Statistical methods for improving low frequency variant calling in cancer genomics

by
Brian K. Mannakee

A Dissertation Submitted to the Faculty of the
Mel and Enid Zuckerman College of Public Health
In Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy
In the Graduate College
The University of Arizona
2019

2

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Brian Mannakee
titled:

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Ryan N Gutenkunst — Date: Oct 31, 2019

Denise Roe — Date: Oct 31, 2019

Justina Dolorita McEvoy — Date: Oct 31, 2019

Edward John Bedrick — Date: Oct 31, 2019

Megha Padi — Date: Dec 4, 2019

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Ryan N Gutenkunst — Date: Oct 31, 2019
Molecular and Cellular Biology

Acknowledgments

I am deeply grateful to the many people who have supported and encouraged me on this journey. My adviser, Dr. Ryan Gutenkunst has given me the gift of his mentorship and guidance for nearly 9 years. When I joined his lab as an undergraduate I had no idea of the incredible journey we would take together. It has been my privilege and delight to work with Ryan, and the best decision I ever made.

The generous members of my dissertation committee, Drs. Bedrick, Roe, Padi, and McEvoy, provided invaluable guidance and encouragement. I would like to particularly thank Dr. Joanna Masel, whose lab is jointly located with ours, and who provided invaluable training in the rigorous analysis of evolutionary processes that formed much of the basis for my thinking about cancer evolution. Over the years, the members of the Gutenkunst and Masel groups, too many to list here, have been wonderful and gracious colleagues who helped me hone my craft and provided a comfortably and collegial work environment.

My parents Bruce and Sandra Mannakee, and Lee and Nancy Mills, provided financial and emotional support that made this journey possible. Without their generosity and encouragement I would never have been able to pursue the incredible opportunities that have come to me over the last 8 years.

Table of Contents

TABLE OF CONTENTS—*Continued*

List of Figures

List of Figures—*Continued*

List of Figures—*Continued*

List of Tables

Abstract

Cancer is not a single disease, but a family of genomic diseases characterized by a set of initiating genomic variants accumulated in a single cell that allows that cell to begin dividing uncontrollably. Tumors grow by cell division, and each cell division generates a new set of variants that are passed along to its offspring. As a result, at the time of diagnosis a typical tumor of approximately 100,000,000 cells contains hundreds of millions of genomic variants, whose frequency in the population is a function of the time that they arose. Mutation accumulation through both inheritance and de novo variant production results in a final tumor in which the vast majority of variants are present at low frequency. Current methods used to identify variants have difficulty identifying low frequency variants. Here I will describe two algorithms aimed at improving low frequency variant calling in two settings.

Patient-Derived Xenografts (PDXs) serve as avatars for individual patient disease as well as invaluable models for studying basic cancer biology. Molecular characterization of PDXs is common, but the extensive homology between human and mouse genes present special challenges in sequencing tumors grown in mice. In Chapter 2 I describe an algorithm and R implementation called MAPEX that allows labs studying PDXs to use commercial sequencing technologies and locally filter false positive variants caused by sequence homology.

Detecting somatic mutations within tumors is key to understanding treatment resistance, patient prognosis, and tumor evolution. In Chapter 3 I present BATCAVE (Bayesian Analysis Tools for Context-Aware Variant Evaluation), which extends current state-of-the-art statistical models for tumor variant calling. I also present an R implementation of the algorithm, and show using simulations that the BATCAVE algorithm improves variant detection.

Chapter 1
Introduction

1.1 Background

1.1.1 Cancer genomics

Cancer is not a single disease, but a family of genomic diseases characterized by a set of initiating genomic variants accumulated in a single cell that allows that cell to begin dividing uncontrollably (Nowell PC., 1976; Fearon & Vogelstein, 1989). Following initiation, tumors grow through the process of cell division, and during this expansion each cell division generates a new set of variants that are passed along to its offspring (Bozic et al., 2016; Williams et al., 2018a). As a result, at the time of diagnosis a typical tumor of approximately 10^8 cells contains in the millions of genomic variants, whose frequency in the population is a function of the time that they arose. Early mutations are present in a large proportion of cells, and late mutations are present in only a small proportion (Bozic et al., 2016; Williams et al., 2018a). This process of mutation accumulation by processes of both inheritance and de novo variant production results in a final tumor in which the vast majority of variants are present at low frequency (Williams et al., 2016). Cancer genomics is the study of the variants present in cancer, and variants are identified by examining the DNA of tumors and determining what variants are present. The nature of the methods used in cancer genomics is such that high frequency variants are relatively easy to identify while low frequency variants are difficult (Cibulskis et al., 2013).

Large numbers of tumors are sequenced to generate catalogues of mutations that occur frequently in cancers, either arising from a specific tissue, or generally across tumors (Pleasance et al., 2010). This allows research towards identifying common cancer drivers which may be broadly useful as drug targets (Bailey et al., 2018), as

well as the discovery of genomic biomarkers that identify potential treatments (Way et al., 2018) or likely outcomes (Liu et al., 2018) for particular genomic profiles. The International Cancer Genome Consortium has a propagated a target for mutation catalogues, suggesting they use computational pipelines that identify 80% of mutations present in every sample added to the catalogue (recall), and that 95% of all variants added are truly present (precision). Sequencing depths as high as 10000X would be required to meet the recall goal (Williams et al., 2018a), but false positive rates increase as sequencing depth increases (Cibulskis et al., 2013), and there is no current variant calling algorithm that can meet the precision portion of the goal at that depth (Griffith et al., 2015). In addition, sequencing individual tumors can be used in clinical applications to identify tumors with validated genomic biomarkers for treatment or prognosis (Liu et al., 2018). Since most variants present in a tumor are at low frequency, clinical sequencing will benefit greatly from identifying low frequency mutations (Jacobs et al., 2018). In both uses of tumor sequencing reliable identification of low frequency variants will be crucial going forward.

The earliest examples of cancer genomics studies were simply microscopic examinations of leukemia samples with specially stained chromosomes to identify large scale chromosomal aberrations such as fusions and translocations Mardis (2018). This led to the discovery of the several activating gene fusions that led to transformation and uncontrolled growth in blood cells in leukemia patients, and the development of targeted drugs that shut off the activity of these fused genes Mardis (2018). The completion of the human reference genome in 2004 generated a DNA template allowing the identification of the location of specific genes in the genome, and led to the use of targeted polymerase chain reaction (PCR) amplification of specific genes, and the enumeration of their actual sequences by Sanger sequencing. These studies facilitated annotation of the majority of the human exome, and when carried out in cancer samples led to the discovery of driver mutations in many cancers that were subsequently shown to be treatable with targeted drugs Mardis (2018). While these technologies

led to early successes, they were limited to identifying genomic aberrations present in the very large majority of cells, and could not provide a complete picture of the genomic aberrations in any individual tumor. Massively parallel sequencing , often called next generation sequencing (NGS), was developed in the mid-2000's and revolutionized cancer genomics. An excellent review of the history of cancer genomics is provided by (Mardis, 2018).

1.2 Cancer variant types

Genomic alterations in cancer fall into three general categories. Single nucleotide variants (SNVs) are single base alterations in which one nucleotide is exchanged for another. Small insertions and deletions (INDELs) are insertions or deletions in the genome of the tumor, typically including changes spanning up to 50 base pairs (Carvalho & Lupski, 2016). Structural variants are large scale genomic rearrangements, including copy number variants, gene fusions, deletions, amplifications, inversions, and translocations (Carvalho & Lupski, 2016). Variant callers typically focus on either small variants (SNVs and INDELs), or structural variants, but typically not both. Structural variants are the primary driving mutations in a large number of cancers (Zhang et al., 2018), and there are many structural variant callers that focus on different types of structural variants (Cameron et al., 2019). In this work I focus exclusively on SNVs, and the types of sequencing experiments and variant callers used to identify them are described in detail below.

