCDS coordinates; V2, Nimblegen V2 solution capture probes targeting RefSeq coordinates ; PE-76, paired-end sequencing using 76 bp reads; PE-100 paired-end sequencing using 100 bp reads.

Sample ID	Capture Method	Positions called on CCDS target at threshold quality (8x coverage q30)	% of CCDS target bases (26.6 Mb) covered at threshold quality	Positions called on Refseq target at threshold quality	% of Refseq target bases (36.6 Mb) covered at threshold quality
LuCaP 23.1	V2	25266029	90.8%	34300618	94.8%
LuCaP 23.12	**V1**	**23964797**	**90.8%**	**25031070**	**69.2%**
LuCaP 23.1AI	**V1**	**23863880**	**90.4%**	**24930663**	**68.9%**
LuCaP 35	**V1**	**23361988**	**88.5%**	**24328027**	**67.3%**
LuCaP 35V	**V1**	**23278126**	**88.2%**	**24248217**	**67.1%**
LuCaP 49	**V1**	**25220912**	**95.5%**	**26517043**	**73.3%**
LuCaP 58	V2	25441803	96.4%	34690100	95.9%
LuCaP 70	V2	24851268	94.1%	33696934	93.2%
LuCaP 73	V2	24958389	94.5%	33822311	93.5%
LuCaP 77	V2	25395101	96.2%	34608983	95.7%
LuCaP 78	V2	25310106	95.9%	34511587	95.4%
LuCaP 81	V2	24991004	94.7%	34081896	94.3%
LuCaP 86.2	**V1**	**25382463**	**96.1%**	**26759534**	**74.0%**
LuCaP 92	V2	24487956	92.7%	33372489	92.3%
LuCaP 93	V2	25341234	96.0%	34543362	95.5%
LuCaP 96	**V1**	**24092606**	**91.3%**	**25169928**	**69.6%**
LuCaP 96AI	**V1**	**23983100**	**90.8%**	**25045910**	**69.3%**
LuCaP 105	V2	25176799	95.4%	34299929	94.9%
LuCaP 115	V2	25470993	96.5%	34719468	96.0%
LuCaP 136	V2	25167391	95.3%	34292172	94.8%
LuCaP 141	V2	25345809	96.0%	34580774	95.6%
LuCaP 145.2	V2	24958089	94.5%	33921712	93.8%
LuCaP 147	V2	25510370	96.6%	34777527	96.2%

Table A.2: Coverage across CCDS coordinates and Refseq coordinates of the 23 LuCaP xenograft samples sequenced. Bolded entries indicate samples that had been captured using Nimblegen V1 probes. More than 90% of the CCDS coding regions were able to be called in most samples (8x coverage and *Samtools*-derived *phred* quality score > 30). Within samples captured using V2 probes, more than 92% of the RefSeq target bases were able to be called in all samples using the same criteria.

Sample ID	On-target (36.6 Mb RefSeq) variants in dbSNP:	On-target variants not in dbSNP:	Number of variants post filtering w/ ~2000 Exomes and 1000 Genomes Data
LuCaP 23.1	19647	2518	643
LuCaP 23.12	13473	1031	153
LuCaP 23.1AI	13169	1061	186
LuCaP 35	12909	938	179
LuCaP 35V	12892	966	203
LuCaP 49	15229	1680	328
LuCaP 58	21013	6326	**4067**
LuCaP 70	19608	2302	389
LuCaP 73	21414	5273	**2972**
LuCaP 77	20900	2388	451
LuCaP 78	21373	2671	372
LuCaP 81	21122	1870	304
LuCaP 86.2	15187	1704	326
LuCaP 92	19333	1732	313
LuCaP 93	20526	2485	487
LuCaP 96	13614	1166	282
LuCaP 96AI	13441	1077	240
LuCaP 105	19942	1853	374
LuCaP 115	20685	2213	380
LuCaP 136	20115	1941	382
LuCaP 141	20482	2651	701
LuCaP 145.2	20522	2515	538
LuCaP 147	21784	4957	**2714**

Table A.3: Frequency of somatic coding variants across prostate cancers sequenced. A majority of the variants identified by exome sequencing were present within dbSNP. After removing from consideration all variants that were observed in the pilot dataset of the 1000 Genomes Project [50,51] as well as any variants present in any of ~2,000 additional exomes sequenced at the University of Washington, the number of variants remaining in 20/23 samples was reduced to ~350. Three xenografts, LuCaP 58, LuCaP 73 and LuCaP 147, (highlighted in gray) have a large number of variants compared with other xenografts.

Sample ID	Number of Filtered Variants (nov-SNVs)	Number of Genes	Nonsynonymous variants (nov-nsSNVs)	Missense variants	Nonsense variants	Splice variants	Ti:Tv ratio of filtered variants	Overall Ti:Tv ratio of all variants
LuCaP 23.1	643	407	285	274	8	3	1.19	2.70
LuCaP 23.12	153	145	89	84	5	0	1.64	3.15
LuCaP 23.1AI	186	178	112	104	8	0	1.45	3.12
LuCaP 35	179	174	110	101	8	1	2.09	3.16
LuCaP 35V	203	183	116	106	9	1	1.9	3.15
LuCaP 49	328	313	197	183	11	3	0.81	2.8
LuCaP 58	**4067**	**3232**	**2393**	**2283**	**75**	**35**	**4.62**	**3.03**
LuCaP 70	389	299	185	178	5	2	1.7	2.81
LuCaP 73	**2972**	**2449**	**1777**	**1681**	**78**	**18**	**4.28**	**2.99**
LuCaP 77	451	345	210	201	5	4	1.62	2.76
LuCaP 78	372	289	176	169	5	2	1.27	2.73
LuCaP 81	304	266	173	166	4	3	1.71	2.96
LuCaP 86.2	326	311	221	214	7	0	1	2.77
LuCaP 92	313	263	158	150	6	2	1.68	3
LuCaP 93	487	383	245	237	5	3	1.31	2.74
LuCaP 96	282	270	178	169	8	1	0.97	3.04
LuCaP 96AI	240	235	159	152	6	1	1.11	3.06
LuCaP 105	374	308	193	187	3	3	1.48	2.89
LuCaP 115	380	313	178	170	6	2	1.47	2.8
LuCaP 136	382	335	189	181	7	1	1.32	2.9
LuCaP 141	701	591	326	312	11	3	2.23	2.92
LuCaP 145.2	538	413	269	254	12	3	1.24	2.74
LuCaP 147	**2714**	**2236**	**1549**	**1452**	**76**	**21**	**4.18**	**2.97**

Table A.4: Characteristics of novel Single Nucleotide Variants (nov-SNVs) across prostate cancers sequenced. While Transition to Transversion (Ti:Tv) ratios of nov-SNVs were less than two for most samples, the hypermutated samples (LuCaP 58, LuCaP 73 and LuCaP 147) exhibit a much larger fraction of transition events (Ti:Tv of >4). The overall Ti:Tv ratio of all variants (unfiltered) within coding regions approximated 3 in most samples.

# of samples seen out of 16	Gene ID	Gene Name	Estimated P-value of being germline	Estimated Singleton P- value	Individual mutations seen
6	MUC16	mucin 16, cell surface associated	0.00255	0.10777	LuCaP105(HIS8238GLN), LuCaP81(ALA7199THR), LuCaP105(VAL6677PHE), LuCaP93(SER6247PHE), LuCaP115(SER1250ILE), LuCaP78(THR35LYS), LuCaP70(SER2402ILE)
5	TP53	tumor protein p53 (Li-Fraumeni syndrome)	< 0.00005	0.00375	LuCaP73(ARG306GLN), LuCaP136(ARG280stop), LuCaP23.1AI(CYS238TYR), LuCaP92(GLU198stop), LuCaP73(ARG175CYS), LuCaP70(TYR163HIS), LuCaP77(PRO278SER)
4	UBR4	E3 ubiquitin-protein ligase	0.0027	0.04129	LuCaP147(MET4596ILE), LuCaP58(TYR4372CYS), LuCaP58(MET4015VAL), LuCaP73(ARG2272HIS), LuCaP136(SER2168TYR), LuCaP23.1AI(GLU2014LYS), LuCaP35V(LEU1894ILE), LuCaP73(ALA1831VAL), LuCaP58(ARG901CYS), LuCaP93(LEU865VAL), LuCaP58(LEU811PRO)
4	SYNE1	spectrin repeat containing, nuclear envelope 1	0.03025	0.08365	LuCaP86.2(ILE6941VAL), LuCaP136(ARG6817GLN), LuCaP86.2(SER5858THR), LuCaP145.2(ARG5432TRP), LuCaP141(LEU5357PRO), LuCaP73(SER1212ARG), LuCaP73(ARG444GLN), LuCaP136(VAL431ILE)
3	SDF4	stromal cell derived factor 4	< 0.00005	0.00214	LuCaP105(ASP276ASN), LuCaP78(GLY76SER), LuCaP115(ALA9SER)
3	PDZRN3	PDZ domain containing RING finger 3	< 0.00005	0.00858	LuCaP23.1AI(ARG727CYS), LuCaP105(GLY570SER), LuCaP73(ARG463CYS), *LuCaP92(ILE331LEU)
3	DLK2	delta-like 2 homolog	0.00005	0.00483	LuCaP70(ARG371HIS), LuCaP145.2(SER361ARG), LuCaP96AI(HIS280GLN)
3	FSIP2	fibrous sheath interacting protein 2	0.00005	0.0059	LuCaP81(LYS22ASN), †LuCaP92(THR698ILE), LuCaP136(GLN1526HIS)
3	NRCAM	neuronal cell adhesion molecule	0.00015	0.00736	LuCaP115(MET1094ILE), LuCaP86.2(LYS645GLU),

					LuCaP145.2(SER329CYS)
3	MGAT4B	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme B	0.0002	0.00697	LuCaP105(ALA504THR), LuCaP23.1AI(ARG168CYS), LuCaP136(VAL150MET)
3	GLI1	glioma-associated oncogene homolog 1 (zinc finger protein)	0.0003	0.00858	LuCaP86.2(ARG20TRP), LuCaP78(ARG81GLN), LuCaP96AI(PRO210THR)
3	PCDH11X	protocadherin 11 X-linked	0.0003	0.00736	*LuCaP145.2(VAL38PHE), LuCaP58(MET867VAL), LuCaP105(VAL1007ILE), LuCaP49(THR1296ASN)
3	KDM4B	Lysine-specific demethylase 4B	0.00035	0.01072	LuCaP73(ALA265VAL), LuCaP105(ARG534TRP), LuCaP35V(ALA555VAL), LuCaP73(ALA827VAL), LuCaP86.2(SER1036CYS)
3	SCRIB	scribbled homolog (Drosophila)	0.00295	0.02145	LuCaP141(ASP1430ASN), *LuCaP92(VAL797LEU), LuCaP35V(ARG513TRP)
3	NPHS1	nephrosis 1, congenital, Finnish type (nephrin)	0.0033	0.01864	LuCaP141(THR1182MET), LuCaP49(ASP713GLU), LuCaP92(ALA658THR)
3	MYO7A	myosin VIIA	0.0069	0.02788	LuCaP77(GLU680GLY), *LuCaP145.2(GLY1222SER), LuCaP136(ASN1411LYS), LuCaP73(ARG1621HIS)
3	FBN2	fibrillin 2 (congenital contractural arachnodactyly)	0.00975	0.03056	*LuCaP92(ASN2110SER), LuCaP49(ARG1832CYS), LuCaP70(GLY364ASP)
3	APOB	apolipoprotein B (including Ag(x) antigen)	0.01425	0.03592	LuCaP136(ASP4457ASN), LuCaP58(THR4037ALA), LuCaP93(ARG3059HIS), LuCaP77(VAL2446MET)
3	LRP1B	low density lipoprotein-related protein 1B (deleted in tumors)	0.0149	0.03592	LuCaP49(GLY4341ARG), LuCaP136(CYS4311TYR), LuCaP73(ASP3341ASN), LuCaP115(PRO3301THR), LuCaP58(ARG2430TRP), LuCaP58(TYR1593stop), LuCaP49(ARG1550SER), †LuCaP147(THR1442ILE), LuCaP73(GLY1202GLU), †LuCaP147(GLN371stop)
3	COL7A1	collagen, type VII, alpha 1 (epidermolysis bullosa, dystrophic, dominant and recessive)	0.01805	0.04021	LuCaP73(ARG2927HIS), LuCaP73(ILE2708THR), LuCaP141(GLU2699LYS), LuCaP96AI(ARG1751GLN), LuCaP73(GLY1329TRP), LuCaP73(ARG793TRP), LuCaP93(ASP149GLU)
3	RYR1	ryanodine receptor 1 (skeletal)	0.03645	0.05094	LuCaP49(ARG1016stop), LuCaP58(MET1285VAL), LuCaP86.2(VAL1551ILE), LuCaP35V(MET3081VAL),

					LuCaP147(splice)
3	MACF1	microtubule-actin crosslinking factor 1	0.0539	0.06005	LuCaP92(VAL935ILE), LuCaP73(GLU2802LYS), LuCaP147(GLN3505ARG), LuCaP81(ASP3830HIS), LuCaP141(GLU4128LYS)
2	DKK1	dickkopf homolog 1 (Xenopus laevis)	< 0.00005	0.00107	LuCaP92(GLU151GLN), LuCaP93(SER244TYR)
2	RAB32	RAB32, member RAS oncogene family	0.00005	0.00161	LuCaP93(VAL66ILE), LuCaP141(SER109stop)
2	PLA2G16	phospholipase A2, group XVI	0.00015	0.00161	LuCaP115(SER85LEU), LuCaP35V(PRO19HIS)
2	TFG	TRK-fused gene	0.00015	0.00161	LuCaP23.1AI(ASN134HIS), LuCaP141(GLN318stop), LuCaP147(TYR319stop)
2	TBX20	T-box 20	0.0002	0.00161	LuCaP77(ARG437HIS), LuCaP96AI(ALA52SER)
2	ZNF473	zinc finger protein 473	0.00025	0.00214	LuCaP105(VAL465ILE), LuCaP115(GLY652ARG)
2	SF3A1	splicing factor 3a, subunit 1, 120kDa	0.0006	0.00268	LuCaP70(PRO558LEU), LuCaP96AI(VAL479ILE)
2	NMI	N-myc (and STAT) interactor	0.00075	0.00245	LuCaP141(ILE302ARG), LuCaP86.2(GLN101ARG)
2	IKZF4	IKAROS family zinc finger 4 (Eos)	0.0008	0.00322	LuCaP93(ASP106ASN), LuCaP81(ASP498ASN)
2	BDH1	3-hydroxybutyrate dehydrogenase, type 1	0.00095	0.00322	LuCaP73(VAL190ILE), LuCaP23.1AI(THR176MET), †LuCaP147(VAL142ILE), LuCaP115(HIS74TYR), †LuCaP147(ALA50VAL)
2	RNF220	ring finger protein 220	0.001	0.00375	LuCaP78(PHE120LEU), LuCaP115(ASP218TYR), LuCaP73(ARG365TRP)
2	CHST6	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 6	0.001	0.00322	LuCaP141(ALA192THR), LuCaP49(SER169ARG)
2	FRMD3	FERM domain containing 3	0.00125	0.00375	LuCaP93(SER202ASN), LuCaP141(MET105ILE)
2	TAF6	TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 80kDa	0.0014	0.00375	LuCaP136(ALA596VAL), LuCaP23.1AI(ARG63TRP)
2	XKR6	XK, Kell blood group complex subunit-related family, member 6	0.0015	0.00294	LuCaP78(VAL497MET), *LuCaP145.2(ALA486VAL)
2	GPC6	glypican 6	0.0018	0.00441	LuCaP141(ARG205GLN), LuCaP105(ALA414THR), LuCaP58(ALA466ASP)
2	SLC22A7	solute carrier family 22 (organic anion transporter), member 7	0.00185	0.00429	LuCaP145.2(ARG172TRP), LuCaP78(THR515MET)
2	GPC5	glypican 5	0.00205	0.00536	LuCaP86.2(ARG199TRP), LuCaP93(TRP521GLY)
2	GPR45	G protein-coupled receptor 45	0.00205	0.00536	LuCaP77(ALA153THR), *LuCaP145.2(VAL175MET)
2	IL1RAP	interleukin 1 receptor accessory protein	0.00225	0.00483	LuCaP105(splice), LuCaP49(SER317ARG)
2	WDR83	WD repeat domain 83	0.00245	0.00536	LuCaP92(ARG103HIS), LuCaP70(ARG141GLY)

2	CHIT1	chitinase 1 (chitotriosidase)	0.00255	0.0059	LuCaP35V(GLY373ASP), LuCaP105(LYS199ARG)
2	GRB10	growth factor receptor-bound protein 10	0.00275	0.00536	LuCaP115(ASP341ASN), LuCaP136(ALA166THR)
2	TBX19	T-box 19	0.0028	0.00536	*LuCaP92(MET158ILE), LuCaP86.2(ALA430VAL)
2	NXF5	nuclear RNA export factor 5	0.0029	0.0059	LuCaP115(ASP136GLY), *LuCaP92(ARG3TRP)
2	EDEM3	ER degradation enhancer, mannosidase alpha-like 3	0.00355	0.0059	LuCaP70(HIS545TYR), LuCaP49(ARG516ILE), LuCaP70(THR348ILE)
2	BICD1	bicaudal D homolog 1 (Drosophila)	0.00395	0.00643	LuCaP35V(GLN138HIS), LuCaP77(PRO975HIS)
2	ITGAL	integrin, alpha L (antigen CD11A (p180), lymphocyte function-associated antigen 1; alpha polypeptide)	0.00395	0.00643	LuCaP86.2(ARG133HIS), *LuCaP92(ILE927ASN)
2	WDFY1	WD repeat and FYVE domain containing 1	0.0044	0.00697	LuCaP93(MET104VAL), LuCaP141(VAL97ILE)
2	MYO3B	myosin IIIB	0.00445	0.00697	LuCaP141(splice), LuCaP77(ARG1104GLN)
2	CDH10	cadherin 10, type 2 (T2-cadherin)	0.00445	0.00697	LuCaP73(ILE714VAL), LuCaP35V(LYS53ARG), *LuCaP145.2(ARG40HIS)
2	TRAK2	trafficking protein, kinesin binding 2	0.00455	0.00643	LuCaP115(GLY877ASP), LuCaP78(ALA120VAL)
2	EHBP1	EH domain binding protein 1	0.00475	0.00697	LuCaP23.1AI(ILE78THR), LuCaP58(TYR701CYS), LuCaP105(MET891VAL)
2	ZNF556	zinc finger protein 556	0.0048	0.00687	LuCaP141(ARG186LEU), LuCaP77(SER215PHE)
2	SORCS3	sortilin-related VPS10 domain containing receptor 3	0.00525	0.00751	LuCaP81(ASP533GLU), LuCaP73(PRO817THR), LuCaP23.1AI(ALA971SER)
2	INTS2	integrator complex subunit 2	0.00535	0.00804	LuCaP115(GLY1172VAL), LuCaP105(ASN1082SER)
2	NDST2	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 2	0.00585	0.00804	LuCaP136(PRO119ARG), *LuCaP145.2(ARG12TRP)
2	SLC25A34	solute carrier family 25, member 34	0.0061	0.00858	LuCaP49(CYS95stop), LuCaP78(ARG277CYS)
2	RPGRIP1	retinitis pigmentosa GTPase regulator interacting protein 1	0.00615	0.00804	LuCaP93(ARG43GLN), LuCaP78(GLU744GLN)
2	C1orf107	chromosome 1 open reading frame 107	0.00615	0.00804	LuCaP58(GLY555ASP), LuCaP141(ARG626HIS), LuCaP115(PHE631LEU)
2	SLC45A2	solute carrier family 45, member 2	0.0066	0.00912	LuCaP86.2(ARG102TRP), LuCaP78(TYR49PHE)
2	PDE4C	phosphodiesterase 4C, cAMP-specific (phosphodiesterase E1 dunce homolog, Drosophila)	0.00675	0.00785	LuCaP23.1AI(ASP548HIS), LuCaP73(HIS511ARG), LuCaP70(MET29THR)
2	FAM171A1	hypothetical protein	0.00725	0.00912	LuCaP105(LEU344HIS), LuCaP141(GLN343LEU), LuCaP73(ALA43THR)
2	CNTROB	centrobin, centrosomal BRCA2 interacting protein	0.0077	0.00912	LuCaP136(ARG202CYS), LuCaP81(GLU290LYS)
2	SFRS14	splicing factor, arginine/serine-rich 14	0.00795	0.00912	LuCaP141(GLY687ARG),

					LuCaP78(GLU556LYS)
2	VWA3B	von Willebrand factor A domain containing 3B	0.00805	0.00965	LuCaP105(ARG802GLN), LuCaP81(LYS1006ASN)
2	ACTN2	actinin, alpha 2	0.00825	0.00965	LuCaP70(HIS474TYR), †LuCaP92(ARG851HIS), LuCaP58(PRO855SER)
2	ERCC2	excision repair cross-complementing rodent repair deficiency, complementation group 2 (xeroderma pigmentosum D)	0.00835	0.00834	LuCaP81(GLY593ARG), LuCaP147(ARG335GLN), LuCaP136(LYS228ARG)
2	WDR4	WD repeat domain 4	0.00865	0.01019	LuCaP73(GLN587HIS), LuCaP58(VAL622MET), LuCaP86.2(ILE633THR), LuCaP58(PRO750HIS), LuCaP115(ALA1141VAL)
2	HAP1	huntingtin-associated protein 1 (neuroan 1)	0.00865	0.01019	LuCaP49(VAL384ASP), LuCaP58(CYS311TYR), LuCaP78(CYS311ARG)
2	AKAP8	A kinase (PRKA) anchor protein 8	0.009	0.01019	LuCaP70(VAL581ALA), LuCaP93(GLN449LYS)
2	CARD11	caspase recruitment domain family, member 11	0.00975	0.01072	LuCaP115(GLY705SER), LuCaP105(ARG688GLN), LuCaP58(ARG386stop)
2	ATG2B	autophagy related 2 homolog B	0.01075	0.01019	LuCaP78(ARG1920HIS), LuCaP93(ARG1602SER)
2	MORC1	MORC family CW-type zinc finger 1	0.01085	0.01126	LuCaP81(ALA949THR), LuCaP141(SER391CYS)
2	NLRP8	NOD-like receptor family pyrin domain containing 8	0.0109	0.01072	LuCaP96AI(CYS612SER), LuCaP105(TRP692ARG)
2	TET1	tet oncogene 1	0.01105	0.01126	LuCaP58(HIS385ARG), LuCaP58(ARG590GLN), LuCaP78(GLY1929CYS), LuCaP77(PRO2032ALA)
2	RRBP1	ribosome binding protein 1 homolog 180kDa (dog)	0.01115	0.01126	LuCaP92(ARG364HIS), LuCaP49(ILE160PHE)
2	C20orf117	chromosome 20 open reading frame 117	0.01115	0.01126	LuCaP145.2(ALA258SER), LuCaP141(ARG234stop)
2	PROM1	prominin 1	0.01175	0.0118	*LuCaP92(splice), LuCaP141(SER324GLY)
2	TRERF1	transcriptional regulating factor 1	0.012	0.01233	LuCaP136(LEU390PHE), LuCaP105(ARG317GLN)
2	FGD6	FYVE, RhoGEF and PH domain containing 6	0.012	0.01233	LuCaP96AI(SER1156ARG), LuCaP81(VAL993ALA), LuCaP73(ARG752CYS), †LuCaP147(ARG729GLN)
2	COL1A2	collagen, type I, alpha 2	0.0121	0.0118	LuCaP141(VAL1245MET), LuCaP115(ARG1258CYS)
2	CECR2	cat eye syndrome chromosome region, candidate 2	0.0128	0.01126	LuCaP78(LEU17MET), LuCaP141(SER150GLY)
2	RAG1	recombination activating gene 1	0.0128	0.01233	†LuCaP145.2(ARG160TRP), LuCaP35V(MET1019THR)
2	ASXL1	additional sex combs like 1 (Drosophila)	0.01315	0.01233	LuCaP73(THR82MET), LuCaP115(TYR449CYS), LuCaP96AI(LEU739ILE)

