Title
Accurate Identification of Somatic Mutations in Clinical Tumor Specimens

Permalink
https://escholarship.org/uc/item/8nh6b89b

Author
Yost, Shawn Eric

Publication Date
2013

Peer reviewed|Thesis/dissertation
Copyright

Shawn Eric Yost, 2013

All rights reserved.
The Dissertation of Shawn Eric Yost is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

Co-Chair

Chair

University of California, San Diego

2013
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signature Page ................................................................................................. iii</td>
</tr>
<tr>
<td>Table of Contents ................................................................................................ iv</td>
</tr>
<tr>
<td>List of Figures ........................................................................................................ vi</td>
</tr>
<tr>
<td>List of Tables .......................................................................................................... viii</td>
</tr>
<tr>
<td>Acknowledgments .................................................................................................... ix</td>
</tr>
<tr>
<td>Vita ........................................................................................................................ x</td>
</tr>
<tr>
<td>Abstract of the Dissertation .................................................................................. xi</td>
</tr>
<tr>
<td>Introduction ............................................................................................................ 1</td>
</tr>
<tr>
<td>Chapter 1: Identifying somatic DNA mutations in sequencing data from formalin fixed paraffin embedded (FFPE) tumor specimens ......................................................................................................................... 10</td>
</tr>
<tr>
<td>1.1 BACKGROUND .................................................................................................. 10</td>
</tr>
<tr>
<td>1.2 METHODS ........................................................................................................ 11</td>
</tr>
<tr>
<td>1.3 RESULTS .......................................................................................................... 22</td>
</tr>
<tr>
<td>1.4 DISCUSSION .................................................................................................... 31</td>
</tr>
<tr>
<td>1.5 FIGURES .......................................................................................................... 36</td>
</tr>
<tr>
<td>1.6 TABLES ............................................................................................................ 46</td>
</tr>
<tr>
<td>Chapter 2: Evaluate the genomic validity of tumor derived pre-clinical models using exome sequencing data ............................................................................................................................... 48</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1.1 ........................................................................................................................................... 36
Figure 1.2 ........................................................................................................................................... 37
Figure 1.3 ........................................................................................................................................... 38
Figure 1.4 ........................................................................................................................................... 39
Figure 1.5 ........................................................................................................................................... 40
Figure 1.6 ........................................................................................................................................... 41
Figure 1.7 ........................................................................................................................................... 42
Figure 1.8 ........................................................................................................................................... 43
Figure 1.9 ........................................................................................................................................... 44
Figure 1.10 ........................................................................................................................................ 45
Figure 2.1 .......................................................................................................................................... 79
Figure 2.2 .......................................................................................................................................... 79
Figure 2.4 .......................................................................................................................................... 81
Figure 2.5 .......................................................................................................................................... 82
Figure 2.6 .......................................................................................................................................... 83
Figure 2.7 .......................................................................................................................................... 84
Figure 2.8 .......................................................................................................................................... 85
Figure 2.9........................................................................................................86
Figure 2.10 .......................................................................................................87
Figure 3.1.........................................................................................................109
Figure 3.2.........................................................................................................110
Figure 3.3.........................................................................................................111
LIST OF TABLES

Table 1.1 .................................................................................................................................................46

Table 1.2 ..................................................................................................................................................47

Table 2.1 ..................................................................................................................................................88
ACKNOWLEDGEMENTS

I would like to acknowledge Professor Kelly Frazer for her support as the chair of my committee.

Chapter 1, in full, is a reprint of the material as it appears in Nucleic Acids Research, 2012, Vol. 40, No 14 e107. Pub online 6 April 2012. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. Shawn E. Yost, Erin N. Smith, Richard B. Schwab, Lei Bao, HyunChul Jung, Xiaoyun Wang, Emile Voest, John P. Pierce, Karen Messer, Barbara A. Parker, Olivier Harismendy, and Kelly A. Frazer. The dissertation author was the primary author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in PLoS ONE 8(2): e56185. doi:10.1371/journal.pone.0056185. Epub 2013 Feb 18. High-Resolution Mutational Profiling Suggests the Genetic Validity of Glioblastoma Patient-Derived Pre-Clinical Models. Shawn E. Yost, Sandra Pastorino, Sophie Rozenzhak, Erin N. Smith, Ying S. Chao, Pengfei Jiang, Santosh Kesari, Kelly A. Frazer, Olivier Harismendy. The dissertation author was the primary author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Bioinformatics. 2013 May 27. [Epub ahead of print]. Mutascope: Sensitive Detection of Somatic Mutations from Deep Amplicon Sequencing. Shawn E. Yost, Hakan Alakus, Hiroko Matsui, Richard B. Schwab, Kristen Jepsen, Kelly A. Frazer, Olivier Harismendy. The dissertation author was the primary author of this paper.
VITA

2005-2008  Bachelor of Science, Bioinformatics, University of California, Santa Cruz
2008-2013  Doctor of Philosophy, Bioinformatics, University of California, San Diego

PUBLICATIONS


FIELDS OF STUDY

Major Field: Bioinformatics
ABSTRACT OF THE DISSERTATION

Accurate Identification of Somatic Mutations in Clinical Tumor Specimens

by

Shawn Eric Yost

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2013

Professor Kelly Frazer, Chair
Professor Vineet Bafna, Co-Chair

Cancer development and progression is driven by genetic alterations. These alterations include somatic DNA sequence changes (somatic SNVs), copy number aberrations (CNAs), chromosomal rearrangements, and epigenetic modification. For over 20 years scientists have been studying cancer genetics to develop new and better treatment options. Within the past few years cancer genetic studies have shifted from single gene studies to whole genome studies with the development of next-generation sequencing (NGS) platforms. One challenge with using NGS platforms to identify somatic mutations in a tumor is determining which somatic mutations are true positives vs. false positives.

Here we developed different methodologies to remove sequencing errors caused by various sample preparations and sequencing platforms in order to identify true
functional somatic variants in clinical samples. We analyzed and evaluated cancer specimens processed in FFPE, pre-clinical tumor models, tumors with low tumor cellularity, and tumor subclones with the use of next generation sequencing data. First, we analyzed whole-genome sequencing data from formalin-fixed paraffin embedded breast cancer samples and developed a new method for filtering false positive somatic mutations caused by formalin damaging the DNA. Second, we evaluated the genomic validity of tumor derived pre-clinical models using exome sequencing data. Finally, we developed a new software suite called Mutascope that is designed to analyze PCR amplicon sequencing data and identify somatic variants that are present in only 1% of the tumor.
Introduction

Cancer is known as unregulated cell growth in the human body. About 500,000 people died due to cancer related illnesses in the United States alone in the past year. Over a hundred years ago Theodor Boveri was the first to propose that cancer is caused by changes in the human genome\(^1\). It wasn’t until 1970’s that this concept was proved by showing the existence of mutated cancer-causing genes\(^2,3\). There are currently two different types of cancer causing genes, oncogenes and tumor suppressor genes. These genes can be activated/suppressed by either single nucleotide variants (SNVs), small insertions or deletions (indels), copy number aberrations (CNAs), chromosomal rearrangements, and epigenetic modifications (RP1). In the mid-1980s scientists started to believe that they would need the complete sequence of the human genome in order to fully study cancer\(^4\).

The cancer community wished to have a draft of the human genome so that they could use it to help understand how cancer develops from a normal cell, progresses into different stages, becomes resistant to various treatments, and how cancer metastasizes. Understanding these different aspects of cancer can lead to new methods on how to prevent and treat cancer. They believed that identifying somatic mutations in the DNA would lead them to biological pathways and finally to a drug target within the pathway. Researchers have now established 6 different biological processes in a cell that if a combination of these are disrupted then a normal cell to could change into a cancer cell. These mechanisms or “Hallmarks of Cancer” are resisting cell death, sustaining proliferative signaling, evading growth suppressors, activating invasion and metastasis,
enabling replicative immortality, and including angiogenesis. One example is the multiple targeted cancer drug therapies targeting the fusion gene BCR-ABL in Leukemia. This is just one example of how scientists have begun to scratch the surface of cancer genetics.

Finally in 1990 the research community got behind the idea of sequencing the human genome and thus started the Human Genome Project (HGP). In 2000 the HGP released the first draft of the human genome and in 2004 the first finished-grade human genome sequenced was released. The cost of the first draft of the human genome was about 3 billion dollars. Due to this enormous price tag of sequencing a single genome scientists next put their efforts into decreasing the cost of sequencing a genome. Within a few years the first wave of new sequencing platforms were available. These platforms are referred to as next-generation sequencing (NGS) and are a lot more cost effective. With the cost of sequencing a single genome now down to $10,000, scientists started using NGS platforms to study cancer genomics. Soon after, various large consortiums of scientists came together to create large cancer genomic projects, such as “The Cancer Genome Atlas” (TCGA) and the international Cancer Genome Consortium.

With researchers growing understanding of cancer genetics, clinics are now starting to treat a patient’s cancer based on the somatic events in that cancer. Using NGS data scientists have discovered different subgroups of cancers of various cancer types. For example, in glioblastoma researchers have found that patients with a hypermutator phenotype should be treated differently in the clinic. The advancements in cancer research and treatment has started a new wave of personalized medicine for patients with cancer. Clinics have already started to include biomarker tests for early cancer detection.
and estimation of the disease risk\textsuperscript{15}. Some of the biomarkers are used to classify different cancer subtypes to improve the patients survivability and well-being\textsuperscript{16,17}. While clinics have started to implement some personalized medicine, most cancer patients are still treated with standard chemotherapy and radiation.

While a lot of what we now about cancer has been established from hypothesis-driven research, the cancer research field has started moving towards systematic studies like the TCGA. As mentioned before these large scale studies are using NGS platforms to generate terabytes of cancer genomics data. However there are multiple challenges to using this data including determine which somatic mutations are true somatic mutations or errors and how to interpret the identified true somatic mutations. The first challenge in understanding somatic mutations found in a cancer sample is which mutation is a somatic mutation that drives the cancer (driver) and which somatic mutations arose during the progression of the disease as a byproduct (passenger). Then once we know which somatic mutations are drivers vs. passengers we then need to understand how these effect survival of a patient and response to cancer therapies. These are just a few challenges to understanding the results found in these large scale cancer genetic studies.

However, before we can try to understand the results of these studies we first need to be confident that what we believe are somatic mutations in the tumors are true somatic mutations. DNA sequencing data from any NGS platform contains errors and researchers need to take into account these errors when analyzing the data\textsuperscript{18}. These errors are caused by random substitutions introduced during the library’s PCR step, mixed clusters/beads reducing the image clarity, neighbor interference, DNA molecules from the same DNA fragment not all extended in every cycle (phasing), and more. NGS platforms try to take
these errors into account by assigning a base quality (BQ) score to each sequenced base that are the probability that the base was incorrectly called. While this does improve the overall usefulness of the data, not all of the error types are captured with the BQ score. Errors can also be introduce when aligning the NGS data back to the genome\textsuperscript{19}. Most of the error rates are caused by trying to align reads to repetitive regions in the human genome, both short simple repeats/homopolymer runs and two or more homologous regions in the genome. All of these different types of error rates can lead to false-positive somatic event calls. This may not cause a major problem in analyzing data from 1,000 different patient’s cancers but is very important when trying to identify a simple patient’s tumor’s somatic events to determine their best treatment option.

With taking these drawbacks of using NGS data into account, cancer genetic studies have shifted from single gene studies\textsuperscript{20-22} to whole genome studies\textsuperscript{23-25} within the past few years. Various studies have used NGS technologies to identify genes and processes driving cancer, from sequencing the coding regions of the genome (exome sequencing)\textsuperscript{26,27} to sequencing the entire genome (whole-genome sequencing)\textsuperscript{28-31}. All of these studies have used tumor DNA isolated from flash frozen samples or cancer cell lines. While using flash frozen DNA for sequencing is ideal, the standard clinical procedure to process tumor resections is to formalin fix and paraffin embed the tumor cells. Due to this it usually not possible to obtain flash frozen DNA for sequencing. Since formalin fixation and paraffin embedding (FFPE) tumor samples has been the standard sample preparation for pathologists for decades, there are numerous large tumor archives with patient history available to researchers. However FFPE processed samples are not the ideal samples to use for genetic studies due to the formalin fixation damaging the
DNA. Very few researchers have sequenced FFPE cancer samples with NGS platforms due to the increased error rates in the resulting data. For this reason removing the additional errors in NGS data caused by FFPE is essential for these samples to be used in cancer studies.

Unlike formaline that damages the DNA that lead to false positive mutation calls, xenograft samples cause false positive mutation calls by mouse DNA contamination in the sequenced sample. Many research labs are using cancer derived cell lines and/or cancer cells grafted to immuno-compromised mice$^{30,32}$. Researchers use these pre-clinical tumor models to examine the effects of new cancer treatment. While they have been using these pre-clinical tumor models for years to test new drugs, only a few studies using pre-clinical tumor models have examined the validity of these tumor models at the genetic level$^{24,25,33,34}$. The TCGA has identified potential drug target pathways for GBM and before testing new treatments on patients they will be tested on pre-clinical tumor models. Some of the problems with trying to identify the validity of the tumor models in order to test new potential drug targets are the errors caused by mouse DNA contamination in the NGS data, how to compare two tumor samples derived from the same tumor cells, and more. Therefore developing methods that are designed to analyze NGS data from pre-clinical tumor models and test their validity is necessary to the development of new treatments for cancers like GBM.

After determining that the cancer treatments developed from pre-clinical tumor models should work on the primary tumor the treatment then moves into the clinic. In order to obtain accurate results from clinical trials of the new cancer treatments the clinic needs to first determine if the patient’s cancer is right for the new treatment. The clinic
determines which treatment option is optimal based on if the genes/pathways altered by
the somatic events in the patient’s tumor match those targeted by the new treatment.
However determining all of the actionable genes/pathways that are altered by somatic
events in a patient’s tumor is currently not done in the clinic. There are biomarkers that
are tested for some of the actionable genes in the clinic but most of the tests only check
for somatic events in one gene at a time. On the other hand, current NGS methods,
whole-exome and whole-genome sequencing, look for somatic events in most if not all of
the actionable genes. However these methods a very expensive and fail to identify
somatic events that are in 1-10% of the tumor cell population. Unlike the samples used to
identify the new drug targets/treatments, patients in the clinic often have low tumor
cellularity and tumor subclones. These sub-clones could contain somatic events that make
the tumor cells susceptible or resistant to a given cancer treatment. Therefore the clinic
needs a new single test that can identifying somatic events in 1% of the tumor cell
population and targets most if not all of the actionable genes.

With the progress of NGS technologies, genomics, and targeted therapies, an
increasing number of cancer somatic mutations are becoming clinically actionable:
predictive of drug sensitivity or resistance. However as mentioned above, clinical
samples are often suboptimal for their comprehensive detection. There are currently
multiple methods to extract multiple small segments of the genome for sequencing. One
method is Ultra-deep targeted sequencing (UDT-Seq) is an assay combining microdroplet
PCR amplification of exonic sequences followed by a direct, oriented sequencing at high
depth of coverage, therefore allowing the detection of low prevalence mutations\textsuperscript{35}. This
method can be used in the clinic for minimal cost to detect low frequency somatic events
in actionable genes instead of using multiple tests for only a few actionable genes. The drawback for using these new methods in the clinic is that current sequencing analysis tools were developed to analyze whole genome and exome shotgun sequencing and not amplicon sequencing data. These tools do not take advantage of UDT-Seq’s and other DNA amplicon sequencing data’s unique data structure; such as knowing that each sequencing read originated from a known strand and location.

With all of the limitations of the current analysis tools developed for non-clinical samples, I have developed a set of analysis software, computational pipelines, and methods to help researchers/clinicians call somatic events in clinical samples. Here I present my work on the analysis of whole genome sequencing of FFPE samples and how the analysis differs from that of frozen samples. The extent of DNA damage caused by formalin varies from sample to sample. I have established a method to determine the extent of DNA damage caused by formalin fixation. With this method, researchers can sequence a small number of reads, as little as 1 million reads, in order to determine the extent of DNA damage caused by FFPE. For researchers who still wish to sequence samples with high amount of DNA damage caused by FFPE, I have also developed a filter designed to remove false positive somatic mutations caused by FFPE DNA damage. Thus the analysis outlined in Chapter One can be used to generate more accurate results of sequencing data from clinical samples stored in FFPE blocks.

In Chapter Two I will present my work on improving the analysis of DNA sequencing data from tumor models. I have found that while most open-source commonly used analysis tools of NGS DNA sequencing data are very robust, however, they still contain mistakes in the code that can lead to inaccurate analysis results. For
example, due to alignment artifacts there are some PCR duplicates that remain after using the available PCR duplicate removal software. These PCR duplicates can lead to false positive somatic variant calls in NGS data. I also propose using a statistical approach to compare the somatic variants between two samples derived from the same tumor to determine if a somatic variant is shared between samples. Furthermore I developed a method to remove mouse contamination from xenograft DNA sequencing reads to reduce the errors in NGS data from xenograft samples. While the methods below may seem common knowledge, at the time there were no published/available software to perform these types of analysis. Therefore the methods proposed in Chapter Two are imperative to the analysis of pre-clinical tumor models.

Finally, in Chapter Three I propose a complete analysis package (Mutascope) dedicated to the analysis of ultra-deep targeted sequencing (UDT-Seq) or similar high-depth amplicon sequencing data. This lab technique can be used in the clinic to sequence a set of actionable genes in order to identify somatic variants in as low as 1% of the tumor sample. Mutascope takes advantage of the current sequencing analysis tools while improving upon the areas that are specific to amplicon sequencing data. For example, after alignment to the genome Mutascope separates the sequencing reads with respect to the amplicon and strand of origin. This allows the experimental measurement of an error rate along the amplicons, which is used to improve the variant detection in clinical samples. In Chapter Three I will also demonstrate that Mutascope’s sensitivity and positive predictive value (PPV) is better than the current analysis software available. Mutascope offers a full analysis pipeline specifically dedicated to the identification of low prevalence somatic mutations in high-depth amplicon sequencing data. As a result
Mutascope increases the overall technical performance of such approaches that are currently being implemented in clinical diagnostics laboratories.

Overall I propose different methodologies to remove sequencing errors caused by various sample preparation and sequencing sources in order to identify functional somatic variants in clinical samples. I analyze and evaluate cancer specimens processed in FFPE, pre-clinical tumor model specimens, and low tumor cellularity and/or tumor subclones with the use of next generation sequencing data. With the study designs proposed, I hope to optimize analytical methods to improve the data processing of cancer study using next-generation sequencing data.
Chapter 1: Identifying somatic DNA mutations in sequencing data from formalin fixed paraffin embedded (FFPE) tumor specimens

1.1 BACKGROUND

To date massively parallel sequencing of cancer genomes has largely been performed using flash frozen tissue or immortalized cancer cell lines\(^{25,28-31,36}\). These studies have provided tremendous insight into the types of mutations and genomic rearrangements that occur in cancer cells. However, limiting sequencing to flash frozen tissues restricts the types of important clinical questions that can be addressed\(^37\). Since formalin fixation and paraffin embedding (FFPE) has been the standard sample preparation for pathologists for decades, the ability to perform massively parallel sequencing of FFPE samples would open up large archived tumor specimen collections. As these large archives frequently have historical records of patient progression and outcome, this would allow for powerful retrospective studies exploring DNA changes that influence disease progression.

RNA isolated from FFPE samples is commonly used for genome-wide expression studies\(^{38-40}\), however performing whole-genome analyses of DNA isolated from FFPE samples has two major application-specific challenges. First, is the fact that tumors of biological and clinical interest stored in blocks are often contaminated with normal stroma, and thus dissection, which is not easy to perform in blocks, is required to enrich for tumor material. A second challenge is the fact that formalin-fixed tissues exhibit a higher frequency of non-reproducible DNA sequence alterations than frozen tissues. This is likely due to formalin cross-linking of cytosine nucleotides on either strand, resulting
in Taq polymerase during PCR not recognizing the cytosine and incorporating an adenine in place of a guanosine causing an artificial C>T or G>A mutation \(^{41,42}\). Previous studies have successfully isolated DNA from FFPE tissue samples stored in paraffin blocks and performed targeted sequencing of single genes\(^{43,44}\) or whole exome sequencing\(^ {45}\). In a few instances, sequencing was extended to the whole genome but was limited to copy number analysis or high-level mutational profile analysis\(^ {46,47}\). In the Kerick study artificial mutations resulting from the formalin fixation process were observed in the sequence data by comparison to matched frozen tissues but methods for removing these false positive calls in the analysis steps of the sequence data were not presented.

In the work presented below, we sequence DNA isolated from two FFPE triple-negative breast tumors archived as 5 µm sections as well as their matched germline DNA. As the tumor was mounted on slides it was straightforward to identify and isolate DNA from areas containing greater than 80-85% malignant cells. By characterizing the patterns of DNA mismatches in the FFPE tumor sequencing reads, we determined that one of the samples was more heavily damaged by fixation than the other and propose guidelines for a rapid FFPE integrity test. We then call somatic variants and implement original filters to remove false positive calls specifically resulting from the formalin fixation process, thus leading to a set of high-confidence somatic mutations in each of the tumors. Finally, we identify a set of mutations of potential functional importance in the progression of the disease (or lack thereof) in each of the two cases.

1.2 METHODS

**Patient Information:**
From the Women’s Healthy Eating and Living (WHEL) cohort we identified two female non-Hispanic white patients (06408 and 02542) diagnosed with Stage III histologic Grade III infiltrating ductal triple negative breast cancer in 1999 and 1995 at the ages of 38 and 30, respectively. All patients provided written informed consent for enrollment in the WHEL Study and for related genomic studies. Triple negative breast cancer indicates that the estrogen and progestosterone receptor staining on tumor tissue was negative and Her2neu over-expression was not observed. Both patients received adjuvant chemotherapy and local regional radiation therapy. Patient 02542’s tumor metastasized 18 months after initial diagnosis and she died shortly afterwards. Patient 06408 is still alive without recurrence as of 2006. The patients underwent curative intent surgical resection and breast tumor material not needed for diagnosis was formalin fixed, embedded in paraffin, sectioned at 5 µm thickness and mounted on slides. Germline DNA was extracted from peripheral blood mononuclear cells (PBMC).