1.2.1 Next generation sequencing (NGS)

Next generation sequencing allows for the comparison of tumor genomes with their corresponding patient genomes and thus the identification of the differences between the tumor genome and the normal genome of the patient. The NGS workflow consists of an automated series of sequential processes. The cells of a tumor sample are lysed

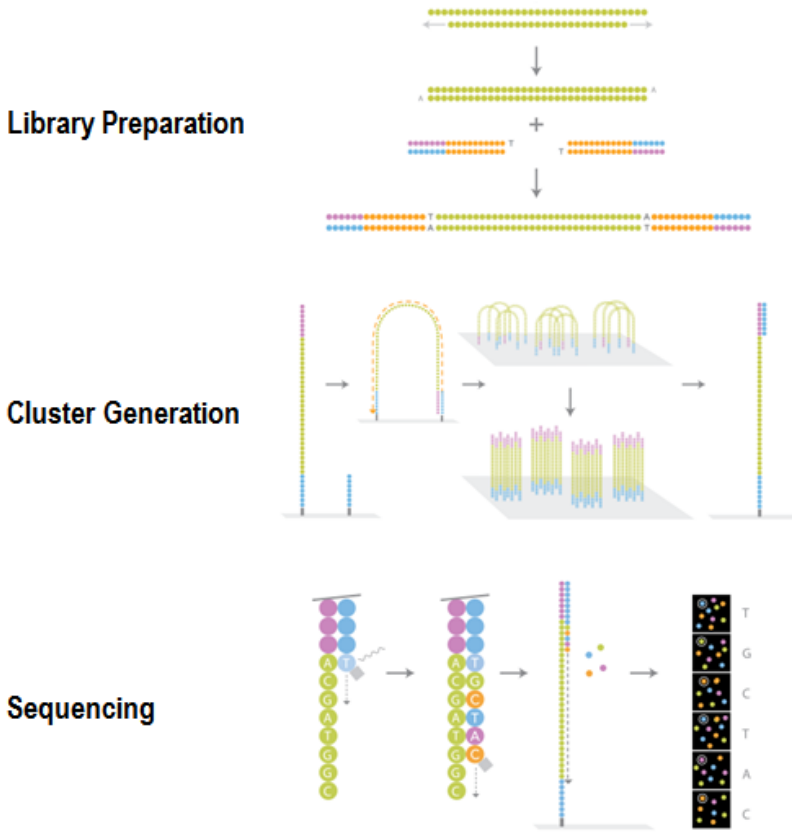

FIGURE 1.1. **Illustration of NGS.** From https://gsc.ku.edu/about

and DNA is separated out then sonically fragmented. Enzymatic reactions convert the fragments to a uniform length and synthetic adapters of known sequence are ligated to each end (Fig. 1.1). This process generates a library of 10s to 100s of millions of fragments that are then washed onto a micro-fluidic flow cell where the adapter on each individual fragment hybridizes to the surface. These individual reads are then amplified generating clusters of identical copies of each library fragment across the flow cell. Fluorescently tagged bases are then washed across the flow cell in cycles, and bases are incorporated where they are complementary. Following each cycle the flow cell is illuminated and the incorporated base at each cluster by recording the fluorescence color emanating from each cluster. Cycles continue until reads reach the desired length (Fig. 1.1). Full length read sequences are then compiled computationally to generate 10s to 100s of millions of individual sequencing reads. Crucially, since all of the DNA present in the sample is used to generate the library, and each cluster is amplified from a single DNA fragment, the fraction of reads that carry a genomic variant relates exactly to the fraction of cells in the sample that carry that variant depending on the ploidy of the cells. The sequencing machine generates a quality score for each sequenced base which is derived from an experimentally generated error model. The quality, or (q), scores typically range from 0 to 50 and the probability that a base has been called in error is equal to $10^{-q/10}$. In downstream processing a minimum base quality score filter is generally applied such that the error rate is no more than a specified number, often $0.1 - 1.0\%$.

1.2.2 NGS read alignment

Following sequencing read generation, reads must be aligned to a reference genome. Fig. 1.2 is an illustration of aligned genomic reads for both a tumor and normal sample from the same patient. Both tumor and normal are aligned to the same reference genome. Every alignment algorithm assigns a quality score for each read

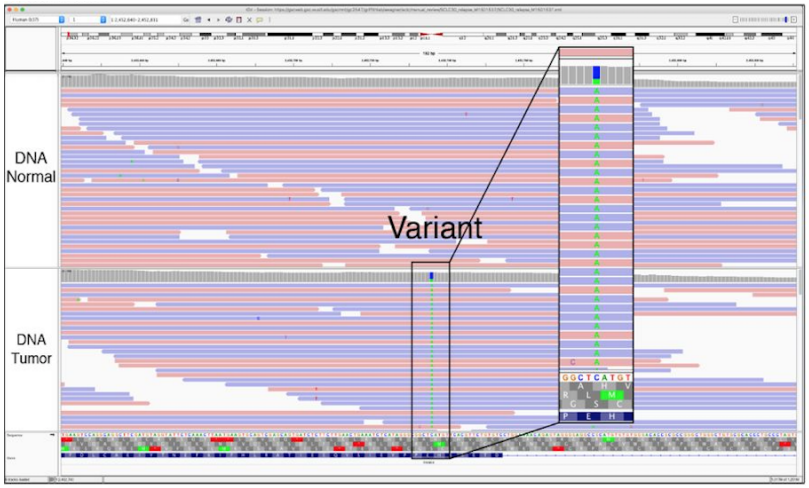

FIGURE 1.2. **Illustration of aligned reads.** Integrated Genome Viewer output for aligned reads from a tumor and normal sample from the same patient from (Barnell et al., 2019).

aligned that represents the same quantity as the q score from the sequencer except that this is the probability that the read has been aligned in error. The number of reads that align to a given base in the genome is called the coverage at that genome. Sequencing experiments typically aim for an average coverage across the genome. For tumors this is typically a minimum of 30X for whole genomes and often greater than 100X for whole exomes.

1.2.3 Variant calling

The data at each site in the genome is often referred to as a *pile* of base calls and their associated quality and alignment scores. Variant callers walk along the reference and generate a pileup at each site consisting of the reference genome location, the identify of the reference base, and the identity and associated base and alignment quality scores of every base aligned to that site. All variant callers apply a set of heuristic filters at this stage to remove base calls with a high error probability. Alignment-specific heuristic filters include those that remove reads with low alignment

quality, reads with many variants close together suggesting problems with aligning highly homologous regions of the genome, and reads with excessive cropping that may be due to insertions of deletions Cibulskis et al. (2013); Kim et al. (2018). There are also heuristic filters that filter variants found only on one DNA strand, or only one component of a paired-end read (Cibulskis et al., 2013). In combination these heuristic filters remove variants that result from known failure modes in sequencing and alignment, and are an essential component in accurate variant calling (Cibulskis et al., 2013). After heuristic filtering, the filtered read pileup is the raw material used to call variants. Statistical variant callers fall into five main categories; heuristic thresholds, joint allele frequency callers, joint genotype callers, machine learning and ensemble callers, and deep neural network classifiers (Xu, 2018).

Heuristic thresholds Varscan was the first somatic mutation calling algorithm in wide use (Koboldt et al., 2012). Varscan generates a pileup of the reads at every sequenced location in the genome for both the tumor and normal sample and uses a heuristic test to determine the genotype of both the tumor and normal sample at each site. For sites where the genotypes don't match Varscan performs Fisher's exact test on a two-by-two table where the entries are the number of reference-supporting and variant-supporting in the tumor and normal samples. If the p-value of the test exceeds a specified threshold – default 0.10 – the site is classified as a somatic variant, and otherwise the site is classified as germ-line. Characteristics of the output from this method are discussed in detail in Chapter 2.

Joint allele frequency callers Joint allele frequency callers model the probability that a mutation to one of the three non-reference bases has occurred at a given site at an allele frequency equal to the fraction of all base calls representing the non-reference base, given the pileup in both the tumor and normal sample (Cibulskis et al., 2013; Saunders et al., 2012; Wilm et al., 2012; Shiraishi et al., 2013; Gerstung et al., 2012; Carrot-Zhang & Majewski, 2017; Kim et al., 2018). These callers test the hypothesis

$$\mathbf{H_0}: \quad \text{Alt allele} = m; \quad \nu = 0 \tag{1.1}$$

$$\mathbf{H_1}: \quad \text{Alt allele} = m; \quad \nu = \hat{f}, \tag{1.2}$$

where m is the identify of one of the three non-reference bases and $\hat{f}$ is the maximum likelihood variant allele frequency calculated as the proportion of reads supporting the alternate allele. Each allele-frequency based variant caller uses a different functional form for both the prior probability of a particular hypothesis and the likelihood of the hypothesis given the pileup. One of the most widely used somatic variant callers, MuTect (Cibulskis et al., 2013), is of this type. A detailed description of the statistical model used by MuTect is given in Chapter 3.