2	COL4A2	collagen, type IV, alpha 2	0.01445	0.01287	LuCaP136(GLY2ARG), LuCaP141(PRO437SER)
2	MYBPC3	myosin binding protein C, cardiac	0.01465	0.01287	LuCaP136(THR885MET), LuCaP70(VAL321MET)
2	AKAP2,PALM2-AKAP2	PALM2-AKAP2 readthrough transcript	0.01525	0.01394	LuCaP86.2(LYS406GLU), †LuCaP92(ALA633THR)
2	ANKRD26	ankyrin repeat domain 26	0.0154	0.0134	LuCaP81(TYR1559CYS), LuCaP115(LEU1537PHE), LuCaP73(HIS1172ARG)
2	EMR2	egf-like module containing, mucin-like, hormone receptor-like 2	0.0159	0.0134	LuCaP70(LEU598ILE), LuCaP35V(VAL597ILE)
2	C20orf26	chromosome 20 open reading frame 26	0.0159	0.0134	LuCaP73(VAL762MET), LuCaP35V(GLU1001stop), LuCaP81(ASP1029GLU), LuCaP73(ARG1115CYS)
2	ADAM12	ADAM metallopeptidase domain 12 (meltrin alpha)	0.0161	0.0134	LuCaP81(SER902PHE), LuCaP93(ALA730SER)
2	SEZ6L	seizure related 6 homolog (mouse)-like	0.01645	0.01394	LuCaP96AI(ARG361GLN), LuCaP58(GLN704ARG), LuCaP81(PHE1007VAL)
2	MYH9	myosin, heavy chain 9, non-muscle	0.01705	0.01324	LuCaP93(GLY1942VAL), †LuCaP147(LEU1619PRO), LuCaP141(ALA1469VAL)
2	DIP2B	DIP2 disco-interacting protein 2 homolog B (Drosophila)	0.01865	0.01501	LuCaP145.2(VAL1284MET), LuCaP105(ARG1443GLN)
2	UGGT2	UDP-glucose glycoprotein glucosyltransferase 2	0.02045	0.01555	LuCaP49(SER933THR), LuCaP96AI(GLY660ASP)
2	PRSS7	protease, serine, 7 (enterokinase)	0.0207	0.01555	LuCaP93(PRO599SER), LuCaP96AI(THR385SER)
2	FAM65C	amily with sequence similarity 65, member C	0.0218	0.01609	LuCaP81(ARG535GLN), LuCaP35V(ARG123CYS)
2	LAMC1	laminin, gamma 1 (formerly LAMB2)	0.02325	0.0152	LuCaP49(ASN306ASP), LuCaP81(ASN794SER)
2	MED13L	thyroid hormone receptor associated protein 2	0.024	0.01662	LuCaP105(VAL531MET), LuCaP77(THR28MET)
2	KIAA1244	KIAA1244	0.0267	0.01877	LuCaP58(VAL266MET), LuCaP58(ALA1420VAL), LuCaP96AI(GLY1436ALA), LuCaP93(VAL1510ILE), LuCaP58(ALA2018THR), †LuCaP147(ARG2088stop)
2	ATM	ataxia telangiectasia mutated (includes complementation groups A, C and D)	0.0292	0.01877	LuCaP141(ARG686GLN), †LuCaP145.2(CYS989ARG)
2	HEATR1	HEAT repeat containing 1	0.0294	0.01877	LuCaP81(ASP2058ASN), LuCaP141(VAL1721LEU)
2	CENPE	centromere protein E, 312kDa	0.03265	0.0193	LuCaP73(ASN2429ILE), LuCaP81(ALA2322THR), LuCaP58(LYS1524ARG), LuCaP115(GLU915LYS)
2	TDRD6	tudor domain containing 6	0.038	0.02198	LuCaP136(GLU1193GLY), LuCaP58(ASP1283GLU), LuCaP93(ARG1953HIS)
2	LOXHD1	lipoxygenase homology domains 1	0.0387	0.02252	LuCaP93(ALA1241THR), LuCaP77(ASP669ASN)
2	SCN11A	sodium channel, voltage-gated, type XI,	0.03955	0.02198	LuCaP70(VAL1359MET),

		alpha			LuCaP105(VAL926ILE)
2	NSD1	nuclear receptor binding SET domain protein 1	0.03975	0.02011	LuCaP58(SER520GLY), LuCaP78(ARG1206SER), LuCaP49(MET1390THR)
2	PRUNE2	Protein prune homolog 2	0.0417	0.02306	LuCaP70(LEU2195VAL), LuCaP93(PRO2030LEU)
2	C9orf79	chromosome 9 open reading frame 79	0.04385	0.02306	LuCaP96AI(GLN557GLU), LuCaP86.2(HIS1275GLN)
2	ATP10A	ATPase, Class V, type 10A	0.04735	0.02413	LuCaP58(PRO1391LEU), LuCaP58(GLN1327ARG), LuCaP58(SER1115TYR), LuCaP96AI(GLY419GLU), LuCaP35V(GLU221LYS)
2	CELSR2	cadherin, EGF LAG seven-pass G-type receptor 2 (flamingo homolog, Drosophila)	0.06555	0.02949	LuCaP58(LEU40VAL), LuCaP141(THR314MET), LuCaP73(PRO376LEU), LuCaP105(HIS716ARG), LuCaP58(CYS1243ARG)
2	MLL2	myeloid/lymphoid or mixed-lineage leukemia 2	0.0774	0.03164	LuCaP77(ARG4825GLN), LuCaP92(THR2524MET)
2	LYST	lysosomal trafficking regulator	0.08225	0.03324	LuCaP86.2(GLU1715ALA), *LuCaP92(ARG141stop)
2	SACS	spastic ataxia of Charlevoix-Saguenay (sacsin)	0.0916	0.03237	LuCaP23.1AI(ALA2154SER), LuCaP70(ILE1731ARG), LuCaP147(ARG129CYS)
2	DCHS2	dachsous 2 (Drosophila)	0.09605	0.037	LuCaP77(ASP2912GLU), LuCaP136(GLN1585HIS)
2	FAT3	FAT tumor suppressor homolog 3 (Drosophila)	0.10521	0.03861	LuCaP73(splice), LuCaP145.2(PRO1718ALA), LuCaP73(ALA3036VAL), LuCaP70(ARG4233CYS)
2	SH3TC1	SH3 domain and tetratricopeptide repeats 1	0.11601	0.04075	LuCaP77(LYS428GLU), *LuCaP145.2(LEU1057VAL)
2	LRP2	low density lipoprotein-related protein 2	0.13291	0.04071	LuCaP141(GLN3650LYS), †LuCaP147(GLU2910GLY), LuCaP58(SER2734GLY), LuCaP86.2(ALA1901THR), LuCaP141(GLY1852ARG), LuCaP73(ALA1582THR), LuCaP58(CYS115ARG)
2	SPTBN5	spectrin, beta, non-erythrocytic 5	0.13966	0.04169	LuCaP77(THR1950MET), LuCaP78(TRP1401stop)
2	DNAH5	dynein, axonemal, heavy chain 5	0.17896	0.05308	LuCaP141(ARG4358GLN), LuCaP70(ARG196CYS)
2	FAT2	FAT tumor suppressor homolog 2 (Drosophila)	0.19621	0.05523	LuCaP70(LYS3586ASN), LuCaP23.1AI(ARG2835LYS), LuCaP73(GLU2378LYS), LuCaP58(PRO2120SER)
2	FCGBP	Fc fragment of IgG binding protein	0.20261	0.05791	LuCaP81(ARG5160HIS), LuCaP73(SER4595ASN), LuCaP58(ASP1052GLY), LuCaP49(TYR938stop), †LuCaP147(ARG576GLN)
2	HMCN1	hemicentin 1	0.20281	0.05737	LuCaP70(THR4329ILE), LuCaP49(VAL4450ILE)

2	SYNE2	spectrin repeat containing, nuclear envelope 2	0.23286	0.05787	LuCaP141(THR1357MET), LuCaP49(LEU6417VAL)
2	RELN	reelin	0.26526	0.06229	LuCaP73(THR3342MET), †LuCaP147(ARG2639HIS), LuCaP73(PRO2296LEU), LuCaP73(ARG2211CYS), LuCaP49(TYR2108stop), LuCaP73(HIS2064TYR), LuCaP58(ASP1044ASN), †LuCaP147(PRO562SER), LuCaP141(GLY63ASP)
2	MUC17	mucin 17, cell surface associated	0.36422	0.08686	LuCaP141(ARG3567CYS), LuCaP115(THR3862ILE)
2	PKD1	polycystic kidney disease 1 (autosomal dominant)	0.40672	0.09491	†LuCaP147(VAL3320ILE), †LuCaP147(ALA2135VAL), *LuCaP145.2(PRO1786LEU), LuCaP141(SER1679ARG)
2	AHNAK2	AHNAK nucleoprotein 2	0.79314	0.1882	LuCaP93(GLU3878LYS), LuCaP141(TRP979ARG)

Table A.5: Genes with recurrent novel, nonsynonymous alterations**.** The left hand column indicates the number of samples that have a recurrent nov-nsSNV in the gene excluding LuCaP 73, LuCaP 147 and LuCaP 58 as well as the castration sensitive lines LuCaP 35, LuCaP 96 and LuCaP 23.1 and LuCaP 23.12. P-values were estimated by randomly sampling from 1,865 other exomes sequenced at the University of Washington to estimate the probability of recurrently observing nov-nsSNVs in a given candidate gene. Samples with a germline P-value of < 0.00005 were not observed to be recurrent in any of 20,000 iterations. Singleton P-value is defined as the probability of seeing a nov-nsSNVs within a gene in one of 1,865 exomes sequenced. This table is sorted by the probability of observing recurrent novel nonsynonymous alterations within a set of 16 individuals. * indicates a position that was determined to be a rare germline variant after sequencing of three normal tissues corresponding to LuCaP 92, LuCaP 147 and LuCaP 145.2. † indicates a position that was determined to be a somatic mutation after sequencing of normal tissues corresponding to LuCaP 92, 147 and 145.2.

Sample ID	Gene ID	Gene Name	Type of Mutation	Position
LuCaP 23.1AI	PRDM2	PR domain containing 2, with ZNF domain	missense	SER1083TYR
LuCaP 23.1AI	ATP13A2	ATPase type 13A2	missense	GLY325VAL
LuCaP 23.1AI	GPSM2	G-protein signalling modulator 2 (AGS3-like, C. elegans)	missense	GLY99CYS
LuCaP 23.1AI	RRP12	ribosomal RNA processing 12 homolog	missense	VAL682GLU
LuCaP 23.1AI	WNT5B	wingless-type MMTV integration site family, member 5B	missense	GLN234LYS
LuCaP 23.1AI	IVD	isovaleryl Coenzyme A dehydrogenase	missense	ALA62VAL
LuCaP 23.1AI	TTLL12	tubulin tyrosine ligase-like family, member 12	missense	CYS254PHE
LuCaP 23.1AI	MGAT4B	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme B	missense	ARG168CYS
LuCaP 23.1AI	FZD6	frizzled homolog 6 (Drosophila)	missense	ARG509GLN
LuCaP 23.1AI	C12orf48	chromosome 12 open reading frame 48	missense	ALA253SER
LuCaP 23.1AI	C21orf58	chromosome 21 open reading frame 58	nonsense	GLN218stop
LuCaP 23.1AI	FAM47C	family with sequence similarity 47, member C	nonsense	GLU345stop
LuCaP 35V	ZNF566	zinc finger protein 566	missense	ILE164PHE
LuCaP 35V	SCP2	sterol carrier protein 2	missense	ARG12LEU
LuCaP 35V	BTBD10	BTB (POZ) domain containing 10	missense	HIS444TYR
LuCaP 35V	RAG1	recombination activating gene 1	missense	MET1019THR
LuCaP 35V	OR5R1	olfactory receptor, family 5, subfamily R, member 1	missense	ARG20GLN
LuCaP 35V	HRASLS3	HRAS-like suppressor 3	missense	PRO19HIS
LuCaP 35V	RAB38	RAB38, member RAS oncogene family	missense	SER204ARG
LuCaP 35V	TPH2	tryptophan hydroxylase 2	missense	HIS318TYR
LuCaP 35V	PRKCA	protein kinase C, alpha	missense	PRO291GLN
LuCaP 35V	EXOC3L2	exocyst complex component 3-like 2	missense	ALA350SER
LuCaP 35V	ZFP28	zinc finger protein 28 homolog (mouse)	missense	TRP172SER
LuCaP 35V	CRYGB	crystallin, gamma B	missense	TYR29SER
LuCaP 35V	WNT6	wingless-type MMTV integration site family, member 6	missense	ARG83HIS
LuCaP 35V	C20orf160	chromosome 20 open reading frame 160	missense	ASP52GLY
LuCaP 35V	ALDH1L1	aldehyde dehydrogenase 1 family, member L1	missense	GLU804ASP
LuCaP 35V	PKD2	polycystic kidney disease 2 (autosomal dominant)	missense	GLU879ALA
LuCaP 35V	PTPRD	protein tyrosine phosphatase, receptor type, D	missense	ARG1487CYS
LuCaP 35V	SLC9A7	solute carrier family 9 (sodium/hydrogen exchanger), member 7	missense	ARG499CYS
LuCaP 35V	ODZ1	odz, odd Oz/ten-m homolog 1(Drosophila)	missense	ASP2709ASN
LuCaP 35V	ITGA7	integrin, alpha 7	missense	ARG347CYS
LuCaP 35V	KIF2B	kinesin family member 2B	missense	ARG539LYS
LuCaP 35V	BMPR2	bone morphogenetic protein receptor, type II	nonsense	GLU368stop

		(serine/threonine kinase)		
LuCaP 35V	SLC4A4	solute carrier family 4, sodium bicarbonate cotransporter, member 4	nonsense	TYR506stop
LuCaP 35V	MDN1	MDN1, midasin homolog (yeast)	nonsense	TRP3032stop
LuCaP 96AI	SARDH	sarcosine dehydrogenase	missense	VAL132MET
LuCaP 96AI	MED18	mediator of RNA polymerase II transcription, subunit 18 homolog (S. cerevisiae)	missense	TYR146ASP
LuCaP 96AI	MSH4	mutS homolog 4 (E. coli)	missense	ASN420HIS
LuCaP 96AI	SLAMF9	SLAM family member 9	missense	SER39ARG
LuCaP 96AI	CDC73	cell division cycle 73, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)	missense	ARG171THR
LuCaP 96AI	HIST3H3	histone cluster 3, H3	missense	GLU98LYS
LuCaP 96AI	ADD3	adducin 3 (gamma)	missense	GLU253ASP
LuCaP 96AI	C11orf40	chromosome 11 open reading frame 40	missense	SER81GLY
LuCaP 96AI	MADD	MAP-kinase activating death domain	missense	GLY1101GLU
LuCaP 96AI	ADAMTS20	ADAM metallopeptidase with thrombospondin type 1 motif, 20	missense	ARG418SER
LuCaP 96AI	SLC39A5	solute carrier family 39 (metal ion transporter), member 5	missense	SER229PHE
LuCaP 96AI	WIF1	WNT inhibitory factor 1	missense	HIS347GLN
LuCaP 96AI	KCNC2	potassium voltage-gated channel, Shaw-related subfamily, member 2	missense	LEU394TRP
LuCaP 96AI	THTPA	thiamine triphosphatase	missense	THR18PRO
LuCaP 96AI	KIAA1622	KIAA1622	missense	ARG722SER
LuCaP 96AI	MAPKBP1	mitogen activated protein kinase binding protein 1	missense	THR40ILE
LuCaP 96AI	ADAL	adenosine deaminase-like	missense	LEU102ARG
LuCaP 96AI	ADAM10	ADAM metallopeptidase domain 10	missense	CYS607PHE
LuCaP 96AI	UBN1	ubinuclein 1	missense	VAL591LEU
LuCaP 96AI	C16orf45	chromosome 16 open reading frame 45	missense	ASP86GLY
LuCaP 96AI	EPAS1	endothelial PAS domain protein 1	missense	ALA698ASP
LuCaP 96AI	SERTAD2	SERTA domain containing 2	missense	ARG8TRP
LuCaP 96AI	HK2	hexokinase 2	missense	GLN580PRO
LuCaP 96AI	RNF103	ring finger protein 103	missense	GLY58SER
LuCaP 96AI	DPP10	dipeptidyl-peptidase 10	missense	TRP336ARG
LuCaP 96AI	MYT1	myelin transcription factor 1	missense	PRO776THR
LuCaP 96AI	PRSS7	protease, serine, 7 (enterokinase)	missense	THR385SER
LuCaP 96AI	SF3A1	splicing factor 3a, subunit 1, 120kDa	missense	VAL479ILE
LuCaP 96AI	COL7A1	collagen, type VII, alpha 1 (epidermolysis bullosa, dystrophic, dominant and recessive)	missense	ARG1751GLN
LuCaP 96AI	RPL24	ribosomal protein L24	missense	ALA136THR
LuCaP 96AI	GSK3B	glycogen synthase kinase 3 beta	missense	TYR216PHE

LuCaP 96AI	RAB28	RAB28, member RAS oncogene family	missense	SER165TYR
LuCaP 96AI	LOC91431	LOC91431	missense	ALA709PRO
LuCaP 96AI	FAT4	FAT tumor suppressor homolog 4 (Drosophila)	missense	ARG2557ILE
LuCaP 96AI	F11	coagulation factor XI (plasma thromboplastin antecedent)	missense	LYS112ARG
LuCaP 96AI	SGTB	small glutamine-rich tetratricopeptide repeat (TPR)-containing, beta	missense	GLY257ARG
LuCaP 96AI	C6orf138	chromosome 6 open reading frame 138	missense	VAL640ALA
LuCaP 96AI	KCNQ5	potassium voltage-gated channel, KQT-like subfamily, member 5	missense	GLY349CYS
LuCaP 96AI	NCOA7	nuclear receptor coactivator 7	missense	ASP413TYR
LuCaP 96AI	KIAA1244	KIAA1244	missense	GLY1436ALA
LuCaP 96AI	FAM71F1	FAM71F1	missense	ARG145LYS
LuCaP 96AI	NRG1	neuregulin 1	missense	SER47LEU
LuCaP 96AI	BAG4	BCL2-associated athanogene 4	missense	SER100ALA
LuCaP 96AI	ZNF75D	zinc finger protein 75D	missense	ASN461TYR
LuCaP 96AI	ITGA7	integrin, alpha 7	missense	VAL165LEU
LuCaP 96AI	ICAM4	intercellular adhesion molecule 4 (Landsteiner-Wiener blood group)	missense	ASN190ASP
LuCaP 96AI	ZFP112	zinc finger protein 112 homolog	missense	GLY489ALA
LuCaP 96AI	SILV	silver homolog (mouse)	missense	PRO199ALA
LuCaP 96AI	ZFP112	zinc finger protein 112 homolog	nonsense	GLY489stop
LuCaP 96AI	TGM2	transglutaminase 2 (C polypeptide, protein-glutamine-gamma-glutamyltransferase)	splice-5	none
LuCaP 96AI	SLC35D1	solute carrier family 35 (UDP-glucuronic acid/UDP-N-acetylgalactosamine dual transporter), member D1	splice-5	none

Table A.6: Genes with mutations present in a castration resistant xenograft and not present in a castration sensitive counterpart derived from the same original tumor. Positions were verified manually using IGV to ensure that variants were present uniquely within the castration resistant xenografts.

Sample ID	# of novSNVs	One copy gain	One copy loss	Two or more copy gain	Two copy loss	Total copy number aberrations
LuCaP 58	4067	555	745	117	165	1582
LuCaP 73	2972	668	702	63	144	1577
LuCaP 147	2714	397	692	39	167	1295
LuCaP 70	389	442	686	103	186	1417
LuCaP 92	313	769	741	78	181	1769
LuCaP 145.2	538	830	997	144	158	2129

Table A.7: Hypermutated tumors do not appear to possess more copy number alterations than other tumors. Copy number alterations were measured using Illumina Arrays for three hypermutated tumors (LuCaP 58, LuCaP 73 and LuCaP 147) as well as three randomly selected tumors (LuCaP70, LuCaP92, and LuCaP145.2). The number of novel single nucleotide variants (novSNVs) is shown alongside copy number calls as determined by Biodiscovery Nexus Copy Number 6.0 software. Total copy number in hypermutated tumors is not significantly different in from non-hypermutated tumors (p-value = 0.29, two sided t-test)

Corresponding Xenograft	Type of tissue	Tissue origin	Capture Method	Indexing	Sequencer	Run-type	Positions called on RefSeq target at threshold quality	% of RefSeq target bases (36.6 Mb) covered at threshold quality
LuCaP 92	normal	Normal kidney	V2	no	HiSeq	PE-100	34696005	95.9%
LuCaP 92	tumor	LN metastasis	V2	no	HiSeq	PE-100	34340968	95.0%
LuCaP 145.2*	normal	Normal muscle	V2	no	HiSeq	PE-100	34658720	95.8%
LuCaP 145.2*	tumor	LN metastasis	V2	no	HiSeq	PE-100	34278718	94.8%
LuCaP 147*	normal	Normal liver	V2	no	HiSeq	PE-100	34592876	95.7%
LuCaP 147*	tumor	Lung metastasis	V2	no	HiSeq	PE-100	34383790	95.1%

Table A.8: Coverage statistics for paired tumor and normal tissue corresponding to three xenograft lines. The exomes of normal and non-xenografted tumor tissue were sequenced for three xenograft lines in order to estimate the efficiency of germline filtering for identifying somatic mutations. *In two cases, the original tumor sample used to generate the xenograft line was not available so a neighboring metastasis from the same individual was used. To capture exomes, we used the Nimblegen EZ SeqCap V2 capture probes(targeting the 36.6 Mb RefSeq database). Approximately 95% of the RefSeq target regions were able to be called in most samples (8x coverage and *Samtools*-derived *phred* quality score > 30). PE-100 paired-end sequencing using 100 bp reads.

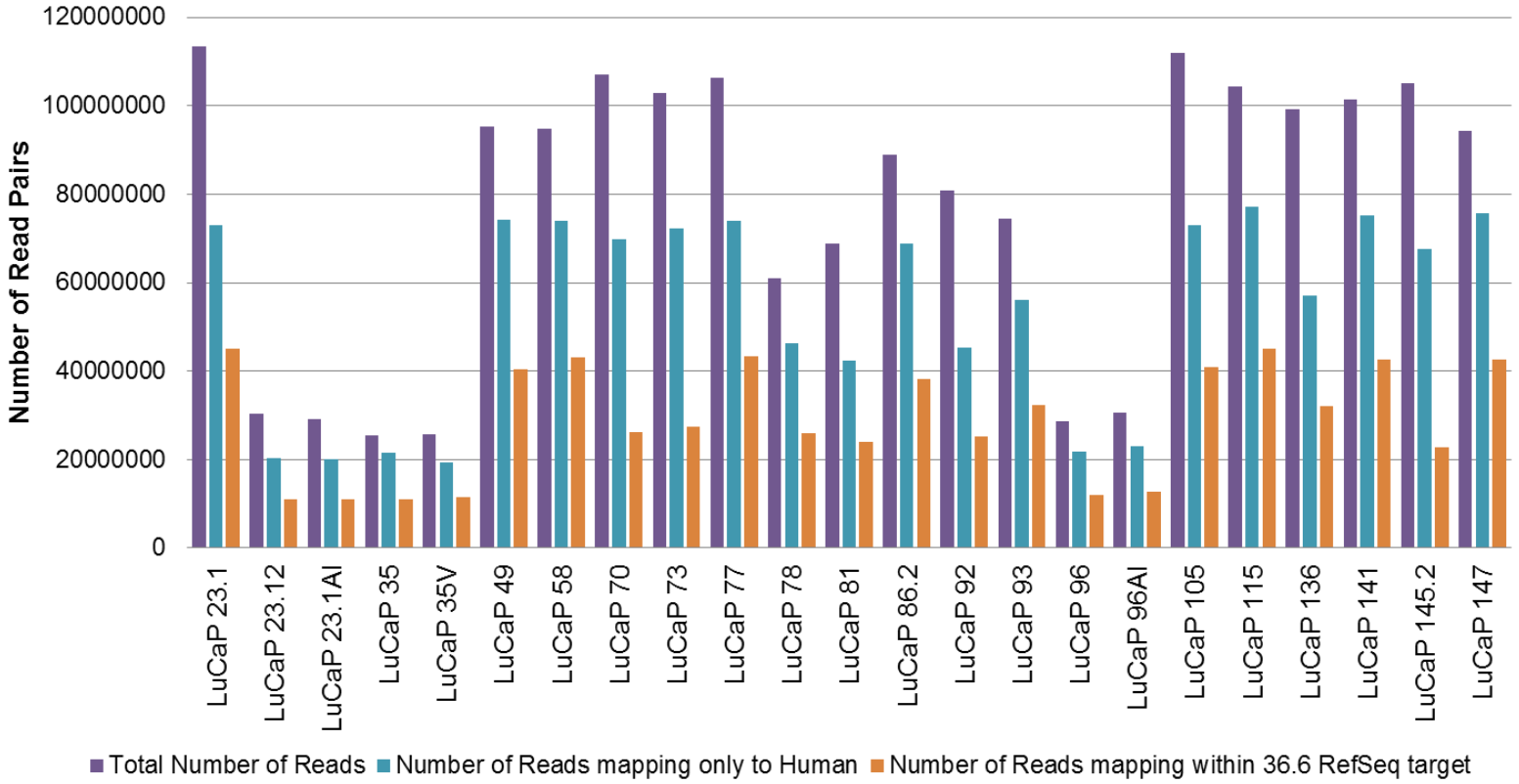


Figure A.1: Summary of mapping statistics across 23 PCa xenograft exomes. LuCaP samples 96, 96AI, 23.12, 23.1AI, 35 and 35V were sequenced using the Illumina GAIIx, which accounts for the smaller number of reads obtained for these samples.

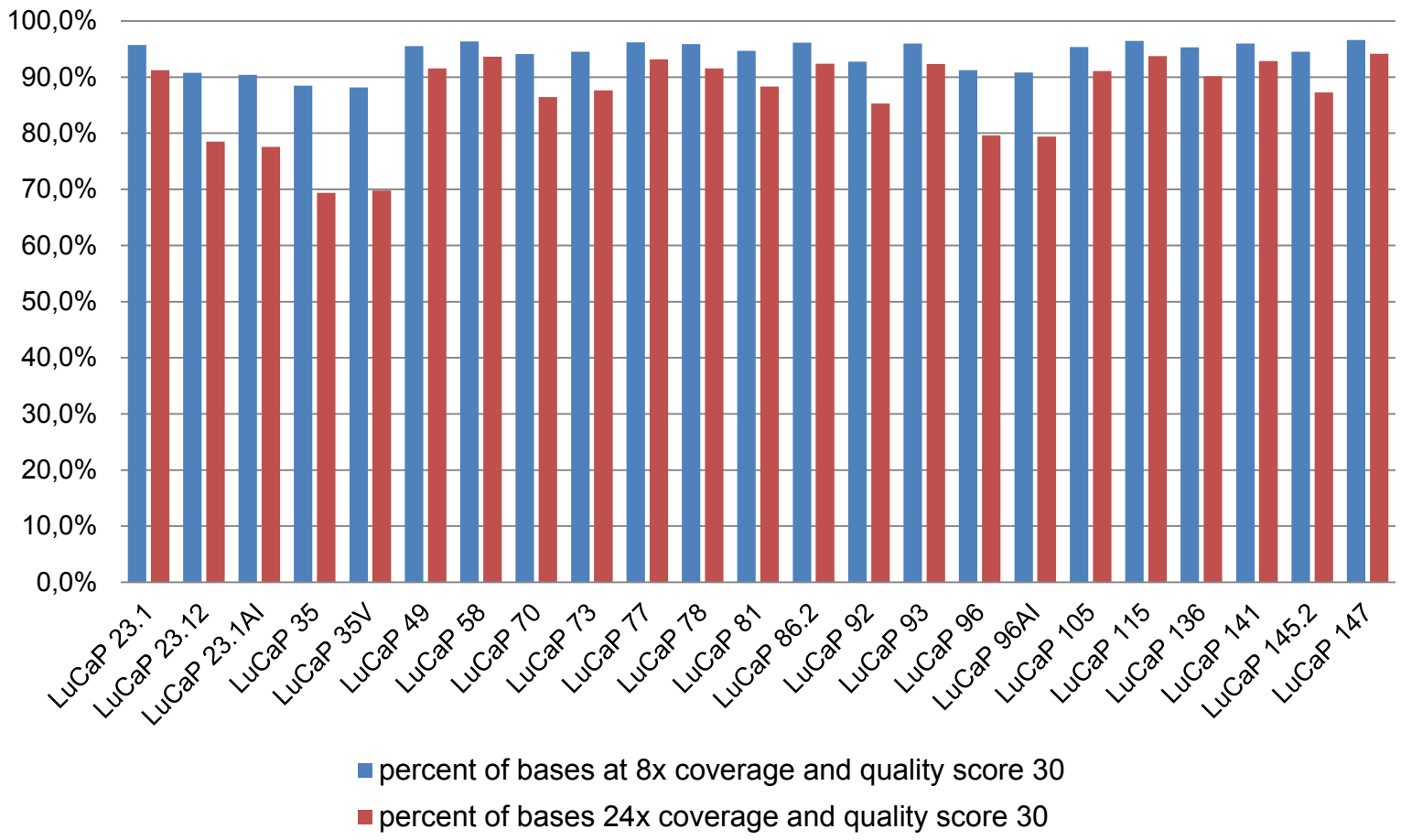


Figure A.2: Fraction of bases in the V1 target definition that were covered to sufficient depth to enable basecalling.