**DNA isolation**

Areas of tumor cells on a hematoxylin and eosin (H&E) stained 5 µm FFPE section were identified and marked by a pathologist allowing the collection of malignant cells with a greater than 80-85% purity (Figure 1.1). Additional tumor material from an adjacent unstained section was isolated by scraping the area corresponding to the marked section with a sterile scalpel. DNA was isolated from the FFPE specimens using BiOstic FFPE Tissue DNA Isolation kit (MO BIO, Carlsbad CA, USA). The samples were heated at 55°C for an hour in an optimized wax melting buffer and Protease K to completely digest the tissue. Samples were heated at 90°C for 1 hour to remove protein-DNA cross-links, purified on a silica spin filter, and eluted with 10mM Tris pH 8.0.
Tumor cell counting

The H&E stained slides were used to estimate the number of tumor cells from which DNA was isolated (Figure 1.1). DNA was isolated from unstained 5 μm thick sections approximately 1.0 cm² and 2.0 cm² areas for samples 06408 and 02542, respectively. We used a Nikon Eclipse E600 microscope to take images of the cells and processed the images with MetaMorph 7.7 (Molecular Devices, Sunnyvale CA, USA). Six random fields within the marked areas were taken. We calculated the number of nuclei in each random field to get an approximate number of cells per slide. To count the cells we first separated the constituent blue, red, and green channels from each of the 24 bit RGB images. Only the blue channel was used to count the number of nuclei in the image. Nuclei were selected by setting the appropriate intensity threshold. The resulting nuclei were filtered by area to remove noise, and counted using the morphometry tool in Metamorph. The number of nuclei was used to calculate the average cell density per image, which was used to extrapolate the number of cells used for sequencing. The area of each image was 1360 pixels by 1024 pixels, with 1 pixel = 0.334355 μm.

Sequencing

Purified tumor and germline DNA were directly used as starting material for SOLiD fragment library preparation (Life Technologies, Carlsbad CA, USA) following manufacturer’s recommendation. DNA was sheared to approximately 150 bp using the Covaris S2 system standard fragmentation conditions recommended in the SOLiD4 Library Prep User Guide. After DNA end-repair, P1 and P2 adaptors were ligated, the adaptor-ligated DNA underwent nick translation and then amplification with 6 and 8 PCR cycles for germline and tumor DNA respectively, using Library PCR primer 1 and 2, and
Platinum PCR amplification mix. Purified library was quantified by TaqMan assay and used for preparing SOLiD templated beads. Each sequencing run resulted in ~500 million raw 50 bp color-space reads per slide. The samples were sequenced over several runs each using both SOLiD3+ and SOLiD4 platforms generating between 1.3 to 3.1 billion total raw reads per sample (Supplemental Table 1).

Genotyping Array Data Generation and Analysis

Germline DNA was genotyped on the Illumina Omni 2.5 M array and processed using GenomeStudio (version 2010.3) using standard methods. Genotypes were exported into reference genome PLUS orientation (build hg19) based on HumanOmni2.5-4v1_D.bpm. As the content on this array contains new SNPs that are not present in dbSNP 132 and were not always named according to dbSNP identifiers, we verified that all positions were present and consistent with dbSNP 132. We converted 1000 Genomes Project SNPs (kgp identifiers) to rsIDs by matching chromosome, position, and alleles in dbSNP132. We excluded 17,959 1000 Genome Project SNPs that were duplicates of SNPs with rsID identifiers, 11,536 SNPs that had more than 2 alleles in dbSNP, and 405,516 SNPs that were not present in dbSNP 132. This resulted in a total of 2,016,729 SNPs. Since the sequencing analysis was performed in the hg18 reference, we converted the positions and orientation of the genotyped SNPs from hg19 to hg18 using the LiftOverVCF.pl script within GATK. The 2,015,517 SNPs with successful coordinate conversion were used in subsequent analysis.

Calculating Concordance between genotyping array and sequencing data

To determine concordance we used the genotypes of the 2,015,517 SNPs described above and the genotypes called in the sequencing data passing Filter 1.1 (see
below) that had at least the indicated coverage (Supplemental Table 4). We calculated the
total number of the genotypes (homozygous reference, heterozygous, and homozygous
alternate) called in the sequencing data that agreed with the genotypes called by the array
and divided by the total number of genotypes called in both data sets.

**Initial Sequence Data Analysis**

*Alignment:* All raw color-space reads were aligned to the human genome
reference sequence (hg18), limited to chromosomes 1-22, X and Y, as well as
mitochondrial genome. The alignment was carried out using BFAST v0.6.1c with default
masks and parameters, except for –M=384 and 10 in the match and local alignment steps,
respectively, and –K=100 in the match step\(^51\). We identified reads originating from
potential PCR duplicate fragments (referred to as duplicate reads) as mapping to the same
location and showing an identical strand orientation and sequence in the first 40
nucleotides. For all duplicate reads, we kept the read with the highest quality score. The
reads were then subjected to local realignment using GATK IndelRealigner\(^50\), to improve
the detection of insertion-deletions (indel) and remove false positive single nucleotide
variants (SNVs) within 200bp of indels.

*Merging of Replicates:* Two independent libraries were generated and sequenced
for both tumor samples 06408 and 02542. The sequences generated from these technical
replicates had similar alignment efficiencies and overall quality metrics (Supplemental
Table 1) without any obvious bias, thus we merged the BAM files resulting from the
alignments and used the consolidated data in the rest of our analysis.

*Coverage:* The coverage was calculated by using SAMTools v0.1.8-13\(^32\) “pileup”
command and custom perl scripts. The normalized coverage was calculated by dividing
the coverage at each base by the average coverage across the genome for each sample (Figure 1.2).

**Mismatches:** To look for potential DNA damage caused by the formalin fixation process, we analyzed the number of mismatches in the mapped reads (Figure 1.3A, Supplemental Table 2a). A mismatch is defined as any base substitution within an aligned read. The number of mismatches within realigned reads was calculated by using the MD field in the SAM file format and custom programs. The MD field characterizes the location, number, and type of mismatches, or differences, a read has with the reference sequence.

**Calculation of Global Nucleotide Mismatch Rates:** We determined the global nucleotide mismatch rate profile for sequencing reads in each tumor sample across all 6 nucleotide substitution types; A·T>C·G, A·T>G·C, A·T>T·A, C·G>A·T, C·G>G·C, and C·G>T·A (Figure 1.4 and Supplemental Table 3). To do this, we investigated a set of high confidence homozygous reference sites, for each patient, derived from a random set of reference loci across the genome. These homozygous reference sites were chosen by first removing all variant positions passing Filter 1.2 in both matched germline and FFPE tumor samples (see below). We then removed all sites that are variant in dbSNP132 and/or the 1000 genomes project. From the remaining homozygous reference loci we randomly selected 4 sets of 100,000 A, T, C, and G sites that had at least 3X coverage in the sample, making a total of 400,000 random loci selected per sample. In each sample, the expected global nucleotide mismatch rate for each substitution type $i \rightarrow j$, $\hat{P}_{ij}$, was then calculated by summing the number of mismatches for a given substitution type and dividing it by the total coverage at the reference site. For example, for the substitution
type A·T > C·G, we summed up the number of times we saw an A>C or T>G substitution, and then divided by the total coverage obtained by summing over all 200,000 reference A and T sites.

**Somatic Variant detection procedure**

**Step1: Variant Calling**

In each sample, we called the variants from the consensus model generated by SAMTools v0.1.8-13 with the following two modifications: 1) to correct for the under calling of homozygous alternate alleles, we set -r to \(7.0 \times 10^{-7}\); and 2) to scale the mapping quality to the BFAST standard, we set -M option to 255.

*Filter 1.1: SAMTools varFilters:* Our first filter removes low confidence variants. Variants were filtered using samtools.pl varFilter command with the following parameters: 1) Minimum Root Mean Square of base quality (RMS) set to 43; 2) Minimum consensus quality set to 20; and 3) the SNP quality set to 50.

*Filter 1.2: Coverage thresholds:* We next filtered to remove false positives caused by too low or too high sequence coverage. To obtain the optimal minimum and maximum coverage thresholds for calling variants we used the set of 2,015,517 loci assayed by the genotyping array to maximize the concordance between the array-based genotype calls and the sequence-based genotype calls, for each patient (both germline and tumor). Due to limited amount of FFPE DNA to carry out genotyping, we compared the tumor FFPE sequencing variant calls to the matched germline array genotypes. The results are presented in Supplemental Table 4a and 4b. We determined that removing positions with less than 5x and 10x for germline and FFPE tumor samples, respectively, and more than
100x depth of coverage optimized the concordance while still being able to call somatic variants in ~80% of the FFPE tumor genomes (Supplemental Tables 4a and 4b).

**Step 2: Identification of Somatic Variants**

We used custom programs to compare the variants called in Step 1 from the germline and FFPE tumor samples for both patients. A variant was called somatic if it passed the following successive filters:

*Filter 2.1: High Quality in matched germline and tumor samples.* This filter removes genomic positions of low quality in either germline or tumor samples. For each subject, we removed the genomic positions that did not pass Filters 1.1 and 1.2 in both germline and tumor samples. This step removes variants that cannot be confidently called somatic due to poor quality or coverage in either sample.

Filters 2.2 and 2.3 below remove potential germline variants.

*Filter 2.2: Novel Variants:* This filter removes previously identified variants present in public databases. We filtered somatic variants in the tumor samples that correspond to known variants present in either dbSNP132 (updated on Mar. 18th 2011)\(^{54}\) or the 1000 genomes project (updated on July 2010)\(^{49}\).

*Filter 2.3: Somatic Variants:* This filter removes variants that either are in the germline sample or have supporting reads in the germline sample: 1) all loci called variant in both the FFPE tumor DNA and the matched germline DNA; and 2) all tumor variants for which 2 or more sequence reads carrying the alternate allele are present in the germline data.

*Filter 2.4: High supporting read diversity:* This filter removes variants with biased read diversity: Duplicate sequencing reads carrying an error can result in false
positive calls. Although duplicate reads were initially removed after alignment, here we increase the stringency for reads supporting alternate alleles in candidate somatic variant positions. Filter 2.4 removes candidate somatic variants supported by reads with less than 3 different start positions.

**Filter 2.5: Normal local mismatch rate:** This filter removes variants in regions with significantly elevated local mismatch rate: The accuracy of Next Generation Sequencing data is very sensitive to sequence context (low-complexity, repeats, di/tri-nucleotide composition) as well as composition (percent GC). We empirically estimated the local mismatch rate at each somatic variant position (see method below) and tested whether the alternate allele frequency supporting the candidate somatic variant was significantly above the expected local mismatch rate (Q score). We removed any variant where a Q score was within the 90th percentile of the Q score distribution of a gold-standard set of heterozygous variants.

**Filter 2.6: Unbiased global nucleotide mismatch profile:** This filter uses the global nucleotide mismatch rates to remove variants supported by significantly biased calls. The formalin fixation introduces a bias in the type of nucleotide substitutions observed\(^1\) (Figure 1.3C). We used the global nucleotide mismatch rate profiles to distinguish candidate somatic variants from random substitutions that result from the fixation procedure. For each genomic position passing Filter 2.5, we calculated a *post hoc* p-value of a \(i \rightarrow j\) substitution using the binomial distribution \(\text{Bin}(x, n, \hat{p}_{(i,j)})\), where \(n\) is the total number of reads covering the position, \(x\) is the number of reads with the alternate allele \(j\), and \(\hat{p}_{(i,j)}\) is the global nucleotide mismatch rate (see *Calculation of Global Nucleotide Mismatch Rates* above) for the given base substitution \(i \rightarrow j\). We
removed all positions where the alternate allele frequency is not significantly different from the expected global nucleotide mismatch rate using ranked p-values corrected for a false discovery rate (FDR) of 0.05 according to Benjamin & Hochberg\textsuperscript{55}.

**Testing for elevated local-mismatch rate**

For the set of candidate somatic variants passing filter 2.4, we calculated the alternate allele frequency (AAF) which is the ratio of alternate allele supporting reads to the total number of reads at that position. We then calculated the local-mismatch rate (LMR) from positions 10 bp upstream and 10 bp downstream of the candidate variant position \( \text{LMR} = \frac{m}{n+m} \), where \( m \) is the number of positions matching the reference and \( n \) the number of mismatched (excluding the candidate variant position itself). Notably, mismatches include nucleotide substitutions, insertions, and deletions. For example, a deletion of 3 bp would result in 3 mismatch counts. Finally, we inferred a Q score = (AAF-LMR) at each position. We generated a gold standard set of heterozygous variant positions by selecting the 1,229,492 and 986,314 heterozygous SNPs from patient 06408 and 02542, respectively that were called in the sequencing data and are present in dbSNP132 and/or the 1000 genomes project. We calculated the Q scores of these gold-standard variants in the tumor FFPE DNA and compared their distribution to the candidate somatic variants Q score (Filter 2.5, Figure 1.5).

**Estimation of alternate allele under-calling:**

To estimate the false negative rate in the sequencing data for each sample, we determined the fraction of genotyping array alternate allele sites not called in the sequencing data that passed Filter 2.1. The numerator (alternate allele sites not called) was calculated by summing the number of sites called as AB by the genotyping array and
as AA in the sequencing data; plus the sites called as BB by the genotyping and AA or AB in the sequencing data. The denominator (number of possible sites with an alternate allele) was calculated by summing all AB and BB sites in the genotyping array excluding sites that were called missed variant (MV) or missed called (MC) in the sequencing data (Supplemental Table 5).

**Annotation of Somatic Variants**

We used the SeattleSeq Annotation server (http://gvs.gs.washington.edu/SeattleSeqAnnotation/) for functional annotation of somatic variants called in FFPE tumor samples 06408 and 05217. To identify genes carrying somatic mutations of potential importance for breast tumor initiation and progression we downloaded the cancer gene census list, updated 2011-03-22, consisting of 457 genes and created a list of DNA damage repair genes from the Gene Ontology database. Briefly, by searching for “DNA damage repair” in the GO terms of “Biological process” we identified 5 GO terms and 1,049 genes.

**Analysis of Illumina sequencing reads for FFPE DNA damage**

We downloaded publicly available Illumina sequence data of 89 FFPE non-small cell lung tumors. The sequencing reads were aligned to the human reference genome (hg19) using BWA v5.9 with default parameters, except for a seed length of 25. BWA is more stringent than BFAST in aligning reads that contain mismatches; therefore samples with high FFPE damage are expected to have fewer Illumina reads aligning to the genome. For this reason, to estimate the extent of DNA damage caused by FFPE we calculated the alignment rate and percent of aligned reads with greater than or equal to 2 mismatches. We used a k-means clustering algorithm on the alignment and mismatch
rates to separate the 89 tumor samples into two groups; one group contained 11 samples and the other contained 78 samples.

1.3 RESULTS

We sequenced two triple-negative breast cancer tumors (WHEL Study samples 06408 and 02542) and matched patient germline DNA. The tumor samples had been formalin fixed and paraffin wax embedded (FFPE) and stored as 5 µm section for 11 and 16 years respectively, before DNA was isolated for our study. DNA was isolated from approximately a 1 cm² area of 85% tumor cellularity containing about 5.4x10⁵ cells from sample 06408 and from approximately a 2 cm² area of 80% tumor cellularity containing about of 1.3x10⁶ cells from sample 02542 (Figure 1.1). We performed technical replicates (DNA isolation, library construction and sequencing) for both tumor samples 06408 and 02542. After read alignment, duplicate reads removal and local realignment, the data resulting from the technical replicates were checked for consistency before being merged into a single dataset for further analysis (Supplemental Table 1). This resulted in a coverage depth respectively of 13X and 23X for patient 06408 germline and FFPE tumor DNA and 12X and 22X for patient 02542 germline and FFPE tumor DNA (Table 1.1). The coverage depth distribution across the genome was similar between FFPE tumor and germline samples (Figure 1.2), indicating that the FFPE process did not create any large-scale bias affecting the ability to examine specific intervals of the genome for somatic variants.

Characterizing formalin fixation induced DNA damage
The DNA damage caused by the FFPE process is expected to lead to a high number of mismatches in the aligned sequencing reads\textsuperscript{41,42} confounding the identification of DNA variants. However, the FFPE damage occurs at different nucleotide positions in different cells of the sample and thus has a random distribution across all DNA sequencing reads. By analyzing the combined signal of mismatches in sequence reads of the FFPE tumor sample, it is possible for the pattern of random FFPE-induced damage to be recognized, and then corrected for in the data analysis. Therefore, in order to comprehensively characterize FFPE induced errors, we analyzed mismatches in each read prior to consensus variant calling. The FFPE tumor DNA showed reduced alignment rates (54-61\%) as compared to the germline (66-67\%) (Supplemental Table 1). Moreover, the proportion of reads with $\geq$1 mismatch was greater in both of the FFPE tumor samples (~32\% in 06408 and ~51\% in 02542) when compared to their corresponding germline samples (~21\%) (Figure 1.3A; Supplemental Table 2). These data are all consistent with formalin fixation induced DNA damage resulting in the FFPE tumor aligned sequence reads having a higher number of mismatches.

Interestingly, FFPE tumor sample 02542 had 1.5 times more reads with $\geq$1 mismatches than FFPE tumor sample 06408. This greater number of mismatches was consistent across technical replicates (Supplemental Table 2), suggesting that the observation was not an artifact of the DNA isolation or library preparation process but that the extent of DNA damage due to formalin fixation is greater in the FFPE tumor sample 02542. Mismatch distribution differences between the two FFPE tumor samples were apparent by examining a random set of 50 million, 5 million, and 1 million non-filtered sequence reads from the germline and FFPE tumor samples of both patients.
(Figure 1.3A and Supplemental Table 2). This implies that by sequencing as few as 1 million reads per sample, one can estimate the extent of DNA damage in a FFPE tumor from the mismatch distribution. To further investigate the ability to assess the extent of DNA damage caused by FFPE in low coverage data we downloaded publicly available Illumina sequence reads from 89 FFPE tumors; each sample has about 1 million reads. We aligned the sequence reads to the human reference genome and then calculated the fraction of reads that aligned and the mismatch rate of the aligned reads. Of the 89 samples, 11 had poor mismatch and alignment rates suggesting that they have a significant amount of DNA damage from FFPE processing (Figure 1.6). The other 78 samples had moderate to good mismatch and alignment rates suggesting that the FFPE DNA damage was minimal. Overall these results suggest that low-coverage datasets can be used to assess the integrity of the FFPE tumor DNA and thus can serve as an important quality control step before performing costly whole genome sequencing.

We next determined the global nucleotide mismatch rate in the DNA sequencing reads (Figure 1.3B), as well as the profile of each of the six different types of substitutions (Figure 1.3C). To estimate the global nucleotide mismatch rate profiles we focused on 4 sets of 100,000 sites each called as homozygous reference A, T, C, and G in each patient’s germline genome (based on random high confidence reference sites across the genome) and had at least 3X coverage in the matched FFPE tumor. While the global nucleotide mismatch rates were similar in the germline DNA of the two patients (~11 x 10^{-3}), the global nucleotide mismatch rates in the FFPE samples were substantially higher (1.6 and 2.9 fold higher than in the germline, for patients 06408 and 02542 respectively). The higher relative global nucleotide mismatch rate in the 02542 FFPE tumor sample
compared to the 06408 FFPE tumor sample is consistent with a greater amount of DNA damage. Across the six substitution types, the FFPE tumor samples have a greater global nucleotide mismatch rate than the germline samples (Figure 1.3C). The increase in the global nucleotide mismatch rate was particularly prominent for C·G>T·A substitutions, which was 1.5 and 1.8 fold higher than the other substitution types in tumor samples 06408 and 02542, respectively. This is consistent with the types of DNA sequence read mismatches expected to result from formalin induced cross-linking of cytosine nucleotides. The atypical global nucleotide mismatch rate profiles of the FFPE tumor sample suggests that the majority of the DNA sequence read mismatches are due to the formalin fixation process rather than the oncogenic process. Consequently, we used the atypical global nucleotide mismatch rate profiles in the FFPE tumor samples to better distinguish high-confidence somatic variants from formalin fixation induced mismatches (see Filter 2.6 in methods and below).

**Variant calling and initial quality assessment**

As described in Methods we called variants using SAMTools v0.1.8-13\textsuperscript{52} and then applied two filters to remove low confidence variants (Filter 1.1) and to remove false positive variants caused by genomic regions with too low or too high sequence coverage (Filter 1.2). We used the genotype information obtained from the Illumina Omni 2.5 array analysis of each patient’s germline DNA to assess variant calling performance and optimize additional standard and novel filters. After applying Filters 1.1 and 1.2 we called 84-95% of the array’s SNP positions in all four samples using the sequencing data (Supplemental Table 5). Of note, this estimation of variant detection sensitivity is likely an overestimate as variants analyzed on genotyping arrays are easier to detect using next
generation sequencing than variants not amendable to array analysis\textsuperscript{60}. The genotype concordance between the array and germline variants was 96.9% and 96.8% respectively in patients 06408 and 02542. For patient 06408 the corresponding FFPE tumor DNA sample had similar concordance with the genotyping array (96.6%); however for patient 02542 the FFPE tumor DNA sample had lower concordance (92.7%). This higher discordance is primarily the result of under-called alternate alleles, which is more prominent in the 02542 FFPE tumor sample (~21%) than in the matching germline sample (~8%) (Supplemental Table 5). For patient 06408 the rate of undercalling alternate alleles was similar between the FFPE tumor (~9%) and the germline sample (~8%). A variety of reasons likely underlie this increased under-calling of the alternate allele in the 02542 FFPE tumor sample including biological reasons, such as deletions resulting in loss of heterozygosity.