Joint genotype callers Joint genotype callers model the joint probability that the tumor and normal sample together have any of 10 possible diploid genotypes (i.e. AA, AC, AG, AT, CC, CG, CT, GG, TT) (Larson et al., 2012; Roth et al., 2012; Christoforides et al., 2013; Jones et al., 2016; Dorri et al., 2019). As with allele frequency based callers, the functional form of both the prior probability of a genotype and the likelihood of the read data given the genotype differs widely. While joint genotype callers remain under active development, they are used much less broadly than the most popular allele-frequency based callers. Since they are not widely used this type of caller plays no further role in this work, although the statistical models used here are amenable to the treatment described in Chapter 3.

Machine learning and ensemble callers Recent developments in somatic variant calling have been to apply machine learning algorithms to the problem. These include both typical machine learning algorithms such as logistic regression (Ainscough et al., 2018) and random forest/regression trees (Ainscough et al., 2018; Wood et al., 2018) as well as deep convolutional neural networks (Ainscough et al., 2018). Ainscough et al. (2018) trained a logistic regression, random forest, and a convolutional neural network

on the same set of features acquired from a large number of samples in The Cancer Genome Atlas (TCGA). The models were trained on two-thirds of the TCGA corpus and tested on the held-out one-third. Both random forest and the neural network out-performed logistic regression by a significant margin, and their performance was very similar to each other. While the two more sophisticated algorithms performed similarly they relied on substantially different features to make predictions. The most heavily weighted feature in the random forest was the count of variant reads in the tumor, suggesting that it was using the read data as strong evidence for the presence of a variant. The neural network placed the most weight on the type of cancer the sample came from, along with a variety of quality metrics assigned by the alignment algorithm to describe the confidence in the alignment. Tumor variant read count was among the least important of the 30 features for the deep learning method. Allele frequency was a heavily weighted feature in both methods. An important weakness of machine learning models that require training data is that they typically train on data from catalogues of somatic mutations such at TCGA. By construction, these catalogues contain variants that were by callable by another simpler variant calling algorithm, and typically only include the most reliably called mutations, which are typically also the highest frequency. As a result, while they achieve incremental improvement over existing methods on the set of easily callable mutations, it is as yet unclear whether they will continue to perform well at higher sequencing depths and lower frequencies.

Deep neural networks The newest frontier in somatic variant calling is deep learning, applying algorithms originally designed for image classification to NGS data. The leading effort in this direction so far has been from Illumina which recently published their method NeuSomatic, a deep convolutional neural network that takes as input a variety of feature matrices derived from read data (Sahraeian et al., 2019). They find that NeuSomatic performs dramatically better on various synthetic datasets than

other variant callers including Varscan2 and MuTect2, when trained and tested on data split from the same data set. Their final neural network has nearly 800,000 parameters, and the values of these parameters are uninterpretable. On the single real data set they tested, a cell line with validated mutations, they used a network trained on a synthetic dataset from the ICGC-TCGA dream challenge. Performance of the stand-alone neural network was significantly worse than the performance of both MuTect2 and Strelka2 on the same dataset. As with the deep learning methods, at this point there is no indication of whether or not this method will provide improved performance at lower allele frequencies.

1.3 Patient-derived Xenografts

Testing the large number of potentially beneficial new compounds that are in constant development in patients is impossible for safety reasons. Patient-Derived Xenografts (PDX) are real human tumors grown in mice in order to facilitate the testing and development of therapeutic compounds (Witkiewicz et al., 2016). When testing new compounds, it is often advantageous to sequence responding and non-responding tumors to understand the molecular characteristics that lead to drug vulnerability or resistance (Knudsen et al., 2017). A significant difficulty arises when sequencing tumors grown in mice because the stroma and blood supply of these tumors are derived from mouse cells (Mannakee et al., 2018). Even with careful dissection of the tumor sample, a substantial fraction of the DNA isolated from these tumors will be mouse germ-line DNA (Witkiewicz et al., 2015). Because the mouse genome has many genes with very high homology to human genes, alignment algorithms will produce alignments with high quality scores in which mouse reads are aligned to their homologous human genes (Woo et al., 2019). As a result, a standard variant calling pipeline will confidently call as somatic variants places in the genome where the mouse reference differs from the human reference (Woo et al., 2019). In the past, this has necessi-

tated manual review of all called variants along with heuristic filtering based on lists of known potential false positive sites due to mouse contamination (Knudsen et al., 2017). In Chapter 2 I describe the MAPEX algorithm, and an associated R package, which uses the NCBI BLASTN program to filter mouse reads from the alignment, allowing confident calling of human mutations without extensive manual review or heuristic filtering.

1.4 Tumor mutational signatures

Somatic mutations arise as a result of intrinsic errors in the replication process, exposure to mutagens, base modifications that effect replication fidelity, and damage to the DNA repair machinery. Each of these mutational process generates mutations that preferentially occur in particular tri-nucleotide contexts (Alexandrov et al., 2013a; Helleday et al., 2014a). The tri-nucleotide context of a genomic site consists of the identity of the reference base and the 3' and 5' flanking bases. Folding the central base to the pyrimidines, there are two possible bases at the focal site, and there are four possible bases 3' and 5' of the focal site, yielding $2 \cdot 4 \cdot 4 = 32$ possible tri-nucleotide contexts. For each of these 32 tri-nucleotide contexts a mutation can be to any one of the three alternate bases, for a total of 96 substitution types. Figure 1.3 shows a number of mutational signatures and the biochemical processes that generate them.

Every tumor acquires random mutations generated by their exposure to mutational processes. It is possible to generate a tri-nucleotide context mutation profile for an individual tumor, constructed the same way as those in Fig. 1.3, by taking the variants called in the tumor and computing the proportion of all mutations in each of the 96 substitution types. Somatic mutations in every tumor are a random sample of substitution types drawn from a data generating process represented by the mutation profile. Two tumors with exactly the same exposure to the same mutation processes will generate different sets of mutations, but they will have the same mutation profile.