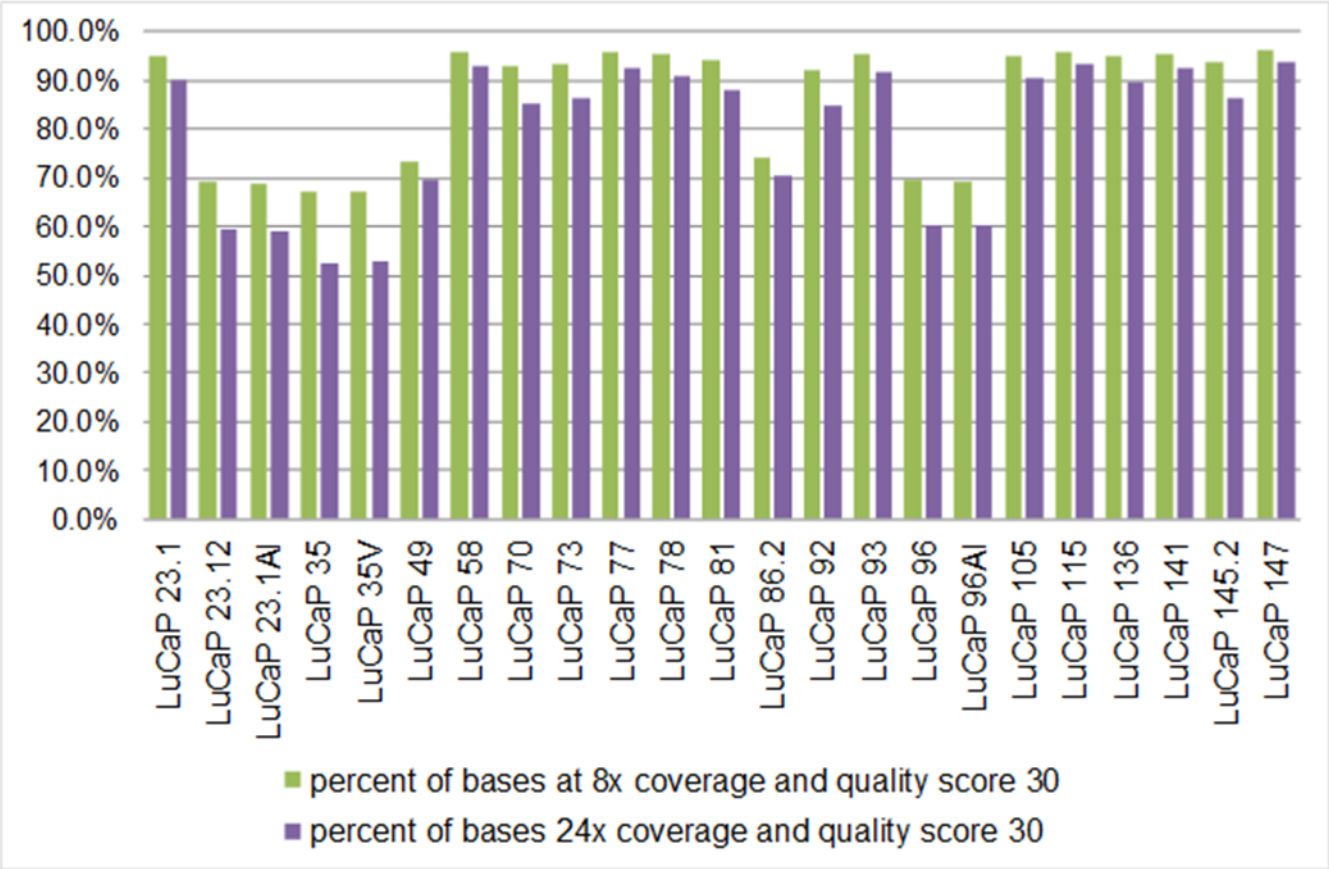


Figure A.3: Fraction of bases in the V2 target definition that were covered to sufficient depth to enable basecalling. Samples LuCaP 96, 96V, 23.12, 23.1AI, 35, 35V, 49 and 86.2 were selected for a smaller (V1) target, which accounts for their relatively lower coverage of these regions.

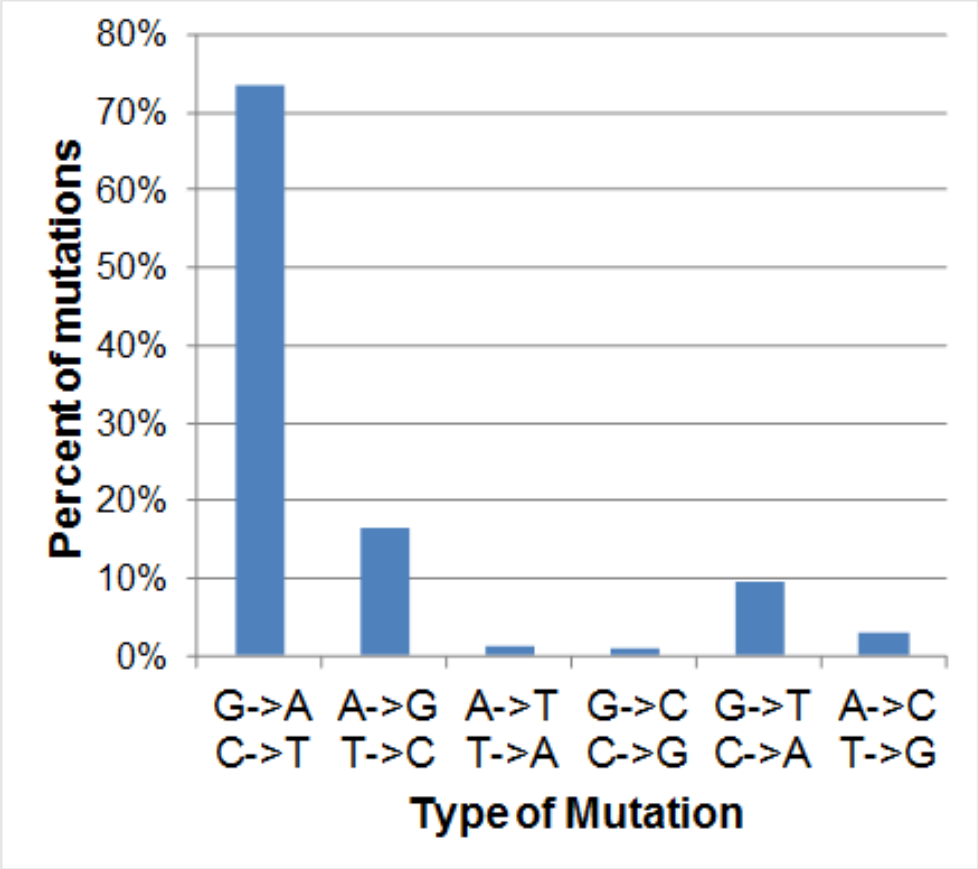
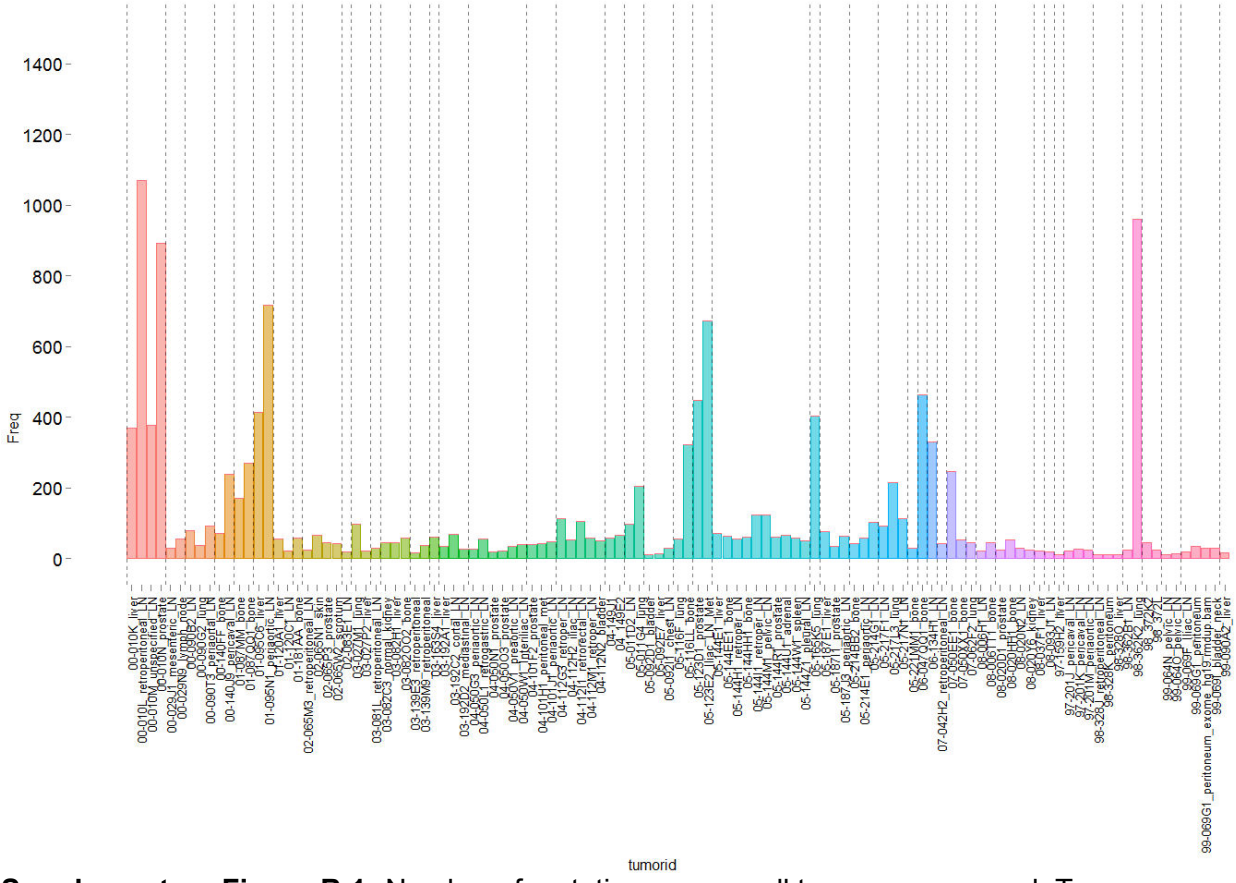
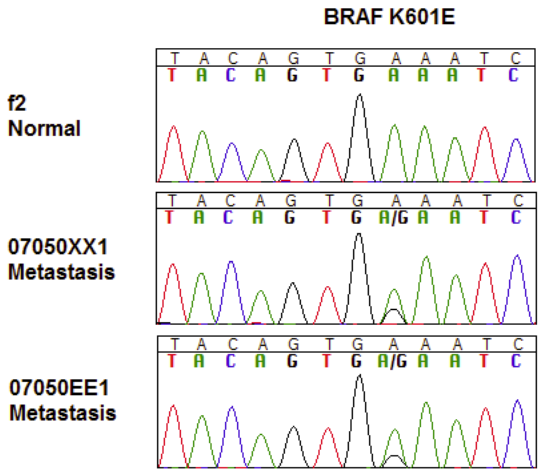


Figure A.4: The mutation profile of a metastasis corresponding to LuCaP 147 as determined by exome sequencing of both normal tissue and metastasis. This Figure A.hows the number of mutations in each of the six possible mutation classes. Transitions are heavily favored in this tumor, making up ~90% of mutations observed (first two bars).

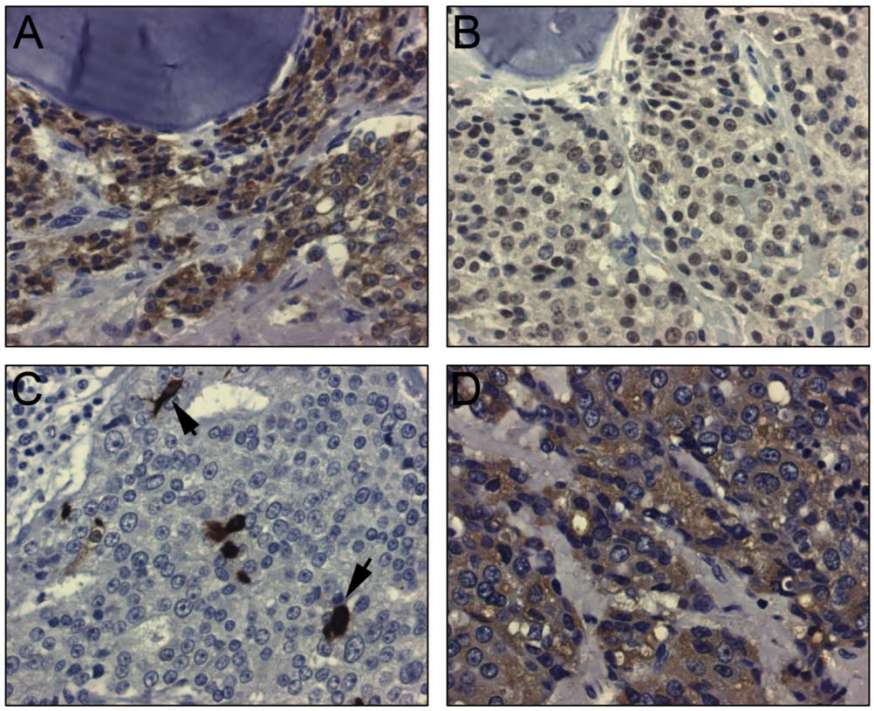
Appendix B- Supplementary Material for Chapter 3



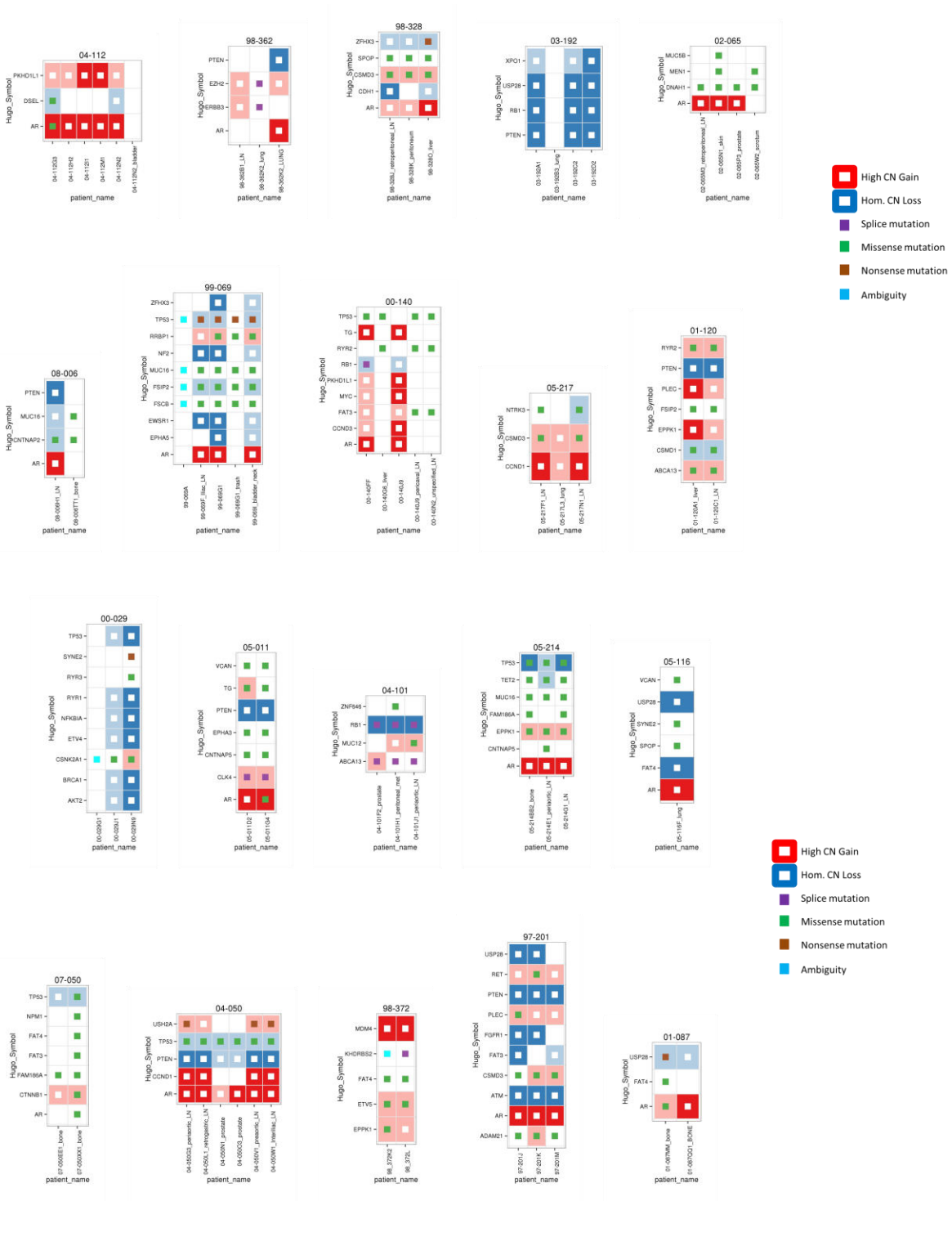
Supplementary Figure B.1: Number of mutations across all tumors sequenced. Tumors are grouped by corresponding patients (dotted lines). Five patients display hypermutation (00-010, 01-095, 05-123, 05-165 and 06-134). 98-362 was likely due to a sample mixup.

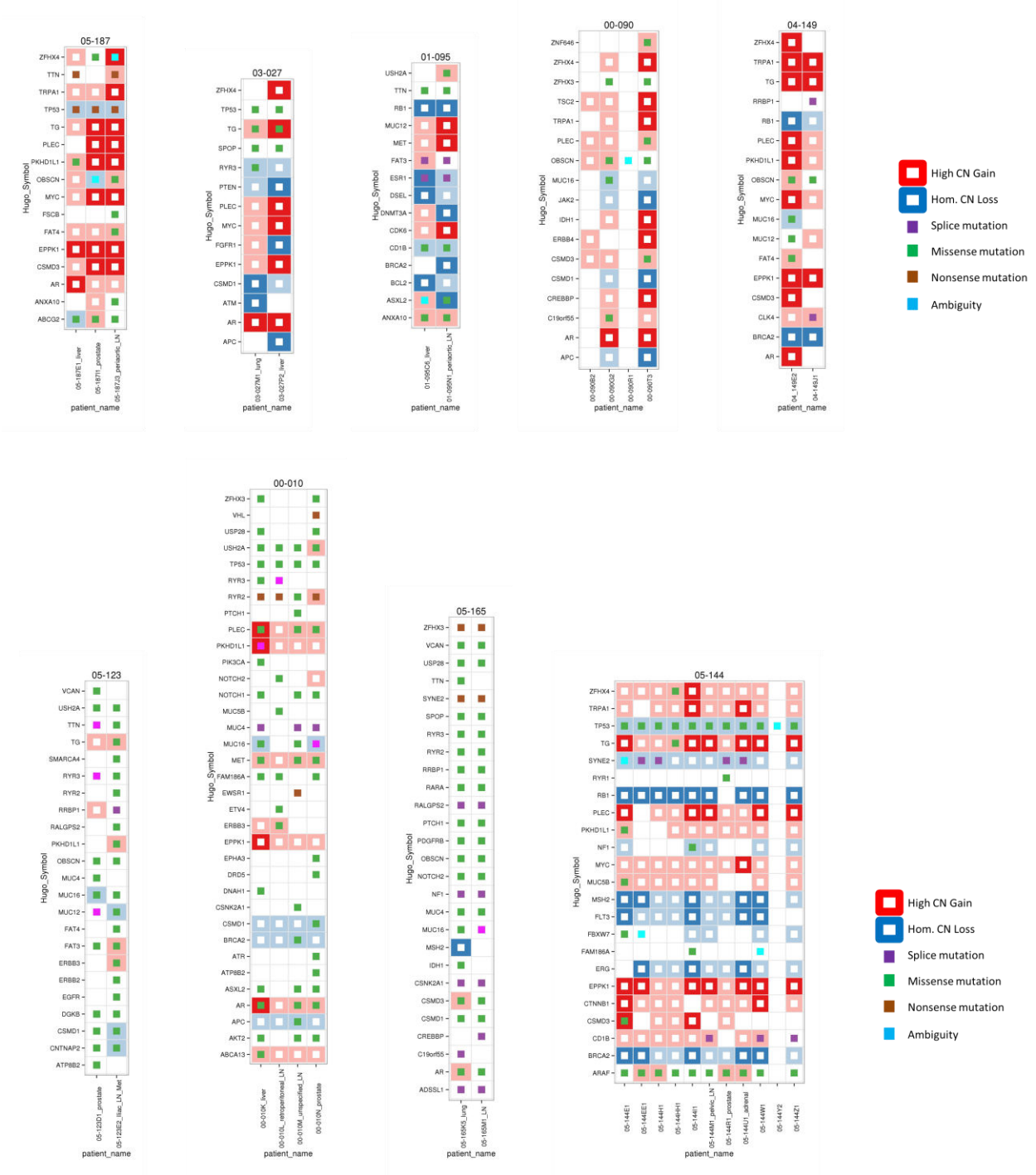


Supplementary Figure B.2: Sanger Traces of the BRAF exon containing the K601E mutation in two representative bone metastases from patient 07-050.



Supplementary Figure B.3: Immunohistochemical stains of bone metastasis 07-050XX display features of prostate cancer. (A) cytokeratins, (B) androgen receptor, (C) chromogranin A, and (D) PSA. Arrows indicate chromogranin A positive cells (400-fold magnification).


04-112
98-362
98-328
03-192
02-065
99-069
08-006
00-140
05-217
01-120
00-029
05-011
04-101
05-214
05-116
07-050
04-050
98-372
97-201
01-087
Hugo_Symbol
patient_name
High CN Gain
Hom. CN Loss
Splice mutation
Missense mutation
Nonsense mutation
Ambiguity



Supplementary Figure B.4: Plots of heterogeneity in somatic mutation within Prostate Cancer metastases. A gene was considered to have a positive hit in a tumor if it either possessed a nonsynonymous point mutation (missense, splice, nonsense) or if it was subject to either homozygous copy number loss or high-level amplification.