Because the amount of DNA isolated from the FFPE tumor samples was low, we examined whether or not contaminating DNA was introduced during the library preparation. For both patients, the FFPE tumor variants were more concordant with the genotyping array results of the matched germline sample than with the other patient’s germline sample (93-97% vs 69%, Supplemental Table 6). These data suggest that a contaminating DNA source was not introduced during library preparation as the cross-sample concordance between the germline array genotypes and the FFPE tumor sequence genotypes would have been lower than what we observed and likely have had an expected inflation of heterozygous calls (Supplemental Table 5). Thus we are confident that we sequenced DNA isolated from the FFPE tumor 5 µm sections.
To characterize the bias in variant calling introduced by the formalin fixation process we compared variants called in the germline and matched FFPE tumors. In each of the four samples, we identified ~1.8-2.1x10^6 variants with high sequence quality (Figure 1.7, passing Filter 2.1). Consistent with the expected findings from the sequencing of a Caucasian individual^50, ~95% of the germline variants have been previously observed and are in public databases (Figure 1.7, passing Filter 2.2). The 02542 FFPE tumor sample had a higher number of novel variants (3.8X) than the 06408 FFPE tumor sample or the matched germline samples. These variant data are in alignment with the observed higher global nucleotide mismatch rate suggesting that the 02542 FFPE tumor sample has extended damage from formalin fixation. Additionally, it is important to consider the fact that these higher number of novel variants may partially be due to an increased number of somatic mutations in the 02542 FFPE tumor sample. We also observed a marked difference in the distribution of the six nucleotide substitution types of the variants passing Filter 2.1 in the 02542 FFPE sample as compared to the matched germline and the 06408 FFPE tumor sample (Figure 1.4).

While the variant substitution profiles in the 06408 FFPE tumor DNA is largely similar to that of the matched germline DNA for most categories of substitution types (Figure 1.4 and Supplemental Table 3), we noted a highly biased profile in the novel heterozygous variants present only in the FFPE tumor DNA of patient 02542; the proportion of C·G>T·A substitutions is 1.9 times higher than that observed in the matched germline and 2.7 fold higher than what is observed in 06408 FFPE tumor DNA. This biased C·G>T·A variant substitution rate is consistent with our previous observation of the increased global nucleotide mismatch rate profiles in the 02542 FFPE tumor (Figure
1.3C). We note that the transition to transversion ratio of the paired known variants (~2.2) is close to the expected value\textsuperscript{50} whereas heterozygous novel variants that are uniquely present in the tumor samples have a substantially lower value (~0.8 -1.7), indicative of a low-confidence for this latter class of variants (Figure 1.4).

**Somatic Variant Calling and Filtering**

Following the above quality assessment, we devised several successive filters to derive a set of high-confidence somatic variant calls. After removing germline variants (Figure 1.7, Filter 2.3) there are 55,551 and 290,341 candidate somatic variants for tumor samples 06408 and 02542, respectively, which is substantially higher than previous reports in breast cancer\textsuperscript{25,33}. Despite removal of duplicate sequencing reads after alignment we noticed that a significant proportion of candidate somatic variants were supported by reads with fewer than 3 different start positions. We believe that the initial filter did not remove all duplicate reads due to the presence of variable insertions and deletions. A more stringent filtering of these duplicate reads (Figure 1.7, Filter 2.4) resulted in a further reduction in the number of somatic variants.

False positive mutations as well as real cancer somatic mutations are generally expected to be heterozygous in the tumor DNA. To further enhance our detection of high-confidence somatic mutations, we compared the alternate allele read frequency at all somatic variant positions to a standard set of germline heterozygous variants from the same patient. The alternate allele read frequency of germline heterozygous variants had a median of 42%, while the candidate somatic mutations alternate allele read frequency was 20-26% (Figure 1.5A). Upon closer inspection, we noticed that the somatic mutations with relatively low alternate allele read frequencies were frequently located in regions
with elevated local mismatch rates, probably resulting from alignment or sequencing artifacts that were not corrected through local realignment (Figure 1.5B). We confirmed that the local mismatch rate is higher for somatic variants than the standard set of variants (4.6 vs 3.1 X 10^{-2} on average). We filtered the candidate somatic mutations for which the alternate allele frequency was not significantly higher than the local mismatch rate calculated 10bp upstream and downstream from the position considered (Figure 1.7, Filter 2.5). After applying this filter, the alternate allele read frequencies and local mismatch rates for somatic mutations in both FFPE tumor samples is closer to the standard set of heterozygous variants (Figure 1.5). This filtering step resulted in 19,750 and 35,733 candidate somatic variants in patient 06408 and 02542, respectively. Finally, Filter 2.6 takes advantage of the biased global nucleotide mismatch rate profiles that we observed in the FFPE tumor DNA sequence reads (Figure 1.3C) to identify a set of high-confidence somatic variants. Here we filtered candidate somatic variants for which the alternate allele read frequency is not significantly different from the global nucleotide mismatch rate. This resulted in 19,176 and 22,524 high-confidence somatic variants in sample 06408 and 02542 respectively. Tumor samples are typically heterogeneous composed of a mixed population of different clones. Given that the minimum alternate allele frequency needed to call a high-confidence somatic variant after applying Filter 2.6 is ~18-21% with the mean around 32-34%, thus in the best case scenario we would be able to call a heterozygous mutation found in about 50% of tumor cells in a sample with intra-tumor heterogeneity.

Examining the six substitution types (Figure 1.8), reveals that this specific filter diminished the C·G>T·A substitution bias characteristic of formalin fixation induced
DNA damage and resulted in a distribution of substitution types similar and more balanced for the 02542 FFPE tumor sample. While sample 06408 only had 3% of its candidate somatic variants filtered by Filter 2.6, 37% of candidate somatic variants in sample 02542 were removed (Figure 1.7). This supports our previous statement that sample 02542 had greater FFPE induced DNA damage causing an increase in the number of false positive somatic variants. On the other hand, both FFPE tumor samples had more than 50% of their candidate somatic variants filtered by Filter 2.5 which removes false positive variants caused by sequencing and alignment errors.

To further examine the effects of Filters 2.4, 2.5, and 2.6 on the total number of candidate somatic mutations and the distribution of substitution types we applied these filters in different combinations and determined that all three filters are necessary (Figure 1.9). These results show the importance of Filters 2.4, 2.5 and 2.6 as FFPE tumor samples have increased alignment errors compared to matched germline samples most likely due to both somatic mutations and formalin fixation induced DNA damage. The succession of filters (2.4 to 2.6) removed ~65% and 92% of the candidate somatic variants in 06408 and 02542 respectively (Figure 1.7). In a recently published framework for somatic variant calling proposed by the Broad Institute, 62% of novel variants were filtered\textsuperscript{50}. The higher fraction of candidate somatic variants filtered in our study is expected, as our goal is to filter out false positive calls due to the formalin fixation induced DNA damage in both FFPE tumors samples.

**Somatic Coding Variation**

The final set of high-confidence somatic mutations contained 19,176 and 22,524 variants in tumor samples 06408 and 02542, respectively. Of those, 268 and 423 variants
were coding or affect splice sites (Figure 1.10 and Supplemental Table 7&8). These numbers are in agreement with previously sequenced whole genomes of breast cancer\textsuperscript{25,33}, which suggests our filtering process has adequate stringency.

We examined 457 genes from the Cancer Gene Census\textsuperscript{57} and 1,049 genes involved in DNA damage repair for somatic coding variants. Sample 06408 had 8 high-confidence somatic mutations in 8 genes (1 nonsense and 7 missense) whereas sample 02542 had 16 high-confidence somatic mutations in 16 genes (1 nonsense, 12 missense, and 3 coding-synonymous) (Table 1.2). A number of these changes are of potential biological interest. Both patients carry variation in TP53: sample 06408 carries a heterozygous nonsense mutation in TP53, suggesting the inactivation of one copy of this tumor suppressor gene and sample 02542 carries a somatic missense mutation. Sample 06408 also carries a heterozygous missense mutation in NOTCH1 which has been shown to be a reoccurring mutation in chronic lymphocyte leukaemia, lung squamous cell carcinoma, and breast cancer\textsuperscript{29,61,62}. The nonsense mutation in TP53 together with the missense mutation in NOTCH1 could be driver mutations for sample 06408’s tumorigenesis. Sample 02542 carries missense mutations in both MLL2 and MLL3 which together were recently found as significantly mutated in 16% of childhood medulloblastoma cases\textsuperscript{63}.

1.4 DISCUSSION

Genomic translational research faces a scarcity of properly stored and annotated clinical samples. Archived formalin fixed tissues in paraffin blocks offer a unique opportunity to study thousands of samples with extensive clinical records and follow-up information. In our study, we show that it is possible to obtain enough DNA from a single
5 μm FFPE slide (approximately 1–2 cm²) to perform whole genome sequencing of sufficient coverage depth to identify potentially important mutations. The FFPE process combined with long storage times is known to result in DNA fragmentation. We show that for the two breast tumor samples analyzed DNA fragmentation did not produce large biases in coverage depth distribution (Figure 1.2). However, we observed a higher global nucleotide mismatch rate within aligned reads from FFPE tumor samples when compared to matched germline (Figure 1.3A) and a higher base substitution rate across all 6 different substitution types (Figure 1.3C). Consistent with damage due to formalin fixation, we observed this increase was biased towards C·G>T·A mismatches. Interestingly the two samples studied were differentially affected by the formalin fixation, tumor 02542 showing a 1.8-fold increase in the global nucleotide mismatch rate and greater C·G>T·A bias compared to tumor 06408. This discrepancy can be explained by the absence of strict standards in the formalin fixation step, where tissue samples are routinely fixed between 24 and 48hrs but sometimes can be fixed for considerably longer times. The time of the formalin fixation step is not known for the studied samples and not generally included in pathology reports. Another possible explanation could be the size of the tumor tissue, or its density, which also affects the fixation procedure. As formalin fixation induced DNA damage could potentially be so great as to inhibit the ability to analyze an FFPE sample by next generation sequencing we have established a relatively simple test to assess the integrity of FFPE samples. By simply sequencing from 500,000 to 1 million raw reads from a single FFPE tumor one can determine the extent of DNA damage and identify the best preserved samples to conduct larger, more expensive whole genome sequencing (Figure 1.3A; Figure 1.6).
Using a set of innovative filters (Filter 2.4-2.6), we establish a successful method for filtering false positive somatic variants caused by the FFPE damage to the tumor DNA, thus increasing our confidence in the final set of called somatic mutations. It is important to compare our novel filters to existing post-alignment filtering methods such as GATK\textsuperscript{50}. Existing methods filter for poor base quality with a stringent threshold; this is due to the fact that incorrectly called variants are typically caused by low quality sequence data. The fact that FFPE causes random damage, the “errors” do not have poor base quality. Our method filters on the alternate allele frequency without using a threshold for all substitution types; but rather it uses a mismatch error rate across the genome of the given sample. This is important as the amount of FFPE DNA damage varies from sample to sample. To achieve the same goal as our novel post-alignment filters, one could propose applying more stringent criteria to align the reads. Aligners that trim the reads when their mismatch rate becomes too high have been implemented\textsuperscript{64,65}. As a result, the global nucleotide mismatch rate would improve, but at the cost of a reduced effective sequencing coverage depth. Such strategies could also remove \textit{bona fide} somatic mutations surrounded by extensive DNA damage therefore limiting the sensitivity to call variants. A second potential alternate approach for achieving a set of high-confidence somatic mutations in FFPE samples would be to sequence to greater coverage depth. Since formalin fixation is performed on the resected tumor sample and will generally randomly affect different DNA locations in different cells, elevated global nucleotide mismatch rates in DNA sequencing reads should still lead to accurate variant calls at sufficiently high sequencing coverage depth. In our study, the global nucleotide mismatch rate was indeed higher than the variant calling rate,
especially in FFPE tumors (18-32x10^{-3} vs 10-11x10^{-4}). In a recent study of whole-exome sequencing of FFPE tumors, 40-fold coverage was insufficient to filter false positives due to formalin fixation DNA damage identified by the substitution profile and discordance with matched frozen tissue^{45}. Indeed the authors estimate that 80X coverage is required to obtain accurate variant calling in the presence formalin fixation DNA damage. However, for samples such as 02542 in our study with substantial amounts of formalin fixation induced DNA damage, the coverage depths required to overcome the global nucleotide mismatch rates in the sequencing reads to achieve accurate variant calls could be even greater. Thus applying our series of standard and novel filters will likely have utility for identifying high-confidence somatic mutations in FFPE tumor samples even when there is relatively low sequence coverage depth.

In our study we have not analyzed the tumors for somatic events such as chromosomal translocations or large copy number alterations (CNA). Methods developed for this purpose^{66-68}, rely more on the correct mapping of read pairs than accurate sequence. We have only sequenced single reads, and were thus not able to perform this analysis. We believe that the vast majority of the reads mapped in our FFPE tumor samples are mapped at the correct location. However, it is possible that the sensitivity of translocation or CNA detection would be affected as a greater number of reads might have ambiguous mappings due to the mismatches introduced by the FFPE damage. Various distributions of insert size in read pairs, especially large ones (1-10kb) obtained through mate-pair libraries, can also improve the sensitivity of the detection of large deletions. However, the FFPE process fragments the DNA and therefore would not be adequate for such studies.
Overall, our study demonstrates that a methodical characterization and analysis of the sequencing data can reduce the noise resulting from formalin fixation induced DNA damage and lead to calling a high-confidence set of somatic mutations. This opens up the possibility of sequencing huge archives of stored clinical FFPE samples of a variety of cancers. Furthermore, we demonstrate that a limited amount of DNA can be used for a genome-wide deep sequencing analysis, which enables studies on small clusters of tumor cells such residual cancer after treatment or dormant metastases.

Chapter 1, in full, is a reprint of the material as it appears in Nucleic Acids Research, 2012, Vol. 40, No 14 e107. Pub online 6 April 2012. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. Shawn E. Yost, Erin N. Smith, Richard B. Schwab, Lei Bao, HyunChul Jung, Xiaoyun Wang, Emile Voest, John P. Pierce, Karen Messer, Barbara A. Parker, Olivier Harismendy, and Kelly A. Frazer. The dissertation author was the primary author of this paper.
Figure 1.1. Images of H&E stained 5 micron sections of FFPE tumor samples. The areas of high tumor density (80-85%) were marked by a pathologist and used as guidelines for isolating cells on the adjacent unstained slide for DNA isolation.
Figure 1.2. Normalized genomic coverage distribution. The distributions for germline and FFPE tumor reads from patients 09515 and 05127, using non-duplicate reads after local realignment and excluding chromosome Y.
Figure 1.3. (A) Frequency of mismatches within sequencing reads for germline and FFPE tumor samples. The distribution of reads with 0, 1, 2, or ≥ 3 mismatches to the reference genome is shown for all sequencing data (All) and a random subset of 50M, 5M and 1M sequencing reads. (B) Read based global nucleotide mismatch rate for all base substitutions. (C) Read based global nucleotide mismatch rate for each substitution type.
Figure 1.4. Distribution of substitution types for variants passing Filter 2.1 in germline (G) and FFPE tumor (T) samples and called homozygous alternate (Alt) or heterozygous (Het). Variants identified in public SNP repository (Known) or novel for both patients in this study (Novel) or passing in both germline and FFPE tumor samples (Paired) or only in one sample (Unique) are distinguished. The fraction of novel heterozygous variants (C·G>T·A) called between the tumor and germline samples of patient 02542 is substantially different.
Figure 1.5. Effect of Filter 2.5 (Remove variants in regions of high local mismatch rate) on (A) alternate allele read frequency and (B) local mismatch rates. Shown are the gold standard set of heterozygous variants (Standard) and candidate somatic variants (Somatic) ‘Before’ and ‘After’ Filter 2.5, for patients 09515 and 05127. Boxes are the 25 percentile and whiskers extend to 99.95 percentile.
Figure 1.6. Analysis of Illumina sequencing reads of 89 FFPE tumors. A k-means clustering of the alignment rates and the fraction of aligned reads with ≥ 2 mismatches separated the samples into two groups indicated by the red and blue dots. Samples in the red group likely have a greater amount of FFPE DNA damage than the samples in the blue group.
Figure 1.7. Flow diagram describing the number of variants passing each filtering step for both patients 06408 (blue) and 02542 (red).
Figure 1.8. Filters 2.5 and 2.6 remove false positive somatic variants due to formalin fixation and other systematic and random errors in the process. Shown is the fraction of substitution types for somatic variants after Filter 2.4, after Filter 2.5, and after Filter 2.6 for 06408 and 02542 FFPE tumors. After Filter 2.6 the novel somatic variants of substitution type C·G>T·A called in 02542 tumor have a similar profile to that observed for novel germline variants in the matched sample (Figure 1.2).
Figure 1.9. The inclusion of Filters 2.4, 2.5 and 2.6 are necessary to remove potential false positive somatic mutations. Shown are different combinations of Filters 2.4, 2.5 and 2.6. The top line shows the number of somatic mutations present after applying the indicated filters. The total number of candidate somatic mutations and the fraction of substitution types differ depending on the combinations of Filters 2.4, 2.5, and 2.6 applied to patient 06408 and 02542 FFPE tumors. All three filters are needed to obtain the appropriate number of somatic mutations and distribution of substitution types.
Figure 1.10. Annotation of high-confidence somatic variants for A) patient 09515 and B) patient 05127.
Table 1.1. Sequencing Statistics

<table>
<thead>
<tr>
<th>Patient</th>
<th>Sample 06408</th>
<th>Sample 02542</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>Germline</td>
<td>FFPE tumor</td>
</tr>
<tr>
<td></td>
<td>1,352,676,08</td>
<td>2,823,592,37</td>
</tr>
<tr>
<td>Raw color-space reads</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Fraction of reads aligned to hg18</td>
<td>67.2%</td>
<td>59.3%</td>
</tr>
<tr>
<td>Fraction of uniquely(^1) aligned reads (^2)</td>
<td>70%</td>
<td>63%</td>
</tr>
<tr>
<td>Average haploid coverage</td>
<td>12.6X</td>
<td>23.4X</td>
</tr>
<tr>
<td>Fraction of genome covered</td>
<td>88%</td>
<td>89%</td>
</tr>
<tr>
<td>Fraction of genome with (\geq)3X coverage</td>
<td>85%</td>
<td>86%</td>
</tr>
</tbody>
</table>

\(^1\) Reads with only one possible mapping location

\(^2\) Reads after mapping, duplicate removal, local-realignment, and merging technical replicates; excluding chrY
### Table 1.2. High-confidence FFPE tumor coding somatic variants within cancer associated genes and/or DNA damage repair genes

<table>
<thead>
<tr>
<th>Patient</th>
<th>Gene</th>
<th>Chr</th>
<th>Position (hg18)</th>
<th>Germline</th>
<th>Tumor</th>
<th>Mutation type</th>
<th>Amino Acid change</th>
</tr>
</thead>
<tbody>
<tr>
<td>06408</td>
<td>ATRX</td>
<td>chrX</td>
<td>76735852</td>
<td>A/A</td>
<td>A/C</td>
<td>missense</td>
<td>L2027R</td>
</tr>
<tr>
<td></td>
<td>ELN</td>
<td>chr7</td>
<td>73109920</td>
<td>G/G</td>
<td>G/C</td>
<td>missense</td>
<td>A458P</td>
</tr>
<tr>
<td></td>
<td>KIAA154</td>
<td>chr7</td>
<td>138253476</td>
<td>T/T</td>
<td>T/C</td>
<td>missense</td>
<td>Q429R</td>
</tr>
<tr>
<td></td>
<td>MYH9</td>
<td>chr22</td>
<td>35040266</td>
<td>T/T</td>
<td>T/A</td>
<td>missense</td>
<td>K475M</td>
</tr>
<tr>
<td></td>
<td>NOTCH1</td>
<td>chr9</td>
<td>138520141</td>
<td>G/G</td>
<td>G/A</td>
<td>missense</td>
<td>A1343</td>
</tr>
<tr>
<td></td>
<td>NUMA1</td>
<td>chr11</td>
<td>71417948</td>
<td>C/C</td>
<td>C/G</td>
<td>missense</td>
<td>V27L</td>
</tr>
<tr>
<td></td>
<td>NUP214</td>
<td>chr9</td>
<td>132998395</td>
<td>A/A</td>
<td>A/G</td>
<td>missense</td>
<td>D270G</td>
</tr>
<tr>
<td></td>
<td>TP53</td>
<td>chr17</td>
<td>7517747</td>
<td>G/G</td>
<td>G/A</td>
<td>nonsense</td>
<td>R306*</td>
</tr>
<tr>
<td>02542</td>
<td>AKT1</td>
<td>chr14</td>
<td>104312544</td>
<td>A/A</td>
<td>A/G</td>
<td>missense</td>
<td>F161L</td>
</tr>
<tr>
<td></td>
<td>BLM</td>
<td>chr15</td>
<td>89105082</td>
<td>T/T</td>
<td>T/A</td>
<td>missense</td>
<td>F492Y</td>
</tr>
<tr>
<td></td>
<td>CREBBP</td>
<td>chr16</td>
<td>3772787</td>
<td>G/G</td>
<td>G/A</td>
<td>missense</td>
<td>P453L</td>
</tr>
<tr>
<td></td>
<td>EXT1</td>
<td>chr8</td>
<td>118886256</td>
<td>G/G</td>
<td>G/T</td>
<td>missense</td>
<td>D647E</td>
</tr>
<tr>
<td></td>
<td>GNA11</td>
<td>chr19</td>
<td>3070205</td>
<td>A/A</td>
<td>A/G</td>
<td>missense</td>
<td>N246S</td>
</tr>
<tr>
<td></td>
<td>JARID1A</td>
<td>chr12</td>
<td>297581</td>
<td>G/G</td>
<td>G/C</td>
<td>missense</td>
<td>T950R</td>
</tr>
<tr>
<td></td>
<td>LPP</td>
<td>chr3</td>
<td>190066803</td>
<td>G/G</td>
<td>G/T</td>
<td>missense</td>
<td>G511V</td>
</tr>
<tr>
<td></td>
<td>MLL2</td>
<td>chr12</td>
<td>47722022</td>
<td>T/T</td>
<td>T/C</td>
<td>missense</td>
<td>K2043</td>
</tr>
<tr>
<td></td>
<td>MLL3</td>
<td>chr7</td>
<td>151504320</td>
<td>G/G</td>
<td>G/C</td>
<td>missense</td>
<td>Q3051</td>
</tr>
<tr>
<td></td>
<td>PDGFRA</td>
<td>chr4</td>
<td>54824777</td>
<td>G/G</td>
<td>G/A</td>
<td>missense</td>
<td>G185E</td>
</tr>
<tr>
<td></td>
<td>RET</td>
<td>chr10</td>
<td>42921884</td>
<td>G/G</td>
<td>G/T</td>
<td>missense</td>
<td>G308W</td>
</tr>
<tr>
<td></td>
<td>RPN1</td>
<td>chr3</td>
<td>129823703</td>
<td>G/G</td>
<td>G/T</td>
<td>nonsense</td>
<td>C545*</td>
</tr>
<tr>
<td></td>
<td>RUNX1</td>
<td>chr21</td>
<td>35181094</td>
<td>C/C</td>
<td>C/T</td>
<td>coding-</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>STK11</td>
<td>chr19</td>
<td>1171708</td>
<td>G/G</td>
<td>G/A</td>
<td>coding-</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>TP53</td>
<td>chr17</td>
<td>7519259</td>
<td>C/C</td>
<td>C/A</td>
<td>missense</td>
<td>K132N</td>
</tr>
<tr>
<td></td>
<td>ZNF521</td>
<td>chr18</td>
<td>21060818</td>
<td>C/C</td>
<td>C/G</td>
<td>coding-</td>
<td>NA</td>
</tr>
</tbody>
</table>
Chapter 2: Evaluate the genomic validity of tumor derived pre-clinical models using exome sequencing data