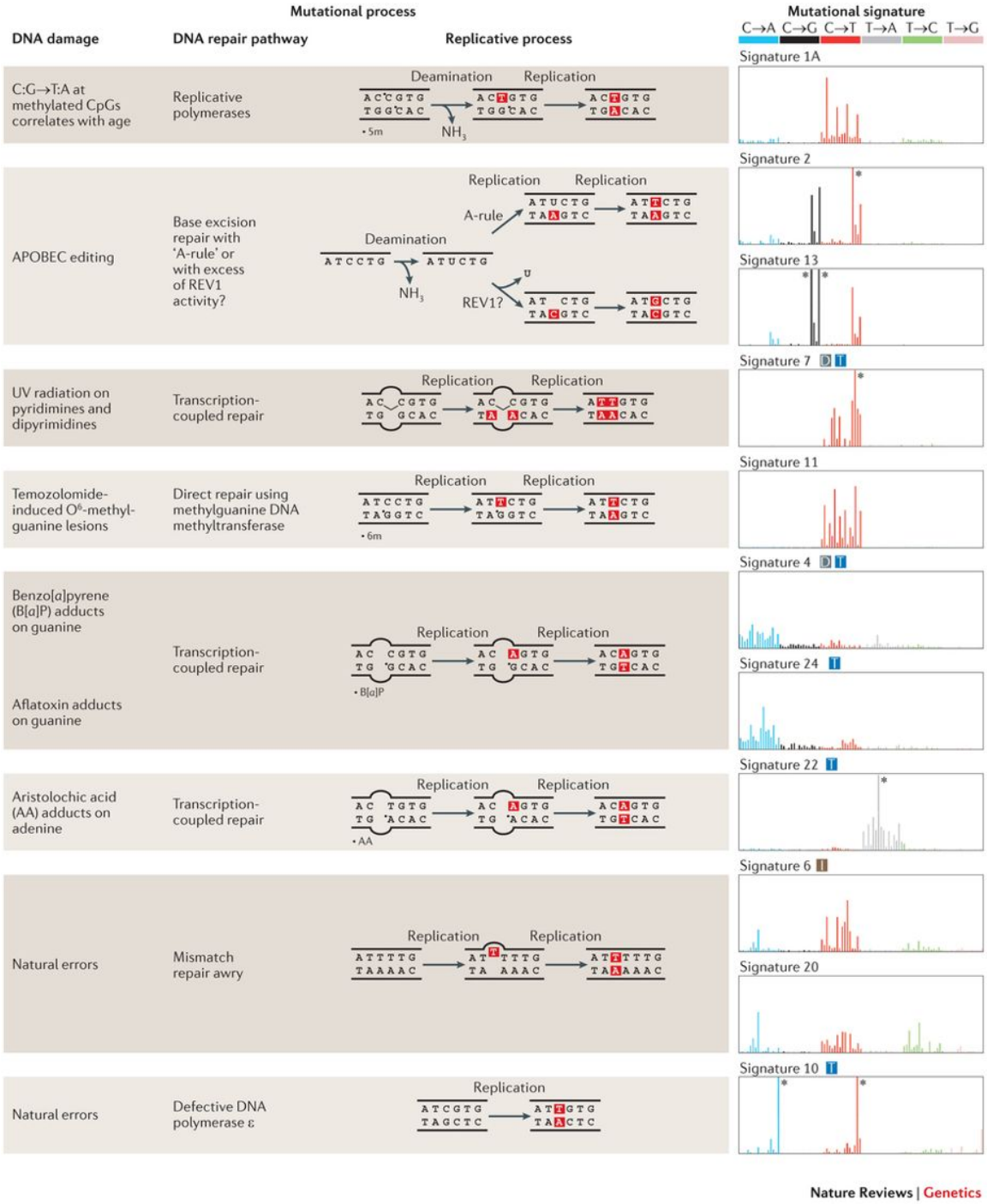

FIGURE 1.3. **Mutation Signatures.** Illustration of the characteristic tri-nucleotide context signatures of different mutational processes (Helleday et al., 2014a)

The probability that a new somatic mutation will be of a particular type is the proportion assigned to that type in that tumor's mutation profile. Thus, the mutation profile of a tumor provides a powerful tool to sharpen the statistical model used to call tumor variants. For all of the variant calling algorithms described above, the model for the posterior probability that a variant is present is the product of the prior probability of a mutation at that site, and the likelihood of the mutation given the read data. Every current variant caller assumes that this value is the same at every site in the genome, which ignores the biology of the mutational processes driving mutations. In Chapter 3 I describe a modification of an allele-frequency based statistical model which uses the mutation profile of the tumor to generate a tumor-and-site-specific prior probability of mutation at every site in the genome.

Chapter 2

Sensitive and specific post-call filtering of genetic variants in xenograft and primary tumors

Originally published as: Brian K Mannakee, Uthra Balaji, Agnieszka K Witkiewicz, Ryan N Gutenkunst, Erik S Knudsen, Sensitive and specific post-call filtering of genetic variants in xenograft and primary tumors, Bioinformatics, Volume 34, Issue 10, 15 May 2018, Pages 1713 – 1718, https://doi.org/10.1093/bioinformatics/bty010

2.1 Abstract

Motivation: Tumor genome sequencing offers great promise for guiding research and therapy, but spurious variant calls can arise from multiple sources. Mouse contamination can generate many spurious calls when sequencing patient-derived xenografts (PDXs). Paralogous genome sequences can also generate spurious calls when sequencing any tumor. We developed a BLAST-based algorithm, MAPEX, to identify and filter out spurious calls from both these sources.

Results: When calling variants from xenografts, MAPEX has similar sensitivity and specificity to more complex algorithms. When applied to any tumor, MAPEX also automatically flags calls that potentially arise from paralogous sequences. Our implementation, `mapexr`, runs quickly and easily on a desktop computer. MAPEX is thus a useful addition to almost any pipeline for calling genetic variants in tumors.

Availability: The `mapexr` package for R is available under the MIT license at `https://github.com/bmannakee/mapexr`.

Contact: rgutenk@email.arizona.edu

Supplementary information: Supplementary data are available at *Bioinformatics*

online.

2.2 Introduction

Molecular characterization of tumors is an important tool in cancer research, and the large-scale sequencing of cancer genomes has led to a deeper understanding of many aspects of the biology of cancer (Stratton MR, 2011). It is now common to sequence tumors from large cohorts of patients, as well as patient-derived xenograft (PDX) models from individual patients. Such sequencing enables identification of mutational signatures (Alexandrov et al., 2013a), functionally important variants (Ding et al., 2012) and evolutionary history of the tumor (Nik-Zainal et al., 2012b; Carter et al., 2012). These genetic features are relevant in evaluating etiological mechanisms (Yachida et al., 2010), prognostic subtypes (Shah et al., 2009; Park et al., 2010), and acquired therapeutic resistance (Witkiewicz et al., 2015). All these applications of tumor sequencing depend on sensitive and specific characterization of low-frequency mutations, and as a result may be biased by spurious variant calls. Here we focus on two specific sources of spurious calls, mouse cell contamination in PDX tumors and mis-alignment of paralogous sequences.

PDX models serve as avatars for individual patient tumors when studying intra-tumor heterogeneity and metastasis and when screening anti-cancer compounds (Dawson et al., 2012; Day et al., 2015; Bruna et al., 2016; Allaway et al., 2016; Knudsen et al., 2017). The primary difficulty in sequencing these models is that mouse stroma is present in all PDX tumors. The high genetic similarity between mouse and human then causes bias when variants are called using bioinformatic pipelines originally developed for primary tumors (Rossello et al., 2013; Tso et al., 2014). Several methods have been developed to facilitate the accurate calling of variants in PDX models. Experimentally, human-specific fluorescence tags can be used to label and isolate human cells prior to DNA extraction (Schneeberger et al., 2016). Bioinformatically,

sequence reads can be aligned to both human and mouse reference genomes, either separately (Conway et al., 2012; Khandelwal et al., 2017) or simultaneously (Bruna et al., 2016), to filter out mouse reads prior to variant calling. Although these approaches greatly improve the reliability of variant calls from PDX models, they entail substantial experimental or bioinformatic burdens. Here we describe a lightweight filtering algorithm that achieves equivalent reliability and can be easily added to standard bioinformatic pipelines, because it uses the same reference genome for alignment as primary tumors.

Many human genes have highly similar paralogous sequences in the genome. Spurious variant calls arising from such paralogs have been recognized as an important source of false positives in the study of rare disease-associated germline variants (Ng et al., 2010; Jia et al., 2012; Zhou et al., 2015; Mandelker et al., 2016). Similarly, paralogs have led to false positives in the study of cancer, including TUBB in non-small cell lung cancer (Kelley et al., 2001), PIK3CA in hepatocellular carcinoma (Tanaka et al., 2006; Müller et al., 2007), and MLL3 in myelodysplastic syndrome (Bowler et al., 2014). To address the paralog problem, some variant callers, such as MuTect2 (currently in beta but included in the Genome Analysis Toolkit (GATK; McKenna et al. (2010))), filter clustered variants, which often result from mis-alignment of paralogous sequences. Many labs also keep lists of suspect genes that tend to suffer from paralog problems and simply ignore any variants called in these genes. These approaches introduce their own biases. Our approach automatically identifies potential spurious calls from paralogs and enables flexible evidence-based filtering.

Here we fully describe and characterize MAPEX (the Mouse And Paralog EXterminator), a BLASTN-based algorithm for filtering variants that was previously introduced by Knudsen et al. (2017). We also present `mapexr`, a fast and lightweight implementation in R. The MAPEX algorithm is aimed at three use cases:

1. Labs that sequence PDX tumors using services that align to the human reference

can easily and accurately filter mouse contamination with `mapexr`

2. Bioinformatically sophisticated labs could align against both the human and mouse genomes to use other filtering approaches, but `mapexr` enables additional variant-level assessment of results

3. Any tumor genomics lab can use `mapexr` as a lightweight approach to identify potentially spurious variants created by paralogous sequences.

We show that, when applied to PDX samples, MAPEX generates calls that are highly similar to other methods, without the need to perform special alignments. We also show that, when applied to primary samples, MAPEX effectively filters paralogs while avoiding biases of existing heuristics. MAPEX is thus a useful addition to many tumor variant calling pipelines.

2.3 Approach

2.3.1 Workflow

The MAPEX algorithm is a post-variant-calling filter designed to fit into a standard tumor variant calling pipeline and flag variants which may arise from mis-alignment of mouse reads or from paralogous sequences (Fig. 2.1). The input for MAPEX is a BAM file containing tumor reads aligned to the human reference genome and a variant callset generated from that alignment. Variant-supporting reads are then BLASTed against the appropriate reference genome(s). Variants are scored by the fraction of supporting reads that align to the called site of the variant in the human genome.