Patient	Block	Tissue	Met type	Exome Sequencing	aCGH
00-010	00-010C	Normal kidney	Normal	Y	Y
00-010	00-010K	liver	LIVER	Y	Y
00-010	00-010L	retroperitoneal LN	LN	Y	Y
00-010	00-010M	unspecified LN	LN	Y	Y
00-010	00-010N	prostate	CAP	Y	Y
00-029	00-029G1	Normal kidney	Normal	Y	Y
00-029	00-029J1	mesenteric LN	LN	Y	Y
00-029	00-029N9	lymph node	LN	Y	Y
00-090	00-090B2	LN	LN	Y	Y
00-090	00-090G2	Lung	LUNG	Y	Y
00-090	00-090R1	Normal kidney	Normal	Y	Y
00-090	00-090T3	Periaortal LN	LN	Y	Y
00-140	00-140B1	Normal kidney	Normal	Y	Y
00-140	00-140FF	bone	BONE	Y	Y
00-140	00-140G6	liver	LIVER	Y	N
00-140	00-140J9	pericaval LN	LN	Y	Y
00-140	00-140N2	unspecified LN	LN	Y	N
01-002	01-002R	Normal kidney	Normal	Y	Y
01-002	01-002V	LN	LN	Y	Y
01-087	01-087D	Normal kidney	Normal	Y	Y
01-087	01-087MM	bone	BONE	Y	Y
01-095	01-095A1	Normal kidney	Normal	Y	N
01-095	01-095C6	liver	LIVER	Y	Y
01-095	01-095N1	Periaortic LN	LN	Y	Y
01-120	01-120A1	liver	LIVER	Y	Y
01-120	01-120C1	LN	LN	Y	Y
01-120	01-120H1	Normal spleen	Normal	Y	Y
01-181	01-181A2	Normal kidney	Normal	Y	Y
02-065	02-065A2	Normal spleen	Normal	Y	Y
02-065	02-065M3	retroperitoneal LN Met	LN	Y	Y
02-065	02-065N1	Skin	SKIN	Y	Y
02-065	02-065P3	Prostate	CAP	Y	Y
02-065	02-065W2	Scrotum	SCROTUM	Y	Y
02-083	02-083D3	lung	LUNG	Y	N
02-083	02-083E1	LN	LN	Y	Y
02-083	02-083G2	Normal kidney	Normal	Y	Y
02-142	02-142E2	Normal spleen	Normal	Y	Y

03-027	03-027G1	Normal kidney	Normal	Y	Y
03-027	03-027M1	lung	LUNG	Y	Y
03-027	03-027P2	liver	LIVER	Y	Y
03-081	03-081H2	Normal liver	Normal	Y	Y
03-081	03-081L	retroperitoneal LN	LN	Y	Y
03-082	03-082C3	Normal kidney	Normal	Y	Y
03-082	03-082H1	liver	LIVER	Y	Y
03-130	03-130B1	Normal spleen	Normal	Y	Y
03-130	03-130L PERI1MET2	retroperitoneal	RETROPERITONEAL	Y	Y
03-130	03-130M4	liver	LIVER	Y	Y
03-139	03-139B3	Normal kidney	Normal	Y	Y
03-139	03-139E3	retroperitoneal	RETROPERITONEAL	Y	Y
03-139	03-139M9	retroperitoneal	RETROPERITONEAL	Y	Y
03-163	03-163C3	Normal kidney	Normal	Y	Y
03-163	03-163S4	liver	LIVER	Y	Y
03-192	03-192A1	liver	LIVER	Y	Y
03-192	03-192B3	lung	LUNG	Y	N
03-192	03-192C2	cortal LN	LN	Y	Y
03-192	03-192D2	mediastinal LN	LN	Y	Y
03-192	03-192H3	Normal kidney	Normal	Y	N
04-050	04-050C2	Normal kidney	Normal	Y	Y
04-050	04-050G3	periaortic LN	LN	Y	Y
04-050	04-050L1	retrogastric LN	LN	Y	Y
04-050	04-050N1	prostate	CAP	Y	Y
04-050	04-050O3	prostate	CAP	Y	Y
04-050	04-050V1	preaortic LN #5	LN	Y	Y
04-050	04-050W1	Interiliac LN #6	LN	Y	Y
04-101	04-101B3	Normal kidney	Normal	Y	Y
04-101	04-101F2	prostate	CAP	Y	Y
04-101	04-101H1	peritoneal met	PERITONEAL	Y	Y
04-101	04-101J1	periaortic LN	LN	Y	Y
04-101	04-101L3	retromediastinal LN	LN	Y	Y
04-112	04-112	Normal kidney	Normal	Y	Y
04-112	04-112G3	retroperitoneal LN	LN	Y	Y
04-112	04-112H2	iliac LN	LN	Y	Y
04-112	04-112I1	retrorectal LN	LN	Y	Y
04-112	04-112M1	retroperitoneal LN	LN	Y	Y
04-112	04-112N2	prostate bladder	CAP	Y	Y
04-149	04-149A2	Normal liver	Normal	Y	Y
04-149	04-149E2	unspecified LN	LN	Y	Y
04-149	04-149J1	Prostate	CAP	Y	Y

05-011	05-011A2	Normal liver	Normal	Y	Y
05-011	05-011D2	LN	LN	Y	Y
05-011	05-011G4	lung	LUNG	Y	Y
05-092	05-092A3	Normal liver	Normal	Y	Y
05-092	05-092D1	prostate bladder	CAP	Y	Y
05-092	05-092E7	liver	LIVER	Y	Y
05-092	05-092I1	Chest LN	LN	Y	Y
05-116	05-116A2	Normal liver	Normal	Y	Y
05-116	05-116F	lung	LUNG	Y	Y
05-123	05-123B3	Normal kidney	Normal	Y	Y
05-123	05-123D1	Prostate	CAP	Y	Y
05-123	05-123E2	Iliac LN Met	LN	Y	Y
05-144	05-144E1	liver	LIVER	Y	Y
05-144	05-144EE1	bone	BONE	Y	Y
05-144	05-144H1	retroperitoneal LN	LN	Y	Y
05-144	05-144HH1	bone	BONE	Y	Y
05-144	05-144I1	retroperitoneal LN	LN	Y	Y
05-144	05-144M1	pelvic LN	LN	Y	Y
05-144	05-144R1	prostate	CAP	Y	Y
05-144	05-144U1	L adrenal	ADRENAL	Y	Y
05-144	05-144W1	spleen	SPLEEN	Y	Y
05-144	05-144Y2	Normal kidney	Normal	Y	Y
05-144	05-144Z1	pleural LN	LN	Y	Y
05-148	05-148B1	Normal skin	Normal	Y	Y
05-148	05-148E3	liver	LIVER	Y	Y
05-165	05-165A1	Normal liver	Normal	Y	Y
05-165	05-165C3	Normal muscle	Normal	Y	Y
05-165	05-165K5	lung	LUNG	Y	Y
05-165	05-165M1	LN	LN	Y	Y
05-165	05-165O	adrenal	ADRENAL	Y	Y
05-187	05-187A3	Normal liver	Normal	Y	Y
05-187	05-187E1	liver met	LIVER	Y	Y
05-187	05-187I1	prostate	CAP	Y	Y
05-187	05-187J3	periaortic LN	LN	Y	Y
05-214	05-214A2	Normal liver	Normal	Y	Y
05-214	05-214BB2	bone	BONE	Y	Y
05-214	05-214E1	periaortic LN	LN	Y	Y
05-214	05-214G1	LN	LN	Y	Y
05-217	05-217A3	Normal liver	Normal	Y	Y
05-217	05-217F1	LN	LN	Y	Y
05-217	05-217L3	Lung	LUNG	Y	Y

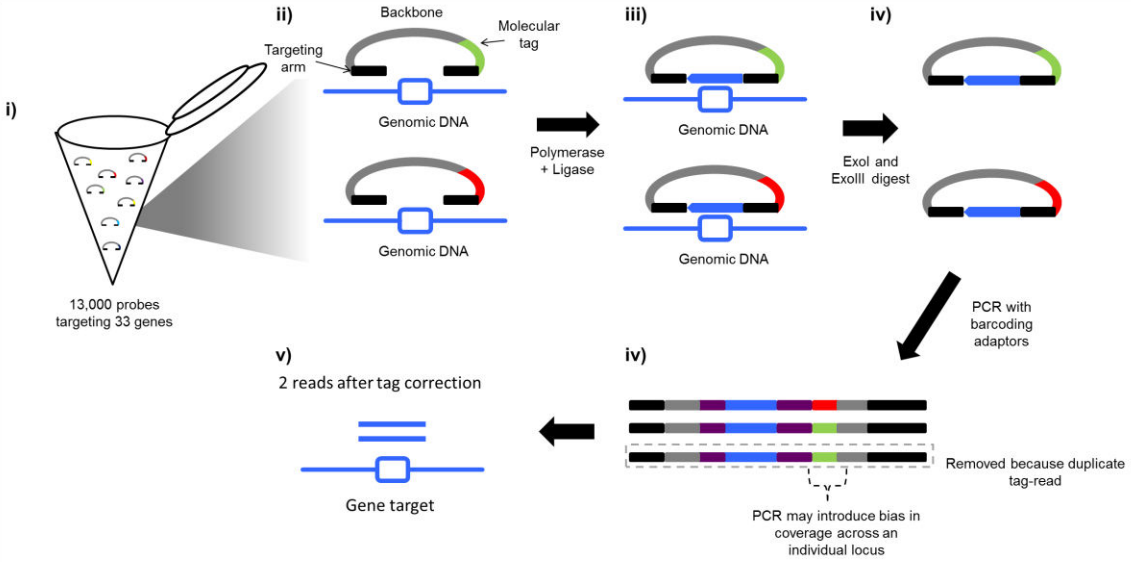
05-217	05-217N1	LN	LN	Y	Y
05-221	05-221A	Normal liver	Normal	Y	Y
05-221	05-221MM1	bone	BONE	Y	Y
06-047	06-047A3	Normal liver	Normal	Y	Y
06-081	06-081B2	Normal skin	Normal	Y	Y
06-081	06-081D3	prostate	CAP	Y	Y
06-081	06-081E4	M liver	LIVER	Y	Y
06-081	06-081F2	rt pelvic LN	LN	Y	Y
06-081	06-081H5	rt pelvic LN	LN	Y	Y
06-081	06-081I1	lung	LUNG	Y	Y
06-081	06-081J1	mediastinal LN	LN	Y	Y
06-081	06-081K2	periaortic LN	LN	Y	Y
06-127	06-127A3	Normal liver	Normal	Y	Y
06-127	06-127H3	iliac LN	LN	Y	Y
06-131	06-131A1	Normal liver	Normal	Y	N
06-131	06-131BB1	bone	BONE	Y	N
06-134	06-134A1	Normal liver	Normal	Y	Y
06-134	06-134H1	LN	LN	Y	Y
07-042	07-042A2	Normal liver	Normal	Y	Y
07-042	07-042H2	medial retroperitoneal LN	LN	Y	Y
07-050	07-050EE1	bone	BONE	Y	Y
07-050	07-050F2	Normal spleen	Normal	Y	Y
07-050	07-050XX1	bone	BONE	Y	Y
07-062	07-062A1	Normal liver	Normal	Y	Y
07-062	07-062F2	lung	LUNG	Y	Y
08-006	08-006A2	Normal liver	Normal	Y	Y
08-006	08-006H1	LN	LN	Y	Y
08-006	08-006TT1	bone	BONE	Y	N
08-020	08-020C1	Normal muscle	Normal	Y	Y
08-020	08-020D1	prostate	CAP	Y	Y
08-020	08-020N2	LN	LN	Y	Y
08-020	08-020Y6	kidney	KIDNEY	Y	Y
08-037	08-037A1	Normal liver	Normal	Y	Y
08-037	08-037F1	liver	LIVER	Y	Y
08-093	08-093C3	Normal muscle	Normal	Y	Y
08-093	08-093J1	LN	LN	Y	Y
09-006	09-006C1	Normal muscle	Normal	Y	Y
09-006	09-006F1	Renal	RENAL	Y	Y
10-013	10-013C1	Normal muscle	Normal	Y	Y
10-013	10-013F1	lymph node	LN	Y	Y
10-013	10-013G1	retroperitoneal	RETROPERITONEAL	Y	Y

10-013	10-013J2	lymph node	LN	Y	Y
10-039	10-039C2	Normal muscle	Normal	Y	Y
10-039	10-039E3	lung	LUNG	Y	Y
11-028	11-028C1	Normal muscle	Normal	Y	Y
11-028	11-028G3	adrenal	ADRENAL	Y	Y
11-028	11-028L1	lung	LUNG	Y	Y
12-005	12-005C1	Normal muscle	Normal	Y	Y
12-005	12-005I2	liver	LIVER	Y	Y
12-005	12-005K1	lung	LUNG	Y	Y
12-011	12-011C1	Normal muscle	Normal	Y	Y
12-011	12-011D26	prostate	CAP	Y	Y
12-011	12-011I7	lymph node	LN	Y	Y
12-011	12-011J1	bladder	CAP	Y	Y
12-021	12-021C1	Normal muscle	Normal	Y	Y
12-021	12-021H4	liver	LIVER	Y	Y
97-159	97-159F	Normal muscle	Normal	Y	Y
97-159	97-159H2	liver	LIVER	Y	Y
97-201	97-201B	Normal liver	Normal	Y	Y
97-201	97-201J	r pericaval LN	LN	Y	Y
97-201	97-201K	r pericaval LN	LN	Y	Y
97-201	97-201M	r periaortic LN	LN	Y	Y
98-328	98-328C	Normal liver	Normal	Y	Y
98-328	98-328H	Appendix	APPENDIX	Y	Y
98-328	98-328J	retroperitoneal LN Met	LN	Y	Y
98-328	98-328K	Peritoneum	PERITONEUM	Y	Y
98-328	98-328O	liver	LIVER	Y	Y
98-362	98-362B1	LN	LN	Y	Y
98-362	98-362E	Normal liver	Normal	Y	Y
98-372	98-372D	Normal kidney	Normal	Y	Y
98-372	98-372K2	Prostate (bladder neck)	CAP	Y	Y
98-372	98-372L	retrocaval LN Met	LN	Y	Y
99-064	99-064I	Normal spleen	Normal	Y	Y
99-064	99-064N	Pelvic LN Met	LN	Y	Y
99-064	99-064O	Pelvic LN Met	LN	Y	Y
99-069	99-069A	Normal kidney	Normal	Y	Y
99-069	99-069F	lung	LUNG	Y	Y
99-069	99-069G1	Peritoneum	PERITONEUM	Y	Y
99-069	99-069I	Bladder (LT-neck/prostate)	CAP	Y	Y
99-090	99-090A2	liver	LIVER	Y	Y

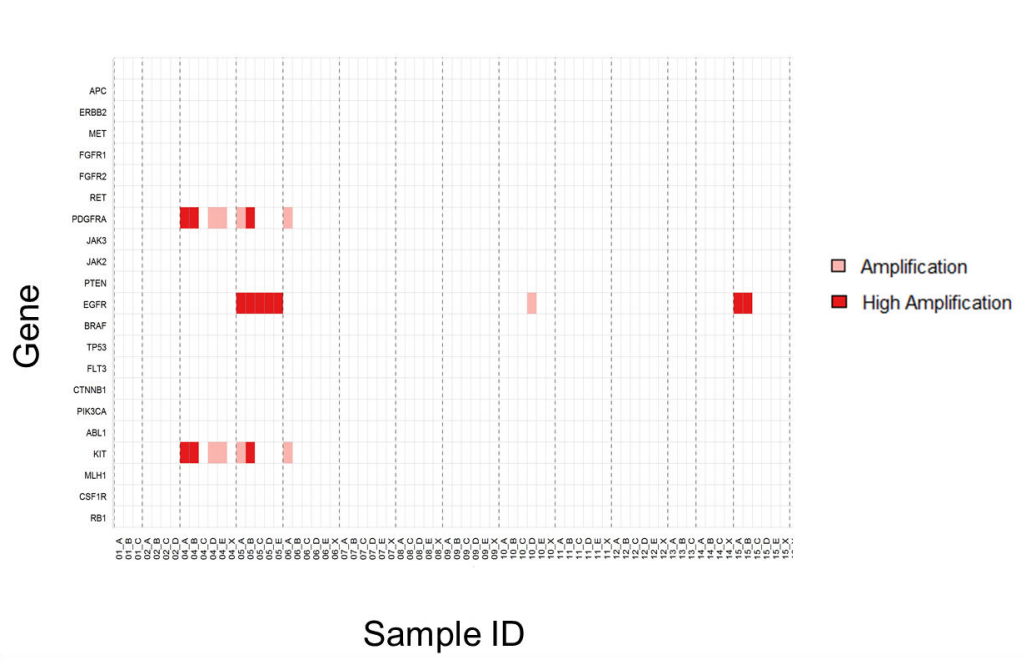
99-090	99-090J2	Normal kidney	Normal	Y	Y
99-091	99-091C	liver	LIVER	Y	Y
99-091	99-091G1	Normal kidney	Normal	Y	N
99-091	99-091I	LN	LN	Y	Y
99-091	99-091J	LN	LN	Y	Y
99-091	99-091K	LN	LN	Y	Y
99-091	99-091N	LN	LN	Y	Y

Supplementary Table B.1: Exome and aCGH status of all samples investigated in this study. We used two versions of Nimblegen EZ SeqCap capture probes in this study. Samples were first captured using V2 probes (targeting the 36.6 Mb RefSeq database), while patients sequenced in the latter half of the study were captured using the V3 probes (targeting a 50 Mb target containing exons and 3' UTR regions). Samples were sequenced on the Illumina HiSeq 2000 using either the PE-50 or PE-100 sequencing protocols.

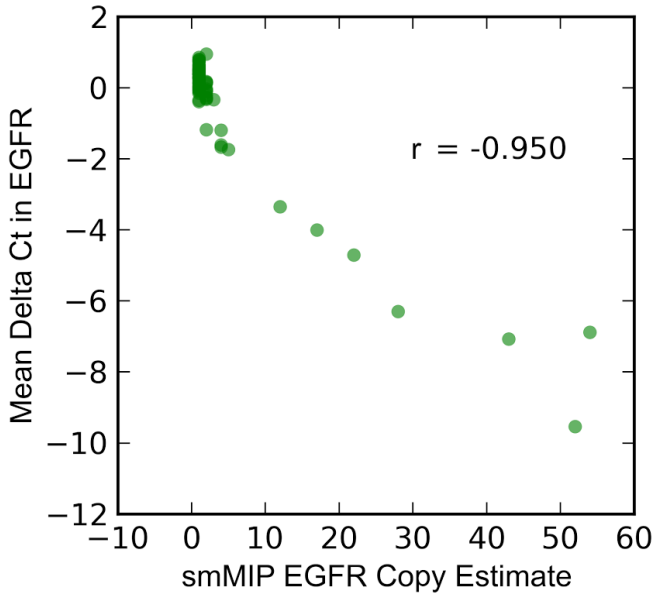
Appendix C- Supplementary Material for Chapter 4



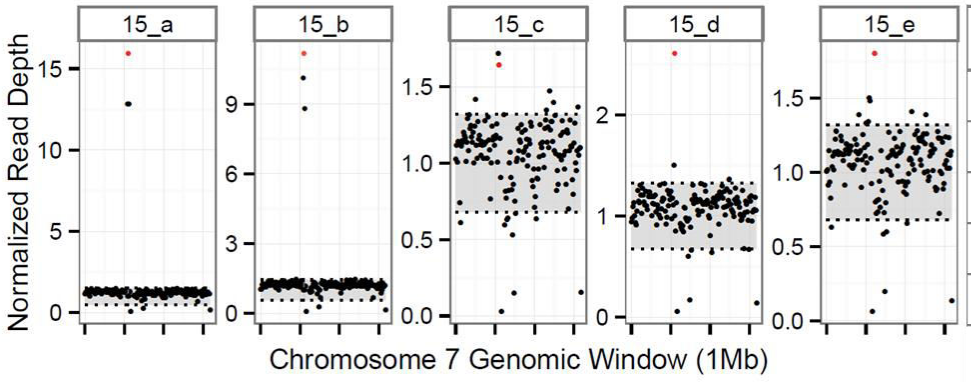
Supplementary Figure C.1: Overview of Molecular Inversion Probe Strategy (detailed). i) genomic DNA from each region is added to a tube containing a mixture of probes targeting the exons of 33 genes. For clarity, we focus on one targeted region. ii) A single molecule Molecular Inversion Probe (smMIP) consists of two regions complementary to a target of interest, a common backbone sequence and a 12bp molecular tag (used in error-correction). iii) After a polymerase gap-fill and ligation, each target sequence is captured to create a circular molecule of DNA. iv) Exonuclease digest removes remaining genomic DNA template and single stranded smMIP probes. v) After inverse PCR (with barcoding adaptors) against the common backbone, some targets are nonuniformly amplified. These instances are removed after tag-correction. Barcode sequences (not shown) allow capture products from multiple individual tumors or regions to be pooled on a single sequencing lane.



Supplementary Figure C.2: Copy number calls in replicate MIP captures. We used matched control tissue for samples BI04, BI06, BI07, BI08, BI09, BI10, BI11, BI12, BI14 and BI15, and a universal control (BI12) for samples BI01, BI02, BI05 and BI13 (as for the latter, matched control tissue was not available). Amplification indicates genes with coverage three-fold higher than median coverage across a sample. High Amplification indicates genes with coverage six-fold higher than median coverage across a sample. Notably, this analysis does not detect EGFR amplification in regions A, B and C of BI10 and region D of tumor BI15, events that were detected with the use of the universal control (BI12; see **Figure C.2** of main text). Careful review indicates this is likely due to increased tumor contamination within the respective control tissues (**Supplementary Table C.3**). For this reason, we chose to rely primarily on the results of analysis when all tumors were matched with BI12 as a universal control.

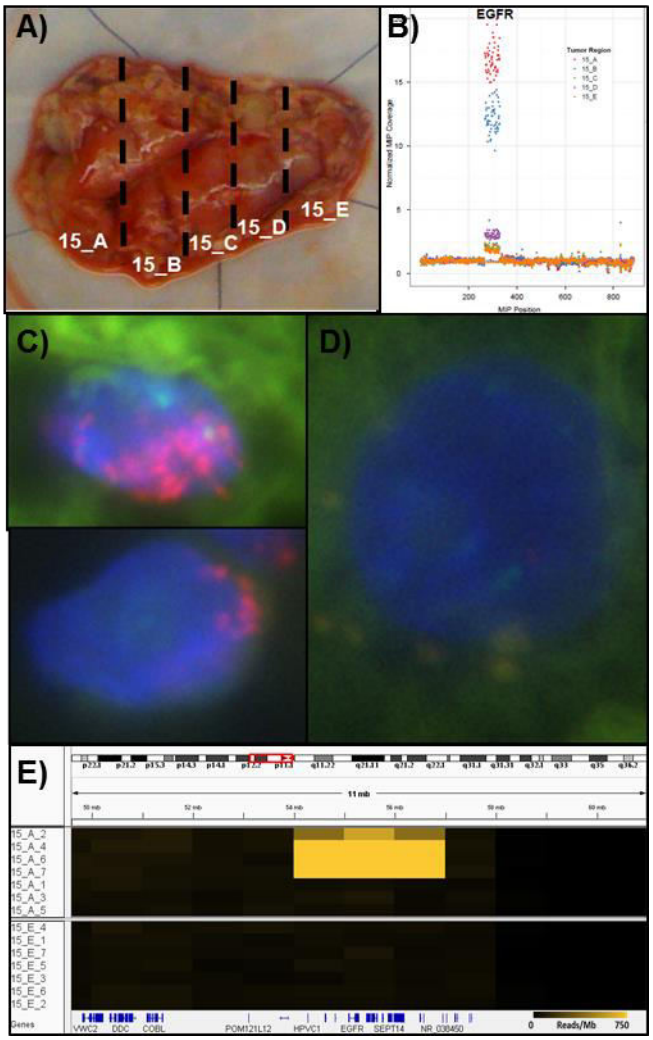


Supplementary Figure C.3: Validation of EGFR gene estimates. Correlation with of copy number estimates from smMIP vs. Taqman for EGFR**.** Taqman experiments were performed in duplicate for EGFR across all 62 regions investigated in this study. smMIP and Taqman copy number estimate were highly correlated with an R^2 of .90. Importantly, all high-level amplifications of EGFR (delta Ct ≤ -2) were identified by the smMIP assay.

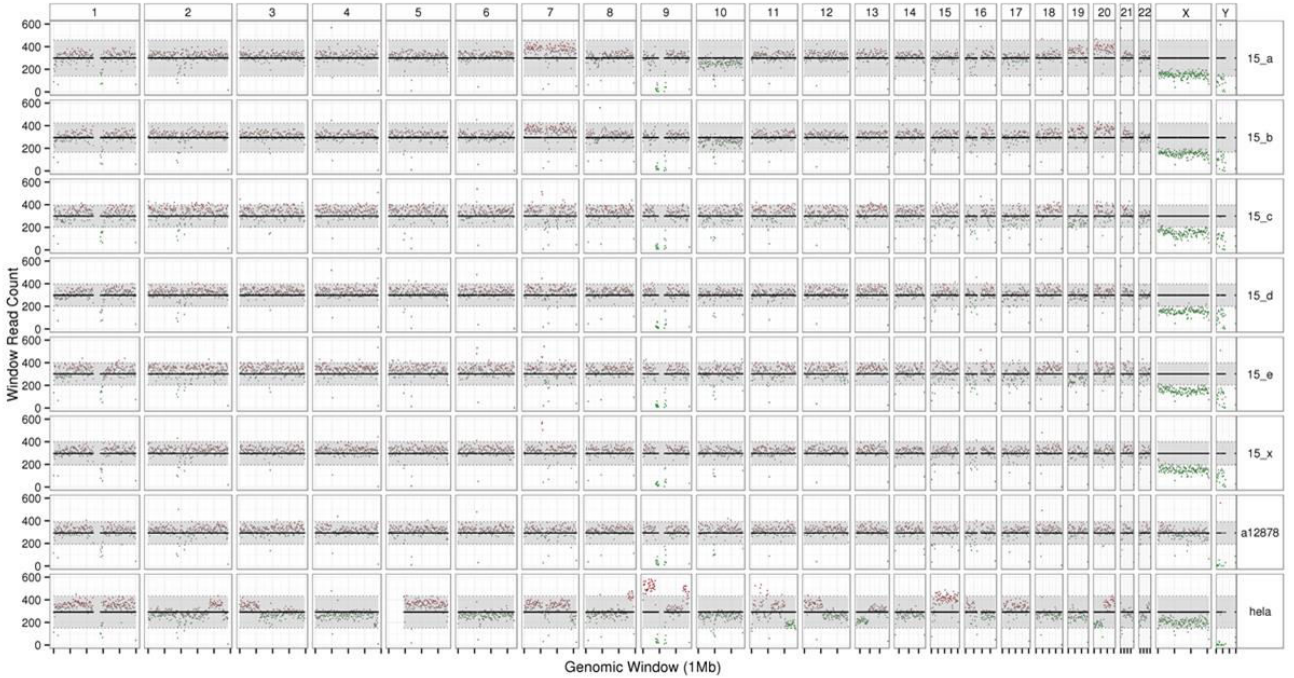


Tumor	MIP_CN	WG_CN
15_a	17	15.9
15_b	12	11.2
15_c	2	1.6
15_d	3	2.6
15_e	2	1.5

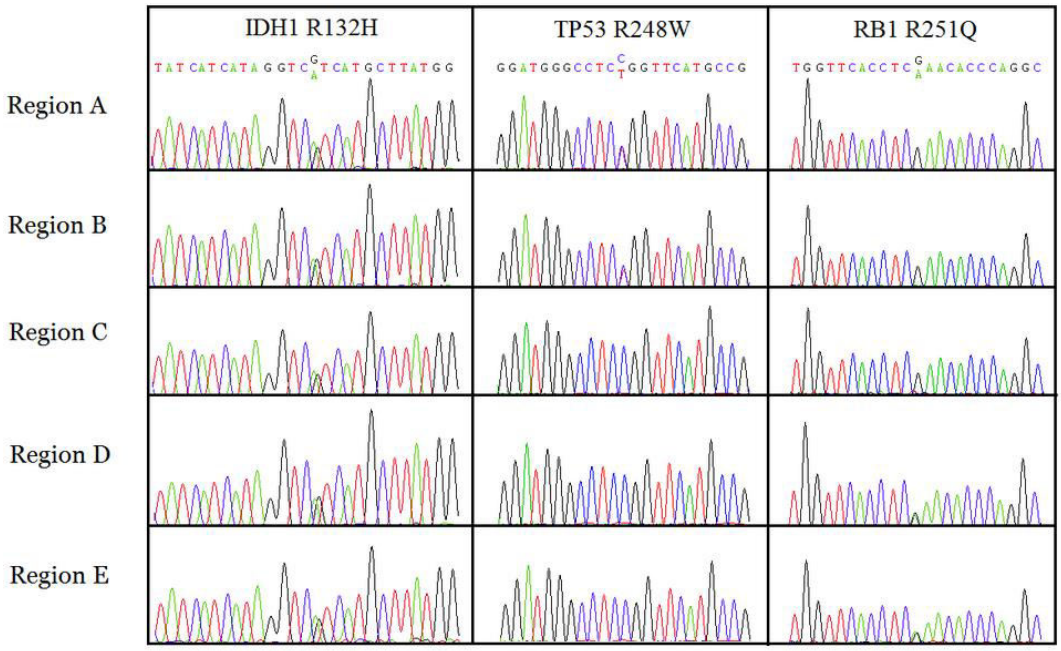
Supplementary Figure C.4: Validation of EGFR copy number by low-pass whole genome sequencing. DNA isolated from regions A-E in BI15 were subjected to light genome sequencing on the Illumina Miseq. Read depth within 1 Mb intervals across Chromosome 7 is normalized with respect to mean read depth across all chromosomes within each sample (see **Supplementary Methods**). Normalized read depth from whole genome sequencing within the 1 Mb region containing *EGFR* is highlighted in red within CN plots. Estimates of *EGFR* copy number (WG_CN) was compared with MIP copy number estimates (MIP_CN). Regions A and B contain high-level amplification in the region containing EGFR while a similar amplification is not seen within regions C, D and E.