2.1 BACKGROUND

The discovery of cancer specific somatic DNA mutations has led to the development of highly effective therapies targeting the corresponding altered protein via monoclonal antibodies or specific inhibitors. These therapies have enhanced activity with a reduced toxicity for the patient in comparison to cytotoxic agents. Recent advances in high throughput sequencing now drives the discovery of actionable mutations in driver genes\textsuperscript{27,69,70}, genes mediating drug sensitivity\textsuperscript{71-73} or supports new indications for targeted therapies\textsuperscript{74}. In 2011, several clinical trials have resulted in the approval of therapies for BRAF\textsuperscript{+} Melanoma (vemurafenib), ALK\textsuperscript{+} Non-Small cell lung cancer (crizotinib) and JAK2\textsuperscript{+} myelodysplasia (ruxolitinib). It is anticipated that the number of clinically approved targeted therapies will increase in the future with our improved ability to discover targets and the effective repurposing of existing drugs for new indications.

Glioblastoma (GBM) is one of the most devastating cancers: it is the most common primary brain malignancy in adults, accounting for over 14,000 deaths per year in the United States\textsuperscript{75}. The standard of care includes surgery followed by chemoradiotherapy. Unfortunately, these treatments are rarely curative and the vast majority of tumors recur locally within a few months. A recent integrated multidimensional genomic analysis has shown that the genetic landscape of glioblastoma is rather heterogeneous with 80\% of the patients affected in one of three main signaling pathways, \textit{TP53}, \textit{PIK3CA} and \textit{RB}\textsuperscript{76}. Importantly, the specific alterations affect different genes in these pathways through various somatic events such as point mutations, copy
number aberrations or transcriptional deregulation. These molecular profiles led to a classification of GBM tumors which have already proven useful in designing more rationalized targeted therapies. An example is the discovery of \textit{IDH1} as a promising new target for younger GBM patients\textsuperscript{77,78}. Current clinical trials in GBM targeting \textit{EGFR}, \textit{VEGF}, \textit{PDGFRA} are all leveraging recent molecular genetic information of GBM\textsuperscript{79}.

The description of the molecular aberrations in tumors has become so comprehensive that we can rationalize the development of novel targeted therapies. Patient derived pre-clinical tumor models are the optimal tools to understand the mode of drug action as well as resistance mechanisms. In most cancers, stable cell lines cultured \textit{in vitro} show a distinct expression profile from primary GBM tumors\textsuperscript{80} therefore raising concerns about their validity as clinical models. In GBM studies, neurospheres cultures grown in growth factor supplemented serum-free media are closer to the primary tumor than serum-fed cell cultures\textsuperscript{81}, specifically from a cellular and transcriptional perspective. Finally, \textit{in vivo} (xenograft) expansion of various glioma cell lines leads to more consistent and more physiologic transcriptional profiles than \textit{in vitro}, with an advantage for intra-cranial over heterotopic mouse xenografts\textsuperscript{82}. Therefore, there are significant phenotypic and transcriptional differences between the various tumor models and primary tumors. These differences are seriously undermining our ability to measure and understand candidate drug effects in pre-clinical studies. However, beyond the phenotypic resemblance, the genetic validity of the tumor model is equally important. Indeed most recent therapies in oncology drugs are designed for particular genetic indications, targeting mutated genes, and the corresponding mutations need to be present and maintained in the pre-clinical model to ensure their utility. It is still unclear whether
pre-clinical models maintain faithfully the entire DNA mutational profile, including the clonal heterogeneity sometimes found in primary tumors, and can therefore be used to develop and study DNA-guided targeted therapies.

Intra-tumor heterogeneity results from the appearance of distinct mutations in different clones of the tumor, and their subsequent evolutionary selection, as the disease progresses or responds to treatment\textsuperscript{83-85}. This heterogeneity is a major cause of resistance to standard treatments. Indeed, single agent therapies do not address the molecular heterogeneity, and the process of tumor evolution, which frequently leads to the recurrence of the tumor. Intra-tumor heterogeneity has only been recently studied at the molecular level, through deep sequencing\textsuperscript{83,84}, genomic profiling of large tumor sections\textsuperscript{86,87} or even single cell analysis\textsuperscript{88}. In GBM, specific investigations of tyrosine kinase receptors amplifications have revealed the presence of independent events in different cells of the same samples\textsuperscript{89-91}. These observations have important implications on the interpretation of whole-sample genomic studies and their applications to investigate cancer progression and drug sensitivity. For this reason, the development of proper pre-clinical models that can recapitulate and maintain the clonal structure found in primary tumors is critical to generate the knowledge required for the development of meaningful treatment combinations. Because these models, such as cell-lines or mouse xenografts, are generated, grown and maintained in experimental conditions different from the primary tumor physiological conditions, they can themselves undergo clonal selection. An initial selection or a genetic drift can both be detrimental to the utility of these models. The potential variability in mutational profile, including mutation type and prevalence, between primary tumors and pre-clinical models has been only partially
investigated. Copy number studies of matched GBM primary and xenograft tumors has provided an estimate of the global genetic validity of the model\textsuperscript{92,93}. However, large genetic differences between a primary GBM tumor and derived model have also been observed. The well studied glioma cell line U87 for example, shows extensive DNA alterations, which likely resulted from \textit{in vitro} clonal selection leading to a mutational profile clearly different a GBM primary tumor\textsuperscript{30}. Similarly, genetic drift has been observed after expansion of clonal cell populations \textit{in vitro}\textsuperscript{94}. Elsewhere, it has been observed that \textit{in vitro} growth of GBM cells selects against \textit{EGFR} amplification and mutations, in contrast to \textit{in vivo} xenograft models\textsuperscript{95}. In breast cancer, whole genome sequencing of matched primary and xenografts are in good agreement, however there is some evidence of clonal evolution in the xenograft, suggesting that additional work is needed to understand the origin and significance of these differences\textsuperscript{33}.

Here we describe the results of a whole exome sequencing (WES) of four patient’s primary glioblastoma and their respective tumor models: one neurosphere culture, one laminin cell culture and two xenografts (Figure 2.1). We identify both somatic mutations and copy number aberrations by comparisons with normal DNA obtained from the patient’s white blood cells. We present an extensive comparison of the primary and model tumor genetic profiles. We develop original analysis methods to perform accurate comparisons and overcome technical variability. Our results illustrate the heterogeneity of the disease from the molecular standpoint and suggest that the pre-clinical models studied maintain their respective parental tumor genetic profiles, regardless of known expression differences\textsuperscript{82}. This work therefore confirms, at a
resolution superior to previous reports, that pre-clinical models can support laboratory investigations and testing of DNA-guided therapies for the treatment of GBM.

2.2 METHODS

**Human Tumor Collection**

Human tissue samples were obtained from 4 newly diagnosed glioblastoma patients. UCSD Institutional Review Board approved the study (RB#100936). All patients signed a written consent form approved by the Institutional Review Board. No treatment was administered prior to obtaining tissue samples. The samples were de-identified, banked as frozen tissue and used to extract DNA (DNAeasy kit QIAGEN) for the present study. Fresh tumor tissues were used to generate tumor sphere cultures and xenografts as described below.

1) Sample SK01600 was resected from a 57-year-old female presented with a large right frontal mass. The pathology showed classical features of glioblastoma. Molecular biomarkers detected in SK01600 include trisomy of chromosome 7, EGFR amplification (FISH) and overexpression (IHC), PTEN and RB1 hemizygous loss and c-MET gain (+1) (FISH).

2) Sample SK00115 was resected from a 64-year-old male, who presented with a right inferior frontal mass. Pathology showed dense hyper cellularity with astrocytic morphology with small monomorphic and anaplastic cells, florid glomeruloid microvascular proliferation, and pseudopalisading necrosis. Immuno-histochemistry analysis reveals loss of p16, PTEN, and p53 as well as a wild type PDGFR-A. Wild type EGFR and cMyc copy number was confirmed by fluorescent in situ hybridization.
3) Sample SK00102 was resected from a 47-year-old male with a right frontal mass. Pathology showed moderate to high cellularity, widespread microvascular proliferation, geographic zones of necrosis and infiltration into white matter.

4) Sample SK00072 was resected from a 60-year-old male who presented with a left occipital mass. Pathology showed moderate to focally high cellularity, pleomorphic astrocytic tumor cells, necrosis, and infiltration into white matter. Molecular biomarkers detected in SK00072 include trisomy of chromosome 7, loss of chromosome 10, EGFR amplification (FISH) and over-expression (IHC), RB1 and CDKN2A deletions (FISH), PDGFR-A and –B overexpression (IHC).

**Short-term Tumor Sphere Cultures**

SK01600 and SK00115 GBM cells were derived from above described primary GBM tissues as follows. Tumor specimens were washed in HBSS and mechanically minced, then dissociated using the MACS Neural Tissue Dissociation Kit (Miltenyi). Cells were subsequently washed, filtered through a 40-μm strainer and plated in low-attachment plates and grown as neurosphere (SK01600) in NeuroCult NS-A proliferation media (Stemcell Technologies) supplemented with 10 ng/mL rhbFGF (StemGent) and 20 ng/mL rhEGF (Stemcell Technologies)\(^{81}\). Alternatively, tumor cells were plated in laminin-coated plates (SK00115) and grown in adhesion using the above-indicated media\(^{96}\). Tumor cells were incubated at normal oxygen levels, at a temperature 37.0 °C and 5% CO2. Samples were collected at passage 6 (SK01600) and passage 3 (SK00115). The DNA was extracted using DNAeasy DNA extraction kit (QIAGEN).

**Xenograft Model**
SK00072 and SK00102 primary GBM tissues were directly passaged in vivo as mouse xenografts. Fresh tumor tissues were washed in HBSS and mechanically minced. Tissue aggregates were suspended in HBSS and mixed one to one with Matrigel (BD Biosciences) for injection. Six to eight week-old immuno-compromised NSG mice (The Jackson Laboratory) were injected at the flank. Tumors were removed when size reached 1-1.5cm³. DNA extraction was performed using DNAeasy DNA extraction kit (QIAGEN). For in vivo tumor maintenance, part of the tumor was mechanically dissociated as described above and reinjected subcutaneously into mice. The xenografts studied were collected at the first passage in vivo. All in vivo experiments were conducted under approved UCSD IACUC (Institutional Animal Care and Use Committee) protocol (S10005).

**Exome Library Generation**

The sequencing libraries were prepared and captured using SureSelect Human All Exon 50Mb kit (Agilent Technologies) following the manufacturer’s instructions. Briefly, 3 µg of genomic DNA from each sample was fragmented by Adaptive Focused Acoustics (Covaris S2, Covaris, Inc., Woburn, MA) to produce an average size of ~175 bp. After repairing the ends, SOLiD sequencing adaptors (Life technologies) were ligated to the fragments and purified with Ampure Beads (Agencourt) for size selection and free adapter removal. In some cases the beads were left in the buffer after the elution step to maximize the recovery rate. Each library was nick-translated, amplified by 6-cycles of PCR, and quantified by Bioanalyzer (Agilent). SureSelect solution hybridization was used to enrich the 50Mbp whole exome: 500 ng of each library was hybridized to the SureSelect baits during a 48 hrs. at 65°C. The RNA-DNA duplex were then captured
using MyOne Streptavidin beads (Life technologies) and the eluted material was further amplified using 8-cycles of PCR to add barcode tags.

**Sequencing**

Prior to sequencing, the libraries were quantified by qPCR using the SOLiD Library TaqMan Quantitation Kit (ABI) and the ABI 7900HT instrument (ABI). The barcoded samples were pooled in equimolar amounts in groups of six and the final pools were diluted to 500pM. The SOLiD EZ BeadTM system (ABI) was used to generate templated beads from the 500pM library pools. Approximately 900 million templated beads were deposited on a full slide. Sequencing was performed using the SOLiD 4 system, generating 50bp forward and 35bp reverse reads. One or two libraries were prepared for each sample, and sequenced across one to three runs, pooling together indexed libraries for the same patients on the same slide of the same run.

**Genotyping Array Data Analysis**

Omni2.5-Quad IDAT intensities were processed to genotypes using GenomeStudio (version 2010.3) using default cluster positions (HumanOmni2.5-4v1_D.egt) and the default GenCall score cutoff of 0.15 for Infinium arrays. Genotypes were exported in reference genome PLUS orientation (build hg19) based on HumanOmni2.5-4v1_D.bpm. We converted 1000 Genomes Project SNPs (kgp identifiers) to rsIDs by matching chromosome, position, and alleles in dbSNP132. We restricted SNPs to those that are present and biallelic in dbSNP132, and did not evaluate indels. We excluded 17,959 SNPs that were present in duplicate on the chip, 11,536 SNPs that had more than 2 alleles in dbSNP, and 405,516 SNPs, which were not reported in dbSNP132. This resulted in a total of 2,016,730 SNPs. Due to questionable strand
orientation, a previously reported problem, we additionally filtered 61,690 A/T and C/G SNPs that overlapped with the targeted region of the sequencing.

**Alignment**

All raw paired-end color-space reads were aligned to the human genome reference sequence (hg19), limited to chromosomes 1-22, X and Y, as well as mitochondrial genome. The alignment was carried out using BioScope v1.3.1 with default parameters for paired-end reads and allowing for insertions and deletion within a read. BioScope was chosen as the aligner since it had the highest percentage of properly-paired reads and aligned 35bp reads. Since SOLiD reads are in color-space, the only other aligners tested were BFAST\(^{51}\) and BWA\(^{53}\). BioScope uses a seed and extend method to align the reads. BioScope first aligns the 50bp read and then using its alignment location to align the 35bp read within the proper distance, 50-250bp, and the correct orientation, opposite strand. If it can’t find a properly-paired alignment location it then tries to align the 35bp read elsewhere in the genome. To align a read, BioScope starts by seeding the first 25bp of a read, allowing for 2 mismatches, and extend. If it can’t seed the first 25bp of a read it then tries to seed the last 25bp of a read. During the extend step, if extending the alignment drives down the alignment score it stops extending and hard-clips (trims) the rest of the bases of that read.

After aligning the reads I realigned the reads around indels using GATK’s IndelRealigner\(^{98}\). GATK’s IndelRealigner requires a minimum of 4 reads and at least 1 read with an indel to consider a location to be realigned. Due to these cutoffs, realigning each sample’s reads individually could result in missed indels and false positive SNVs
caused by misaligned reads. For example, if one sample from a patient has 0 reads with an indel and another sample, from the same patient, has multiple reads with an indel, only one sample’s reads will be realigned to incorporate the indel. However, if I combine all the reads from those two samples then the reads from both samples will be realigned to incorporate the indel. Thus combining the reads would result in an increase in the sensitivity of called indels and reduced the amount of false positive SNVs around indels. Therefore I combined the reads on a per-patient base (germline, primary tumor, and tumor model reads) and realigned them together and then split them back into the germline, primary tumor and tumor model samples.

After realignment I removed PCR duplicates with Picard Tools v1.53. One of the steps to generate a sequencing library is to amplify the DNA with PCR. This then creates duplicate DNA fragments that are then sequenced and result in identical reads. These identical reads do not add more information to the data and can cause false positive variants due to errors being added in during the PCR step. However, duplicate reads can only be detected if they are aligned to the genome with the same coordinates. If the entire 50 and 35 bp reads were aligned to the correct coordinates this would make it easy to identify all PCR duplicates. Yet aligners like BioScope, the aligner I used, trim the ends of some aligned reads. For example, if three reads are actual PCR duplicates but one read is not trimmed, the first 10 bp of another read is trimmed, and the last 15 bp of the last read is trimmed then the three reads would have different alignment coordinates. The current tools developed to remove PCR duplicates would miss these PCR duplicates because they have different alignment coordinates, even though all three are PCR duplicates. Therefore I developed my own method to remove
PCR duplicate reads for unpaired reads (see Removal of non-paired PCR duplicate reads). This new method would remove two of the three reads in the above example and keep the read with the highest mapping score. Finally, to improve the base quality scores of the remaining reads I used GATK’s TableRecalibration tool to recalibrate the reads’ base quality scores to match the observed mismatch rates.

**Coverage**

On target, near target, and off target bases, respectively, exome capture targeted areas, +/- 250bp from on target bases, and all other bases coverage was calculated using Picard Tools CalculateHsMetrics. The normalized coverage was calculated by dividing the coverage at each base by the average coverage across the genome for each sample.

**Removal of Missed Non-paired PCR Duplicate Reads**

During the analysis I found evidence of false-positive variants caused by PCR duplicates. Upon further inspection I realized that BioScope trimming reads caused Picard Tools to miss some of the PCR duplicate reads, specifically the reads with only one end aligning. All published tools that remove duplicate reads use the reads starting and ending coordinates to determine which reads are duplicates. As a result they remove reads that have identical mapped external coordinates. For paired-end reads they remove reads that have identical external coordinates at both ends of the read. For most alignments these tools remove all PCR duplicates. However, for alignments that have reads with at least one end trimmed, like BioScope alignments, these tools do not remove all PCR duplicates. I developed a method to remove the PCR duplicate reads missed by the standard tools. The method uses the CIGAR string to extract the coordinate based...
true-end of trimmed reads. The true-end of a trimmed read corresponds to the coordinate of the read-end if it hadn’t been trimmed during alignment. For example, if a read aligned to chromosome 1 at position 105 and had a CIGAR string of 5H45M (H for trimmed and M for aligned) my method would replace the starting alignment position, 105, with the true-end, 100. Lastly I removed all reads with only one end aligned and that had the same true-end coordinates, except for the read with the highest average base quality score.

Alignment to Mouse Genome

To remove mouse contamination in xenograft samples, see below, we needed to align the reads to the mouse genome. I downloaded mm9 reference from the UCSC genome browser and converted the fasta file into the reference file formats required for BioScope. Reads were aligned to mm9 using Bioscope with the same parameters previously set for the alignment to the human genome. For patients with a xenograft sample, we aligned the reads from all samples (Germline, Primary and Xenografts) from that patient to mm9 using BioScope.

Xenograft Filter

In order to identify potential mouse DNA contamination of the xenograft tumor sample, the reads from xenograft samples were aligned to the mouse genome (UCSC mm9) using Bioscope with the same parameters used to align reads to the human genome (see above). We then compared the alignment to the human and mouse genome to remove reads belonging to the mouse genome (Figure 2.2). All reads can fall in the two following categories: 1) Reads aligned only to one genome: the reads aligned uniquely to the human genome only are kept. Reciprocally, the reads aligning uniquely to the mouse
genome are discarded. 2) **Reads aligned to both genomes**: we looked at their respective matched (paired) read from the same fragment, which defined the following three sub-categories: 2a) the reads not properly paired in any genome are discarded; 2b) the reads properly paired in one genome only are kept (human) or discarded (mouse); 2c) the reads properly paired in both genome are then evaluated for their matching (M) score. The M score is calculated by summing up the number of perfectly matching bases (X) and subtracted the total number of mismatch bases (Y), \( M = X - Y \). When comparing the read’s M scores between the human and mouse genomes, prioritizing, for each fragment, the 50bp read over the 35bp read. The read is kept (respectively discarded) if the 50bp read’s M score is higher in the human (respectively mouse) genome. In the case of an identical score from the 50bp read comparison, we apply the same criteria to the 35bp read. Reads from fragment with identical M scores on both reads in both genomes were kept.