2.3.2 Algorithm

Each read supporting a variant is BLASTed against the appropriate reference genome for the application. For PDX applications, this is the combined human/mouse refer-

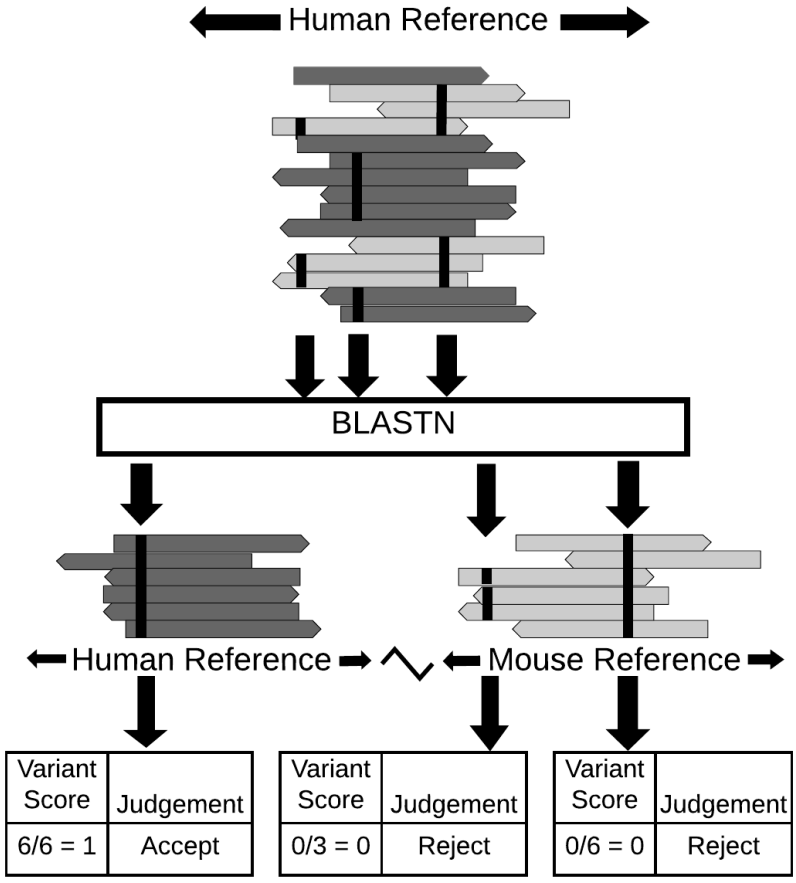

FIGURE 2.1. **Illustration of MAPEX applied to a PDX sample.** MAPEX begins with variants called from tumor reads aligned to the human genome. For each variant, the supporting reads are BLASTed against the combined human and mouse reference genomes. Variants are then scored by the fraction of supporting reads that align to the called site of the variant in the human genome.

ence, and for primary tumor applications, this is just the human reference. The best hit for each read is determined by bit score. Reads for which the best hit overlaps the called variant location are classified as "on target" and assigned a score of 1. Reads for which the best hit is a different region of the human genome or a region of the mouse genome are classified as "off target" or "mouse", respectively, and assigned a score of 0. Reads from genes with close paralogs in the human genome may generate multiple best hits (ties). In this case, the read score is averaged over all best hits, and the read is classified based on the most common result from the best hits. Each variant is then assigned a score that is the average score of all reads supporting that variant and is classified based on the most common classification of the supporting reads.

2.3.3 Implementation

We have implemented the MAPEX algorithm as an R package (mapexr). The package leverages the Bioconductor packages `Rsamtools`, `GenomicAlignments`, and `GenomicRanges` for fast and memory-efficient BAM file handling and read sequence extraction (Lawrence et al., 2013; Morgan et al., 2017). The package requires a local BLASTN installation and a BLAST database constructed from either a combined human/mouse reference genome or a human reference genome, depending on the application.

2.4 Methods

2.4.1 Samples

To characterize the performance of MAPEX, we used whole exome sequence trimmed fastq reads obtained from pancreatic ductal adenocarcinoma (PDAC) samples described previously by Knudsen et al. (2017) (PDX) and Witkiewicz et al. (2015)

(primary). For the PDX analysis, we analyzed a total of 34 PDXs derived from 9 primary tumors, sequenced to mean coverage depth of 124x. For the paralog analysis, we analyzed 93 primary tumors sequenced to a mean coverage depth of 40x.

2.4.2 Alignments and variant callers

All alignments were done using `bwa-mem` with default parameter settings (Li & Durbin, 2009). For initial variant calling, we aligned all reads in the samples to the human reference genome `GRCh37`. We then called variants using MuTect version 1.1.1 (Cibulskis et al., 2013), MuTect2 (as part of the GATK version 3.6, McKenna et al. (2010)), and Varscan 2 (Koboldt et al., 2012), all with default parameters. MuTect 1 and 2 variant calls were used without any post-filtering, but for Varscan 2 we used the built-in `processSomatic` and `fpfilter` functions with default parameters to generate a set of high-confidence variant calls. Variants were annotated with Oncotator (Ramos et al., 2015) and the annotation database `oncotator_v1_ds_April052016`. We considered only non-synonymous single nucleotide variants when comparing between methods. For paralog filtering, we used a conservative variant score cutoff of 0.8.

For comparison with Bruna et al. (2016), we aligned reads to a combined human/mouse reference genome `GRCh37/mm9` and called variants using MuTect 1.1.1. We calculated the fraction of mouse contamination using the method described in Bruna et al. (2016). Briefly, they generated data comparing the fraction of mouse cells in a sample with the fraction of total reads aligned to the mouse portion of a combined reference genome. We used this data to fit a LOESS regression model for contamination fraction vs fraction aligned, and used this to predict mouse contamination based on the fraction of reads aligned to the mouse genome in our samples.

For comparison with `bamcmp` (Khandelwal et al., 2017), we aligned reads separately to the human and mouse reference genomes and ran `bamcmp` with default parameters. The output of `bamcmp` includes alignment files for reads that aligned to only the human

reference and that aligned to both references but with a higher human alignment score. We merged these two alignments, performed indel realignment and base score recalibration using the GATK, and used the merged alignment to call variants with Mutect version 1.1.1. All scripts (doi:10.5281/zenodo.1112101) and the version of `mapexr` (doi:10.5281/zenodo.1112234) used to conduct the analysis have been archived with Zenodo.

2.5 Results & Discussion

2.5.1 Methodological

MAPEX is a lightweight filtering algorithm that adds little overhead or complexity to existing tumor variant-calling pipelines. The runtime for `mapexr` is linear in the number of variants to be filtered, processing roughly 250 variants per minute on a 4-core machine (Figure S1).

MAPEX has only one tunable parameter, the minimum mapping quality score required for a variant read. The default minimum score is 1, which includes all reads with an unambiguous best mapping. In pipelines in which a minimum mapping quality score is used for variant calling, that score should also be supplied to `mapexr`, to prevent evaluating reads that were not used by the variant caller. The output from `mapexr` is an R data frame with four columns – chromosome, start location, variant score, and variant classification – and one row for each variant evaluated. Users may also optionally provide a file path to `mapexr` which will generate a tab-delimited file with BLAST results and scores at the read level. The user can choose the variant score threshold used to classify variants as human- or mouse-derived. Here we use a threshold of 0.5, so that a variant is flagged as spurious if less than half of the supporting reads BLAST as "on target". In practice, the distribution of variant scores is bimodal and highly concentrated at 0 and 1, so results are insensitive to the exact threshold (Fig. S2).

2.5.2 Filtering mouse calls from PDX samples

One important use case for MAPEX is as a post-variant-calling filter for PDX samples that have been aligned to a human reference genome. To test the precision of MAPEX, we compared variant calls from aligning reads to the human reference and filtering with MAPEX to calls from two other methods. The first alternate method is to align reads to a combined human and mouse reference and then call variants (Bruna et al., 2016), which we refer to as the "combined reference" method. This requires similar CPU time to using MAPEX. The second method is to align reads separately to human and mouse references and call variants using only those reads that align better to the human reference, which is the method implemented in `bamcmp` (Khandelwal et al., 2017). This requires twice as much CPU time for alignment as MAPEX, and the post-alignment step is typically faster for MAPEX, although it can be longer for samples with very high mouse contamination (Fig. S3). For three representative PDX tumors, all three methods yield similar callsets (Figure 2.2A). The differences are primarily confined to low-frequency variants, and almost all high-frequency variants are called by all three methods (Figure 2.2B). MAPEX might reduce power to identify low-frequency subclonal variants, if some of the few reads supporting a variant BLASTed to incorrect locations. This would yield an intermediate variant score. Because variant scores are strongly bimodal (Fig. S2), we expect that MAPEX causes little to no reduction in power. Across 34 PDX tumors, all three methods yield a similar dramatic reduction in called variants (Figure 2.2C).