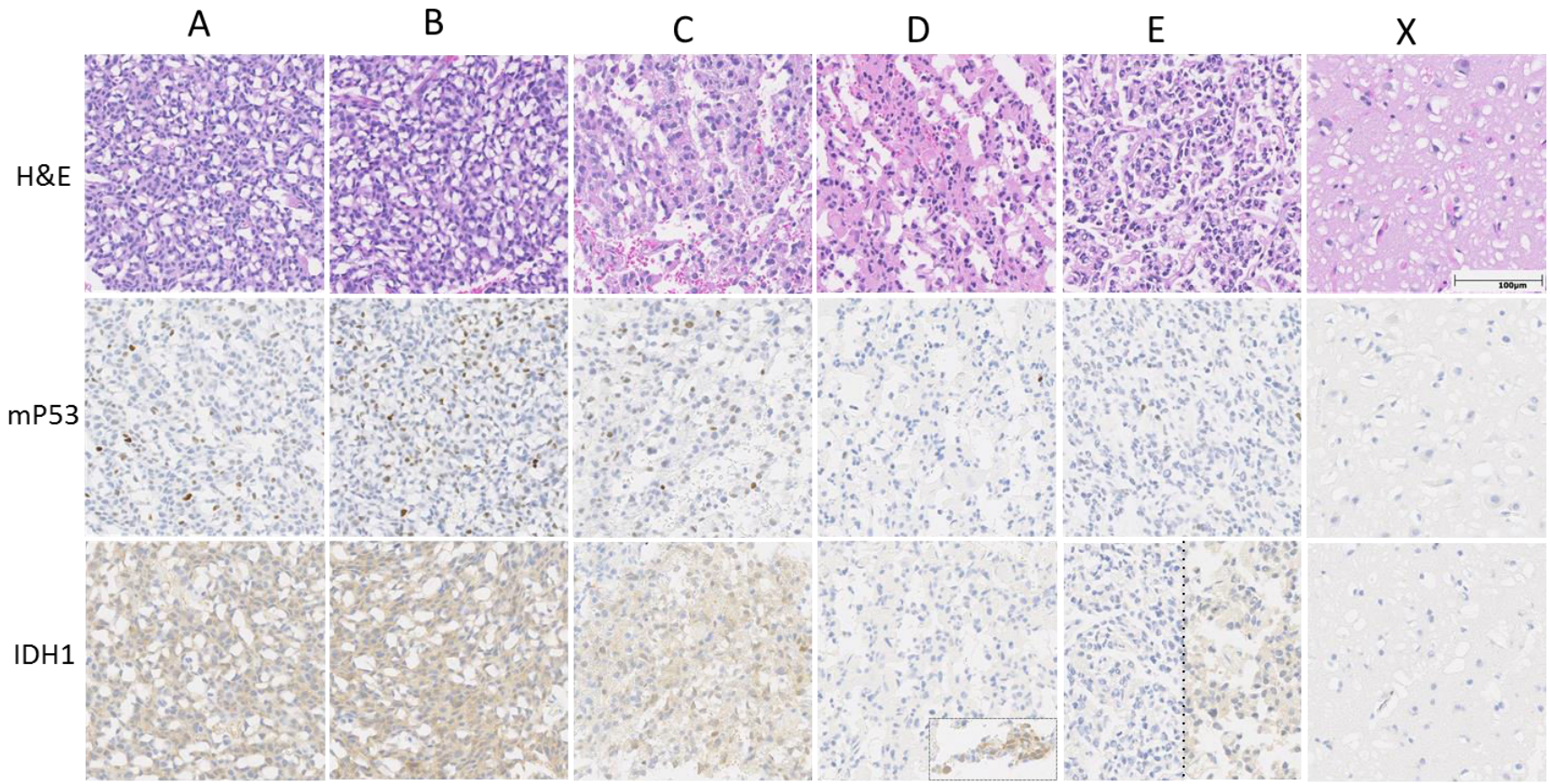
Supplementary Figure C.5: Measured *EGFR* amplification heterogeneity a result of varying levels of stromal contamination in BI15. a) GBM tumor used in dissection. **b)** Copy number estimates based on smMIP probe data. *EGFR* amplification (labeled) was called in regions A and B with only mild amplification detected in region C, D and E. Histologic examination and whole genome sequencing (**Supplementary Figure C.6**) suggested a marked decrease in tumor cellularity in regions C, D and E, likely accounting for the difference in copy number. **c)** and **d)** show representative FISH detection of *EGFR* amplification in region A (left images) and its absence in region E, respectively. Unprocessed images were obtained using a dual pass filter for spectrum orange and spectrum green and spectrum blue (DAPI)**. e)** Validation of *EGFR* amplification in region A using single cell sequencing. Single cells from regions A and E were flow sorted, amplified and sequenced on the Illumina Miseq, resulting in 100,000 reads per sample. Copy number profiles were created by plotting read depth across the genome in 1 Mb intervals, with color of each genomic region corresponding to the number of mapping reads per interval. Four of seven cells from region A (15_A_2, 15_A_4, 15_A_6 and 15_A_7) have high level *EGFR* amplification while zero of seven cells in region E have similar amplification.



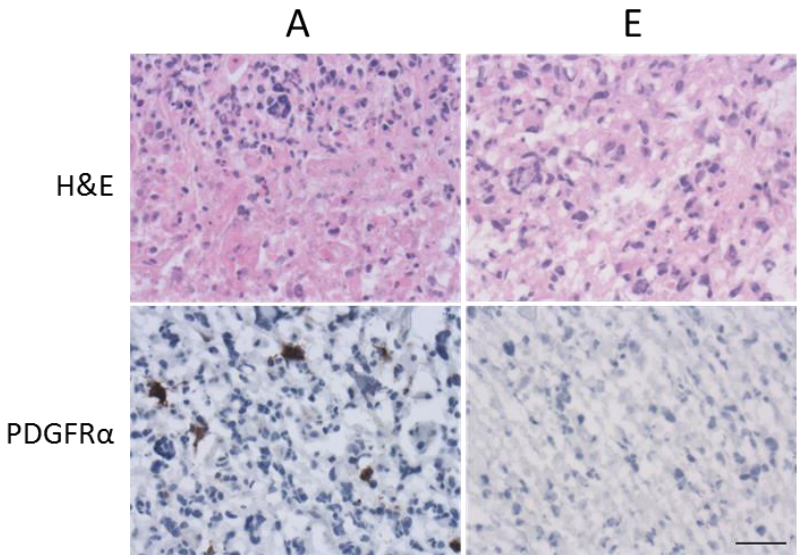
Supplementary Figure C.6: Whole genome copy number profiles of regions A-E in BI15. To identify other possible copy number alterations that may be shared across all tumor sections, DNA isolated from regions A-E in BI15 were subjected to light genome sequencing on the Illumina Miseq. 500,000 reads per sample were aligned to the hg19 reference and copy number is shown across the genome in 1 Mb intervals. Regions A and B of BI15 share a gain of chromosome 7 and loss of chromosome 10, but no gross chromosomal aberration was shared across all tumor regions. Black line corresponds to the mean coverage across all 1 Mb windows in autosomes. Shaded regions correspond to the region 1 S.D. below and above mean coverage for each sample. Two cell lines (12878 and HeLa) derived from female individuals are shown for comparison. Chromosome X appears as lost in all regions (including control) of the tumor as it was derived from a male patient.



Supplementary Figure C.7: Sanger validation of *TP53* and *RB1* heterogeneity in tumor BI09. Tumor regions A and B possess a mutation in *TP53*, while regions D and E possess a mutation in *RB1*. All five regions (A-E) share mutations in *IDH1*.



Supplementary Figure C.8: H&E and immunohistochemical (IHC) staining of p53 and IDH1 in tumor BI09. Differential staining of p53 and IDH1 in sections A-E and X is present and corroborates the intratumoral heterogeneity identified with sequencing. Partitioning of IDH1 photographs D and E illustrates that IDH1 heterogeneity was present within sections.



Supplementary Figure C.9: H&E and PDGFRα IHC staining of regions A and E in tumor BI05. IHC of regions A and E reveals differential staining of PDGFRα and corroborates the intratumoral heterogeneity identified from sequencing. Original magnification 40x. Scale bar indicates 30 microns. EGFR IHC revealed robust expression across all regions.

Sample Name	Tumor Type*	Grade*	Age at Surgery	IDH1 status	p53 status	1p19q status	Control Tissue	# Samples (Mutation)	# samples (CN)
BI_01	Ependymoma	III	33	mutant	NA	NA	N	3: A,B,C	3: A,B,C
BI_02	GBM	IV	29	mutant	mutant	NA	N	4: A,B,C,D	4: A,B,C,D
BI_04	GBM	IV	67	wt	NA	NA	Y	5: A,B,C,D,E	5: A,B,C,D,E
BI_05	GBM	IV	60	wt	NA	NA	N	5: A,B,C,D,E	5: A,B,C,D,E
BI_06	GBM	IV	63	mutant	NA	NA	Y	5: A,B,C,D,E	5: A,B,C,D,E
BI_07	GBM	IV	62	mutant	NA	NA	Y	5: A,B,C,D,E	5: A,B,C,D,E
BI_08	Astrocytoma	III	35	mutant	mutant	Not-del	Y	4: A,C,D,E	4: A,C,D,E
BI_09	AO	III	72	mutant	mutant	Co-del	Y	5: A,B,C,D,E	5: A,B,C,D,E
BI_10	GBM	IV	73	wt	NA	NA	Y	5: A,B,C,D,E	5: A,B,C,D,E
BI_11	GBM	IV	41	wt	NA	NA	Y	5: A,B,C,D,E	5: A,B,C,D,E
BI_12	GBM	IV	63	wt	NA	NA	Y	5: A,B,C,D,E	5: A,B,C,D,E
BI_13	GBM	IV	68	wt	NA	NA	N	3: A,B,C	3: A,B,C
BI_14	Astrocytoma	II	19	mutant	NA	NA	Y	3: A,B,C	3: A,B,C
BI_15	GBM	IV	57	wt	NA	NA	Y	5: A,B,C,D,E	5: A,B,C,D,E

Supplementary Table C.1: Tumors investigated in this study. Our sample set includes nine cases of GBM. GBM, Glioblastoma Multiforme; wt, wild-type; CN, copy number. IDH1 status was measured by IHC against IDH-R132H mutant. AO: Anaplastic Oligodendroglioma

*Diagnosis based on neuropathology report, which is based on the highest degree tumor available at the time

Gene	# of Targeted Coding Bases	Median % bases >30x coverage
ABL1	3613	96.5%
AKT1	1501	84.5%
AKT2	1561	96.7%
APC	8828	96.4%
BRAF	2389	92.3%
CDK4	947	98.3%
CDKN2A	1293	73.9%
CSF1R	3082	98.0%
CTNNB1	2386	97.6%
EGFR	4288	97.8%
ERBB2	3832	89.0%
FGFR1	2776	98.1%
FGFR2	2915	98.0%
FGFR3	2679	87.9%
FLT3	3075	95.5%
HRAS	722	97.2%
JAK2	3511	97.8%
JAK3	3421	92.9%
KIT	3042	98.3%
KRAS	739	99.1%
MET	4285	98.9%
MLH1	1963	97.9%
MYC	1403	97.6%
NRAS	595	98.8%
PDGFRA	3356	98.8%
PIK3CA	3267	96.5%
PTEN	1257	88.8%
RB1	2838	94.5%
RET	3599	95.4%
SRC	1666	91.7%
STK11	1377	89.9%
TP53	1448	94.6%
VHL	510	94.9%

Supplementary Table C.2: Genes Targeted by single molecule Molecular Inversion Probe (smMIP) Assay with Capture Efficiency. Our probes target the coding sequence of 33 cancer related genes. Here we list gene names as well as the number of coding bases targeted and capture efficiency of each gene. Median % bases >30x coverage represents the median percent of targeted coding bases that have greater than 30x coverage across all samples captured.

					Allele Balance within each sample						Coverage (tag-corrected)					
Tumor	**Chr**	**Position**	**Gene**	**Mutation type**	**X**	**A**	**B**	**C**	**D**	**E**	**X**	**A**	**B**	**C**	**D**	**E**
BI01	4	55593464	*KIT*	missense	NA	0.5093	0.5064	0.5039	NA	NA	NA	699	1637	772	NA	NA
BI01	17	7577121	*TP53*	missense	NA	0.9442	0.9688	0.9606	NA	NA	NA	1325	2661	635	NA	NA
BI02	4	55593464	*KIT*	missense	NA	0.5181	0.5226	0.5074	0.544	NA	NA	1355	597	808	897	NA
BI02	17	7577124	*TP53*	missense	NA	0.8952	0.8488	0.8918	0.9006	NA	NA	1680	820	915	1147	NA
BI04	17	7577535	*TP53*	missense	0.1792	0.7706	0.642	0.2302	0.5122	0.7401	1557	1308	1014	1690	1478	1443
BI05	13	28626716	*FLT3*	missense	NA	0.709	0.7306	0.9126	0.9792	0.8601	NA	409	657	206	192	243
BI05	17	37866342	*ERBB2*	missense	NA	0.1538	0.0805	0.0345	0	0.1481	NA	65	87	58	33	54
BI06	10	89711874	*PTEN*	splice	0	0.5143	0.7172	0.5348	0.3033	0.6885	125	175	396	230	244	183
BI07	10	89624273	*PTEN*	missense	0.1956	0.5303	0.3749	0.5211	0.4	0.6341	675	413	939	545	215	246
BI08	17	7577120	*TP53*	missense	0.4149	0.7721	NA	0.6099	0.73	0.814	1157	1009	NA	1033	3834	645
BI08	17	7578204	*TP53*	missense	0.0009	0	NA	0	0.1362	0.0747	1090	787	NA	867	3341	629
BI09	13	48936984	*RB1*	missense	0	0	0	0	0.4451	0.4333	763	369	363	387	1813	630
BI09	17	7577538	*TP53*	missense	0.0069	0.4975	0.4342	0.0045	0.0005	0.0007	1732	1598	1428	1539	4208	1507
BI10	17	7577022	*TP53*	stop-gained	0.0611	0.8765	0.8404	0.6253	0.645	0.9228	1440	502	1247	1073	1690	557
BI12	3	41266113	*CTNNB1*	missense	0	0.3684	0.3737	0.4512	0.4455	0.3493	568	1056	1132	1106	918	355
BI12	3	178921435	*PIK3CA*	missense	0	0.2032	0.142	0.2953	0.1928	0.2808	721	1629	1676	1497	1385	616
BI12	5	112102891	*APC*	missense	0	0.3954	0.3882	0.4678	0.4406	0.3416	448	951	930	902	799	363
BI12	5	112177421	*APC*	missense	0.0023	0.374	0.3708	0.4224	0.4614	0.3527	1318	3032	2964	2919	2484	1069
BI12	7	116340036	*MET*	missense	0	0.0726	0.0212	0.0637	0.1243	0.0226	629	1294	1322	1302	1038	443
BI12	9	5055669	*JAK2*	missense	0	0.3764	0.3005	0.3973	0.4524	0.366	858	2115	2020	1787	1786	817
BI12	10	89685290	*PTEN*	missense	0	0.0009	0.106	0.0363	0.001	0	1401	2124	2255	2206	1917	886
BI12	10	89692790	*PTEN*	missense	0	0.5278	0.4972	0.6395	0.5754	0.5084	668	1061	1090	1057	862	419
BI12	10	89711910	*PTEN*	stop-gained	0	0	0	0.0056	0.1242	0	320	446	755	536	451	182
BI12	12	25380240	*KRAS*	missense	0.0011	0	0.0026	0.0062	0.0974	0.0013	887	2119	1919	1940	1684	766
BI12	13	28588626	*FLT3*	missense	0.0028	0.3483	0.3383	0.6072	0.4048	0.2731	358	669	677	499	583	260
BI12	13	28610078	*FLT3*	missense	0	0.3484	0.3467	0.5534	0.4138	0.3245	327	620	672	506	580	265
BI12	13	28636062	*FLT3*	missense	0	0.4107	0.4206	0.6316	0.5	0.3111	63	112	126	114	90	45
BI12	13	49039143	*RB1*	missense	0	0.3867	0.3962	0.2616	0.4737	0.3333	87	150	212	172	152	66

BI12	17	7578199	*TP53*	missense	0	0.3466	0.3272	0.3974	0.4112	0.3009	452	929	865	780	766	319
BI12	17	7579355	*TP53*	missense	0	0.403	0.3103	0.4209	0.4333	0.2818	188	263	290	297	270	110
BI14	5	149449827	*CSF1R*	missense	0.4326	0.3538	0.4058	0.3844	NA	NA	675	277	855	588	NA	NA
BI14	17	7578263	*TP53*	stop-gained	0.0487	0.012	0.4054	0.3719	NA	NA	719	249	782	691	NA	NA
BI15	17	37866422	*ERBB2*	missense	0.506	0.5057	0.5036	0.4346	0.5371	0.484	747	350	417	260	391	281

Supplementary Table C.3: Protein-altering candidate somatic mutations. Allele balance of protein-altering candidate somatic mutations across all tumor regions are shown. Candidate mutations were not previously observed in a database derived from >5,000 exomes from the Exome Sequencing Project (ESP) that had been modified to remove positions also found in COSMIC.

Supplementary methods:

Single cell sequencing

Single nuclei copy number analysis was performed as previously described [117] with two modifications. Briefly, individual nuclei were either isolated from tumor regions A and E from BI15 by mincing tumor tissue in nuclei lysis buffer or isolated from a HapMap cell line GM12878 (Coriell) directly. Suspended nuclei were passed through a 0.2 um filter and sorted on an FACS Aria cell sorter. Sorted cells were placed into individual tubes, amplified using the PicoPLEX (Rubicon Genomics) single cell amplification kit and prepared for sequencing using the Nextera library preparation kit (Illumina). Libraries were sequenced on the Illumina Miseq using paired-end 100 bp reads. 100,000 reads from each single cell library were mapped to the human hg19 reference. Genomic copy number profiles were created by plotting the number of reads mapping across 1 Mb intervals across the reference genome. While the number of reads made identification of smaller amplifications/deletions difficult, cells with EGFR amplification also appeared to have a deletion of chromosome 10.

Whole genome sequencing

Light whole genome sequencing was performed on DNA isolated from multiple regions of BI15 as well as DNA extracted from the Coriell cell line 12878. Purified DNA was fragmented by sonication with the Covaris S2 instrument. Shotgun sequencing libraries were prepared using the KAPA library preparation kit (Kapa Biosystems) with sample barcoding following manufacturer's instructions. All libraries were sequenced on Miseq instruments (Illumina) using paired-end 100-bp reads. Copy number profiles were generated as described in "Single Cell Sequencing".

Appendix D- Inherited BTNL2 Variant in Aggressive Prostate Cancer.

Note: This chapter is based on the following published paper:

Liesel M. FitzGerald, Akash Kumar, Evan A. Boyle, Yuzheng Zhang, Laura M. McIntosh, Suzanne Kolb, Marni Stott-Miller, Tiffany Smith, Danielle Kayardi, Elaine A. Ostrander, Li Hsu, Jay Shendure, and Janet L. Stanford. Germline missense variants in the BTNL2 gene are associated with prostate cancer susceptibility. Cancer Epidemiol. Biomarkers Prev. 22:1520-8, 2013

Bold face indicates equal contributors.

This study was led by Liesel FitzGerald and Janet Stanford. I performed analysis of exome data in conjunction with investigators at the Center for Inherited Diseases Research (CIDR) at Johns Hopkins. With Evan Boyle, I designed, performed and analyzed MIP capture experiments in this study. Exome libraries were prepared and sequenced by CIDR. Taqman validation was performed by Elaine Ostrander, Tiffany Smith and Danielle Kayardi. Statistical interpretation was performed by Yuzheng Zhang and Li Hsu. Liesel FitzGerald and Janet Stanford wrote the majority of the manuscript, with my contribution being the methods and results sections for the exome and MIP analyses.

D.1- ABSTRACT

Background: Rare, inherited mutations account for 5%−10% of all prostate cancer (PCa) cases. However, to date, few causative mutations have been identified.

Methods: To identify rare mutations for PCa, we performed whole-exome sequencing (WES) of 91 subjects from 19 hereditary prostate cancer (HPC) families characterized by aggressive or early onset phenotypes. Candidate variants (*n* = 130) identified through family- and bioinformatics-based filtering of WES data were then genotyped in an independent set of 270 HPC families (*n* = 819 PCa cases; *n* = 496 unaffected relatives) for replication. Two variants with supportive evidence were subsequently genotyped in a population-based case-control study (*n* = 1,155 incident PCa cases; *n* = 1,060 age-matched controls) for further confirmation. All participants were men of European ancestry.

Results: The strongest evidence was for two germline missense variants in the *butyrophilin-like 2 (BTNL2)* gene (rs41441651, p.Asp336Asn and rs28362675, p.Gly454Cys) that segregated with affection status in two of the WES families. In the independent set of 270 HPC families, 1.5% (rs41441651; *P* = 0.0032) and 1.2% (rs28362675; *P* = 0.0070) of affected men, but no unaffected men, carried a variant. Both variants were associated with elevated PCa risk in the population-based study (rs41441651: OR = 2.7; 95% CI, 1.27−5.87; *P* = 0.010; rs28362675: OR = 2.5; 95% CI, 1.16−5.46; *P* = 0.019).

Conclusions: Results indicate that rare *BTNL2* variants play a role in susceptibility to both familial and sporadic prostate cancer.

Impact: Results implicate *BTNL2* as a novel PCa susceptibility gene.

D.2- INTRODUCTION

PCa is a complex and heterogeneous disease that has a strong genetic component to its etiology, with an estimated 42% of disease incidence attributed to heritable factors [118]. Genome-wide association studies of PCa have identified over 70 common low-penetrance single nucleotide polymorphisms (SNPs) that are confirmed to be associated with weak to modest alterations (average per allele ORs = 1.1−1.3) in disease risk [119,120], and which taken together may explain up to 30% of the genetic risk for PCa. In addition, genome-wide linkage studies of HPC families have searched for genomic regions that harbor rare, moderate- to high-penetrance mutations. These linkage studies have discovered more than two dozen putative susceptibility loci [121-123]; but only a few candidate genes underlying these loci have been proposed, and to date, even fewer rare, genetic mutations for PCa have been confirmed [124-128]. Recently, a targeted next-generation sequencing study of candidate genes across a linkage region on 17q21-22 identified a rare germline *HOXB13* mutation (G84E) in four HPC families of European descent [127]. Subsequent studies confirmed that the mutation (rs138213197) was carried by 2.4% of affected members from 1,892 independent HPC families tested [128], and was present in about 1% of PCa cases ascertained from the general population [129-131].

In order to find novel germline mutations for PCa, we completed one of the first WES studies of 19 HPC families in which multiple affected men per family with an aggressive or early onset phenotype were selected for sequencing. Candidate variants were then genotyped in an independent set of 270 HPC families and a population-based, case-control study for further confirmation.

D.3- METHODS

Study Populations

Participants selected for WES are members of 19 selected families chosen from a larger dataset of 289 HPC families of European ancestry [132]. Each of the 19 families has five or more affected men with at least three diagnosed with a more

aggressive phenotype and/or early onset prostate cancer based on the median age at diagnosis (i.e., 65 years) of cases from the 289 HPC families. From the 19 families two to six affected men (*n* = 80) and, where possible, one older, unaffected, PSA screened negative male relative (*n* = 11) were sequenced (**Table D.1**). The majority of affected men sequenced were diagnosed with more aggressive disease features (i.e., Gleason score 8−10 or regional/distant stage: *n* = 43 men) or at earlier ages (≤ 65 years: *n* = 55; mean age = 62 years), or both (*n* = 23). To decrease the likelihood of identifying false-positives due to inheritance identical-by-descent (IBD), the majority of affected men selected are 2nd- or 3rd-degree relatives. The eleven unaffected male relatives are older (mean age = 82 years) and thus are presumed less likely to develop HPC due to even moderately penetrant mutations. All 91 men sequenced were previously genotyped with the Illumina Linkage IVb panel [132].