I prioritized the paired-end information over the M score for the following reasons: all reads with both ends aligned, should be properly paired, unless the read spans a breakpoint of a large chromosome rearrangement (insertion, deletion, translocation, and/or inversion). For the cases that the read spans a breakpoint, it is likely that the read will also not be properly paired in the mouse genome. Therefore, if a read is properly paired in one genome and not in the other, it is likely that the non-properly paired genome is an incorrect alignment. For the cases were a read is properly paired in both genomes, I would expect the alignment with the least amount of mismatches is the correct alignment.

**Calling Variants**
We used VarScan v 2.5.5\textsuperscript{101} to determine which sites in the genome are reference sites or non-reference sites, variants. VarScan uses the tumor and matched germline Pileup files created from SAMTools\textsuperscript{100} and a Fisher exact test to determine whether a position is variant. VarScan calls multiple types of variants, germline (inherited), somatic, loss of heterozygosity (LOH), and unknown, from both categories: single nucleotide variants (SNVs) and small insertions and deletions (indels). Unknown variants are sites which the germline sample was called homozygous alternate and the tumor sample is called homozygous reference or two different alternate alleles were called in the matched samples. To determine which sites are variants, VarScan first ignores any site in the genome where either sample does not meet the minimum coverage. For each site meeting the minimum coverage filter, VarScan determines the genotype of the germline and tumor sample independently. The site’s genotype is called heterozygous or homozygous alternate if there were at least 2 reads containing the alternate allele and the alternate allele frequency was ≥ 10%; else it is called homozygous reference. After calling a genotype for each position VarScan compares the genotypes called from each sample. If the genotypes are the same they are called either reference (the genotype matches the reference allele) or germline (the genotype doesn’t match the reference allele). In the case that the genotypes are different it then determines the significance of the difference in alternate allele frequencies using a right-tailed Fisher exact test. Significant variants (p<0.05 default) are then assigned to the three categories: somatic if the normal genotype equals the reference, LOH if the normal is heterozygous, unknown otherwise. All remaining variants not significantly different between tumor and germline are called germline.
When running VarScan I altered the default parameters. I set the minimum coverage to 1X and minimum average quality score to 0. I did not use the minimum average quality score cutoff because I created my own method to remove variants with low average alternate allele quality scores (see Filter1.3). For all further analysis, LOH variants are treated as germline variants because I did not use VarScan’s LOH calls to determine LOH regions. Note that a LOH variant is the same as a germline variant except that the alternate allele frequency increased or decreased in the tumor sample. By default, unknown sites are filtered out by VarScan.

**Variant Filtering Steps:**

Only bases within the targeted regions of the SureSelect kit were considered when calling variants.

*Filter 1: Filtering low quality gaps:* Calling small insertions and deletions, indels, is more difficult than calling SNVs with next-generation sequencing data. Some of the sequencing and alignment errors have been shown to be caused by single nucleotide repeats in the genome. I applied additional filters to potential indels (gaps) to remove these low quality gaps caused by sequencing and alignment errors. For a gap to be considered an indel we required a minimum coverage of 10X, a minimum of 3 reads supporting the indel, and a maximum of 5% of germline reads supporting the indel for Somatic indels. To improve the confidence in a called indel, I applied the minimum coverage and minimum number of supporting reads. I applied the maximum fraction of reads in the germline sample to remove potential germline indels that were missed in the germline sample.
Filter 2: Default VarScan filters: We applied the remaining default VarScan filters to the variants called by VarScan after removing low quality gaps. This filter removes variants that met any of the following criteria. 1) The variant is within 3bp of indels, to remove false-positive variants caused by indels. 2) Three or more variants within 10bp of each other are removed to reduce false positives caused by missed indels or alignment errors. 3) A minimum alternate allele frequency of 20%. A lower alternate allele frequency decreased the concordance with the genotyping array (data not show).

Filter 3: Remove low average alternate allele base quality variants: This filter removes potential false positive variants caused by sequencing errors. After applying filters 1.1 and 1.2 I realized that the overlap with known variants was lower (90%) than expected (95%)\(^{50}\). After investigating the difference between the known and novel variants I found that only the average alternate allele quality scores differed between the samples. The average alternate allele quality scores are higher in the set of known variants when compared to the novel variants. To remove variants with a low average alternate allele quality score I first calculated the average alternate allele’s base quality score (AQ) for all SNVs. Next I calculated the minimum AQ of a SNV (see Calculating the Minimum AQ) and removed germline, LOH, and somatic with a smaller AQ then the calculated minimum AQ score.

Filter 4: Remove low quality somatic variants: To remove germline variants incorrectly called somatic, we requested a maximum of 5% alternate allele frequency in the normal DNA. Low confidence somatic variants were finally removed by applying a minimum right-tailed Fisher exact P-value of 0.05 as determined by VarScan.
After filtering we identified, in the targeted regions, an average of 27,934 germline variants and ~233 somatic mutations in the 8 tumor samples (Table S11)

**Calculating the Minimum AQ**

To determine the minimum AQ score to use we started by clustering the SNV’s AQs using MCLUST\textsuperscript{102}. MCLUST selects the optimal model according to Bayesian information criterion (BIC) for expectation maximization (EM) initialized by hierarchical clustering for parameterized Gaussian mixture models. Based on the observed bimodal distribution of the average alternate allele’s base quality score, we ran MCLUST with the default parameters except for $G = 2$, where $G$ is the number of clusters for which the BIC is to be calculated. This resulted in two $Z$ scores ($Z_1$ and $Z_2$) for each SNV. $Z_1$ and $Z_2$ are the probabilities of the SNV belonging to the first cluster and the second cluster, respectively ($Z_1 = 1 - Z_2$). The first and second clusters refer to the SNV clusters containing the smaller and larger AQ scores, respectively. We used the smallest AQ with $Z_2 \geq 0.95$ as the minimum AQ for a variant to be accepted.

**Somatic Variant Annotations**

SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation/) was used to annotate somatic variants for function. To find relevant somatic variants, we compared our set of somatic variants to the COSMIC database version 55\textsuperscript{103}. We considered only mutations in the COSMIC database that met the following criteria: 1) the mutations were described in a published literature (PubMed ID), 2) the mutations had valid genomic coordinates, and 3) the mutant allele was reported. We compared our set of discovered
somatic mutations to the 151,502 entries of 4,542,623 total COSMIC entries that met these three criteria.

**Comparing Primary Tumor and Tumor Model**

Similar to the comparison between germline and primary tumor samples to identify somatic variants (implemented in VarScan, see above), a two-sided Fisher exact test was used to determine if a somatic variant found in the primary tumor and/or the tumor model was unique to that sample or shared by both samples. We performed the test on the union of the somatic variants identified in the primary and model tumors for each patient. By comparing allele read support in each sample, the two-sided Fisher exact test was used to calculate the probability, P-value, that the variant is unique to a sample. To correct for multiple testing, the data was permuted 100 times by randomly shuffling the sample a read belongs to and using the randomly assigned reads to calculate a random P-value. Lastly, an FDR of 0.05 was used to filter the somatic variant’s based on their p-value. The resulting non-significantly different variants were called ‘shared constant’. The variants that were determined as significantly different were further split into ‘unique’ somatic variants (mutant allele frequency<5% in one sample) or ‘shared changing’ somatic variants (mutant allele frequency ≥5% in both samples).

**Estimation of the primary tumor purity**

In order to estimate the amount of normal DNA in the tumor DNA, we compared, for each patient, the mutant allele frequencies of somatic variants with ≥ 30X coverage that were shared between primary and model. Assuming that the model is 100% tumor, then, \( mAF_p = \frac{1}{1+c} \times mAF_m \) where c is the fraction of normal DNA in the primary tumor and \( mAF_p \) and \( mAF_m \) are the mutant allele frequencies for the same somatic variant in the
primary tumor and tumor model samples, respectively. We estimated \( c \) using a linear regression model between \( m \text{AFp} \) and \( m \text{AFm} \) assuming a null y-axis intercept. We estimated that SK01600, SK00115, SK00102 and SK00072 primary tumors have 9%, 11%, 25% and 41% purity.

**Identification of Copy Number Aberrations (CNAs)**

We used ExomeCNV version 1.2\textsuperscript{104} to call CNAs and LOH regions throughout the genome. The B allele frequencies inputted into ExomeCNV were on target germline heterozygous variants passing Filters 1-3. To improve the power of ExomeCNV’s CNAs calling algorithm, we used regions of DNA corresponding to bases located on target or near target (+/- 250bp). To minimize sequencing noise in the CNA calls caused by library or sequencing biases, we created a high confidence “normal” coverage by pooling all male normal samples (SK00115, SK00102, and SK00072) using ExomeCNV’s pool.coverage. CNAs were called by comparison to this normal coverage using the default parameters except for setting the contamination level (\( c \)) to 0.1, minimum coverage to 5X, and the read length (\( l \)) to 43. For calling LOH regions we used the “two.sample.fisher” method and the suggested parameters except for an alpha of 0.999 in each iteration of the CBS algorithm in the multi.LOH.analyze step. We then removed any copy number call (neutral, deleted, amplified) that did not meet the two following filters: 1) the CNA segment was < 100 Kb in length, 2) two targets within the CNA segment were > 10 Mb apart. This prevents spurious CNA calls due to the scattered and uneven distribution of the exome capture baits. Using this methodology, we noticed an excess of amplification in SK00115 neurospheres calls when compared to the matched germline. After close inspection, we observed that the sequencing of this sample led to a specific
coverage bias, not found in the other samples. Specifically, the distribution of coverage depth as a function of the GC content of the probe showed a marked difference between SK00115 neurospheres and all other samples (Figure 2.3), although all experiments were performed in parallel. After correction by statistical regression, we were able to improve the calls (Methods). This observation indicates that calling CNAs using coverage information is highly sensitive to slight experimental variations despite the care taken to process matched samples identically.

A chromosome arm was called deleted or amplified if the sum of the size of the segments called deleted or amplified, respectively, represented ≥ 20% of the arm total length. To identify high confidence focal CNAs we first excluded CNAs segments present within a chromosome arm deletion or amplification. We then called high-confidence amplified (respectively deleted) the segment with a logR ratio higher (respectively lower) than the 95th (respectively 5th) percentile of the logR ratio of copy neutral segments.

Cancer Gene Census Copy Numbers: We extracted the copy number calls made by ExomeCNV for all 405 genes from the cancer gene census (02/12/2012 [7]) and removed any gene that was called copy number neutral in all 8 samples. For each of the 283 remaining genes, we calculated the average log R ratio as

\[
\text{avg LRR}_g = \frac{\sum_{s=1}^{n} (\text{LRR}_s \times O_{gs})}{\sum_{s=1}^{n} O_{gs}}.
\]

Where s is the copy number segment outputted by ExomeCNV, LRR_s is the LRR for a given segment s, n is the total number of segments overlapping with gene g, and O (g,s) is the number of base pairs overlapping between segment s and gene g.

Comparing CNA Calls between Samples
I developed a method to compare the CNA calls between two samples. Thus I method called positions ‘uncalled’ if ExomeCNV didn’t give it a copy number or it was filtered out by the Filtering CNA Calls method. Next I went base by base comparing the copy number call (1, 2, 3, 4, 5, or uncalled) between the two samples. We then summed up the number of bases with X and Y copy number calls.

2.3 RESULTS

We performed whole exome capture using hybrid selection of 12 samples (4 blood, 4 primary and 4 models) from 4 GBM patients (SK01600, SK00115, SK00072 and SK00102). High-throughput sequencing resulted in ~69% of the targeted bases covered at 10x or more, for an average on-target coverage of 59x across all 12 samples, therefore allowing accurate base calling at the majority of the coding portion of the genome (Table S1 and Figure 2.4). We first assessed the quality of the resulting calls by analyzing germline single nucleotide variants (SNV) comparing the results of sequencing and microarray genotyping at 62,550 positions investigated by the two methods. Of those, 52,905 were confidently called by sequencing of patient SK00072 germline DNA, passing our quality review. Ninety-seven percent of them had a consistent call between the two methods. Across all four patients, we estimated that ~90% of the germline SNVs identified are present in dbSNP(132) (Table S2), which is slightly lower than expected (95%) for Caucasian patients. A close inspection of the novel SNVs reveals a bi-modal distribution of the quality score of the alternate allele, contrasting with the single distribution at known SNPs (Figure 2.5). This indicates that a subset of the novel SNPs is of lower confidence and possibly resulting from sequencing errors. We separated the two
distributions using normal mixture modeling\textsuperscript{102}. The resulting set of high quality SNPs are now ~95% in dbSNP, and their transition to transversion ratio is ~3.1 (Table S2), closer to expected\textsuperscript{50}, therefore indicating an improvement in the accuracy of the SNV calls. Learning from this analysis of germline variants, we subsequently applied this strategy to filter all somatic calls.

**Primary Tumors Mutational Landscape**

We compared the variant calls between tumor and normal DNA\textsuperscript{101} restricting our analysis to the positions located on the capture targets. We identified a total of 682 somatic mutations across all patients, ranging from 130 to 191 per patient (Table S3). Of these, 384 are located in coding exons or a predicted splice site with 234 (61\%) missense, 12 (3\%) nonsense, 3 (<1\%) frameshift, 1 (<1\%) in-frame deletion, 5 (1\%) splicing and 129 (34\%) synonymous mutations (Figure 2.6A). This distribution leads to a non-synonymous to synonymous ratio of 1.98 consistent with the positive selection of driver mutations. These numbers and distributions are also in agreement with the mutational profile observed in exomes of GBM and other solid tumors\textsuperscript{76,106}. In order to determine which of these mutations are more likely to play a role in GBM progression, we used information from larger repositories such as COSMIC\textsuperscript{103} or TCGA\textsuperscript{76}. Ten of the 250 non-synonymous mutations have been previously identified in cancer samples (Table 2.1), among these, three \textit{PTEN} mutations in two patients were previously found in gliomas and three \textit{TP53} mutations in one patient were previously identified in various cancer types\textsuperscript{103}. We also identified one patient with an EGFR-C326S mutation, a position previously seen mutated in glioblastoma\textsuperscript{12}, as well as one patient with a NRAS-Q61K mutation, common in melanoma but never seen in gliomas. Expanding our investigation to 2,850 genes
known to be mutated in gliomas\textsuperscript{76}, we note a total of 59 non-synonymous or splice-site mutations in 53 genes (Tables S4 and S5). Apart, from \textit{PTEN}, \textit{GPR98} is the only recurrently mutated gene. This gene spans more than 600 kb and mutations in its sequence are more likely to be passenger. Therefore, except for mutations in \textit{PTEN}, the four patients seem to have mostly divergent sets of mutations contributing to the genetic make-up of their cancer.

Large chromosomal aberrations such as chromosome 7 trisomy or a loss of chromosome 10 are an important characteristic of glioblastoma mutational landscape. Although traditionally assayed through cytogenetic assays or Comparative Genomic Hybridization (CGH) and more recently with next generation sequencing\textsuperscript{101,104}, copy number aberrations (CNAs) can also be identified via exome sequencing strategies, using notably coverage differences, between tumor and normal DNA as well as evidence of loss of heterozygosity. Applying ExomeCNV\textsuperscript{104}, a segmentation strategy to evaluate the copy number and Loss of Heterozygosity (LOH) status of consecutive exons, we were able to call 32 large (chromosome arm level) CNAs in the four primary tumors (Figure 2.6B, 2.6C). All 4 patients showed an loss of chromosome 10 and an amplification of chromosome 7. Half of the patients also show evidence of a loss of one allele in chromosome 6q, 13q or 15q. Deletions of 12p, 14q, 17p or amplification of 2p and chromosome 19 were also observed each in a single case. The majority of these large CNAs are consistent with the most frequently recurring CNAs in glioblastoma\textsuperscript{76}. We also identified 23 high confidence focal CNAs (8 amplification and 15 deletions) in regions outside of large CNAs. Six of them encompass genes of the Cancer Gene Census\textsuperscript{57} (Table S6). Patient SK00115 shows a 4.3Mb deletion around \textit{PIK3R1} (Figure 2.6D). \textit{PIK3R1}
has been identified as a candidate cancer driver gene and is mutated in ~9% of GBM patients\textsuperscript{76,93}, but the loss of one allele, as seen here, has not been reported. Other CNAs deleted or amplified more than 2 fold in one or more sample are affecting 72 cancer genes (Figure 2.7C) and correlate well with array CGH diagnostic results obtained on three patients (Table S7). Notably, we could verify the 4-6-fold amplification of \textit{EGFR} locus in SK01600 primary tumor, encompassing a 5 MB segment (Figure 2.6E). This focal high-level amplification occurs in 40% of glioblastoma conjointly with the more common trisomy of chromosome 7. It is important to note that, in contrast with CGH and whole genome sequencing strategies, whole exome sequencing can introduce some bias in the estimation of CNAs: exons are not evenly distributed along the genome, which can lead to issues in resolving focal amplifications using whole exome sequencing coverage data\textsuperscript{101}. Our results suggest however that exome-based CNA calls are a good indicator of the presence of CNAs genome-wide and can therefore be used in cases where the amount of available DNA is scarce, a frequent situation in oncology translational studies.

Taken together our results reveal common molecular markers of GBM primary tumors, including nucleotide substitutions, small insertions and deletions as well as CNAs. Our results of both gene mutations and copy number alterations illustrate the heterogeneity observed across 4 patients, which is typical of the diversity of glioblastoma seen in the clinic. Some of these mutations are considered clinically actionable, such as alterations in the PI3K pathway (loss of \textit{PTEN} or \textit{PIK3R1}, amplification of \textit{PIK3CA}) for which targeted therapies are currently in clinical trials in several cancer types. Having established a comprehensive mutational profile of the primary tumor, we can now use the
same, high-resolution assessment, to study the maintenance of this profile in the corresponding pre-clinical models.

**Comparison to the tumor model mutational profile**

We applied the strategy described above for the primary tumors to identify somatic mutations in each tumor model derived from the four patients. The number of mutations in the SK01600 cells and SK00115 neurospheres was 184 and 194 respectively in close agreement with the findings in the primary tumor (165 and 196 respectively – Figure 2.8). In contrast, we noticed 1.8 and 3.6 fold excess in the number of mutations in the two xenografted tumors, SK00102 and SK00072 respectively. We suspected that mouse DNA contaminated the xenograft tumor samples, which led to their unspecific capture and alignment to the human genome, especially at genes of strong orthology. Using mouse to human alignment comparison, we were able to identify the most likely species of origin of each sequenced fragment (Methods and Figure 2.2). As expected, the resulting filter does not significantly change the number of somatic mutations identified in SK00102 and SK00072 primary tumors – from 191 and 130 to 189 and 121, respectively – whereas it significantly decreases the number of somatic mutations in the xenograft samples – from 338 and 468 to 201 and 220 respectively (Table S8). Thus, in all the subsequent analysis, we used reads filtered for murine contamination for all SK00102 and SK00072 samples (germline, primary and xenografts). Comparing the total number of somatic variants identified in all 8 samples, we note that SK00072 primary tumor shows fewer somatic mutations (Figure 2.9A) and a reduced mutant allele frequency (p<7 10^{-4}) as inferred from coverage depth (Figure 2.10). These observations suggest that SK00072 primary tumor DNA sample contains normal DNA leading to a
reduced sensitivity to detect somatic mutations. This conclusion is confirmed by the histological analysis of the tissue, indicating parenchymal infiltration within this tumor specimen (Methods). Contamination of the tumor DNA with normal DNA is a recurrent challenge for the sensitive detection of somatic mutations via high throughput sequencing. Therefore, it is important to know whether the derivation of pre-clinical models, in addition to preserving the tumor clonal heterogeneity, can have a purifying effect by selecting tumor cells only.

The total number of somatic mutations is in agreement between all three types of tumor models and their respective tumor of origin, suggesting an equivalent mutational load and the absence of hyper-mutator phenotype acquired during the derivation of the model. In order to refine this vision, and detect rare somatic differences between primary and model, we implemented a strict statistical comparison of the fraction of mutant allele supporting reads at all positions identified as somatic mutations in each set of matched samples. We were able to distinguish between shared mutations showing no significant changes in frequencies (referred to as shared constant), shared mutations with a significant change in frequency (referred to as shared changing), and mutations identified only in the model or the primary (referred to as unique), at a false discovery rate of 0.05. On average across all 4 pairs, 98% of the mutations identified in the primary were shared with the model (Figure 2.9B). Reciprocally, only ~2% of the mutations identified in the model were unique and not found in the primary, with the exception of patient SK00072 for which 11% (23/213) of the mutations are unique to the model. This result is not surprising given the lack of sensitivity to detect somatic mutations in SK00072’s primary tumor due to normal DNA contamination. This observation supports the idea of a
purifying process during the derivation of the xenograft, either through the preparation of
the sample or during its expansion in vivo. For the remaining three pairs of samples, a
discrepancy of ~2% between primary and model is within the range observed when
comparing mutations detected in control split-sample experiments (Table S8), and below
our FDR threshold, therefore pointing to a systematic bias rather than true genetic
differences.

Out of the 1005 mutations shared between primary and model, 293 were covered
by 30 reads or more in both samples and showed a high correlation in mutant allele
frequency between primary and model (r>0.7) (Figure 2.9C-F). Furthermore, 48 out the
49 of the mutations covered at 30x with a significant change in mutant allele frequency
(FDR=0.05) show a unidirectional change, a modest enrichment in the model, suggesting
a higher purity of these samples when compared the primary rather than true allele
frequency differences due to clonal selection. These results suggest therefore the absence
of strong clonal selection using either in vitro or in vivo models. Using the differences in
allele frequency, and assuming the purity of cancer cells in the tumor models, we can
establish that the primary tumors were contaminated with 9%, 11%, 25% and 41% of
normal DNA in patient SK01600, SK00115, SK00102 and SK00072, respectively, which
is consistent with the reduced sensitivity in detecting mutations in SK00072 primary
tumor.