To further validate MAPEX, we compared PDX variant calls before and after filtering to the primary tumor from which the PDX was derived, where mouse contamination is not an issue. Across 34 PDX tumors derived from 9 primaries, MAPEX dramatically enriches PDX calls for variants that were also found in the primary tumor and removes few PDX calls that were found in the primary tumor. Among variants in the PDXs, only 0.3% to 10% called before MAPEX filtering were also found in

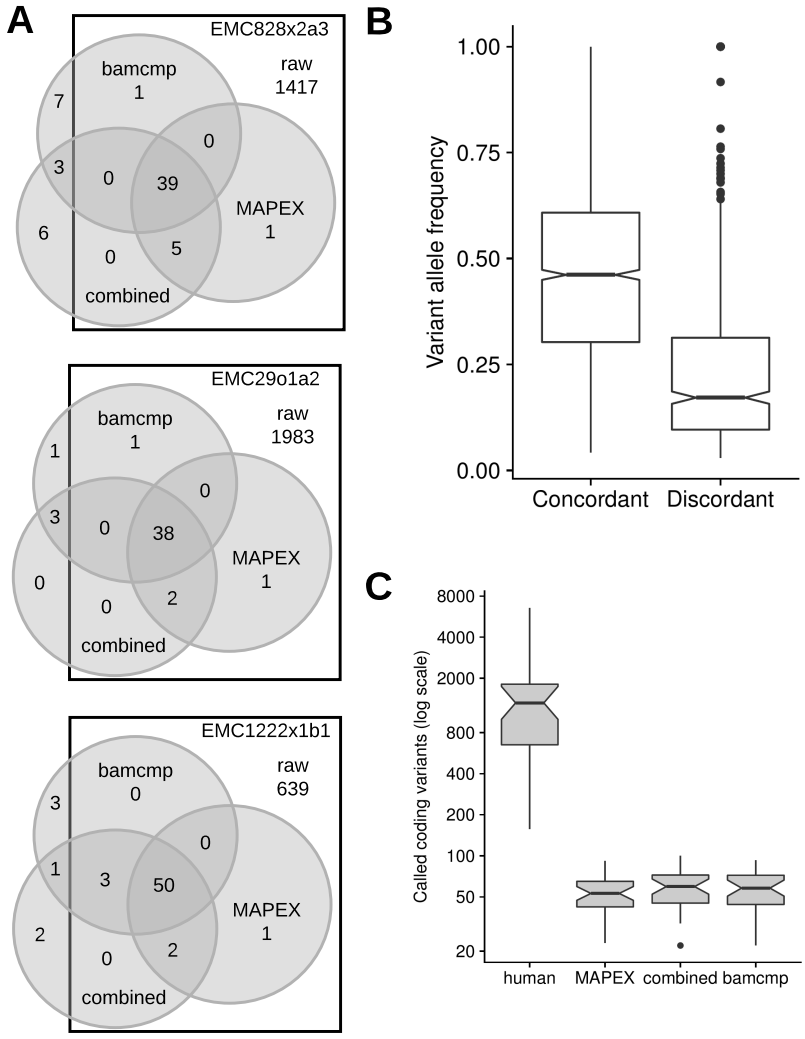

FIGURE 2.2. Comparison of MuTect 1.1.1 variants calls between MAPEX, combined reference, and `bamcmp` methods. **A:** Detailed breakdown of variant call overlap between the unfiltered human alignment (squares), MAPEX filtered human alignment (right circles), bamcmp filtered human alignment (top circles) and unfiltered combined alignment (bottom circles) for representative PDXs created from three different primary tumors. **B:** Variant allele frequencies for calls in 34 PDX samples that are concordant (n=1663 variants) and discordant (n=552 variants) between the methods. **C:** Comparison of total calls between the methods, n=34 PDX samples. Boxplots depict 25th and 75th percentile with 1.5×IQR whiskers. Notches are Median ± 1.58×IQR/sqrt(n), and represent rough estimates of 95% confidence interval around the median.

the primary tumor, but 23% to 90% of variants called after MAPEX filtering were found in the primary tumor (Table S1). This suggests that MAPEX enriches strongly for true variants. Among variants found both in the primary after MAPEX filtering and in the PDX before MAPEX filtering, 97% to 100% were retained in the PDX after MAPEX filtering (Table S1). Only one variant identified in each of two primary tumors was filtered by `mapexr` in a derived PDX. In primary tumor EMC1222, only 60% (slightly above the 50% cutoff) of variant reads mapped on-target for the primary variant (suggesting that it may be a spurious variant caused by a paralogous sequence), while in the PDXs only 20-45% (slightly below the cutoff) of variant reads mapped on-target. In EMC226, the variant appears to be from human wild-type to mouse wild-type, so 55% of variant reads (in a PDX with 57% mouse contamination), mapped to the mouse genome. Together these results suggest that MAPEX removes few true variants.

To validate the usefulness of MAPEX in practice, we focused on calls within known cancer-associated genes, using the COSMIC database . Among the pancreatic ductal adenocarcinoma (PDAC) samples in COSMIC, 34 genes are mutated in more than 3% of samples. Before filtering with MAPEX, 910 variants were found in these genes among the 34 PDXs we studied. After filtering with MAPEX, only 70 variants were retained.

These results suggest that MAPEX removes many false positives, dramatically simplifying variant interpretation. Of particular interest are KRAS, TP53, and SMAD4, which are the most commonly mutated genes in PDAC (Table 2.1). All of the KRAS mutations filtered by MAPEX are I187V mutants, which result from aligning wild-type mouse KRAS reads to human KRAS, and all 34 PDXs retained the KRAS mutation found in their primary tumor. All of the SMAD4 and TP53 mutations that were retained by MAPEX in the PDXs also appeared in the corresponding primary tumors, and all of those filtered were not found in the corresponding primary tumors. ARID1A is particularly susceptible to spurious variants caused by

TABLE 2.1. Variants detected in PDX samples for important PDAC genes.

	before MAPEX		after MAPEX	
Gene	Total variants	Samples with a variant	Total variants	COSMIC prevalence
KRAS	56	34	34	0.64
TP53	9	9	7	0.39
SMAD4	5	5	5	0.14
SYNE1	3	3	0	0.05
CSMD3	96	25	0	0.05
GNAS	6	6	6	0.05
HMCN1	10	5	0	0.04
APC	12	11	0	0.04
NEB	31	17	0	0.04
WDFY4	6	4	1	0.04
LRP1B	32	18	1	0.04
ARID1A	131	33	1	0.04

mouse contamination; only one of the 133 variants originally called in ARID1A was retained by MAPEX. We confirmed that the single retained variant was found in the primary tumor from which the PDX was derived, while none of the 132 rejected variants were found in their corresponding primaries.

2.5.3 Effects of variant call filters on PDXs

We carried out our primary analyses with the variant caller MuTect 1.1.1, but to test the performance of MAPEX with other variants callers, we also considered MuTect2 and Varscan 2.

If mouse contamination were perfectly filtered, the number of called variants should not depend on the level of mouse contamination. For all three variant callers the number of raw calls was strongly correlated with estimated mouse contamination (Fig. 2.3A-C), although MuTect2 and Varscan2 did produce substantially fewer calls overall than MuTect 1. After filtering with MAPEX, the numbers of variants called by all three callers was not significantly correlated with the level of mouse contamination (Fig. 2.3D-F).