The remaining independent set of 270 HPC families (described in (15)) was used to determine the frequency and distribution of candidate variants (*n* = 130) discovered in the 19 families in a larger representative group of HPC families. A total of 869 affected men and 519 unaffected male relatives with DNA samples are included in the confirmation genotyping effort.

The population-based, case-control study was used to estimate risk of PCa associated with genetic variants (*n* = 2) with supportive evidence from analyses of HPC families. Participants are from two population-based studies of PCa conducted in residents of King County, Washington [133,134]. For this genotyping effort, only men of European ancestry with DNA available are included (*n* = 1,155 incident cases; *n* = 1,060 age-matched controls).

This study was approved by the Fred Hutchinson Cancer Research Center's Institutional Review Board, and informed consent was obtained from all study participants. Genotyping of the case-control study samples was also approved by the Institutional Review Board of the National Human Genome Research Institute.

Whole-Exome Sequencing (WES) in 19 HPC Families

A total of 10μg of genomic DNA per subject was sent to the Center for Inherited Disease Research (CIDR) for sequencing. For quality control and inheritance checks, all samples were first run on the OmniExpress Array (Illumina, Inc.). Once initial quality control checks were completed, 3μg of DNA per subject was sheared, underwent library construction and was hybridized to the SureSelect Human All Exon 50Mb Array (Agilent). The captured library was PCR amplified, indexed and loaded on the HiSeq 2000 (Illumina, Inc.) for 75 bp paired-end sequencing.

WES Data Quality Control and Analyses

Sequencing reads were de-multiplexed at CIDR and fastq files were created for each sample. The Burrows-Wheeler Aligner (BWA) [135] was used to align reads to the hg19 reference genome, and GATK [136] was used for local realignment. Molecular duplicates were marked using Picard, and SAMtools [76] was used to sort, index and generate pileup files for variant calling. Sequencing coverage statistics, bases on target, transition/transversion ratios (Ti/Tv), variant/reference base ratios for heterozygous single nucleotide variants (SNVs), and concordance between the OmniExpress and sequencing data were calculated. Variant files containing SNVs and insertions or deletions (indels) were annotated using SeattleSeq and ANNOVAR, respectively, after filtering using the SAMtools.pl varFilter (all defaults except for minimum coverage of 8-fold and D=20,000) [16,137].

SNVs were filtered on a family-level basis using four different methods, some allowing for incomplete penetrance and phenocopies. In all instances, the minor allele frequency (MAF) of SNVs was determined using a subset of exomes sequenced as part of the NHLBI/NIH Exome Sequencing Project (ESP). The four filtering approaches were as follows: 1) MAF <0.02, present in all affecteds, not present in the unaffected male relative if available, and not present in any other unaffected males; 2) MAF <0.02, present in all but one of the affecteds, not present in the unaffected male relative if available, and not present in any other unaffected

males; 3) MAF <0.01, present in all affecteds, present in the unaffected male relative if available, and not present in any other unaffected males; and, 4) MAF <0.01, present in all but one of the affecteds, present in the unaffected male relative if available, and not present in any other unaffected males. Filtering methods 2 and 4 allowed for phenocopies, and methods 3 and 4 allowed for incomplete penetrance. These filters highlighted 1,459 SNVs that were further prioritized according to the following information: type (nonsense < splice site < missense); prediction scores based on the evolutionary conservation of the reference base and the impact the variant would have on the resulting amino acid change using GERP (≥ 5) [138], PolyPhen (probably damaging) [139], SIFT (0.00−0.10) [140], Grantham (≥ 151) [141], PhyloP (≥ 3) [138], likelihood ratio test (damaging) [142], and BLOSUM62 (−2 to −4) [143]; gene information contained in the UCSC Genome Browser [144] and PubMed; and, presence within a previously identified linkage region (i.e., a dominant or recessive LOD ≥ 1.86) from an earlier genome-wide linkage scan [132].

Indels were also filtered on a family-level basis, but different methods were used to prioritize candidates. The four filters applied to the data allowed for none, one, two, or three phenocopies, respectively. Indels were removed if they were observed in any of the unaffected men. The 2,510 filtered indels (Supplementary Figure D.1) were then prioritized according to the following information: type (frameshift < UTR < non-frameshift); location (exonic < intronic); number of families in which the indel was observed, with greater weight placed on those that were seen in fewer families; gene information contained in the UCSC Genome Browser and PubMed; and, presence within a previously identified linkage region (i.e., a dominant or recessive LOD ≥ 1.86) from an earlier genome-wide linkage scan [132].

Genotyping of Candidate Variants in HPC Families

The molecular inversion probe (MIP) assay [145] was used to genotype candidate variants identified from the WES analyses. The protocol used was similar to that of O'Roak et al. [146]. Briefly, 70 bp oligonucleotide inversion probes (Integrated

DNA Technologies, Inc.) were designed against 196 candidate SNVs and indels. These oligonucleotides were 5' phosphorylated and added to ~200 ng of germline DNA at a ratio of 200:1 MIPs to template. The probe/DNA mixture was incubated with ligase, polymerase and nucleotides for 48 hours, resulting in targeted regions being "captured" within single-stranded circular DNA. After exonuclease removal of non-circularized DNA, captured products were amplified using PCR with barcoded primers containing adaptor sequences. The amplified products were pooled and sequenced on the HiSeq Illumina platform. Sequencing data were aligned to the human hg19 reference genome using BWA [135], and variant calls were made using SAMtools [76]. A position was considered to possess a variant if it was covered to a minimum 8x depth and had at least 20% of reads supporting the variant allele.

Genotyping Data Quality Control

To identify variants with a high probability of being artifacts, a comparison of the MIP and WES data was undertaken in members of the 19 sequencing families with both types of data. A second method of identifying probable artifacts looked for significant differences in variant allele frequencies in the 5,379 ESP exomes from individuals of European ancestry ($P < 0.05$ based on a binomial distribution). From the total of 196 candidate SNVs ($n = 174$) and indels ($n = 22$) selected for follow-up genotyping, 66 were excluded for the following reasons: MIP design failure ($n = 6$); low call rate within genotyped subjects (> 15% missing, $n = 55$); and, probable artifacts discovered through a comparison of MIP and WES data (< 95% concordance, $n = 5$).

Genotyping Data Analyses in HPC Families

The PedGenie program [147] was used to assess the association of 130 candidate variants with affection status in the 270 independent HPC families. This program can handle pedigrees of arbitrary size and structure and provides valid statistical inference by gene dropping to generate the null distribution. Each SNV or indel was coded as a binary variable with 0 and 1 indicating the absence/presence of

the candidate variant, respectively. Statistical significance was determined by the Monte Carlo approach to account for potential correlation of genotypes within a family and rarity of the SNVs and indels. A total of 100,000 simulated datasets were generated to form the null distribution. A one-sided *P*-value was calculated by dividing the χ^2 *P*-value by two for candidate variants that were observed more frequently in men with, compared to without, PCa, and one minus the χ^2 *P*-value divided by two for candidate variants that were observed less frequently in affected men than in unaffected men. A *P*-value < 0.05 was considered statistically significant in testing for confirmation of candidate variants.

Genotyping of *BTNL2* Candidate Variants in Case-Control Samples

A custom designed TaqMan SNP genotyping assay (Applied Biosystems, Foster City, CA) was used to genotype the two *BTNL2* candidate variants (rs41441651 and rs28362675) on the ABIPrism 7900HT sequence detection system according to the manufacturer's instructions.

Genotyping Data Analysis in the Case-Control Study

Unconditional logistic regression was used to estimate the odds ratio (OR) and 95% confidence interval (CI) as a measure of association between the two *BTNL2* candidate variants and PCa [148], as implemented in STATA version 11.0 (Stata Corp). Potential confounding factors, including age at reference date, PCa screening history, and first-degree family history of PCa, were examined to see if such factors changed the risk estimates by 10% or more. After these analyses, only age at reference date was included in the final models. Regression models were also used to generate ORs and 95% CIs for the association between SNV genotypes in men stratified by family history (yes vs. no). A product term between SNV genotypes and family history was included in logistic regression models, and a log-likelihood ratio test was used to compare logistic models with and without the product term to test whether the effects of SNV genotypes differ by family history.

D.4- RESULTS

WES data were available for 91 men from 19 HPC families (**Table D.1**). In the 89 individuals for whom WES data passed quality control, an average of 70x (range: 20x to 132x) coverage of the target was achieved (Agilent SureSelect Human All Exon 50 Mb, Illumina HiSeq 2 x 75 bp) with ~88% of target bases having at least 8x coverage. Concordance between genotyping data from the OmniExpress array and WES data was 99.9%. Family- and bioinformatics-based filtering of the WES data prioritized 174 SNVs and 22 indels as candidate variants for follow-up genotyping in 270 independent HPC families. After quality control (see Materials and Methods), data for 130 of these candidates remained for analysis.

For the 130 candidate variants, the average concordance between 15 blind duplicates was 99.5% (non-reference concordance was 82.5%), and 99.9% (non-reference concordance was 99.3%) for individuals who had both WES and MIP data available. A total of 1,388 men from the 270 HPC confirmation families were genotyped; 73 individuals who were missing > 15% of the 130 MIP genotypes were excluded, leaving 819 affected and 496 unaffected men in the analysis.

Family-based association analysis of the 270 independent HPC families provided evidence (i.e., higher MAF in affected vs. unaffected men, Monte-Carlo based one-tailed $P < 0.05$) for two rare variants in *BTNL2* (**Table D.2**). These missense variants, rs41441651 (D336N) and rs28362675 (G454C), are not present in HapMap and therefore a formal test for linkage disequilibrium (LD) was not possible, however they are located within a single haplotype block of eight Kb. This and the fact that all but two of the 22 affected carriers with data were concordant for both variants (two men had poor coverage for rs28362675 and thus have unknown carrier status) suggest that these variants are in strong LD. The smallest *P*-value observed was for rs41441651 ($P = 0.0032$).

The rs41441651 variant segregated with affection status in two of the 19 WES families (Figure D.1) and was present in 10 of 12 genotyped affected men.

(Sanger sequencing was used to confirm the carrier status of the female in Family 11). The affected carriers had an average age at diagnosis of 62.6 years, and 60% had regional stage PCa at diagnosis. Gleason scores ranged from 5−9 (average 6.5). Of the unaffected genotyped men in these two families, none carried either candidate variant.

In the 270 independent HPC families evaluated, 3.3% and 2.9% had one or more affected members who carried the rs41441651 or rs28362675 variant, respectively. In total 12 (1.47%) of 819 genotyped affected men, but none of the 496 genotyped unaffected men carried the rs41441651 candidate variant. However, in these families, which contained fewer affected men and were of smaller size than the 19 families selected for WES, there was not clear evidence of co-segregation with disease state.

The two *BTNL2* candidate variants genotyped in the case-control dataset had distributions among controls that were consistent with Hardy-Weinberg equilibrium ($P > 0.05$). For rs41441651, 26 (2.3 %) cases and 9 (0.9%) controls carried the missense variant; for rs28362675, 24 (2.1 %) cases and 9 (0.9%) controls were carriers (**Table D.3**). Both candidate variants were associated with statistically significant increases in the risk of PCa (rs41441651: OR = 2.7; 95% CI, 1.27−5.87; $P = 0.010$; rs28362675: OR = 2.5; 95% CI, 1.16−5.46; $P = 0.019$). These risk estimates did not differ by family history of PCa, but this subgroup analysis had limited power. The mean age at PCa diagnosis was 59.0 years for carriers of one or both variants and was 59.8 years for non-carriers ($P = 0.6$).

D.5- DISCUSSION

WES of 91 men in 19 HPC families, followed by replication (*n* = 130 candidate variants) in an independent set of 270 HPC families and further testing of candidate variants with replication support (*n* = 2) in a population-based case-control study, provides compelling evidence that rare germline variants in *butyrophilin-like 2* (*BTNL2*) are associated with genetic susceptibility to PCa. These rare missense variants, rs41441651 (exon 5; D336N) and rs28362675 (exon 6; G454C), occur in the same haplotype block on chromosome 6p21.32, and were observed to be in strong LD among controls in the case-control dataset (r^2 = 0.99). This is the first WES study focused on aggressive or early onset PCa phenotypes and the first to implicate rare germline *BTNL2* variants as predisposing to familial and sporadic PCa.

There is some prior suggestive evidence for a role of *BTNL2* in PCa. A recent exome sequencing study of prostate tumor tissue from 50 patients with lethal PCa identified one patient who had a somatic *BTNL2* mutation (c.709+9T>G)[86]. Also, Acevedo et al.[149] found that BTNL2 protein was significantly over-expressed in advanced PCa tumor tissue relative to normal prostate tissue in a mouse model of the disease[86]. In comparing the protein-encoding transcriptomes of 79 different normal human tissues, Su and colleagues [150] found that BTNL2 mRNA expression in the prostate was above the median expression level observed in other tissues.

Butyrophilin-like (BTNL) molecules are thought to play a role in immune regulation and have been functionally implicated in T cell inhibition and modulation of epithelial cell-T cell interactions[151]. *In vitro* mouse studies indicate that the BTNL2 protein is a negative regulator of T cell proliferation and cytokine production[152,153].

Genetic polymorphisms in *BTNL2* have been associated with several immunological diseases. Studies have reported significant associations between *BTNL2* SNPs and the inflammatory autoimmune diseases sarcoidosis[154,155] and rheumatoid arthritis[156], as well as inflammatory bowel disease and ulcerative colitis [157,158]. None of the affected carriers of *BTNL2* variants in our study population

reported a history of these inflammatory conditions. In addition, no participants reported a family history of sarcoidosis or any other autoimmune diseases among close family members. Interestingly, one of the SNPs previously associated with ulcerative colitis, rs9268480[157], is located only 44 bp from the *BTNL2* rs41441651, and is within the same haplotype block. Given the biological activity of BTNL2, our results provide further support for the role of the inflammation pathway in the development of PCa[159,160].

Four of seven functional and conservation prediction scores suggest that rs28362675 is damaging, although current evidence is equivocal for rs41441651; both variants change the encoded amino acid. The rs41441651 appears to be present within a cluster of CpG dinucleotides that are either heavily methylated or unmethylated according to the cell line assayed[161]. This SNV may therefore disrupt methylation at this site. It is possible, however, that the *BTNL2* variants we describe here are not causative, rather they are in LD with a yet undiscovered functional variant. This is a formal consideration as the haplotype block in which they are located extends into the 3' regulatory region of the gene, which had limited sequencing coverage.

Among the potential HPC variants highlighted by this study, the *BTNL2* variants were notable in that they were observed only in affected men in both the WES families and the 270 HPC family replication dataset. The variants segregated with disease in two of the WES families, and although this was not the case in the 270 HPC family replication dataset, about 3% of the latter families had affected carriers of one or both variants. These variants were also observed in 2.3% (rs41441651) and 2.1% (rs28362675) of sporadic PCa cases and 0.9% of the population-based controls. These observations are similar to those seen for the *HOXB13* mutation, rs138213197[127]; two unaffected males were observed to carry the *HOXB13* mutation in the four HPC discovery families and a carrier frequency of 0.1% was observed in 1,401 controls. Further, studies of both familial and case-control

datasets have indicated that while rs138213197 is significantly associated with PCa risk, it rarely segregates perfectly with disease in HPC families and it is seen at a low frequency in controls[128-131,162].

Allele frequency data for the two *BTNL2* variants are available from several recent sequencing efforts. In the large NHLBI GO Exome Sequencing Project[6749] consisting of mixed race U.S.-based studies, the MAF for both variants is reported as 0.5% (chromosomes = 4542−4550). In the ClinSeq Project[163] consisting of individuals of European ancestry, the minor allele frequency (MAF) for both variants is 1.5% (chromosomes = 1310−1323). Finally, in a pilot 1000 Genomes Project[164] population of Chinese and Japanese, the MAF for both variants is 15% (chromosomes = 120). From previous analyses of linkage data, we confirmed that the 289 HPC families in this study are of European and not Asian or African descent. Therefore, the ClinSeq population is most representative of our HPC and case-control study populations, and the average MAF of the case (2.2%) and control (0.9%) samples from our population-based dataset is similar to that of the ClinSeq study. PCa is a prevalent disease and the PCa status of the ClinSeq male population is not publicly available, so it is possible that the MAF in the ClinSeq data is inflated due to the inclusion of affected men. Regardless, the MAF in our cases is higher than that in the population controls and the ClinSeq population.

There were a number of SNVs and indels highlighted in the 19 WES HPC families that were only observed once or not at all in the 270 replication HPC families. Due to the rarity of these potential variants, additional follow-up in a larger set of HPC families will be needed to confirm these associations and determine the proportion of HPC that may be attributable to these other rare variants. In addition, the 66 candidate SNVs and indels that were unable to be evaluated in the 270 independent HPC families require further study.

Identifying HPC mutations has been challenging due to the genetic heterogeneity of the disease and the phenotypic complexity of PCa. This study is the

first to demonstrate the value of WES in large multiplex HPC families characterized by aggressive or early onset PCa, with replication in an independent HPC family dataset and a population-based case-control dataset. We identified two rare *BTNL2* variants that segregate with disease in two HPC families with sequencing data and that are carried only by affected men, but no unaffected men, in eight (rs28362675) and nine (rs41441651) of the additional 270 HPC families tested. We also found that these two variants are associated with statistically significant 2.5- to 2.7-fold elevations in the relative risk of PCa in the general population, with slightly over 2% of incident sporadic PCa cases carrying at least one of these variants. Larger studies of densely affected HPC families (≥ 5 affected men) and case-control datasets are now needed to establish the significance of these novel *BTNL2* missense variants in further defining PCa genetic susceptibility.

D.6- TABLES

Family ID	No. PCa cases	Mean age at PCa diagnosis	No. WES cases with aggressive PCa[a]	No. WES cases with early-onset PCa[a]	No. WES cases per family	No. WES unaffected men per family
1	9	64.4	2	3	4	1
2	7	62.2	2	3	4	
3	7	69.9	2	2	5	1
4	9	67.3	4	2	5	1
5	8	68.0	3	1	4	1
6	6	60.6	1	5	5	
7	11	64.6	2	2	3	1
8	5	54.0	2	3	3[b]	1
9	7	57.2	3	2	3	
10	6	66.0	3	2	4	
11	7	59.0	5	3	5[b]	
12	9	68.4	2	1	4	1
13	9	60.0	1	5	5	
14	10	65.1	2	6	6	1
15	7	61.9	3	3	4	
16	8	63.4	0	4	4	
17	9	66.2	1	2	2	1
18	10	65.6	2	3	5	1
19	7	67.8	3	3	5	1
Total	151	63.8	43	55	80	11

Table D.1: Characteristics of 19 hereditary prostate cancer (PCa) families with whole-exome sequencing (WES) data.

[a] A total of 23 cases had both aggressive and early-onset PCa.

[b] WES failed or was of low quality for one of the affected men in these families.

						Discovery	Validation			
Gene	Genomic Position (hg19)	Variant and rs ID	Protein	MAF in ESP	MAF in ClinSeq	No. of WES families with carriers (Aff/Unaff)[b]	No. (%) of 270 families with affected carriers	MAF in 819 genotyped affected men[c]	MAF in 496 genotyped unaffected men[c]	*P*-value[d]
BTNL2	Chr6: 32,363,888	C>T rs41441651	Missense: p.Asp336Asn	0.009	0.005	2 (10/0)	9 (3.33)	0.0073	0	0.0032
BTNL2	Chr6: 32,362,521	C>A rs28362675	Missense: p.Gly454Cys	0.008	0.005	2 (10/0)	8 (2.96)	0.0061	0	0.0070

[a]Top ranked ($P < 0.05$) single nucleotide variants identified by whole-exome sequencing (WES) of 19 HPC families.
[b]Aff = number of affected carriers / Unaff = number of unaffected carriers.
[c]The number of affected carriers for rs41441651 and rs28362675 is 12 and 10, respectively, in the 270 independent HPC families.
[d]Monte-Carlo based one-sided *P*-value from the PedGenie chi-square test for association based on the 270 independent HPC families.

Table D.2: Results for single nucleotide variants[a] identified by whole-exome sequencing and genotyped in 270 independent HPC families of European ancestry.

Genotype	Cases (*n* = 1,155) n	%	Controls (*n* = 1,060) n	%	OR[a]	95% CI	*P*-value
rs41441651							
CC	1,129	(97.8)	1,051	(99.2)	1.00	−	
CT or TT [b]	26	(2.3)	9	(0.9)	2.73	(1.27−5.87)	0.010
rs28362675							
CC	1,131	(97.9)	1,051	(99.2)	1.00	−	
CA or AA[b]	24	(2.1)	9	(0.9)	2.52	(1.16−5.46)	0.019

[a] Adjusted for age.
[b] One case is homozygous variant for both SNVs; 22 cases and 9 controls are heterozygous for both SNVs.

Table D.3: Odds ratios (OR) and 95% confidence intervals (CI) for prostate cancer associated with single nucleotide variants in *BTNL2* in European Americans.

D.7- FIGURES

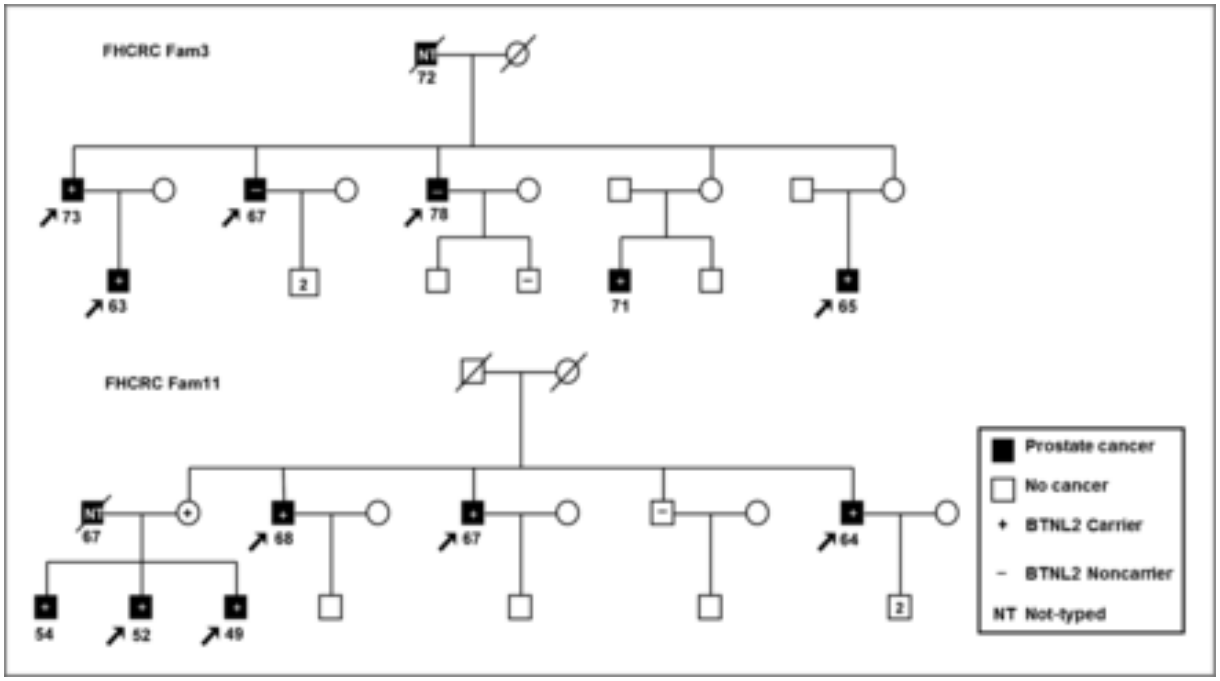


Figure D.1. HPC pedigrees of two families with segregating *BTNL2* variants. Participants selected for whole-exome sequencing in 19 HPC families are indicated by arrows. Affected men indicated by black shading. The remaining symbols are described in the key. Squares indicate males, and circles indicate females. The age at diagnosis of men with prostate cancer is shown under the squares. A slash through the symbol indicates that the individual is deceased. The carrier status of the female in Family 11 was confirmed by Sanger sequencing.

Appendix E- Genome Sequencing of Idiopathic Pulmonary Fibrosis in Conjunction with a Medical School Human Anatomy Course

Note: This chapter is based on a recently submitted manuscript:

Akash Kumar, Max Dougherty, Gregory M. Findlay, Madeleine Geisheker, Jason Klein, John Lazar, Heather Machkovech, Jesse Resnick, Rebecca Resnick, Alexander I. Salter, Faezeh Talebi-Liasi, Christopher Arakawa, Jacob Baudin, Andrew Bogaard, Rebecca Salesky, Qian Zhou, Kelly Smith, John I. Clark, Jay Shendure, Marshall S. Horwitz. Genome Sequencing of Idiopathic Pulmonary Fibrosis in Conjunction with a Medical School Human Anatomy Course.

This manuscript arose out of a course in which students participated in whole genome sequencing of a cadaver used in the gross anatomy dissection lab. The approach and idea for the class came from Marshall Horwitz, Jay Shendure and John Clark. As the teaching assistant for the course, I performed and/or assisted with all experiments and analysis in conjunction with 15 students taking the class. Marshall Horwitz and I supervised the team-writing of the manuscript with contributions made by students and instructors of the class.