We next evaluated whether CNAs were conserved between primary and model.
Restricting the primary-model comparison to high confidence CNA calls, we observed
that 97, 94 and 97% of the base pairs in CNAs are consistently called between primary
and model in samples SK01600, SK00115, SK00102, respectively (Figure 2.7A and
Table S10). In contrast only 77% of the high confidence CNAs base pairs show this level of consistency between the two SK00072 samples, while 14% are called in the primary at a lower copy number than in the model. This result is consistent with the presence of normal cells in SK00072 primary tumor, which affects the sensitivity of CNA detection. Although the global landscape of structural variants is important to study the mechanisms of cancerogenesis and clonal selection, our ability to interpret the biological consequences of CNAs is limited to the coding portion of the genome, where gain and losses of specific alleles have a frequently demonstrated oncogenic potential. In order to validate our method for the accurate estimation of more biologically significant CNAs, we performed a specific inspection of the copy number status at 450 genes from the cancer gene census\(^\text{57}\). We could identify 72 genes with a copy number change of more than 2 fold in one or more samples (Figure 2.7B). Consistent with the previous results, there is a very strong correlation between the copy number observed in the primary and in the tumor model. Interestingly, the copy number estimation in SK00072 primary tumor is lower than in its matched xenograft, again highlighting how contamination of the primary sample with normal DNA can underscore the sensitivity of the mutational analysis. We observed the strong amplification of \(EGFR\) and neighboring gene \(IKZF1\) in the primary tumor of patient SK01600, but it seems to be partially lost in the matched neurosphere culture. This result is consistent with the frequent loss of \(EGFR\) amplification in serum-fed culture\(^\text{95}\), as well as growth factor supplemented serum-free culture\(^\text{107}\).

Overall, using genome-wide high-confidence CNA calls as a molecular signature, we observed that primary and tumor models are more related to one another that to
another patient sample (Figure 2.7C). We do not observe a closer relationship between \textit{in vivo} models and primary than between \textit{in vitro} model and primary therefore showing that, from a genetic perspective, \textit{in vitro} and \textit{in vivo} models are both faithfully matching the primary tumor.

2.4 DISCUSSION

High-throughput DNA sequencing now offers the opportunity to obtain a detailed molecular profile of a biological sample such as the ones we studied here. This recent technology has been evolving at fast pace, and some systematic errors and samples preparation difficulties can make their use challenging. This is especially true for studies aimed at comparing longitudinal samples, between primary tumor and relapse, or between primary and tumor-model. Systematic errors, often platform dependent, can lead to false positive rates sometimes as high as 10\%.\textsuperscript{108} We present here novel analytical methods that can increase our confidence in the mutations detected through exome sequencing, by identifying mouse reads, correcting for mutant allele base quality, accounting for GC bias in CNA calls or performing a strict statistical comparison between two samples instead of relying on simple identification of the mutants. Through these improvements, our results support the global maintenance of the primary tumor genetic profile in the chosen pre-clinical model.

We demonstrate that intra-tumor clonal heterogeneity is conserved in the various models. The maintenance of tumor heterogeneity is important to understand the mechanism of resistance occurring in the patient during therapy. Recent study of post-relapse leukemia have shown the presence of the resistant clone in earlier samples.\textsuperscript{83} Such
studies have yet to be performed in glioblastoma, and xenografts would be an advantageous model for this, as the tumor can be isolated at different stages of the progression. Similarly, the number of passages of *in vitro* and *in vivo* models is thought to cause important genome remodeling, principally through large structural rearrangements and polyploidy. Model cell lines, maintained in the laboratory for decades, such as U87, show highly remodeled genome, with little in common with the genome of primary tumors. Stable late passage models tend to mimic more closely the biology of the primary tissue, but early passage models are highly valuable for identifying the optimal targeted therapy within the lifespan of the patient\textsuperscript{109}. Therefore, although we did not strictly address the genetic drift of the tumor through passages, our results suggest that a moderate number of passages (1 for *in vivo*, 3 to 6 for *in vitro*) do not have detectable consequences on their genomes. One exception seems to be the maintenance of *EGFR* copy number *in vitro*, which is affected by the presence of EGF in the medium. As more inhibitors of growth factors such as nilotinib for *PDGRA*, are being tested in clinical trial, it is crucial to carefully select the laboratory conditions in which pre-clinical experiments are being performed and to ensure that the presence of the marker can be maintained *in vitro*.

Molecular profiling of patients and patient-derived samples has become a central part of personalized medicine. Several centers are promoting the clinical sequencing of patients’ samples to guide treatment\textsuperscript{110}. As an increasing number of targeted therapies are approved, drug resistance will become increasingly problematic and repeated molecular profiling on relapse biopsies will be needed to choose the appropriate second line of therapy and hopefully convert cancer in a manageable disease under surveillance\textsuperscript{111}. For
these reasons, that directly impact the care of the cancer patients, the availability of representative pre-clinical model to study drug sensitivity and resistance becomes crucial. Two recent drug screening studies of fully characterized cell lines illustrate the utility of combining genomic and pharmacological information\textsuperscript{71,72}. This approach has limitations, as it does not recreate the micro-environmental niche in which the primary tumor resides and grows and that may as well affect treatment response\textsuperscript{96,112}. Nonetheless, our results indicate that both patient-derived \textit{in vitro} cultures and \textit{in vivo} xenografts represent robust pre-clinical models systems reflecting the genomic diversity of primary GBMs, and highlight their utility in defining tumor genetics predictors of drug response and drift of tumors from selective pressures (e.g. treatments). Hence, ensuring our comprehensive understanding of the molecular forces at play in these models, will favor the successful translation of these discoveries to the clinical care where molecular profiling will become standard.

Chapter 2, in full, is a reprint of the material as it appears in PLoS ONE 8(2): e56185. doi:10.1371/journal.pone.0056185. Epub 2013 Feb 18. High-Resolution Mutational Profiling Suggests the Genetic Validity of Glioblastoma Patient-Derived Pre-Clinical Models. Shawn E. Yost, Sandra Pastorino, Sophie Rozenzhak, Erin N. Smith, Ying S. Chao, Pengfei Jiang, Santosh Kesari, Kelly A. Frazer, Olivier Harismendy. The dissertation author was the primary author of this paper.
2.5 FIGURES

---

**Figure 2.1. Experimental Design.** The matched blood and primary tumor’s DNA from 4 patients were analyzed in addition to the patient derived neurospheres (SK01600), laminin cell culture (SK00115) or mouse xenografts (SK00102 and SK00072).

---

**Figure 2.2. Filtering of the Mouse contaminating reads.** A succession of filters (green arrow: pass, red arrow: do not pass) compares pairing information as well as matching score to determine the species of origin of each read. (*) Match score \( (M) = \# \text{ of Matches} - \# \text{ of Mismatches} \).
Figure 2.3. SK00115 tumor model shows GC induced bias in the coverage distribution. Normalized average coverage per GC% of targets for all four patients. Germline (black), primary tumor (red), and tumor models (blue) are displayed.
Figure 2.4. Sequencing Quality Assessment. (A) The Reads were sequenced on SOLiD4 instrument and aligned to the reference genome using BioScope. The duplicate reads were identified using Picard MarkDup and custom scripts (Methods). (B) Coverage depth cumulative distribution for all 12 samples (matched germline, primary tumor, and tumor model). (C) Capture enrichment specificity. The fraction of bases sequenced on or near (+/- 250bp) the Agilent SureSelect 50MB kit targets is indicated (Table S12).
Figure 2.5. Alternate allele’s base quality score filtering. (A) Distribution of the average alternate allele’s base quality score for germline variants present in dbSNP132. (B) Same as (A) for novel germline variants. (C) Variants are filtered out (red) when they belong to the lower quality distribution as determined by mixed model deconvolution.
Figure 2.6. Mutational Landscape of the primary tumors. (A) The cumulative distribution of the somatic mutations identified on the targeted exons of the four patients primary tumors is reported as a function of their class and predicted protein changes. (B) Circular diagram [48] representing all 23 chromosomes and their cytogenetic map (outer circle, grey scale bands and red centromeres). The logR tumor/normal coverage ratios (black dots) and the inferred CNA (red: amplification, blue: deletion, blue bars: Loss of Heterozygosity) identified in the 4 primary tumors (from outer to inner circle: SK01600, SK00115, SK00102, SK00072) using whole exome sequencing data are represented. (C) Chromosome-arm level copy number aberrations are observed in the 22 autosomes when >20% of a chromosome arm is reported as deleted (blue) or amplified (red). (D) A focal deletion of ~10Mb (set of blue segments) including a large (4.3 Mb) CNA segment affects PIK3R1 gene in SK00115 primary tumor. The LogR ratio of tumor/normal coverage (x axis) at each exon capture probe (grey dots) allows the identification of DNA segments deleted (blue bars) or amplified (red bars). (E) Similar to (D), a focal amplification of EGFR containing segment (red) is identified in addition to the chromosome 7 trisomy in patient SK01600 primary tumor. Some segments may appear to overlap as a result of the plotting resolution.
Figure 2.7. Comparative evaluation of the CNAs between primary and model tumors (A) The evaluation of the copy number status at all base pairs called in high-confidence CNA segments in both primary and model tumors identifies positions with a consistent (grey), lower (blue) or higher (red) copy number call in the model when compared to the primary tumor (Table S10). (B) Average copy number status (blue-red color scale, log2 ratio) at 72 genes of the cancer gene census showing more than 2 fold copy number difference in one or more sample. (C) Euclidian distance based dendrogram classifying the 8 tumor samples using the logR ratio of high-confidence CNA called in one or more sample
Figure 2.8. Identification of somatic mutations in the tumor models. (A) The number and distribution of mutations in the tumor models matches the primary except for xenograft samples, suggesting mouse DNA contamination. (B) The total number of somatic mutation before filtering of the mouse reads (grey) and after filtering of the mouse reads (black).
Figure 2.9. Comparative evaluation of the somatic mutations between primary and model tumors. (A) The cumulative distribution of the somatic mutations identified on the targeted exons of the four patients’ primary tumors (P) as well as tumor models (N: Neurospheres, C: Cell culture, X: Xenograft) is reported as a function of their class and predicted protein changes. The mutations were identified after excluding mouse reads from patients’ SK00102 and SK00072 data. (B) A statistical comparison of the somatic mutations called between primary and model identifies shared mutations at constant mutant allele frequencies (black), shared mutations with changing mutant allele frequency (red) as well mutations specific to the primary (green) or the tumor model (blue). (C-F) Mutant Allele frequency differences between the primary tumor (x axis) and the model tumor (y axis) of patient SK01600 (C), SK00115 (D), SK00102 (E), SK00072 (F) at all positions identified as somatically mutated in either sample and covered by ≥30 reads. Mutations are classified as shared with constant frequency (black), with changing frequencies (red), specific to the primary tumor (green) or to the tumor model (blue).
Figure 2.10. **SK00072 primary tumor shows a significantly lower mutant allele frequency.** (A) Distribution of the mutant allele frequency at mutations shared between primary and model. (B) Student T-test p-value (red scale –log10 (P-value)) of the 6 possible comparisons from (A), showing SK00072 as significantly lower mutant allele frequency.
## 2.6 TABLES

### Table 2.1. Somatic non-synonymous mutations observed in the primary tumors and overlapping with known COSMIC (v55) entries

<table>
<thead>
<tr>
<th>Patient</th>
<th>NCBI37</th>
<th>Alleles</th>
<th>Number of reads</th>
<th>Type</th>
<th>Codon change</th>
<th>Cancer Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SK01600</td>
<td>chr7:55223610</td>
<td>G/C</td>
<td>7:0</td>
<td>EGFR</td>
<td>missense</td>
<td>C326S</td>
</tr>
<tr>
<td></td>
<td>C/A</td>
<td>217:0</td>
<td>43:45</td>
<td>PTEN</td>
<td>missense</td>
<td>D92E</td>
</tr>
<tr>
<td></td>
<td>AGT/-</td>
<td>14:0</td>
<td>7:4</td>
<td>TP53</td>
<td>In frame deletion</td>
<td>L265</td>
</tr>
<tr>
<td></td>
<td>C/A</td>
<td>16:0</td>
<td>17:7</td>
<td>TP53</td>
<td>missense</td>
<td>S215I</td>
</tr>
<tr>
<td></td>
<td>G/A</td>
<td>18:0</td>
<td>9:3</td>
<td>TP53</td>
<td>missense</td>
<td>R158C</td>
</tr>
<tr>
<td>SK01012</td>
<td>chr1:15256530</td>
<td>G/T</td>
<td>96:0</td>
<td>NRAS</td>
<td>missense</td>
<td>Q61K</td>
</tr>
<tr>
<td></td>
<td>T/A</td>
<td>10:0</td>
<td>2:2</td>
<td>PTEN</td>
<td>missense</td>
<td>H61Q</td>
</tr>
<tr>
<td></td>
<td>G/A</td>
<td>123:0</td>
<td>34:42</td>
<td>PTEN</td>
<td>missense</td>
<td>G165R</td>
</tr>
<tr>
<td></td>
<td>G/A</td>
<td>13:0</td>
<td>19:6</td>
<td>CREBBP</td>
<td>nonsense</td>
<td>R1303*</td>
</tr>
<tr>
<td>SK00072</td>
<td>chr1:159504907</td>
<td>C/G</td>
<td>27:0</td>
<td>OR10J5</td>
<td>missense</td>
<td>K297N</td>
</tr>
</tbody>
</table>

1. Can includes normal DNA contamination and effect of copy number
2. The same position but not the same mutation was found in Glioma
Chapter 3. Mutascope: Sensitive Detection of Somatic Mutations from Deep Amplicon Sequencing

3.1 BACKGROUND
The accurate detection of somatic mutations in tumors is critical for precise diagnostic and selection of targeted therapies\textsuperscript{113} but the low allelic fraction frequently encountered in heterogeneous or poor cellularity clinical specimens renders this task challenging. In current clinical assays, amplicons covering the exons of 10-100 cancer genes are amplified via PCR-based or analogous approaches and sequenced at high depth to identify mutations present in less than 5% of a DNA sample\textsuperscript{35}. Despite high coverage depth, the error rate resulting from systematic sequencing bias\textsuperscript{60}, can hinder the detection of mutations. While experimental\textsuperscript{114} or analytical\textsuperscript{50,98} methods, or comparison to the normal DNA\textsuperscript{101,115} can mitigate this effect, most analysis strategies were developed for sequencing of random shotgun DNA fragments, and thus do not take into account systematic errors specific to amplicon sequencing. In amplicon sequencing, loci are covered by reads with identical genomic starting positions and because the error rate increases along the length of the read (Fig. 1a), a variable consensus error rate exists over the target (Fig. 1b). Analytical strategies specifically designed for amplicon sequencing have the potential to enhance the mutation detection accuracy of current clinical assays, especially at low allelic fraction.

In the past 5 years, there has been a large number of highly sophisticated bioinformatics and statistical tools published to analyze next generation sequencing data, specifically to call DNA variants. Some of these tools have been developed by large consortiums like the 1000 Genomes project (GATK) or The Cancer Genome Atlas.
(Varscan) and are widely used in research settings. As high throughput sequencing matures, it is making its way into many applications, including clinical diagnostics. For that purpose companies like Illumina or Life Technologies are marketing table-top sequencers, with lower yield but faster turn around time. In the clinic, these sequencers are typically used for the targeted sequencing of candidate genes, which are clinically actionable. In many instances, this panel of genes is amplified via multiplex or microfluidic PCR or similar approaches such as Molecular Inversion Probes. In oncology, the mutational status of these actionable genes in the tumor is important to select the appropriate treatment regimen. An increasing number of cancer centers and hospitals are using these technologies to guide clinical care. However, the analysis tools mentioned above were developed for whole genome and whole exome sequencing performed after random fragmentation of the DNA. These tools are not adapted to the high depth and specific error mode of amplicon sequencing, which can result in lower performance, especially for the detection of mutations with low allelic fraction. To our knowledge, there are no variant calling tool specifically designed to analyze high throughput sequencing of amplicons, which was our main rational to develop Mutascope.

The data we present explains the main advantages of such a dedicated analysis tool and demonstrate it superior performance over alternate approaches. In particular we demonstrate that the factory analysis strategy offered by Illumina, and which is more likely to be chosen by end diagnostic users, lacks sensitivity to detect mutations with a low allelic fraction. We think that our study is very timely as an increasing number of laboratories rely on high throughput amplicon sequencing. A recent study published in Nature Biotechnology1, presented MuTect, developed by the Broad Institute. MuTect is
an excellent tool for whole exome and whole genome sequencing, but its statistical framework and filtering strategy is not adapted to amplicon sequencing. Similar to MuTect and in contrast to most other tools available, Mutascope is specifically tuned to identify rare somatic events from a paired sequencing of tumor and normal DNA. This specification is essential since most tumors available in the clinic are heterogeneous and typically contain a low fraction of tumor cells. Mutascope read alignment filtering, variant identification and filtering are all specifically built to optimize the detection of mutations at low allelic fraction from amplicon datasets, resulting in a higher performance.

Here we present Mutascope, a software dedicated to the detection of mutations at low allelic fraction from amplicon sequencing of matched tumor-normal samples pairs. Mutascope determines the amplicon of origin for each read and measures the specific experimental error rate from sequencing the normal DNA. The mutations in the tumor are then identified by comparison to the error rate using a binomial statistics and classified as germline or somatic by comparison to the normal DNA. A set of filters adapted to amplicon sequencing then eliminates false positive calls. We used two experimental datasets, a mixture of 8 normal DNA (MIX) and a set of 80 Tumor-Normal Spiked-In (TNS) pairs derived from 38 different normal germline DNA to measure the performance of the approach in comparison to other mutation callers.

3.2 METHODS

Sample collection: The genomic DNA of 8 samples from the CEPH cohort (Supplemental Table 1) was purchased from the Coriell institute (Camden, NJ). An additional thirty-eight individuals were enrolled in the UC San Diego and UC Irvine
repository (referred to as UC samples), after signing an informed consent for a study approved by the respective Institutional Review Boards.

**Sample preparation:** We generated a pool of germline DNA (MIX) from 8 Coriell samples by mixing variable amounts (Supplemental Table 1). The DNA from the 38 UC samples was extracted from buffy coat using the All Prep DNA extraction kit (QIAGEN) following the manufacturer’s instructions. Both the 8 individual Coriell DNA, the MIX sample and the 38 UC samples were subjected to Ultra-Deep Targeted Sequencing (UDT-Seq) as described below.

**Genomic DNA preparation:** Genomic DNA samples were fragmented by Adaptive Focused Acoustics (E220 Focused Ultrasonicator, Covaris, Woburn, Massachusetts) to produce an average fragment size of 3 kb. Fragmentation of the genomic DNA to 2 to 4 kb allows for optimal template size for performing PCR in droplets. Sheared genomic DNA was purified using MinElute PCR Purification Kit (Qiagen) using the manufacturer’s recommended protocol. Fragmented genomic DNA was analyzed by gel electrophoresis on a 0.8% agarose gel to confirm that the genomic DNA was in the correct size range (~3 kb). To prepare the input DNA template mixture for targeted amplification, 1.5 μg of the purified genomic DNA fragmentation reaction was added to 9.4 μl 10× High-Fidelity Buffer (Invitrogen, 11304-029), 2.5 μl of 50mM MgSO₄ (Invitrogen, 11304-029), 2.5 μl 10 mM dNTP (New England Biolabs, Ipswich MA, USA, NO447S/L), 7.2 μl 4 M Betaine (Sigma, B2629-50G), 7.2 μl of RDT Droplet Stabilizer (RainDance Technologies, 30-00826), 3.6 μl dimethyl sulfoxide (Sigma, D8418-50 ml) and 1.4 μl 5 units/μl of Platinum High-Fidelity Taq (Invitrogen, 11304-
029) and the samples were brought to a final volume of 50 μl with nuclease free water (Teknova-Fisher Hollister CA, USA, 50843418).

**Microdroplet PCR:** RainDance Technologies (Lexington MA) prepared a custom library of droplets for 1736 primer pairs (Supplemental Table 2). The primer droplets were merged with the sample droplets on the RDT1000 (RainDance Technologies, 20-01000) using the manufacturer's recommended protocol. The PCR reactions were carried out on GeneAmp 9700 thermal cycler (Applied Biosystems) as follows: initial denaturation at 94°C for 2 minutes; 55 cycles at 94°C for 30 seconds, 54°C for 30 seconds and 68°C for 30 seconds; final extension at 68°C for 10 minutes, followed by a 4°C hold. Following PCR amplification, the emulsion of PCR droplets was broken to release each individual amplicon from the PCR droplets. For each sample, an equal volume of RDT 1000 Droplet Destabilizer (RainDance Technologies, 40-00830) was added to the emulsion of PCR droplets, the sample was vortexed for 15 seconds and then centrifuged at 12,000 × g for 10 minutes. The aqueous phase of each sample was purified over a MinElute column (Qiagen, 28004) and eluted with 11 μl of the Qiagen EB buffer. The purified DNA was analyzed on an Agilent Bioanalyzer to quantify PCR yield.

**PCR to Add Illumina Indexes:** Four microliters of amplicons (2.5 ng/μl) were combined with 3.25 μl 10× High-Fidelity Buffer (Invitrogen, 11304-029), 0.88 μl of MgSO₄ 50 mM (Invitrogen, 11304-029), 0.88 μl of 10 mM dNTP (New England Biolabs, NO447S/L), 2.5 μl of 4 M Betaine (Sigma, B2629-50G), 2.5 μl of RDT Droplet Stabilizer (RainDance Technologies, 30-00826), 1.25 μl dimethyl sulfoxide (Sigma, D8418-50 ml), 0.5 μl of 5 units/μl of Platinum High-Fidelity Taq (Invitrogen, 11304-029), and 0.5 μM final of a universal forward primer and an index-specific reverse primer.
(Supplemental Table 3). Samples were amplified in a GeneAmp 9700 thermal cycler (Applied Biosystems) as follows: initial denaturation at 94°C for 2 minutes; 10 cycles at 94°C for 30 seconds, 56°C for 30 seconds and 68°C for 1 minute; final extension at 68°C for 10 minutes, followed by a 4°C hold. Each sample was purified using the Agencourt AMPure XP beads (Beckman Coulter, Fullerton, CA, USA) at a ratio of 0.8:1. The purified amplified library was then analyzed on an Agilent Bioanalyzer to quantify final amplicon yield.