Importantly, as a post-variant-calling filter, MAPEX can not evaluate variants

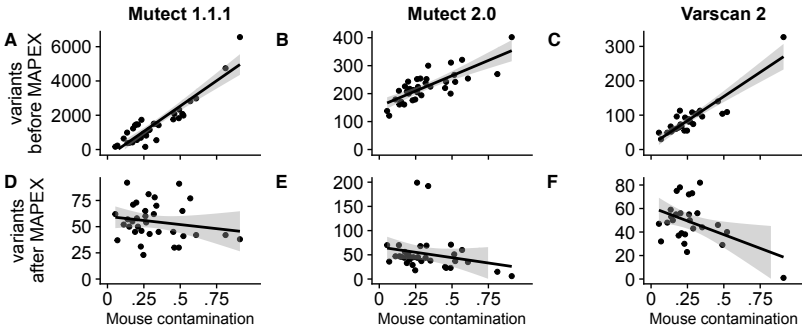

FIGURE 2.3. Effects of variant caller on analyzing xenograft samples with MAPEX. A,B,C: For all three calling algorithms and 34 xenograft samples (black dots), the number of raw variants called was strongly dependent on estimated mouse contamination. D,E,F: After filtering with MAPEX, the number of calls was independent of mouse contamination for all three callers. Lines show linear regressions and shading denotes 95% confidence intervals.

that were not initially called. Filters implemented with a variant caller, generally designed to improve results from primary tumors, can cause problems when using MAPEX. For example, MuTect2 applies a clustered event filter designed to reduce the number of false-positive variant calls due to mis-alignment of highly paralogous sequences. In regions of high similarity between mouse and human, this filter can remove true variants. For instance, Figure 2.4 shows the result of aligning a PDX with modest mouse contamination to the human reference for a small portion of the KRAS oncogene. MuTect 1.1.1 and Varscan 2 both called three variants at this locus, and MAPEX correctly rejected the two spurious variants arising from mouse contamination and retained the true G12D variant. MuTect2 fails to call any of these variants, because they are filtered as likely homologous mapping events, so MAPEX does not see and cannot retain the true G12D variant. In our PDX samples, we found instances of the clustered event filter removing true variants from other PDAC oncogenes, including SMAD4 and TP53.

Overall, the performance of MAPEX does not depend sensitively on the variant caller used, but callers can introduce specific biases. In particular, the default param-

eters for Varscan 2 yield high sensitivity but low specificity, so the use of the built-in post-call variant filters is necessary to prevent excessive false positives (Fig. S4). By contrast, the default parameters for MuTect2 yield much higher specificity, but at the cost of sensitivity in the PDX context. Currently, the clustered event filter cannot be disabled in MuTect2. We thus advise that users pairing MAPEX with MuTect2 be cautious when interpreting callsets from PDX samples in genes with high similarity between human and mouse.

2.5.4 Flagging potential false positives resulting from paralogous sequences

In addition to removing mouse contamination from PDX samples, MAPEX can also filter potential paralogs in primary samples. Across 93 PDAC primary tumors, a mean of 11% of total variant calls were flagged by MAPEX as potential paralogs, with a range of 2-33%. The genes in which variants were most frequently flagged as potentially arising from paralogous sequences include members of large gene families, such as mucins, zinc-finger nucleases, and the PRAME family (Table 2.2). Variants in citrate synthase (CS) were also frequently flagged (Table 2.2). Citrate synthase has a known pseudogene `NCBI: LOC440514` that was responsible for all of the spurious calls. We called variants with MuTect 1.1.1 and filtered with MAPEX, but MuTect2 includes new clustered event and read-mapping quality filters to prevent calling variants caused by paralogs. Using MAPEX yielded call sets that were identical with MuTect2 for all the genes in Table 2.2, with the exception of MUC12 and MUC5B, which differed by 3 variants. MAPEX can thus be efficiently and confidently used to remove variants that likely arise from paralogous sequences, with the additional benefit that the reason for classifying a variant as a potential paralog, as well as the genomic locations of the paralogous sequences, can be investigated.

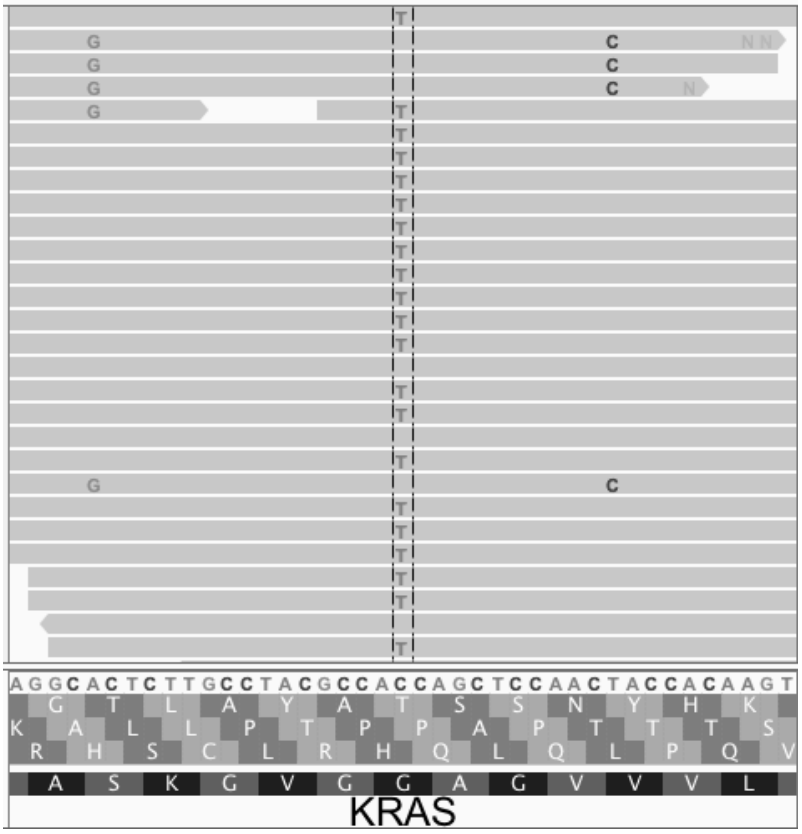

FIGURE 2.4. This Integrative Genomics Viewer (Thorvaldsdottir et al., 2013) window covers a portion of the human KRAS gene. The C>T variant is the classic KRAS G12D mutation that appears in many PDAC tumors. The A>G and T>C variants both result from aligning wild-type mouse reads to the human sequence. When used with MuTect 1.1.1 or Varscan 2, MAPEX correctly retains only the G12D variant. MuTect2, however, filters all three variants, so the G12D variant cannot be retained.

TABLE 2.2. Top genes for which MAPEX flagged variants as potentially arising from paralogs.

Gene	Variants flagged	Samples with a flagged variant
ZNF814	15	15
CS	12	7
IGFN1	8	6
KMT2C	7	7
FRG1	6	6
LILRB3	6	6
MUC12	6	6
RGPD3	6	6
USP6	6	3
FCGBP	5	4
MUC5B	5	5
NBPF1	5	3
PRAMEF11	5	4
PRB4	5	3
RGPD8	5	4

2.6 Conclusion

Genome sequencing is an increasingly important tool in cancer research, but spurious variant calls remain a challenge. MAPEX is an algorithm designed to filter spurious variants caused by mouse reads in patient-derived xenografts (PDXs) and caused by paralogous sequences in primary tumors. We showed that MAPEX is as sensitive and specific as more computationally intensive methods for calling variants from PDX tumors. We also showed that MAPEX successfully flags variant calls in potentially problematic gene families in primary tumors. Our implementation, `mapexr`, fits cleanly into standard tumor variant-calling pipelines and runs quickly on modern desktop computers. MAPEX is thus a potentially useful new component for many tumor variant-calling pipelines.

Funding

This work was supported by the National Science Foundation via Graduate Research Fellowship DGE-1143953 to BKM and by the National Institutes of Health via grants R01CA211878-01 and P30CA023074-36S2 to AKW and ESK.

Chapter 3

BATCAVE: Bayesian Analysis Tools for Context-Aware Variant Evaluation

3.1 Abstract

Detecting somatic mutations withins tumors is key to understanding treatment resistance, patient prognosis, and tumor evolution. Mutations at low allelic frequency, those present in only a small portion of tumor cells, are particularly difficult to detect. Many algorithms have been developed to detect such mutations, but none models a key aspect of tumor biology. Namely, every tumor has its own profile of mutation types that it tends to generate. We present BATCAVE (Bayesian Analysis Tools for Context-Aware Variant Evaluation), an algorithm that first learns the individual tumor mutational profile and mutation rate then uses them in a prior for evaluating potential mutations. We also present an R implementation of the algorithm, built on the popular caller MuTect. Using simulations, we show that adding the BATCAVE algorithm to MuTect improves variant detection. It also improves the calibration of posterior probabilities, enabling more principled tradeoff between precision and recall. We also show that BATCAVE performs well on real data. Our implementation is computationally inexpensive and straightforward to incorporate into existing MuTect pipelines. More broadly, the algorithm can be added to other variant callers, and it can be extended to include additional biological features that affect mutation generation.