E.1- ABSTRACT

Even in cases where there is no obvious family history of disease, genome sequencing may contribute to clinical diagnosis and management. Clinical application of the genome has not yet become routine, however, in part because physicians are still learning how best to utilize such information. As an educational research exercise performed in conjunction with our medical school human anatomy course, we explored the potential utility of determining the whole genome sequence of a patient who had died following a clinical diagnosis of idiopathic pulmonary fibrosis (IPF). Medical students performed dissection and whole genome sequencing of the cadaver. Gross and microscopic findings were more consistent with the fibrosing variant of nonspecific interstitial pneumonia (NSIP), as opposed to IPF *per se*. Variants in genes causing Mendelian disorders predisposing to IPF were not detected. However, whole genome sequencing identified several common variants associated with IPF, including a single nucleotide polymorphism (SNP), rs35705950, located in the promoter region of the gene encoding mucin glycoprotein MUC5B. The *MUC5B* promoter polymorphism was recently found to markedly elevate risk for IPF, though a particular association with NSIP has not been previously reported, nor has its contribution to disease risk previously been evaluated in the genome-wide context of all genetic variants. We did not identify additional predicted functional variants in a region of linkage disequilibrium (LD) adjacent to *MUC5B*, nor did we discover other likely risk-contributing variants elsewhere in the genome. Whole genome sequencing thus corroborates the association of rs35705950 with *MUC5B* dysregulation and IPF. This novel exercise additionally served a unique mission in bridging clinical and basic science education.

E.2- INTRODUCTION

Recently, several studies have supported the value of clinical genome sequencing, particularly when there is diagnostic uncertainty[190-193]. As clinical genome sequencing becomes more widely available, it is likely to provide useful information, even when there is no family history of disease. However, interpretation of genomic data has not yet been widely incorporated into medical school curricula, and how such studies can best be employed to inform medical practice remains a subject of intense interest [194].

To address emerging implications for applying clinical genomics, students earning concurrent M.D. and Ph.D. degrees in the University of Washington Medical Scientist Training Program (MSTP) participate in a new course, also open to a limited number of M.D.-only students, in which a cadaver undergoes whole genome sequencing in association with dissection in the human anatomy lab. The cadaver selected for the inaugural exercise had been an otherwise healthy male with non-familial idiopathic pulmonary fibrosis (IPF).

Interstitial lung diseases can be difficult to clinically and pathologically characterize[195]. Many cases are diagnosed as idiopathic, highlighting a need to develop better understanding of their pathogenesis. Current evidence suggests that IPF follows a "two-hit" disease model where hereditary factors alter underlying disease susceptibility to environmental stressors such as cigarette smoke, asbestos, or silica[196,197]. Familial IPF kindred and genome-wide association studies (GWAS) have identified several genetic variants that alter disease susceptibility[198]. Interestingly, many of the variants described in the literature are known to be involved in host defense, cell adhesion, maintenance of genomic integrity, and preservation of lung architecture. To cite one prominent example, a single nucleotide polymorphism (SNP) in the promoter of the

MUC5B gene was recently reported to up-regulate expression of the gene and is both genetically linked and associated with IPF[199-201].

As an exercise in medical education, we interpret the cadaver's genome sequence in concert with gross and microscopic anatomical examination and discuss its potential relevance for diagnosis and management of IPF.

E.3- METHODS

Ethics. Written informed consent for body donation for research and education was obtained through the University of Washington School of Medicine Willed Body Program. All research was conducting according to Declaration of Helsinki principles.

Genome sequencing. Tissue samples were obtained postmortem from the unembalmed cadaver. A cube of liver tissue approximately 1 cm on a side was dissected from the liver and frozen at -80˚C to be used in the preparation of DNA for sequencing. 17 µg of genomic DNA was extracted using Qiagen DNeasy Blood & Tissue Kit from this sample. Shotgun sequencing libraries were prepared using the KAPA library preparation kit (Kapa Biosystems) following manufacturer's instructions. DNA was sequenced on an Illumina HiSeq 2000 with paired-end 100 bp reads. The raw sequence data was mapped to hg19/GRCh37 and variants were called with GATK using best practices[13,135,202]. Variants were annotated using SeattleSeq[16]. Allele frequencies for the European-American population were derived from the University of Washington Genome Variation Server (http://gvs.gs.washington.edu). Findings discussed in the manuscript were manually evaluated by verifying read data using the Integrative Genomics Viewer (IGV, http://www.broadinstitute.org/igv/).

Analysis of variants in the vicinity of rs35705950 (*MUC5B*). To assess the possibility of causal variants underlying the *MUC5B* GWAS signal in LD with rs35705950, we used the recently described CADD method[203] to generate "C-scores" for each of the patient's variants within 1 Mb of rs35705950 (Chr11:1,241,221). LD data

from the SNP Annotation and Proxy Search (SNAP)[204] did not identify any variants in the CEU population in high LD (r-squared > 0.6) with rs35705950 within 1 Mb. Nonetheless, we ranked variants by C-scores, which are an integrated measure of "deleteriousness" outputted on a "phred-like" scale from 0 to 99[203].

Didactics. A group of 15 first year medical students, including 12 combined M.D./Ph.D. students from the Medical Scientist Training Program (MSTP), plus a more senior MSTP student (AK) who functioned as a teaching assistant, met in the Human Anatomy Lab and classroom for a total of 8 hours dispersed through 5 sessions. In smaller groups, the students, under the supervision of the teaching assistant, prepared DNA samples for genomic sequencing. Students were then paired off to complete remaining bioinformatic analysis and jointly draft the manuscript using a shared document online over the ensuing academic quarter. Instructors and students met as a group in 2 additional sessions spanning a total of 4 hours in order to refine the manuscript.

E.4- RESULTS

Clinical history. The patient was a 61 year-old man from eastern Washington state of self-reported European-American ancestry without significant prior medical history who enjoyed good health until approximately eight months before he expired, when he developed flu-like symptoms marked by progressive dry cough and dyspnea. He first sought medical attention when, two months later, he presented with lower extremity edema. Pulmonary artery catheterization at that time revealed severe pulmonary artery hypertension and other changes consistent with *cor pulmonale*. Subsequently, he developed atrial flutter requiring cardioversion and increasing dependence on supplemental oxygen. X-ray computed tomography (CT) of the chest revealed changes consistent with chronic pulmonary fibrosis, comprised of severe dilation of the main pulmonary artery, diffuse basilar ground-glass opacities, and

subpleural reticular opacities. The patient had no known family history of pulmonary diseases. He was married and a father. He worked as a truck driver and was previously employed in a chemical manufacturing facility. There was no known history of asbestos exposure. He kept a pet bird as a younger adult. He had not resided in regions associated with endemic fungal disease. Substance abuse was confined to a 32 pack-year history of smoking.

The patient's symptoms worsened despite treatment consisting of diuretics, corticosteroids, and sildenafil. Approximately two months prior to death, he was transferred to our institution for evaluation for lung transplant. Imaging studies and routine clinical laboratory analysis revealed no evidence for thromboembolism, infection, connective tissue disease, or underlying immunodeficiency. During this interval, it became increasingly difficult to maintain adequate oxygenation. Radiographic chest imaging showed progressive lobar consolidation. As hypoxia worsened, mental status deteriorated. His family requested that care be limited to comfort measures, and he died shortly afterward. Lung tissue was not obtained for diagnosis prior to death.

Gross anatomic findings. The lungs exhibited smooth pleural surfaces, and sectioning revealed diffuse consolidation, without evidence of accentuated subpleural fibrosis and honeycomb patterns of airspace enlargement. The pulmonary vessels showed focal intimal thickening and plaque formation, consistent with pulmonary hypertension, but no evidence of recent or remote thromboemboli. The hilum and mediastinum contained enlarged, reactive-appearing lymph nodes. The heart was also enlarged and all chambers were hypertrophic. The right ventricle demonstrated marked muscular hypertrophy, and had a wall thickness that equaled the left ventricle, consistent with *cor pulmonale*. There was no significant atherosclerotic coronary artery disease, valvular disease, or evidence of myocardial infarction.

Microscopic findings. Histologic examination of lung tissue revealed diffuse fibrous thickening of alveolar septae (**Figure E.1**). Changes were relatively uniform throughout the lung. Dense bundles of collagen and scant mononuclear inflammatory cell infiltrates existed within thickened septae. Focally, apical subpleural regions exhibited increased fibrosis and remodeling, with associated airspace enlargement. Overall, the pattern of injury was most consistent with the fibrosing variant of nonspecific interstitial pneumonia (NSIP), in contrast to the initial diagnosis of IPF [205]. Some of the pulmonary arteries demonstrated fibrous intimal thickening, and the myocardium showed myocyte hypertrophy. Alveolar hemosiderin-laden macrophages were present within lung sections and most likely reflect pulmonary hemorrhage secondary to pulmonary hypertension. Finally, pathological sections revealed an acute bronchopneumonia, consistent with terminal bronchopneumonia most likely due to aspiration. Lymph nodes showed only nonspecific reactive changes.

Genome sequencing. We performed whole genome sequencing, yielding 1.14 billion read pairs resulting in 52× median coverage across the mappable genome (3.1 Gb) and 44× median coverage across the exome (36.6 Mb). A total of 3.2 million single nucleotide variants (SNVs) were identified across the genome, with a Ti:Tv of 2.12[13]. A total of 29,646 SNVs altering protein-coding were observed, of which 146 were not seen in dbSNP v137. Additionally, 305 novel coding indels not reported in dbSNP v137, were also identified.

Mendelian disorders. Mutations in several genes have been reported to segregate with familial forms of IPF (**Table E.1)**. We searched the patient's genome for rare variants (minor allele frequency (MAF) < 0.01) in protein-coding regions within this set of genes. We did not identify rare variants predicted to alter protein sequence or splicing in these genes, although we did find 3 rare synonymous codon substitutions in each of *TERT*, encoding a component of telomerase; *DSP*, producing desmoplakin, a

component of desmosomes; and *DPP9*, the product of which is a serine protease. While synonymous changes in protein coding regions are increasingly reported to influence heritable susceptibility to disease[206], the significance of these variants remains uncertain. DNA sequence analysis software also detected coding variants in *MUC2*, but upon further scrutiny we interpreted them as artifacts attributable to DNA alignment errors due to sequence similarity among mucin gene family members[207].

GWAS variants. We next cross-referenced the patient's genome with SNPs previously implicated in IPF by GWAS (**Supplementary Table E.1**). The patient was heterozygous for five variants influencing susceptibility to IPF (**Table E.2**). Three of the 5 are associated with elevated risk for IPF. One of these variants (**Figure E.2**), rs35705950, is located within the promoter region of *MUC5B* and has a particularly strong association with IPF, with odds ratio (OR) estimates ranging from 4.5-6.8 for heterozygote carriers. (It also demonstrates linkage in familial forms of IPF.) The presence of other variants also increases risk for IPF, albeit to a lesser extent, including one near *TERC*, encoding the RNA component of telomerase, and one intergenic variant on chromosome 7[201]. Two of the 5 variants are associated with reduced susceptibility to IPF and include SNPs in *OBFC1*, a gene involved in telomere maintenance, and *MAPT*, the gene from which the microtubule-associated protein tau is produced[201].

Analysis of variants in the vicinity of rs35705950 (*MUC5B*). The availability of whole genome sequence allowed us to explore whether the *MUC5B* promoter variant contributes to increased risk for IPF, as opposed to alternatively serving only as a marker in linkage disequilibrium (LD) with other causative variant(s) in the same region. For this analysis we used a recently described approach, Combined Annotation-Dependent Depletion scoring system (CADD)[203], which estimates the relative pathogenicity of variants based on a variety of predicted functional effects. We first used HapMap data for individuals of European ancestry to define LD in the vicinity (1 Mb in

either direction) of rs35705950. The highest scoring SNV out of 2,284 in the region, a nonsense variant within *MUC6* (C-score = 42) failed manual validation, due to likely misalignment, and no other SNVs had C-scores above 23.

Rare coding variants. To assess genes not previously associated with IPF but potentially relevant to observed pathology, we filtered variants for rare protein-altering SNVs. With a MAF threshold of 0.01, we identified 1,291 novel or rare coding SNVs. A subset of 57 coding variants at nucleotide positions demonstrating significant interspecies conservation (conScoreGERP > 5.75) was delineated. GeneCards (http://genecards.org) and literature searches were consulted to determine the function, associated disease, and expression profile of each of these variants. Few of these variants are known to be expressed in lung or specifically in diseased lung tissue from patients with IPF[208]; however, this does not rule out their contribution to disease as the expression profile may be incomplete, the gene's effect on the lung may be indirectly mediated through exogenous inflammatory pathways, or the deleterious effects of the genes may arise from abnormal expression. We compared our list of novel variants to a previously published set of genes that are differentially expressed between IPF patients and controls[208]. 29/146 of the novel variants are included in this set. This list includes nonsense variants in genes *NCKAP5* and *SLC25A25*, which were under-expressed in IPF patients, and one nonsense variant in gene *MNS1/TEX9*, which was over-expressed in IPF patients. Other coding variants for genes in this list have previously been associated with different inherited disorders, but none seem pertinent to the patient's illness.

E.5- DISCUSSION

We present what we believe to be the first human genome sequence performed on an individual carrying a clinical diagnosis of IPF. His otherwise excellent health

affords a unique opportunity to uncover genetic factors specifically contributing to development of pulmonary disease.

Infrequently, mutations in several genes are linked to heritable, highly penetrant forms of IPF. The patient lacked rare coding sequence alterations in any of the previously identified genes, in accord with an absence of a family history of pulmonary disease.

Nevertheless, common heritable variants, in this case not altering protein coding, have been found through linkage analysis and GWAS to contribute to risk for IPF. For those for which published evidence is most robust, the patient had a mixture of both risk-reducing (rs1981997 and rs11191865) and risk-elevating (rs4727443 and rs35705950) alleles[199,201]. Among them, rs35705950, a SNP contained in the promoter of *MUC5B* far outweighs others in markedly predisposing to development of IPF (OR, 4.5-6.8).

MUC5B encodes for mucin 5B glycoprotein, which is expressed in saliva and lung tissue and is thought to have lubricating and viscoelastic properties[209]. Recently it has been shown to play an important role in mucociliary clearance, defense against pulmonary infection, and regulating airway inflammation[210].

A tissue diagnosis was not made during the patient's life. Microscopic analysis of tissue obtained upon gross dissection indicates that the patient's pulmonary disease is more appropriately classified as nonspecific interstitial pneumonia. In distinction with IPF, NSIP tends to occur at a younger age, is associated with a better clinical outcome, and occurs in a wide variety of clinical contexts, sometimes in association with an underlying disorder[205]. However, by history and clinical laboratory examination, a predisposing disorder remains undiscovered. One exception, though, was the patient's extensive smoking history, which is a known risk factor for IPF[211], though only indirectly so for NSIP[212].

It is worth noting that the *MUC5B* variant, while initially detected in genetic studies exclusively investigating IPF, has also been associated with similarly appearing fibrotic lung disease detectable by chest CT imaging[200]. Given a paucity of other risk factors, it seems reasonable to hypothesize that the *MUC5B* variant contributed to development of NSIP in this patient although further studies are certainly required to explore this link.

We also believe that this is the first whole genome sequence completed on an individual with the *MUC5B* variant. We are therefore in a position to address, first, whether rs35705950 is merely in LD with other adjacent variants that may actually be responsible for disease and, second, whether variants at other loci modulate the risk for fibrotic lung diseases associated with this SNP.

With respect to the first question, rs35705950 is located within the promoter region of *MUC5B*, is predicted to disrupt transcription factor binding sites, and is correlated with elevated MUC5B expression[199]. Nevertheless, in contrast to whole genome sequencing, not all variants residing on a common haplotype have necessarily been identified and tested for association with disease. We therefore searched contiguous DNA sequence for additional adjacent variants in the vicinity of rs35705950, but did not find additional variants in apparent LD with rs35705950 and predicted to be highly deleterious. Thus, our data do not detract from the hypothesis that rs35705950 is causative of MUC5B dysregulation and disease association.

Relevant to the second question, we searched for all rare and novel variants in the patient's genome, including those not previously associated with IPF. There were no immediately plausible candidates amongst the hundreds of genes for which the patient, as would be expected for anyone[213], harbored rare and private variants. However, nonsense variants were found in the peripheral clock gene *NCKAP5* and the calcium-binding mitochondrial carrier *SLC25A25*, which are each down-regulated in IPF[208].

The patient's somewhat acute presentation following onset of flu-like symptoms is perhaps consistent with an antecedent viral infection, which is a setting in which NSIP has been known to occur[214]. In principle, genomic sequence analysis could permit identification of pathogens that might have triggered putative immune responses and ultimately set the stage for development of fibrotic lung disease. The DNA sequence read mapping strategy we pursued here, involving alignment to a reference genome, filters away the DNA sequences of other organisms. Moreover, for reason of convenience, DNA was extracted from liver, prior to embalming, whereas DNA extraction from lung tissue or regional lymph nodes would have served better for the purpose of detecting the genomes of pathogens.

In addition to offering insight into the pathogenesis of the patient's lung disease, whole genome sequence information may also help to infer prognosis and guide treatment. For example, although this patient unfortunately suffered a rapidly progressive course, in general, the presence of the *MUC5B* SNP rs35705950 has been recently shown to confer a more favorable prognosis in patients with pulmonary fibrosis[207]. Other potentially identifiable genetic variants are associated with habitual tobacco use and may be used to help guide smoking cessation strategies[215], thus mitigating at least one controllable risk factor for lung disease. Although, again unfortunately, this patient did not survive to lung transplant, and while there are alternative conventional serological approaches available for tissue typing, whole genome sequence information[216] can be used to precisely match blood types, which is strongly preferred for lung transplant[217], as well as refine HLA typing, which, when matched between donor and recipient, improves outcomes[218].

Needless to say, genomic sequence is useful for genetic counseling and providing risk assessment to relatives. In this particular situation—body donation for education and research—there is no intent to communicate findings to loved ones.

Nevertheless, post-mortem genomic analysis could conceivably enhance the educational and research value of autopsy, as well as return risk predictive information to survivors.

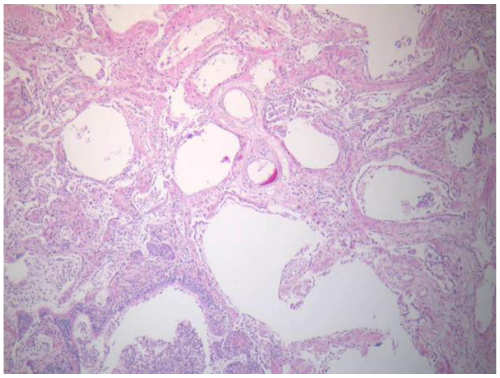
Finally, we wish to comment on the didactic value of this exercise, which, in addition to its research contribution, fulfilled several educational objectives. Detailed medical history of the cadaver is not typically provided to medical students in human anatomy courses[219]. In this case, the students participating in the course received the benefit of a case presentation from a physician who had cared for the patient while hospitalized. It is also unusual to combine the goals of an autopsy with gross dissection[219]. Similarly, histopathologic examination of embalmed tissue is typically not also performed on cadavers in human anatomy courses[219]. In this course, both the gross and microscopic examination of tissues was performed under the guidance of a clinical pathologist, offering a unique opportunity to tie together clinical, anatomic, and cellular findings. Students collectively performed the laboratory and bioinformatic analysis required to assemble the patient's genome and jointly interpreted genetic findings employing literature searches, as well as a variety of databases and computational approaches. Students collectively drafted the manuscript. In summary, this novel case study promoted teamwork and honed clinical, laboratory, computational, writing, and other skills important for career development of physicians and scientists, while contributing genetic insight into a poorly understood disease.

ACKNOWLEDGEMENTS

We deeply appreciate the participation of individuals in the Willed Body Program at the University of Washington. We thank Dr. Daniel Graney for assistance in the planning and anatomy laboratory and instructional phases of the project. We also thank Aaron McKenna and Martin Kircher for their assistance with data analysis.

E.6- FIGURES

A



B

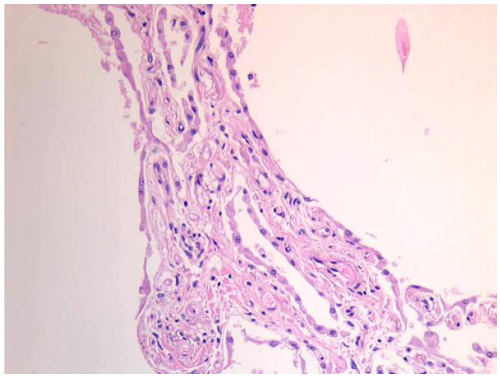


Figure E.1. Histological demonstration of the NSIP pattern of IPF in the patient's lungs. Microscopic examination of the lungs. (**A**) 40×; (**B**) 400×. Note uniform fibrotic thickening of the alveolar septae and type II pneumocyte hypertrophy. There was no histologic evidence of sarcoidosis, hypersensitivity pneumonitis, organizing pneumonia, or diffuse alveolar damage.

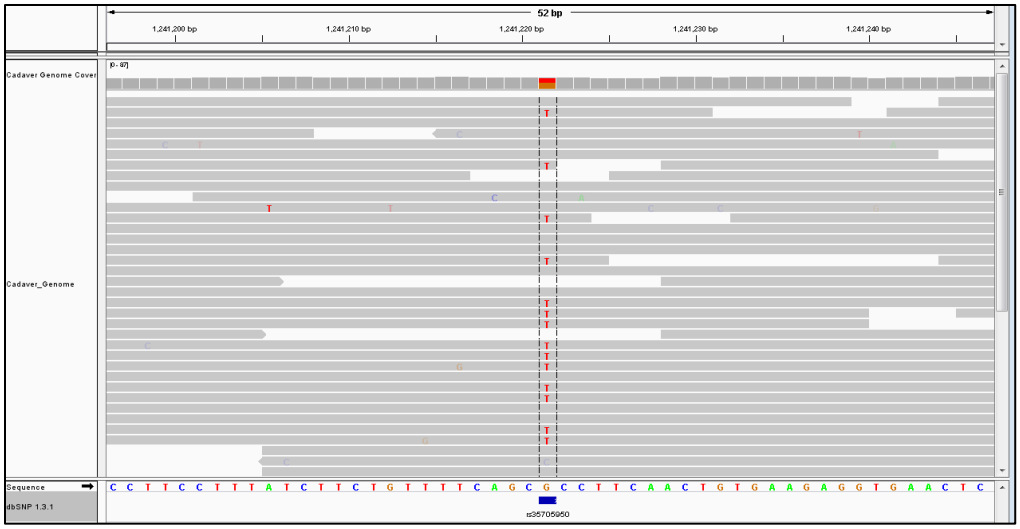


Figure E.2. Integrated Genomics Viewer (IGV) screenshot of the rs35705950 variant.

E.7- TABLES

Gene	Reference
ABCA3	220
MICA	221
SFTPA1	222
SFTPA2	223
SFTPC	220,224
TERC	225
TERT	225

Table E.1. Genes previously associated with familial IPF.

Nearest Gene	SNP ID	Chrom	Position	Variant Type	Minor Allele	Major Allele	Patient Genotype	MAF (1000 Genomes)	OR	Reference
MUC5B	rs35705950	11	1241221	promoter	T	G	T / G	0.052	6.8 [199] 6.3 [200] 4.51[201]	[199-201]
AZGP1	rs4727443	7	99593346	intergenic	A	C	C/A	0.411	1.3[A] 1.11[B]	[201]
MAPT	rs1981997	17	44056767	intronic	A	G	A/G	0.117	0.71[A] 0.67[B]	[201]
OBFC1	rs11191865	10	105672842	intronic	G	A	A/G	0.584	0.8[A] 0.87[B]	[201]
TERC	rs1881984	3	169464459	intergenic	G	A	G/A	0.327		[201]

Table E.2. Variants associated with IPF that were also seen in this individual. Allele frequencies accessed 1/20/2014.
[A]Discovery and [B]replicate GWAS.

E.8- SUPPLEMENTARY INFORMATION

Nearest Gene	SNP ID	Chr	Position	Variant Type	Minor Allele	Major Allele	Patient Genotype	MAF (1000 Genomes)	OR	Reference
MUC5B	rs35705950	11	1241221	promoter	T	G	T/G	0.052	6.8 [9] 6.3 [10] 4.51[11]	[10-12]
AZGP1P1	rs4727443	7	99593346	intergenic	A	C	C/A	0.411	1.3^{A} 1.11^{B}	[12]
MAPT	rs1981997	17	44056767	intronic	A	G	A/G	0.117	0.71^{A} 0.67^{B}	[12]
OBFC1	rs11191865	10	105672842	intronic	G	A	A/G	0.584	0.8^{A} 0.87^{B}	[12]
TERC	rs1881984	3	169464459	intergenic	G	A	G/A	0.327		[12]
LRRC34	rs6793295	3	169518455	Missense	C	T	T/T	0.4261	1.30^{A} 1.39^{B}	[12]
FAM13A	rs2609255	4	89811195	Intronic	G	T	T/T	0.309	1.2^{A} 1.43^{B}	[12]
TERT	rs2736100	5	1286516	Intronic	C	A	C/C	0.4477	0.73^{A} 0.74^{B}	[12,40]
DSP	rs2076295	6	7563232	Intronic	T	**G**	G/G	0.4536	1.43^{A} 1.26	[12]
MUC2	rs7934606	11	1093945	Intronic	C	**T**	T/T	0.205	1.52^{A} 1.56^{B}	[12]
ATP11A	rs1278769	13	113536627	3′ UTR	A	G	G/G	0.2351	0.79^{A} 0.80^{B}	[12]
IVD	rs2034650	15	40717302	Intronic	G	A	A/A	0.4839	0.77^{A} 0.82^{B}	[12]
DPP9	rs12610495	19	4717672	Intronic	G	A	A/A	0.2057	1.29^{A} 1.30^{B}	[12]

Supplementary Table E.1. IPF-associated SNPs investigated in this study. MAF accessed 1/20/2014. Includes data from Table 2. [A]Discovery and [B]replicate GWAS.

REFERENCES

1 Group, U. S. C. S. W. United States Cancer Statistics: 1999–2010 Incidence and Mortality Web-based Report. (U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute, Atlanta, 2013).

2 Mariotto, A. B., Yabroff, K. R., Shao, Y., Feuer, E. J. & Brown, M. L. Projections of the cost of cancer care in the United States: 2010-2020. *Journal of the National Cancer Institute* **103**, 117-128, doi:10.1093/jnci/djq495 (2011).