**Sequencing and analysis using Illumina MiSeq:** The amplified libraries were quantified and pooled in equi-molar amounts. The pool was loaded at between 8 and 11 pM (depending on the run) and sequenced using MiSeq Reagent Kits (300 cycles) (Illumina, MS-102-1001) MiSeq Reagent Kits V2 (300 cycles) (Illumina, MS-102-2002) using custom sequencing primers (Supplemental Table 3). The resulting reads were deconvoluted based on their index. The raw reads are publically available through the Short Reads Archive (SRA) at the NCBI: SRA067609 (Coriell + MIX samples) and SRA067610 (UC samples).

**Generation of the benchmarking datasets**

In order to generate the benchmarking datasets, MIX and Tumor-Normal Spike-In (TNS) samples, we first determined the genotypes of the 8 Coriell samples and 38 UC samples. Genotypes were called using a Bayesian Likelihood genotyping method\textsuperscript{98}. We removed any variant that was within the primer regions unless covered by another amplicon. To improve accuracy, we applied the following set of filters to the called variants; 1) 3 or more SNPs within 10bp of each other; 2) the SNP is within 10bp of an Indel; 3) minimum average mapping quality of 15; 4) minimum percentile coverage of
0.5%; 5) both reference strands have > 1X coverage while only one strand has > 1X coverage for the alternate allele; 6) read-group bias score less than $10^{-15}$, and 7) and a minimum average alternate allele base quality score of 25.

Using the filtered SNVs identified in the 8 Coriell samples, we determined the list and allelic fraction of expected variants in the MIX sample. To test the performance of the variant callers and not that of the assay, we removed any expected variant from the MIX sample that had an observed allelic fraction below 0.005.

Using the filtered SNVs identified in the 38 UC samples, we prepared the TNS pairs. For each pair, we used the reads from the sequencing of 3 different UC sample, referred to as germline (G), spike in source (SIS) and spike in destination (SID), respectively. For each variant position in any of the three samples, we classify the variants in classes labeled $C_{n,p,q}$, where $n$, $p$ and $q$ corresponds to sample G, SIS and SID respectively and $(n,p,q) \in \{0,1\}$: 0 if the position is reference, 1 if the position contains an alternate allele. According to this notation a variant of class $C_{1,0,0}$ is only variant in the G sample and reference in SIS ad SID. The G sample is considered the normal sample, therefore, variants of class $C_{1,0,1}$ are considered germline variants in a TNS pair. To prepare the tumor of each TNS pair, we randomly selected reads from the SIS sample overlapping variants of class $C_{0,1,0}$. The selected reads were spiked-in at specific amounts in the SID sample’s reads to simulate somatic variants at 1%, 3% or 10% allelic fraction.

We selected $R$ reads from the SIS sample so that $R = \frac{f_{PTN} \cdot C}{f_{SIS} - f_{PTN}}$ where $f_{TNS}$ is the desired somatic variant allelic fraction, $C$ is the coverage of the corresponding variant position in the SID sample and $f_{SIS}$ is the allelic fraction of the alternate allele in the SIS sample.
Using 38 samples, we followed the above procedure with 40 unique combinations of G, SIS and SID samples to generate 80 TNS pairs, 40 at each 1% and 10% allelic fraction. Each TNS pair contains between 17 and 55 (mean 34) spiked-in somatic variants.

**Mutascope Analysis**

The analysis is presented following the sequence of Mutascope modules. For more information, refer to Mutascope manual.

*Identification of potential False Positive variants (module makeBlackList):* Non-specific PCR amplification at off-target locations happen at low efficiency, and may lead, after alignment to the targeted region to false positive substitutions. Mutascope makeBlackList predicts the likely false positive substitutions to exclude them from the analysis. For this, the first and reverse-complemented last \(L\) (\(L=151\) for this study) nucleotides of each targeted amplicon are used to generate two fastq file of quality \(Q=33\) mimicking the result of a paired end targeted sequencing of the human reference genome. These reference reads are aligned to the human genome using BWA’s standard alignment method with default parameters except for \(-N\) and \(-n 15\) in the \texttt{aln} step and \(-n 100\) in the \texttt{samse} step. For each amplicon, Mutascope extracts the list of alternate alignment locations using the \texttt{XA} field from the resulting SAM file. Using a Needleman-Wunsch global alignment between the targeted amplicon and each of the alternate alignment locations, Mutascope then identifies the specific mismatches corresponding to substitutions to ignore from the final calls (black list).

*Read Processing (modules runBWA, refinement, groupRealign, and xpileup):* All sequencing reads are initially aligned with BWA\(^{53}\) using the Smith-Waterman alignment algorithm (bwasw) with default parameters except for \(-z\) and \(-s\) both set to 5. After
removing any read that aligns to multiple locations in the genome a minimum Smith-Waterman score (AS field in the SAM file) of 111 is applied to the reads. Next, the reads are grouped based on their amplicon of origin (using alignment coordinates) and read type (forward or reverse). Reads starting more than 2 bp from the expected amplicon boundary are discarded. The soft-clip tag added by BWA is then removed and the interval corresponding to the PCR primer is soft-clipped. All reads from a matched tumor-normal pair are then realigned together following GATK’s local realignment procedure to improve SNP and indels comparative detection between the two samples.

Using the SAMtools mpileup command (-B, -O, and -s -d 5000000 -Q 0), Mutascope then counts the number of occurrence of each allele at each position in each read group before consolidating the results across all read groups into an extended pileup format (xpileup file).

**Measurement of error rate (module calcErrorRates):** Using the non-dbSNP (version 135) locations and non-black list substitutions in the xpileup file from the normal sample sequencing results, we measured the error rate at each sequenced position as

\[ E_{p,r,s_{ij}} = \frac{M_{p,r,s_{ij}}}{N_{p,r,i=s_{j}}} \]

where \( M_{p,r,s_{ij}} \) is the total number of substitution \( s_{ij} \) (i reference, and j alternate allele) that are at a position \( p \) in a read of read type \( r \), and \( N_{p,r,i=s_{j}} \) (respectively \( N_{p,r,i=w} \)) is the total number of C or G (respectively A or T) reference bases at position \( p \) in a read of read type \( r \). When several read groups overlap (overlapping amplicons, overlapping read1 and read2) we combined error rates according to \( Err = \frac{\sum_e N_e e}{\sum_e N_e} \) where \( e \) is \( E_{p,r,s_{ij}} \) and \( N_e \) is the number of bases with the given error rate \( e \).
Calling Variants (module callSomatic): For each position that is covered by at least one read in both tumor and normal sample, Mutascope determines if there is a potential Germline, Loss of Heterozygosity (LOH), or somatic mutations. SNVs are called by first removing any sequenced base that matches a location and substitution in the set of blacklisted substitutions. The alternate allele is then identified as the non-reference allele with the highest number of supporting calls. The number of alternate alleles with Q>10 in reads with mapping quality>1 and which sum represents more than 0.5% of the total coverage at the position are then subjected to our primary binomial classifier (VCF QUAL field corresponding to the –log10 p-value), using the probability given by the error rate matched for position, read type r, and substitution (see above). Somatic variants are then distinguished from germline or systematic noise using a Fisher exact test comparing reference and alternate calls in the matched normal sample. The genotype of the germline variants with frequency above 0.1, is determined using a Bayesian Likelihood method\(^98\). Alternate insertion and deletions (indels) are determined in a similar manner, except using a uniform probability of 0.005 in the binomial classifier.

Filtering Mutascope Variant Calls: To determine the final list of high confidence somatic variants the resulting calls can be filtered further. 1) Low coverage: the coverage is below the 0.5 percentile of all targeted bases with one or more reads covering it. 2) Coverage bias: Both reference strands have > 1X coverage while only one strand has > 1X coverage for the alternate allele. 3) Read group bias: A read group bias score is calculated using a chi-square test comparing calls at overlapping read groups (when several read groups overlap the position). The read group bias score should be higher than 10\(^{-15}\) for both normal and tumor sample. 4) Ambiguity of alternate alleles: the
distance between first and second alternate allele is $DS < 0.5$ where $DS = \frac{N_1 - N_2}{\sum N}$, with $N_1$ is the number of reads supporting the first alternate allele, $N_2$ is the number of reads supporting the second alternate allele, and $\sum N$ is the total number of reads with a non-reference allele. 5) **Low mapping quality**: the average RMS mapping quality is below 15. 6) **Lack of somatic evidence**: the fisher-exact p-value is below $10^{-2}$. 7) **Low alternate allele quality score**: the average alternate allele base quality score (AQ) is below $X$. The value $X$ is determined by using MCLUST\textsuperscript{102} assuming two distributions. This resulted in two z scores ($Z_1$ and $Z_2$) for each variant. $Z_1$ and $Z_2$ are the probabilities of the variants belonging to the first (low AQ) and second distribution (high AQ), respectively. We retained variants with $Z_2 \geq 0.95$. 8) **Proximity to an indel**: the SNV is within 10bp of an Indel. 9) **Variant cluster**: there are 3 or more SNVs within 10bp of each other. 10) **Homopolymer (indel specific)**: the indel is within 2bp of a 5nt long homopolymer. 11) **Allelic ambiguity (indel specific)**: there is more than one potential indel at that site. 12) **Indel Cluster** (indel specific) there are 2 or more Indels within 10bp of each other.

**Benchmarking Analysis**

The benchmarking was performed using ROCR package\textsuperscript{116}. Variants in the MIX sample with expected allelic fraction between 0.01 and 0.1 were used as true positive. All false negative prediction score were set to 0. The MIX variants outside of this expected range of allelic fraction were removed from the analysis. For each tool benchmarked, we analyzed somatic mutations that had an observed or expected allelic fraction between 0.01 and 0.1 (as reported by the variant caller). The prediction score used for the classification corresponded to the binomial p-value (**Mutascop**e), somatic p-value
(VarScan), tumor Fstar LOD score (MuTect), and VCF quality score (LoFreq, Illumina MiSeq Reporter and GATK). All false negative prediction score were set to 0. All benchmarked tools relied on the same alignment performed by Mutascope as described above, except for Illumina MiSeq reporter which performs its own alignment. In order to remove the effect of any prior filtering and compare only the classification, the following tools and parameters were used.

- **VarScan**: VarScan v2.3 Somatic caller. The pileup files were generated using default parameters except for -Q 0, -d 500000, -B, and -q 0. The variants were called using default parameters except for --somatic-p-value 1, --p-value 1 --min-avg-qual 0, and --min-var-freq 0.
- **MuTect**: MuTect v1.1.4 was used to call variants using default parameters. We used all variants called (REJECTED or KEEP) for the ROC analysis.
- **LoFreq analysis**: LoFreq v0.40 lofreq_snpcaller.py was used to call variants with default parameters except for --baq=off, -d 500000, and -s 1.
- **Illumina MiSeq Reporter**: We used Illumina’s MiSeq Reporter (version 2.1.13.0), specifically the CallSomaticVariants v2.1.0.0 to call somatic variants.
- **GATK analysis**: We used GATK v1.6-5-g557da77 UnifiedGenotyper to call variants with default parameters except for -mbq 10, -stand_call_conf 0.0, -stand_emit_conf 0.0, -dcov 500,000, and -glm SNP.

In order to estimate also the effect of the high confidence filtering (Figure 3.1f), we have used the Mutascope filters described earlier (Mutascope Analysis), the VarScan somaticFilter and processSomatic steps with default parameters except for --min-var-freq 0.005 and --min-coverage 100, only using the set of ‘High Confidence’ somatic variants.
as determined by VarScan. MuTect v1.1.4 was run with default filters. However not all MuTect filters are appropriate for amplicon sequencing of tumor-normal pairs. To ensure a fair comparison to Mutascope and Varscan filters, we ignored variant flagged for strand_artifact (directional sequencing without fully overlapping paired-end reads can lead to strand bias), clustered_read_position (in direct amplicon sequencing all mutations are located at the same position in the read), possible_contamination (we expect mutations at low allelic ratio), and fstar_tumor_lod (primary classifier used).

**Coverage Down-sampling:** To determine the impact of coverage depth on the performance we sub sampled reads from the MIX sample. We randomly selected 50% and 10% of the aligned reads to simulate reduced coverage depth (Figure 3.1d). Each random subsample of reads was then analyzed through the identical analysis steps and parameters as the full set of reads.

### 3.3 RESULTS

We benchmarked Mutascope against other mutation callers using sequencing data generated from a mixture of 8 normal DNA samples with known genotypes (MIX sample) resulting in “somatic mutations” at variable allelic fraction. The classification of the 162 somatic mutations at low allelic fraction (0.01-0.1) by Mutascope was more accurate than other standard tools (area under the curve: 0.97 - Fig. 1c). Not surprisingly, tools designed to identify heterozygotes in diploid genomes were missing most mutations (GATK), while tools dedicated to tumor-normal pairs performed better (VarScan and MuTect). In order to estimate the impact of coverage depth, we selected reads from the MIX sample down to 50% or 10% (490x & 98x resp.) of the maximum. As expected, the sensitivity decreased equally for all the tools considered (Fig. 1d).
To expand the performance evaluation to additional mutations and experimental conditions, we prepared a set of 80 Tumor-Normal Spiked-In (TNS) pairs by mixing reads obtained from sequencing 38 normal DNA. Using these, we interrogated 402 unique “somatic mutations” (between 17 and 55 per pair) at an allelic fraction of 0.01 or 0.1 (40 pairs each). Mutascope was more accurate to detect mutations at an allelic fraction of 0.1 rather than 0.01 (Fig. 1e), and in the former case its performance was comparable to MuTect and superior to VarScan or LoFreq.

Finally we tested the effect of the empirical filters applied by each tool after the classification. These filters are important to eliminate false positives resulting from unpredictable sources of error and not accounted for by the statistical model. While Mutascope’s filters, such as the read group bias and non-specific amplification, are specifically compatible with amplicons sequencing, we adjusted the parameters of the other tools to ensure a fair comparison, such as strand bias and minimum alternate allele frequency filters. When applied to the mutations at low allelic fraction in the MIX samples, these filters increase the sensitivity and positive predictive value (Fig. 1f and Supplemental Discussion). The set of high confidence filters from MuTect affects the sensitivity the most. This observation highlights synergies between Mutascope’s two core statistical components: the experimentally driven mutation detection (binomial test) and tumor-normal comparison (Fisher test) resulting in a superior performance.

Therefore, by design, Mutascope specifically optimizes the mutation detection and filtering for deep amplicon sequencing. The resulting higher accurate detection of somatic mutations at low allelic fraction increases utility in cancer molecular diagnostics.
3.4 DISCUSSION

We describe here a comprehensive analysis pipeline, Mutascope, dedicated to the analysis of high throughput sequencing of amplicons. In particular our strategy was entirely geared towards the identification of rare somatic mutations from the analysis of a matched pair of samples from the normal and tumor DNA. Importantly we decided to measure this performance based on experimental data designed specifically for that purpose. We first sequenced a heterogeneous mix of 8 Coriell germline DNA samples, for which we independently identified the genotypes. This MIX sample, which consist of an uneven mixture of germline DNA from 8 different individuals, was considered a “heterogeneous tumor” containing mutations of various frequencies and was compared to the major sample in the mix, considered the “matched normal”. Using the genotypes identified in the individual Coriell samples and the relative mixing amount of DNA used in the MIX sample, we inferred the list of 431 expected variants in the MIX sample as well as their expected frequencies (Figure 3.2).

**Sensitivity and Uniformity**

When developing an optimal mutation-calling tool, it is important to distinguish the performance of the assay itself, from the performance of the analysis. For example, it is possible that some alleles do not get amplified due to stochastic or systematic variation in the DNA purification or amplification. In the case of the MIX sample for example, we identified 55/431 expected variants absent from the raw data, of those 44 were expected at an allelic below 0.5%. It is possible that given the expected low allelic fraction, the total number of amplicons targeted, the amount of input DNA used, and the total number of droplets generated have impacted the sensitivity and increased the rate of experimental
false negative. To focus the study on the performance of the analysis, these experimental false negative were excluded to distinguish them from analytical false negative.

Another important consideration when sequencing multiple amplicons in high throughput is the uniformity of coverage. Due to the multiplexing aspect, different amplicons may have different efficiency. Microdroplet PCR can mediate these effects\textsuperscript{117}, but other targeting method such as Molecular Inversion Probes\textsuperscript{114}, Haloplex\textsuperscript{118}, Ampliseq (Life Technologies, Carlsbad CA) or TruSeq Custom Amplicons (Illumina San Diego CA), can also be subjected to this bias. Design and experimental optimization can improve uniformity. A more uniform targeting will guarantee that for a given total sequencing yield, a larger fraction of the targeted region is covered at a given depth thus directly impacting the sensitivity of the sequencing. Assuming the coverage depth is normally distributed around the average, the ideal cumulative distribution represents a step function (blue curve – Figure 3.3a) whereas less uniform distribution can have a spectrum of profiles and in the worst case lead to uncovered bases (red curve – Figure 3.3a). A summary metric of the uniformity is the fraction of targeted base pair within 2 fold of the mean coverage depth (referred to as Cov2x). A perfectly uniform sequencing would give a Cov2x of 1. Our experience shows that digital droplet PCR as a Cov2x of 0.9\textsuperscript{35}. Given this distribution, it is possible to determine the impact of coverage uniformity, or Cov2x, on the sensitivity, given a total sequencing yield (Figure 3.3b). At a Cov2x of 0.9, one needs 1500x average coverage to ensure that 80% of the targeted bases are covered at \(\geq1000\) fold. In contrast, with a Cov2x of 0.55, one would require 6000x coverage to reach the same coverage rate (Figure 3.3b). These results indicate that
highly uniform targeting methods can result in up to 4-fold economy in sequencing yield to reach the same sensitivity.

**Homologous False-positive filtering**

Direct amplicon sequencing is the most specific method to target sequencing to a region of interest, resulting up to 99% of bases on target\(^{117}\). However, PCR amplification can still occur at off-target locations, especially in highly homologous regions of the genome. This can be mediated by careful design of the primers, but some residual unspecific PCR amplification can still occur, even at low efficiency. If this efficiency is slightly different between the amplification of the normal DNA and the tumor DNA, mismatches between the homologous regions could be interpreted as somatic mutation of low allelic fraction. In anticipation of such cases, we have devised a method to predict false positive variants originating from homologous, unspecific amplicons in order to remove them from the analysis. Briefly, for each targeted amplicon we search for all homologous regions by aligning the first and last 151 nt of the amplicon, simulating to two paired-end sequencing reads. Then we use a global Needleman-Wunsch (NW) alignment of the alternate alignment locations to precisely identify all mismatches. Using a NW global alignment is important to capture the full-length fragments that can have multiple mismatches when compared to the targeted amplicons. We then repeat this procedure for all amplicons. The mismatches represent specific nucleotide substitutions that could lead to false positive somatic variants. For a set of 1736 cancer gene amplicons used in the study, representing 155,116 bp of DNA, we could identify 354/465,348 (0.08%) of such nucleotide substitutions of which 54 were actually identify as somatic in one of 38 pairs of tumor/normal sequencing (Harismendy et al, in preparation), This
observation suggests that this filter is not excessively stringent and quite efficient at capturing potential false positive somatic mutations.

**Sequence based genotyping can detect mutations at extremely low allelic fraction**

High throughput sequencing brings the opportunity to screen a broad region of the genome for the presence of rare DNA variants. This is in stark contrast with locus specific assays developed for single position only. Even the more recently developed digital PCR is limited in breadth and relying on site specific fluorescent probes. The breadth of high throughput sequencing solutions comes however with the cost of specificity. Using the same data however, one can restrict the search for mutations to very specific genomic positions and substitution types with generally an increased power to detect rare mutations. Under such paradigm, true positive mutant alleles can be detected below 0.5\% with high confidence. If this level of detection is not directly useful for clinic care, it is promising for the identification of markers in circulating DNA or for the study of the genetic evolution of blood cancers for example.

**Filtering is required after classification**

Most DNA variant callers, whether they were designed for the identification of inherited variation genotypes from diploid DNA sequence, mixed frequencies from pooled DNA sequencing or somatic mutations from a matched tumor-normal comparison rely on a statistical model. The power of this statistical model is generally limited by the prior assumptions such as heterozygosity rate in the population or confidence in the variant calls, whether collected from the instrument (Q score) or measured from the data. The resulting statistical classification helps discriminate real mutations from artifacts, but is far from perfect. Indeed, most statistical models cannot account for all unpredictable
sources of errors and variation in the data. For that reason, each tool usually recommends the use of empirical filters, applied after the variant calling, to improve the accuracy of the data. These filters can include strand bias, minimal coverage depth, presence of clusters of variants, proximity to an insertion-deletion, sequence context, position in the read, minimal average base or mapping quality. In general such filters are similar across all tools, only the threshold can be adjusted depending on the applications. It is not clear however, how to select the appropriate threshold for users with little experience. For Mutascope benchmarking, we wanted to compare variant callers solely based on the basis of their statistical modeling, trying to identify the best classifier, disregarding their respective posterior filtering thresholds. We demonstrate that the binomial approach used by Mutascope results in a slightly better classification that the tumor-normal Fisher test based comparison used by VarScan and much better classification than other tools which do not use data from a matched normal (Figure 3.1c). This result is not entirely surprising. Indeed the sequencing of a matched normal DNA provides an ideal control, prepared in similar conditions. The systematic errors will likely be identical and can therefore be accounted for, resulting in lower false positive rate. In contrast, the absence of a normal control, and/or the reliance on the sole instrument derived quality scores and coverage depth are more likely to result in false positives. In contrast to VarScan, which uses Fisher test for its primary classification, Mutascope or MuTect can use it after the classification, in order to decrease the false positive rate. When comparing the classification of mutations at low allelic fraction (1-10%) between Mutascope, MuTect and Varscan, we notice that in the absence of filters, Varscan’s fisher test is much more powerful at removing false positive mutations (Figure 3.1f, dotted lines). However, once
we implement the high confidence filters, Mutascope classification is superior to Varscan’s, resulting in both higher sensitivity and positive predictive value (Figure 3.1f, continuous lines). The set of high confidence filters from MuTect, although selected for their compatibility with amplicon data (Supplemental Methods), affects the sensitivity the most. This observation demonstrates that the advantages of Mutascope error measurement to detect the mutations (binomial test) and the statistical tumor-normal classification (Fisher test) are synergetic and result in more accurate results.