3 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

4 Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684-1689 (1990).

5 Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300**, 143-149 (1982).

6 Friend, S. H. *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643-646, doi:10.1038/323643a0 (1986).

7 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).

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

9 Loeb, L. A. Cancer cells exhibit a mutator phenotype. *Adv Cancer Res* **72**, 25-56 (1998).

10 Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732, doi:10.1126/science.1117389 (2005).

11 Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145, doi:10.1038/nbt1486 (2008).

12 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467 (1977).

13 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

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

15 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

16 Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276, doi:10.1038/nature08250 (2009).

17 Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat Meth* **7**, 111-118 (2010).

18 Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods* **6**, 315-316, doi:10.1038/nmeth.f.248 (2009).

19 Kumar, A., Shendure, J. & Nelson, P. S. Genome interrupted: sequencing of prostate cancer reveals the importance of chromosomal rearrangements. *Genome Medicine* **3**, 23, doi:10.1186/gm237 (2011).
20 Turner, E. H., Ng, S. B., Nickerson, D. A. & Shendure, J. Methods for genomic partitioning. *Annu Rev Genomics Hum Genet* **10**, 263-284, doi:10.1146/annurev-genom-082908-150112 (2009).
21 O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* **43**, 585-589, doi:10.1038/ng.835 (2011).
22 O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-1622, doi:10.1126/science.1227764 (2012).
23 Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res*, doi:10.1101/gr.147686.112 (2013).
24 Walsh, T. *et al.* Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proc Natl Acad Sci U S A* **107**, 12629-12633, doi:10.1073/pnas.1007983107 (2010).
25 Wood, L. D. *et al.* The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science* **318**, 1108-1113 (2007).
26 Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807-1812, doi:10.1126/science.1164382 (2008).
27 Swanton, C. Intratumor heterogeneity: evolution through space and time. *Cancer Res* **72**, 4875-4882, doi:10.1158/0008-5472.CAN-12-2217 (2012).
28 Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).
29 Wiegand, K. C. *et al.* ARID1A mutations in endometriosis-associated ovarian carcinomas. *N Engl J Med* **363**, 1532-1543, doi:10.1056/NEJMoa1008433 (2010).
30 Jones, S. *et al.* Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330**, 228-231, doi:10.1126/science.1196333 (2010).
31 Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121-1134, doi:10.1016/j.cell.2012.08.024 (2012).
32 Druker, B. J. *et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med* **2**, 561-566 (1996).
33 Crivellari, D. & Molino, A. Small tumor size and node-negative HER2-positive breast cancer: a step forward for a better treatment? *J Clin Oncol* **28**, e257; author reply e258-259, doi:10.1200/JCO.2009.27.7301 (2010).
34 Bryant, H. E. *et al.* Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**, 913-917, doi:10.1038/nature03443 (2005).
35 Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917-921, doi:10.1038/nature03445 (2005).

36 Pritchard, C. C. *et al.* Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. *J Mol Diagn* **16**, 56-67, doi:10.1016/j.jmoldx.2013.08.004 (2014).
37 Roychowdhury, S. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med* **3**, 111ra121, doi:10.1126/scitranslmed.3003161 (2011).
38 Merlo, L. M., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature reviews. Cancer* **6**, 924-935, doi:10.1038/nrc2013 (2006).
39 Martinez, P. *et al.* Parallel evolution of tumour subclones mimics diversity between tumours. *J Pathol* **230**, 356-364, doi:10.1002/path.4214 (2013).
40 Berger, M. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-220 (2011).
41 Holcomb, I. *et al.* Genomic alterations indicate tumor origin and varied metastatic potential of disseminated cells from prostate cancer patients. *Cancer Res* **68**, 5599-5608 (2008).
42 Robbins, C. M. *et al.* Copy number and targeted mutational analysis reveals novel somatic events in metastatic prostate tumors. *Genome Res* **21**, 47-55, doi:10.1101/gr.107961.110 (2011).
43 Taylor, B. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer cell* **18**, 11-22 (2010).
44 Helgeson, B. E. *et al.* Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer. *Cancer Res* **68**, 73-80, doi:10.1158/0008-5472.CAN-07-5352 (2008).
45 Tomlins, S. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (New York, N.Y.)* **310**, 644-648 (2005).
46 Tomlins, S. A. *et al.* Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595-599, doi:10.1038/nature06024 (2007).
47 Holcomb, I. N. *et al.* Comparative analyses of chromosome alterations in soft-tissue metastases within and across patients with castration-resistant prostate cancer. *Cancer Res* **69**, 7793-7802, doi:10.1158/0008-5472.CAN-08-3810 (2009).
48 van Weerden, W. M., Bangma, C. & de Wit, R. Human xenograft models as useful tools to assess the potential of novel therapeutics in prostate cancer. *Brit J Cancer* **100**, 13-18, doi:DOI 10.1038/sj.bjc.6604822 (2009).
49 Clark, M. J. *et al.* U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet* **6**, e1000832, doi:10.1371/journal.pgen.1000832 (2010).
50 Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
51 Corey, E. *et al.* LuCaP 35: A new model of prostate cancer progression to androgen independence. *The Prostate* **55**, 239-246 (2003).
52 Liu, W. *et al.* Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med* **15**, 559-565, doi:10.1038/nm.1944 (2009).

53 Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153-1157, doi:10.1038/nature04240 (2005).
54 Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994-997, doi:10.1038/nature06611 (2008).
55 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913 (2005).
56 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
57 Speiser, P. *et al.* A constitutional de novo mutation in exon 8 of the p53 gene in a patient with multiple primary malignancies. *Brit J Cancer* **74**, 269-273 (1996).
58 Sánchez-Solana, B. *et al.* The EGF-like proteins DLK1 and DLK2 function as inhibitory non-canonical ligands of NOTCH1 receptor that modulates each other's activities. *Biochimica Et Biophysica Acta* (2011).
59 Scherer, P. E. *et al.* Cab45, a novel (Ca2+)-binding protein localized to the Golgi lumen. *J Cell Biol* **133**, 257-268 (1996).
60 Kang, H., Escudero-Esparza, A., Douglas-Jones, A., Mansel, R. E. & Jiang, W. G. Transcript analyses of stromal cell derived factors (SDFs): SDF-2, SDF-4 and SDF-5 reveal a different pattern of expression and prognostic association in human breast cancer. *International Journal of Oncology* **35**, 205-211 (2009).
61 Filmus, J. Glypicans in growth control and cancer. *Glycobiology* **11**, 19R -23R-19R -23R (2001).
62 Okamoto, K. *et al.* Common variation in GPC5 is associated with acquired nephrotic syndrome. *Nat Genet* **43**, 459-463 (2011).
63 Williamson, D. *et al.* Role for amplification and expression of glypican-5 in rhabdomyosarcoma. *Cancer Res* **67**, 57-65 (2007).
64 Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801-1806, doi:10.1126/science.1164368 (2008).
65 Timmermann, B. *et al.* Somatic Mutation Profiles of MSI and MSS Colorectal Cancer Identified by Whole Exome Next Generation Sequencing and Bioinformatics Analysis. *PloS one* **5**, e15661-e15661 (2010).
66 Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-190 (2010).
67 Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196 (2010).
68 Sun, C. *et al.* Androgen receptor mutation (T877A) promotes prostate cancer cell growth and cell survival. *Oncogene* **25**, 3905-3913, doi:10.1038/sj.onc.1209424 (2006).
69 Loeb, L. A., Bielas, J. H. & Beckman, R. A. Cancers Exhibit a Mutator Phenotype: Clinical Implications. *Cancer Res* **68**, 3551-3557 (2008).
70 Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nat Rev Cancer* **11**, 450-457, doi:10.1038/nrc3063 (2011).

71 Yan, S. Y. *et al.* Three novel missense germline mutations in different exons of MSH6 gene in Chinese hereditary non-polyposis colorectal cancer families. *World J Gastroenterol* **13**, 5021-5024 (2007).
72 Corey, E., Quinn, J. E. & Vessella, R. L. A novel method of generating prostate cancer metastases from orthotopic implants. *The Prostate* **56**, 110-114 (2003).
73 van Weerden, W. M., Bangma, C. & de Wit, R. Human xenograft models as useful tools to assess the potential of novel therapeutics in prostate cancer. *Br J Cancer* **100**, 13-18 (2008).
74 Corey, E. V., RL. in *Contemporary cancer research* (eds Leland W. K. Chung, William Brewster Isaacs, & Jonathan W. Simons) 3-32 (Humana Press, 2007).
75 O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet*, doi:ng.835 [pii]
10.1038/ng.835 (2011).
76 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
77 Jemal, A., Siegel, R., Xu, J. & Ward, E. Cancer statistics, 2010. *CA Cancer J Clin* **60**, 277-300, doi:10.3322/caac.20073 (2010).
78 Schoenborn, J. R., Nelson, P. & Fang, M. Genomic profiling defines subtypes of prostate cancer with the potential for therapeutic stratification. *Clin Cancer Res* **19**, 4058-4066, doi:10.1158/1078-0432.CCR-12-3606 (2013).
79 Kohli, M. & Tindall, D. J. New developments in the medical management of prostate cancer. *Mayo Clin Proc* **85**, 77-86, doi:10.4065/mcp.2009.0442 (2010).
80 Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* **44**, 685-689, doi:10.1038/ng.2279 (2012).
81 Baca, S. C. & Garraway, L. A. The genomic landscape of prostate cancer. *Front Endocrinol (Lausanne)* **3**, 69, doi:10.3389/fendo.2012.00069 (2012).
82 Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677, doi:10.1016/j.cell.2013.03.021 (2013).
83 Kumar, A. *et al.* Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proceedings of the National Academy of Sciences* (2011).
84 Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer cell* **18**, 11-22, doi:10.1016/j.ccr.2010.05.026 (2010).
85 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).
86 Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239-243, doi:D - nlm: nihms368879
D - NLM: PMC3396711 EDAT- 2012/06/23 06:00 MHDA- 2012/08/10 06:00 CRDT- 2012/06/23 06:00 PMCR- 2013/01/12 00:00 PHST- 2012/01/04 [received] PHST- 2012/04/05 [accepted] PHST- 2012/05/20 [aheadofprint] AID - nature11125 [pii] AID - 10.1038/nature11125 [doi] PST - ppublish (2012).
87 Lindberg, J. *et al.* The mitochondrial and autosomal mutation landscapes of prostate cancer. *Eur Urol* **63**, 702-708, doi:10.1016/j.eururo.2012.11.053 (2013).

88 Eisinger-Mathason, T. S. & Simon, M. C. HIF-1alpha partners with FoxA2, a neuroendocrine-specific transcription factor, to promote tumorigenesis. *Cancer cell* **18**, 3-4, doi:10.1016/j.ccr.2010.06.007 (2010).
89 Lohr, J. G. *et al.* Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat Biotechnol*, doi:10.1038/nbt.2892 (2014).
90 Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114-1117, doi:10.1038/nature09515 (2010).
91 Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A* **110**, 4009-4014, doi:10.1073/pnas.1219747110 (2013).
92 Snuderl, M. *et al.* Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer cell* **20**, 810-817, doi:10.1016/j.ccr.2011.11.005 (2011).
93 Nickel, G. C. *et al.* Characterizing mutational heterogeneity in a glioblastoma patient with double recurrence. *PloS one* **7**, e35262, doi:10.1371/journal.pone.0035262 (2012).
94 DNAcopy: DNA copy number data analysis.
95 Watanabe, T., Nobusawa, S., Kleihues, P. & Ohgaki, H. IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas. *Am J Pathol* **174**, 1149-1153, doi:10.2353/ajpath.2009.080958 (2009).
96 Lass, U. *et al.* Clonal analysis in recurrent astrocytic, oligoastrocytic and oligodendroglial tumors implicates IDH1- mutation as common tumor initiating event. *PloS one* **7**, e41298, doi:10.1371/journal.pone.0041298 (2012).
97 Szerlip, N. J. *et al.* Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response. *Proc Natl Acad Sci U S A* **109**, 3041-3046, doi:10.1073/pnas.1114033109 (2012).
98 Ino, Y. *et al.* Molecular subtypes of anaplastic oligodendroglioma: implications for patient management at diagnosis. *Clin Cancer Res* **7**, 839-845 (2001).
99 Okada, Y. *et al.* Selection pressures of TP53 mutation and microenvironmental location influence epidermal growth factor receptor gene amplification in human glioblastomas. *Cancer Res* **63**, 413-416 (2003).
100 Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-220, doi:10.1038/nature11690 (2013).
101 Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).
102 Barbieri, C. E. *et al.* The mutational landscape of prostate cancer. *Eur Urol* **64**, 567-576, doi:10.1016/j.eururo.2013.05.029 (2013).
103 Choi, P. S. & Meyerson, M. Targeted genomic rearrangements using CRISPR/Cas technology. *Nat Commun* **5**, 3728, doi:10.1038/ncomms4728 (2014).
104 Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**, 1173-1175, doi:10.1038/nbt.1589 (2009).

105 Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45**, 1021-1028, doi:10.1038/ng.2713 (2013).
106 Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265-270, doi:10.1038/nbt.2136 (2012).
107 Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc Natl Acad Sci U S A* **110**, E1263-1272, doi:10.1073/pnas.1303309110 (2013).
108 Zhang, Q. *et al.* Acceleration of Emergence of Bacterial Antibiotic Resistance in Connected Microenvironments. *Science* **333**, 1764-1767 (2011).
109 Lambert, G. *et al.* An analogy between the evolution of drug resistance in bacterial communities and malignant tissues. *Nature reviews. Cancer* **11**, 375-382 (2011).
110 Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* **6**, 224ra224, doi:10.1126/scitranslmed.3007094 (2014).
111 Leary, R. J. *et al.* Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med* **4**, 162ra154, doi:10.1126/scitranslmed.3004742 (2012).
112 Welty, C. J. *et al.* Single cell transcriptomic analysis of prostate cancer cells. *BMC Mol Biol* **14**, 6, doi:10.1186/1471-2199-14-6 (2013).
113 Fehm, T. *et al.* Cytogenetic evidence that circulating epithelial cells in patients with carcinoma are malignant. *Clin Cancer Res* **8**, 2073-2084 (2002).
114 Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94 (2011).
115 Navin, N. & Hicks, J. Future medical applications of single-cell sequencing in cancer. *Genome Medicine* **3**, 31-31 (2011).
116 Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005 (2010).
117 Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nat Protoc* **7**, 1024-1041, doi:10.1038/nprot.2012.039 (2012).
118 Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**, 78-85 (2000).
119 Goh, C. L. *et al.* Genetic variants associated with predisposition to prostate cancer and potential clinical implications. *J Int Medicine* **271**, 353-365 (2012).
120 Eeles, R. *et al.* Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet* **45**, 385-391 (2013).
121 Ostrander, E. A. & Stanford, J. L. Genetics of prostate cancer: too many loci, too few genes. *American journal of human genetics* **67**, 1367-1375 (2000).
122 Easton, D. F., Schaid, D. J., Whittemore, A. S., Isaacs, W. J. & ICPCG. Where are the prostate cancer genes?--A summary of eight genome wide searches. *The Prostate* **57**, 261-269 (2003).
123 Schaid, D. J. The complex genetic epidemiology of prostate cancer. *Hum Mol Genet* **13**, R103-R121 (2004).

124 Edwards, S. M. *et al.* Two percent of men with early-onset prostate cancer harbor germline mutations in the BRCA2 gene. *Am J Hum Genet* **72**, 1-12 (2003).
125 Agalliu, I. *et al.* Rare germline mutations in the BRCA2 gene are associated with early-onset prostate cancer. *Br J Cancer* **97**, 826-831 (2007).
126 Kote-Jarai, Z. *et al.* BRCA2 is a moderate penetrance gene contributing to young-onset prostate cancer: implications for genetic testing in prostate cancer patients. *Br J Cancer* **105**, 1230-1234, doi:10.1038/bjc.2011.383 (2011).
127 Ewing, C. M. *et al.* Germline mutations in *HOXB13* and prostate-cancer risk. *N Engl J Med* **366**, 141-149, doi:10.1056/NEJMoa1110000 (2012).
128 Xu, J. *et al. HOXB13* is a susceptibility gene for prostate cancer: Results from the International Consortium for Prostate Cancer Genetics (ICPCG). *Hum Genet* **132**, 5-14 (2013).
129 Akbari, M. R. *et al.* Association between germline *HOXB13* G84E mutation and risk of prostate cancer. *Journal of the National Cancer Institute* **104**, 1260-1262, doi:10.1093/jnci/djs288 (2012).
130 Karlsson, R. *et al.* A population-based assessment of germline *HOXB13* G84E mutation and prostate cancer risk. *Eur Urol*, ePub ahead of print: 20 July 2012, doi:10.1016/j.eururo.2012.07.027 (2012).
131 Stott-Miller, M. *et al. HOXB13* mutations in a population-based, case control study of prostate cancer. *Prostate* **73**, 634-641 (2013).
132 Stanford, J. L. *et al.* Dense genome-wide SNP linkage scan in 301 hereditary prostate cancer families identifies multiple regions with suggestive evidence for linkage. *Human molecular genetics* **18**, 1839-1848 (2009).
133 Stanford, J. L., Wicklund, K. G., McKnight, B., Daling, J. R. & Brawer, M. K. Vasectomy and risk of prostate cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **8**, 881-886 (1999).
134 Agalliu, I., Salinas, C. A., Hansten, P. D., Ostrander, E. A. & Stanford, J. L. Statin use and risk of prostate cancer: results from a population-based epidemiologic study. *Am J Epidemiol* **168**, 250-260 (2008).
135 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
136 McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
137 O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* **43**, 585-589, doi:10.1038/ng.835 (2011).
138 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913, doi:10.1101/gr.3577405 (2005).
139 Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Human molecular genetics* **10**, 591-597 (2001).
140 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* **31**, 3812-3814 (2003).

141 Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-864 (1974).
142 Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553-1561 (2009).
143 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919 (1992).
144 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
145 Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature methods* **6**, 315-316, doi:10.1038/nmeth.f.248 (2009).
146 O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619-1622 (2012).
147 Allen-Brady, K., Wong, J. & Camp, N. J. PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. *BMC Bioinformatics* **7**, 209 (2006).
148 Breslow, N. E. & Day, N. E. *Statistical Methods in Cancer Research, Volume 1-The Analysis of Case-Control Studies*. (International Agency for Research on Cancer, 1980).
149 Acevedo, V. D. *et al.* Inducible FGFR-1 activation leads to irreversible prostate adenocarcinoma and an epithelial-to-mesenchymal transition. *Cancer cell* **12**, 559-571 (2007).
150 Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-6067 (2004).
151 Abeler-Dorner, L., Swamy, M., Williams, G., Hayday, A. C. & Bas, A. Butyrophilins: an emergine family of immune regulators. *Trends in Immunology* **33**, 34-41, doi:10.1016/j.it.2011.09.007 (2012).
152 Arnett, H. A. *et al.* BTNL2, a butyrophilin/B7-like molecule, is a negative costimulatory molecule modulated in intestinal inflammation. *Journal of immunology (Baltimore, Md. : 1950)* **178**, 1523-1533 (2007).
153 Nguyen, T., Liu, X. K., Zhang, Y. & Dong, C. BTNL2, a butyrophilin-like molecule that functions to inhibit T cell activation. *Journal of immunology (Baltimore, Md. : 1950)* **176**, 7354-7360 (2006).
154 Valentonyte, R. *et al.* Sarcoidosis is associated with a truncating splice site mutation in *BTNL2*. *Nat Genet* **37**, 357-364, doi:10.1038/ng1519 (2005).
155 Rybicki, B. A. *et al.* The *BTNL2* gene and sarcoidosis susceptibility in African Americans and whites. *American journal of human genetics* **77**, 491-499, doi:10.1086/444435 (2005).
156 Mitsunaga, S. *et al.* Exome sequencing identifies novel rheumatoid arthritis-susceptible variants in the *BTNL2*. *J Human Genet* **58**, 210-215, doi:10.1038/jhg.2013.2 (2013).
157 Franke, A. *et al.* Sequence variants in *IL10*, *ARPC2* and multiple other loci contribute to ulcerative colitis susceptibility. *Nat Genet* **40**, 1319-1323, doi:10.1038/ng.221 (2008).

158 Silverberg, M. S. *et al.* Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat Genet* **41**, 216-220, doi:10.1038/ng.275 (2009).
159 Coussens, L. M. & Werb, Z. Inflammation and Cancer. *Nature* **420**, 860-867 (2002).
160 Nelson, W. G., De Marzo, A. M., DeWeese, T. L. & Isaacs, W. B. The role of inflammation in the pathogenesis of prostate cancer. *J Urol* **172**, S6-11; discussion S11-12 (2004).
161 ENCODE. http://genome.ucsc.edu/ENCODE/.
162 Breyer, J. P., Avritt, T. G., McReynolds, K. M., Dupont, W. D. & Smith, J. R. Confirmation of the *HOXB13* G84E germline mutation in familial prostate cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **21**, 1348-1353, doi:10.1158/1055-9965.epi-12-0495 (2012).
163 Project, C. http://www.genome.gov/20519355.
164-190 1000 Genomes Project. *1000 Genomes Project*, <http://www.1000genomes.org/page.php> (2008-2011).
191 Dixon-Salazar, T. J. *et al.* Exome sequencing can improve diagnosis and alter patient management. *Sci Transl Med* **4**, 138ra178, doi:10.1126/scitranslmed.3003544 (2012).
192 Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* **362**, 1181-1191, doi:10.1056/NEJMoa0908094 (2010).
193 Ashley, E. A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525-1535, doi:10.1016/S0140-6736(10)60452-7 (2010).
194 Topol, E. J. From dissecting cadavers to dissecting genomes. *Sci Transl Med* **5**, 202ed215, doi:10.1126/scitranslmed.3007091 (2013).
195 Noble, P. W., Barkauskas, C. E. & Jiang, D. Pulmonary fibrosis: patterns and perpetrators. *J Clin Invest* **122**, 2756-2762, doi:10.1172/JCI60323 (2012).
196 Selman, M. & Pardo, A. Idiopathic pulmonary fibrosis: an epithelial/fibroblastic cross-talk disorder. *Respir Res* **3**, 3 (2002).
197 Boucher, R. C. Idiopathic pulmonary fibrosis--a sticky business. *N Engl J Med* **364**, 1560-1561, doi:10.1056/NEJMe1014191 (2011).
198 Herazo-Maya, J. D. & Kaminski, N. Personalized medicine: applying 'omics' to lung fibrosis. *Biomark Med* **6**, 529-540, doi:10.2217/bmm.12.38 (2012).
199 Seibold, M. A. *et al.* A common MUC5B promoter polymorphism and pulmonary fibrosis. *N Engl J Med* **364**, 1503-1512, doi:10.1056/NEJMoa1013660 (2011).
200 Hunninghake, G. M. *et al.* MUC5B promoter polymorphism and interstitial lung abnormalities. *N Engl J Med* **368**, 2192-2200, doi:10.1056/NEJMoa1216076 (2013).
201 Fingerlin, T. E. *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet* **45**, 613-620, doi:10.1038/ng.2609 (2013).
202 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

203 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, doi:10.1038/ng.2892 (2014).
204 Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938-2939, doi:10.1093/bioinformatics/btn564 (2008).
205 du Bois, R. & King, T. E., Jr. Challenges in pulmonary fibrosis x 5: the NSIP/UIP debate. *Thorax* **62**, 1008-1012, doi:10.1136/thx.2004.031039 (2007).
206 Stergachis, A. B. *et al.* Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342**, 1367-1372, doi:10.1126/science.1243490 (2013).
207 Peljto, A. L. *et al.* Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *Jama* **309**, 2232-2239, doi:10.1001/jama.2013.5827 (2013).
208 Yang, I. V. *et al.* Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. *Thorax* **68**, 1114-1121, doi:10.1136/thoraxjnl-2012-202943 (2013).
209 Turner, J. & Jones, C. E. Regulation of mucin expression in respiratory diseases. *Biochem Soc Trans* **37**, 877-881, doi:10.1042/BST0370877 (2009).
210 Roy, M. G. *et al.* Muc5b is required for airway defence. *Nature*, doi:10.1038/nature12807 (2013).
211 Baumgartner, K. B., Samet, J. M., Stidley, C. A., Colby, T. V. & Waldron, J. A. Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* **155**, 242-248, doi:10.1164/ajrccm.155.1.9001319 (1997).
212 Marten, K. *et al.* Non-specific interstitial pneumonia in cigarette smokers: a CT study. *Eur Radiol* **19**, 1679-1685, doi:DOI 10.1007/s00330-009-1308-7 (2009).
213 Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**, 628-640, doi:10.1038/nrg3046 (2011).
214 Poletti, V., Romagnoli, M., Piciucchi, S. & Chilosi, M. Current status of idiopathic nonspecific interstitial pneumonia. *Semin Respir Crit Care Med* **33**, 440-449, doi:10.1055/s-0032-1325155 (2012).
215 Kortmann, G. L., Dobler, C. J., Bizarro, L. & Bau, C. H. Pharmacogenetics of smoking cessation therapy. *Am J Med Genet B Neuropsychiatr Genet* **153B**, 17-28, doi:10.1002/ajmg.b.30978 (2010).
216 Erlich, H. HLA DNA typing: past, present, and future. *Tissue Antigens* **80**, 1-11, doi:10.1111/j.1399-0039.2012.01881.x (2012).
217 Sano, Y., Date, H., Nagahiro, I., Aoe, M. & Shimizu, N. Relationship between anti-ABO antibody production and hemolytic anemia after minor ABO-mismatched living-donor lobar lung transplantation. *Transplant Proc* **37**, 1371-1372, doi:10.1016/j.transproceed.2004.12.205 (2005).
218 Smits, J. M. *et al.* Three-year survival rates for all consecutive heart-only and lung-only transplants performed in Eurotransplant, 1997-1999. *Clin Transpl*, 89-100 (2003).
219 Papa, V. & Vaccarezza, M. Teaching Anatomy in the XXI Century: New Aspects and Pitfalls. *ScientificWorldJournal* **2013**, 310348, doi:10.1155/2013/310348 (2013).