Chapter 3, in full, is a reprint of the material as it appears in Bioinformatics. 2013 May 27. [Epub ahead of print]. Mutascope: Sensitive Detection of Somatic Mutations from Deep Amplicon Sequencing. Shawn E. Yost, Hakan Alakus, Hiroko Matsui, Richard B. Schwab, Kristen Jepsen, Kelly A. Frazer, Olivier Harismendy. The dissertation author was the primary author of this paper.
3.5 FIGURES

Figure 3.1. Mutoscope principle and performance. (a) The sequencing error rate varies based on the read type (blue, red), position in the read (x-axis) or reference base sequenced (lines). (b) Paired reads (red and blue) from shotgun and amplicon sequencing distribute differently over the targeted region (grey box) resulting in different consensus error rates (right panel). (c-e) Comparison of 4-6 tools by ROC analysis showing the classification of mutations at low allelic fraction (1-10%) in the MIX samples (c), after down-sampling reads to 50 or 10% of max. coverage (d), or using 1% and 10% allelic fraction variants from TNS pairs. (f) Evolution of the true positive rate and positive predicted value from the MIX sample low allele frequency variants (1-10%) before (dotted line) and after (continuous line) application of high confidence filters.
Figure 3.2. Cumulative distribution of the expected frequency (0-50% shown) of the 431 variants in the MIX sample.
Figure 3.3. Impact of coverage uniformity on sensitivity. (a) The normalized coverage depth follows a normal cumulative distribution, modeling from high (blue) to low (red) uniformity. (b) Average coverage depth (y axis) required to cover 60% to 90% of the target base pairs at 1000 fold, given decreasing uniformity (Cov2x x axis). Results of a Cov2x of 0.9 and 0.55 for 80% sensitivity are shown with dashed lines.
CONCLUSION

We designed this research to address issues concerning the identification of true somatic mutations in clinical tumor samples. Identifying true somatic mutations can lead to the discovery of new cancer treatments, determining the efficacy of a cancer treatment given a tumor’s molecular profile, and improve the patient’s treatment options in the clinic. While great strides have been made in cancer research most patients are treated with standard chemotherapy and these toxic regimens still remain one of the highest leading causes of death. Researchers around the world are gathering sequencing data of thousands of cancer patients in order to increase our understanding cancer etiology. With the ever growing amount of Next-Generation Sequencing (NGS) data, we will need to continue to develop new analysis methods that are more accurate and efficient. Therefore I have developed a set of analysis methods that can be used to identify true somatic mutations in a variety of DNA sequencing data generated from both formalin fixed paraffin embedded (FFPE) and frozen tumor samples.

The first challenge of analyzing DNA sequencing data was to identify true somatic mutations from false positive somatic mutation calls from DNA isolated from FFPE blocks. The difficulty is that formalin damages the DNA and damaged DNA can imitate true somatic mutations in NGS data. To solve this problem I started by using a set of published software to processes the whole genome sequencing data of two breast tumor samples that have been stored in formalin. Upon examination of the initial results, I realized that the two samples sequenced had different levels of DNA damage. Sample 02542 FFPE sample had 1.6 fold more reads with at least 1 mismatch than sample 06408. Sample 02542 also had a 2.25 fold increase in known type of mutations caused by...
formaldehyde, CG>TA mutations. This led me to the conclusion that some samples have high amounts of DNA damage caused by formaldehyde while other samples have low amounts of DNA damage caused by formaldehyde. I tested this theory by evaluating DNA sequencing data from a set of 89 FFPE tumor blocks. Even though the data was sequenced using a different sequencing platform, each sample only had ~1 million reads, and the data was aligned to the human genome with a different alignment algorithm, I was still able to distinguish the samples with high DNA damage from those with low DNA damage.

Determining the extent of DNA damage caused by the formaldehyde can help researchers choose which samples to include in their studies. However, it is still necessary to be able to distinguish between true somatic mutations and false positive mutations caused by the formaldehyde regardless of the sample chosen. Using the standard analysis methods at the time, SAMTools and BFAST, resulted in over 10 times more somatic mutations called then the expected number of somatic mutations found in breast cancer. In order to reduce the number of somatic mutation calls to reflect the number of expected somatic calls I developed a set of 3 filters. This first somatic mutation filter I developed, “High supporting read diversity,” was designed to remove false positive mutation calls caused by PCR duplicates. The now standard PCR-duplicate removal software, PicardTools MarkDuplicates, performs this same function. Changing the PCR-duplicate removal step in analysis pipeline should remove the same somatic mutation calls as my first mutation filter.

The second novel filter that I developed, “Normal local mismatch rate,” was designed to remove potential false positive somatic variants in regions with a high
number of mismatches. Currently it is believed that the damage to the DNA caused by the formaldehyde is not sequence specific, other than the known CG>TA substitution mechanism. However, I did find that there were small regions in the genome that had a high number of sequenced bases that were called non-reference. Some of these regions could be caused by normal sequencing errors, such as high GC content or low DNA complexity. Yet the average local mismatch rate for the high FFPE damage sample, patient 02542, was twice as high as the low FFPE damage sample suggesting that this could be caused by the formaldehyde. This theory could be tested by sequencing the whole-genome of multiple samples with DNA was isolated from FFPE blocks. If the formalin preferentially damages the DNA in specific regions then the regions with a high local mismatch rate should overlap between the samples and should not overlap any repeat regions.

Whether or not the regions of high-local mismatch rates is caused by the formalin or some other mechanism, another potential method to remove the false positive somatic mutation calls with a high local mismatch rate would be to align the data to the genome using a more stringent alignment algorithm. The algorithm used in this study, BFAST, was designed to align reads to the genome without taking into account the number of mismatches in the resulting alignment. BFAST does this by first finding potential alignment locations and then aligning with a Smith-Waterman algorithm. Using this approach BFAST aligns more reads to the genome than most alignment algorithms but this also causes BFAST to align low quality reads that should not be aligned. In future work with samples stored in formaldehyde I would use a more stringent alignment algorithm that does not align reads with a high number of mismatches in the alignment.
Changing this step in the data analysis pipeline could eliminate the need to apply the second somatic mutation filter. This change to the alignment algorithm used might decrease the sensitivity of the somatic variant caller but it should greatly increase the specificity.

The third and final novel filter developed, “Global nucleotide mismatch profile”, is designed to remove potential false positive somatic variants caused specifically by formaldehyde. The novel concept behind this method was to use the error rates of the sample in order to filter the somatic variant. Using this filter I was able to remove the high CG>TA mutation bias in patient 02542. This was the only somatic mutation filter that was able to completely remove the somatic mutation bias caused by the formaldehyde. Currently there are no other publications on a method to filter out false positive mutations caused by formaldehyde. In order to further test the accuracy of my somatic filtering method I would propose to sequence matched tumor samples from a patient that have either been stored in FFPE blocks or flash frozen. Both the high FFPE-damaged and the low FFPE-damaged samples would be sequenced in order to determine the effect of the global nucleotide mismatch profile filter on samples that have different levels of DNA damage. Using the matched frozen tumor sample as a standard I could determine the sensitivity and specificity of the third somatic mutation filter I developed.

In summary of the first chapter, I have identified some of the effects formalin has on DNA sequencing data and how these effects influence the alignment and variant calling algorithms. I was able to demonstrate that using the alignment and mismatch rates of the NGS data it is possible to determine the extent of DNA damage caused by formalin a sample has. I was also able to remove potential false positive somatic mutations caused
by the formalin damaging the DNA, using my novel global nucleotide mismatch profile filter. Overall I have established that DNA isolated from FFPE tumor blocks can presumably be used to identify true somatic mutations in the tumor if the correct filters are applied.

In the second chapter of my dissertation I investigated the question of whether genetics of a pre-clinical tumor model derived from a primary tumor matches that of the primary tumor from which it was derived. Previous publications addressing this question have disagreed on whether the genetic makeup of the matched tumor and pre-clinical models are the same. In this chapter I took a different approach than the previous studies when trying to answer this question. I chose to accurately identify both somatic mutations and copy number aberrations (CNAs) in the primary tumor. To do this I started by using published methods to process the exome sequencing data. I quickly came to realize that while the published methods did provide a good starting point for my analysis, they needed improvement.

One example of such deficiency was the discovery of false positive somatic mutation calls caused by PCR duplicates in the sequencing data. Although I used the standard PCR duplicate removal software, my data still contained PCR duplicates. These duplicates were missed due to the alignment software hard clipping the sequencing reads. PCR duplicate reads that were hard clipped with varying lengths would have different start and stop coordinates when aligned back to the human genome. I therefore created my own method to take into account the original lengths of the reads before the hard clipping and removed the remaining PCR duplicate reads. I have mentioned this bug to
the developers of some of the PCR duplicate removal software, such as PicardTools and SAMTools, in hopes that they will take this into account in their next software update.

Another method I developed in my second chapter was a filter designed to remove false positive somatic variants based on their average alternate allele base quality (AAABQ) score. During my analysis of the exome sequencing data in my second paper I found that known variants’ AAABQ scores have a binomial distribution with a mean of ~35. While the novel variants’ AAABQ scores have a bimodal distribution with the higher peak overlapping the peak of the known variants. Thus I used MCLUST, a normal mixture modeling for model-based clustering, to determine an AAABQ score that would separate the distribution of the variants. I found that removing variants in the lower distribution increased the percent of known variants and the transition to transversion ratio, which are two metrics used as quality controls for variant calling. While I was unable to validate the filtering method using Sanger sequencing, in the third chapter I applied this same method and found that only false positive somatic mutations were filtered out.

A similar approach to my AAABQ score filter has been applied to the latest versions of some variant callers, such as VarScan2. However, this approach compares the reference’s average base quality score to the alternate allele’s average base quality scores and applies a fixed hard cutoff. The method I developed does not use a fixed cutoff but instead uses a statistical approach that sets the cutoff based on the data’s base quality score. This is an important distinction to make because different sequencing runs have different average base qualities scores and using the same base quality score cutoff for each sample would not take into account the differences in sequencing runs.
After applying all of the somatic variant filters but before comparing somatic variant calls between matched primary tumors and pre-clinical tumor models, I needed to remove the sequencing reads that were generated from mouse contaminating DNA in the xenograft models. At the time there were no published methods to remove mouse DNA sequencing reads that were generated from a sample containing both human and mouse DNA. The method I developed was a straightforward approach that took into account the similarities and differences between the human and mouse genomes. By aligning the reads to both the human and mouse genomes I could determine which genome a read came from. If the read had a better alignment to the human genome then I determined the read came from a human DNA segment. Otherwise, I either could not determine which genome it came from or that it came from the mouse genome. Using this simple method I was able to remove the false positive somatic variants that were caused by contaminating DNA sequencing reads that belonged to the mouse genome. In the future when trying to remove mouse contamination from sequencing Xenograft samples I would suggest the newly published methods designed specifically to address this issue, such as DeconSeq.

Once I had the final set of somatic variants from both the primary tumors and pre-clinical tumor models, I was able to address the issue of determining which of these somatic variants were shared or unique between the matched samples. I did this using a Fisher-exact statistical test, instead of using hard cutoffs that previous papers addressing this issue have done. Using this approach I was able to determine that ~98% or more of the somatic variants found in either of the matched primary tumors or pre-clinical tumor models were shared between the samples from the same patient. Additionally I tested whether the 2% difference between the matched samples was due to sequencing artifacts
or were actual unique variants to the samples. Due to the fact that we generated multiple
independent libraries from the same sample I was able to test this theory by splitting the
sequencing data based on the library the data was generated from. This allowed me to use
the same methods above but instead of using two sets of reads from different samples I
was able to use two sets of reads generated from the same sample but different libraries.
Using this approach I found that there were ~2% of called somatic variants that were
unique to the libraries generated from the same sample. This suggest that most if not all
of the somatic variant differences between the matched primary tumors and pre-clinical
tumor models were most likely due to sequencing artifacts.

To confirm whether or not the somatic mutations that were identified as unique in
a sample are true positive mutations I would use Sanger sequencing to confirm the
mutations. To further the study I would also suggest sequencing matched primary tumor
and pre-clinical tumor models of different cancer types. Most of the studies that disagree
on if the genetic makeup of the primary tumors and pre-clinical tumor models match are
done using different cancer types. There is a possibility that different cancers behave
differently when they are made into pre-clinical tumor models. The next step after this
would be to determine if the methylation and/or transcriptomes are the same in the
primary tumors and preclinical models using either array or sequencing approaches.
Many of the targeted cancer therapies target expressed genes and if the gene expression
levels are different in the pre-clinical models then the pre-clinical trials would not reflect
the outcome in the clinical trials.

I was able to identify and fix bugs in the current published NGS analysis software.
While we were not able to identify a new affected pathway in our Glioblastoma samples,
we were able to identify the commonly mutated genes (PTEN and TP53) and CNAs (deletion of chr10 and amplification of chr7). In this chapter I was able to determine that the genetic makeup of the primary tumors and pre-clinical tumor models I analyzed were almost identical except for a handful of changes.

In my third and final chapter I developed an analysis software called Mutascope that is designed specifically for processing amplicon DNA sequencing data. With the development of desktop sequencers, such as the MiSeq, clinics can now sequence all clinically actionable genes within 24 hours. Currently the easiest way to extract the coding regions of these clinically actionable genes is with PCR amplification of those regions. Within the last year there has been a rising amount of labs/clinics that are generating some sort of amplicon sequencing data. Most of these labs/clinics are using this technology to analyze cancer patients, specifically identifying DNA somatic mutations. Therefore I felt it was necessary to develop an analysis pipeline designed to call somatic mutations in matched tumor-normal samples using amplicon sequencing data.

Amplicon sequencing data has various unique characteristics when compared to regular shotgun sequencing data, such as the data analyzed in chapters one and two. Unlike shotgun sequencing data, amplicon sequencing reads have a fixed start and stop coordinate, fixed directionality, and variants sequenced are always seen at the same position in the reads. The software suite I developed, Mutascope, uses this information to implement novel methods for analyzing amplicon sequencing data. One of the biggest benefits of using Mutascope is the ability to go straight from the raw output of the sequencing platforms (FASTQ files) to a list of germline and somatic variants in the patient (VCF file). This option is essential for labs/clinics that do not have strong
bioinformatics support and only want a list of somatic mutations in the cancer sample. I have worked to analyze the challenges of using amplicon sequencing data and discover solutions to such challenges.

One of the first challenges that I came across when analyzing amplicon sequencing data was that each PCR amplicon was not always specific to one region of the genome. There are numerous genes that have homologous genes in the genome that make it sometimes impossible to design a PCR amplicon that only amplifies one region of the genome. In order to remove false positive mutation calls that can arise from amplification of homologous regions in the genome I developed a Mutascope module called “makeBlackList”. Given a set of amplicons this module will identify homologous regions in the genome for each amplicon, if one exists, and will then identify the possible false positive mutations caused by the off-target amplification of these homologous regions. These potential false positive mutation sites are then ignored when calling variants. One concern about using this module would be that it removes too many possible mutation sites in the targeted regions. However only 0.08% of the potential mutation calls were blacklisted in an amplicon set that contains ~150,000bp of the genome. Thus this module will greatly increase the specificity of the somatic mutation calling while only slightly, if at all, decreasing the sensitivity.

Another feature of Mutascope is that for each sequenced read it can identify the PCR amplicon it originated from. This is done in order for Mutascope to calculate coverage and logR values for each amplicon and most importantly remove any off-target aligned read. The off target reads are removed because amplicon sequencing data should align to the specific regions being amplified. Mutascope also adds a read group to each
read that specifies which amplicon the read belongs to. With this Mutascope can easily identify which amplicon a read came from without having to calculate it. This read group information is used later on in the Mutascope pipeline to help calculate error rates and filter out false positive variants. This method of assigning a read back to where it originated from is unique to Mutascope.

Not only does Mutascope provide unique read processing steps for analyzing amplicon sequencing data but it also has unique variant calling and filtration methods. Mutascope is the first variant caller designed to detect somatic variants that are at 1% prevalence in the tumor sample. As shown in chapter 3, Mutascope has the same positive predictive value (PPV) to detect somatic mutations at a 1-10% prevalence as the current published variant calls but Mutascope’s a sensitivity 1.5 times greater than the other variant callers. This improvement in sensitivity without effecting the PPV is mainly due to the variant probability score calculated for each variant and the amplicon sequencing data specific variant filtration.

Mutascope’s variant probability score calculated for each potential somatic variant does a better job at classifying variants than any other variant caller. The reason it does so much better than other variant callers is that it uses the error rates of the samples to determine the probability of a variant being a true variant. These error rates are calculated by Mutascope reflect the sequencing platforms error rates based on if the base is in the first read sequenced or the second, what position the base is within the read, and the mutation type of the potential somatic mutation. This concept is similar to that of GATK’s base quality score recalibration except that unlike GATK’s algorithm,
Mutascope’s error rate calculation can work with DNA sequencing data of a small region of the genome.

During the variant calling step, Mutascope also calculates a Bayesian likelihood probability for each germline variant. Mutascope is the first variant caller that uses a different statistical test to identify germline and somatic variants. While germline variants should be in 50% or 100% of the sequenced reads, somatic variants can have a wide range of frequency. This is why most somatic variant callers are designed to detect only somatic variants, for example MuTect. MuTect, produced by the Broad Institute, only calls somatic mutations using a matched tumor-normal sample. However the Broad Institute also has a germline variant caller, GATK UnifiedGenotyper, which is dedicated to calling germline variants. Instead of having two separate variant calling algorithms I have implemented both algorithms, a Bayesian likelihood algorithm to call germline variants and a Binomial test to call somatic variants, in Mutascope’s single variant calling module. Doing this allowed me to optimize both the germline and somatic variant calling in Mutascope’s variant caller.

One approach to improve the results of Mutascope’s variant caller was to use as little hard-cutoffs as possible when filtering the variants. The problem with using a strict cutoff to filter a variant is that this cutoff is usually created to work perfectly with a specific sample. However, each sequencing run is different and the optimal filtering cutoff for one sample may not be optimal for another. Taking this into account I use the coverage distribution of a sample in order to determine the minimum coverage required to call a variant. All of the current published variant callers use a default minimum coverage cutoff that is not always an ideal cutoff for samples with either very low or high
coverage. Mutascope also uses the average alternate allele base quality (AAABQ) score filter that I developed in chapter two, which calculates a minimum AAABQ score for each sample independently. These are just a couple examples of how I have developed methods to use information from the data to determine the correct filtering cutoffs for that sample.

While using a statistical approach to determine the cutoffs for filtering variants may be ideal, unfortunately this is not always possible. Some of Mutascope’s amplicon specific filters do use hard-cutoffs, like the distance to the second allele (DS) filter. The DS filter removes potential variant sites that have a high number of non-reference alleles in the sequencing reads. The only problem with this filter is that it removes any variant site that has more than 2 alleles. Some cancer studies have found that cancer cells within a single tumor can have different somatic mutations at the same position in the genome. In the future I would change Mutascope to look for two or more allelic sites in order to further expand the capabilities of Mutascope’s variant caller. Currently there are only a couple of variant callers published that allow the user to look for tri-allelic sites in a cancer sample.

Another feature that I would like to add to Mutascope is a module to analyze a tumor sample without a matched normal sample. After publishing Mutascope I have received various emails from users who do not have a matched normal sample and still wish to identify variants in their tumor sample. The main problem with adding this module to Mutascope is that I wouldn’t be able to determine which variants are somatic and which are germline. I also think that I would need to increase the minimum alternate allele frequency to call a variant from 1% to 5% because I would no longer have the
matched normal sample to help filter systematic errors. I would also like to implement Mutascope in Java so I could upload the program to Illumina’s BaseSpace. BaseSpace is Illumina’s cloud computing and they have opened up their server to allow people to upload software for any Illumina MiSeq user to use and implement locally on their MiSeq. This would greatly increase the user base for Mutascope.

In conclusion I was able to develop a unique software suite called Mutascope that is designed to analyze amplicon sequencing data. Mutascope outputs multiple quality control metrics to determine if there was a problem with the PCR amplification or sequencing run and it can also detect sample mixing or sample swapping. Mutascope is the first variant caller that is designed to detect somatic variants at a 1% allelic frequency in the tumor sample. It is also designed to be user friendly so it can easily be implemented in the clinic. The goal for Mutascope is to be used in clinics around the world to facilitate personalized cancer treatment in the clinic by providing an accurate list of somatic mutations.

In summary, I have worked on developing methods and analysis pipelines to accurately identify somatic mutations in clinical tumor specimens. The methods I have developed have been successful in identifying somatic mutations in tumor DNA collected from FFPE tumor blocks, xenografts, and tumor samples with high levels of normal contamination.
REFERENCES