3.2 Introduction

Cancer develops through the accumulation of somatic mutations and clonal selection of cells with mutations that confer an advantage. Understanding the evolutionary

history of a tumor, including the mutations that drive its growth, the genetic diversity within it, and the accumulation of new mutations, requires accurate variant identification, particularly at low variant allele frequency (Williams et al., 2016; Bozic et al., 2016; Williams et al., 2018b; Shi et al., 2018). Accurate variant calling is also critical for optimizing the treatment of individual patients' disease (Ding et al., 2012; Mardis, 2012; Chen et al., 2013; Borad et al., 2014; Findlay et al., 2016). Low frequency mutations challenge current variant calling methods, because their signature in the data is difficult to distinguish from the noise introduced by Next Generation Sequencing (NGS), and this challenge increases with sequencing depth.

Many methods have been developed for calling somatic mutations from NGS data. The earliest widely used somatic variant callers developed specifically for tumors, MuTect1 (Cibulskis et al., 2013) and Varscan2 (Koboldt et al., 2012), used a combination of heuristic filtering and a model of sequencing errors to identify and score potential variants and set a threshold score designed to balance sensitivity and specificity. Subsequent research gave rise to a number of alternate strategies, including haplotype-based calling (Garrison & Marth, 2012), joint genotype analysis (SomaticSniper (Larson et al., 2012), JointSNVMix2 (Roth et al., 2012), Seurat (Christoforides et al., 2013), CaVEMan (Jones et al., 2016), and MuClone (Dorri et al., 2019)), allele frequency-based analysis (Strelka (Saunders et al., 2012), LoFreq (Wilm et al., 2012), EBCall (Shiraishi et al., 2013), deepSNV (Gerstung et al., 2012), LoLoPicker (Carrot-Zhang & Majewski, 2017), and MuSE (Fan et al., 2016)), and ensemble and deep learning methods (MutationSeq (Ding et al., 2012), BAYSIC (Cantarel et al., 2014), SomaticSeq (Fang et al., 2015), and SNooPer (Spinella et al., 2016)). These methods vary in their complexity and specific focus. But they all implicitly or explicitly assume that the rate of mutation is uniform across the genome.

The mutational processes that generate single nucleotide variants in tumors do not act uniformly across the genome. If fact, even the processes of spontaneous mutation that are active in all somatic tissues depend sensitively on local nucleotide context

(Nik-Zainal et al., 2012a; Alexandrov et al., 2015; Lee-Six et al., 2018). Additional mutational processes are active in tumors, due to mutagen exposure or defects in DNA maintenance and repair, and these processes are also sensitive to local nucleotide context (Alexandrov et al., 2013b; Helleday et al., 2014b; Nik-Zainal et al., 2016; Kandoth et al., 2013; Alexandrov et al., 2016). The specific mutational processes active in a particular tumor generate its unique mutation profile, and differences within and between tumor types are pronounced (Stephens et al., 2005; Burrell et al., 2013; Nakamura et al., 2015; Witkiewicz et al., 2015; Kumar et al., 2016). For example, the mutation profiles differ substantially among the three breast tumors illustrated in Figure 3.1B-D.

Here we present an enhanced variant-calling algorithm that uses the biology of each individual tumor's mutation profile to improve identification of low allelic frequency mutations. Our BATCAVE algorithm first estimates the tumor's mutation profile and mutation rate using high-confidence variants and then uses them as a prior when calling other variants. Our R implementation of the algorithm, `batcaver`, takes output from the MuTect variant caller as input and returns the posterior probability that a site is variant for every site observed by MuTect. Using both simulated and real data, we show that the addition of a mutation profile prior to MuTect produces a superior variant caller. Our algorithm is simple and computationally inexpensive, and it can be integrated into numerous other variant callers. Broad adoption of our approach will enable more confident study of low allelic frequency mutations in tumors in both research and clinical settings.

3.3 Materials and methods

3.3.1 Somatic variant calling probability model

At every site in the genome with non-zero coverage, Next Generation Sequencing produces a vector $\mathbf{x} = (\{b_i\}, \{q_i\}), i = 1 \ldots \mathrm{d}$ of base calls b and their associated

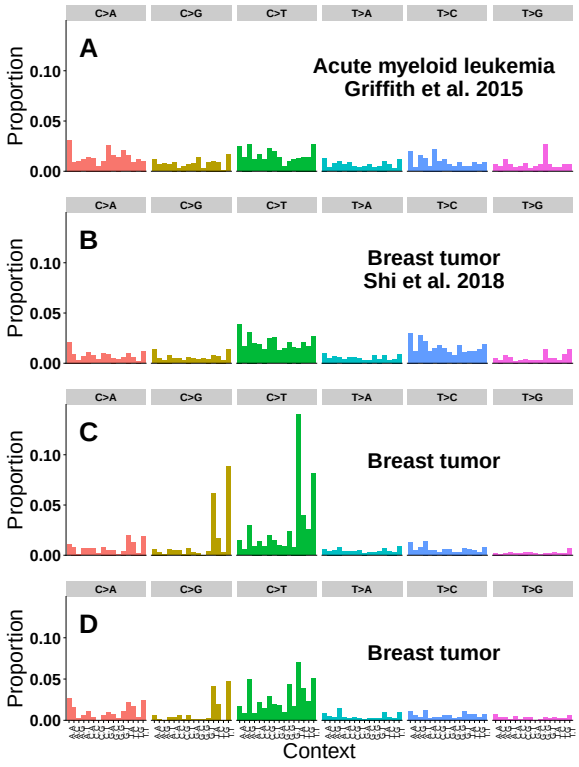

FIGURE 3.1. Real tumor mutation profiles. In each panel, the x-axis corresponds to each of the 96 possible mutation types, and the y-axis is the proportion of total mutations of each type. (A) The observed mutation profile of an acute myeloid leukemia used in our real data analysis (Griffith et al., 2015). (B) The observed mutation profile of a breast tumor used in our real data analysis (Shi et al., 2018). (C)&(D) The observed mutation profiles of two additional breast tumors (Alexandrov et al., 2019).

quality scores q, where d is local read depth. Variant callers use the data $\mathbf{x}$ to choose between competing hypotheses:

$$\mathbf{H_0}: \quad \text{Alt allele} = m; \quad \nu = 0 \tag{3.1}$$

$$\mathbf{H_1}: \quad \text{Alt allele} = m; \quad \nu = \hat{f}. \tag{3.2}$$

Here m is any of the 3 possible alternate non-reference bases and ν is the variant allele frequency. The maximum likelihood estimate of ν is simply $\hat{f}$, the number of variant reads divided by the local read depth. The posterior probability of a given hypothesis, $\mathrm{P}(m, \nu)$, is the product of the likelihood of the data given that hypothesis and the prior probability of that hypothesis. Assuming that reads are independent, this is

$$\mathrm{P}(m, \nu) = \mathrm{p}(m, \nu) \cdot \prod_{i=1}^{\mathrm{d}} \mathrm{f}_{m,\nu}(x_i), \tag{3.3}$$

where $\mathrm{f}_{m,\nu}(x_i)$ is the probability model for reads, and $p(m, \nu)$ is the prior.

Assuming that the identity of the alternate allele and its allele frequency are independent and that ν is uniformly distributed, Eq. 3.3 becomes

$$\mathrm{P}(m, \nu) = \mathrm{p}(m) \cdot \prod_{i=1}^{\mathrm{d}} \mathrm{f}_{m,\nu}(x_i). \tag{3.4}$$

The focus of BATCAVE is to provide a tumor- and site-specific estimate of the prior probability of mutation $\mathrm{p}(m)$.

3.3.2 Site-specific prior probability of mutation

The probability that we have denoted $\mathrm{p}(m)$ in Eq. 3.4 is more precisely the joint probability that a mutation has occurred M and that it was to allele m, which we denote $\mathrm{p}(m, M)$. But $\mathrm{p}(m, M)$ is not uniform across the genome. Rather it depends on the local genomic context C, so its full form is $\mathrm{p}(m, M|C)$ (Buisson et al., 2019). Assuming that m and M are independent conditional on the genomic context,

$\mathrm{p}(m, M \mid C) = \mathrm{p}(m \mid C)\mathrm{p}(M \mid C)$, which we can use Bayes' theorem to further decompose as

$$\mathrm{p}(m, M \mid C) = \mathrm{p}(m \mid C)\mathrm{p}(C \mid M)\frac{\mathrm{p}(M)}{\mathrm{p}(C)}. \tag{3.5}$$

We next show how to estimate the quantities in Eq. 3.5.

3.3.3 Estimation of the mutation profile

Many aspects of genomic architecture can affect the somatic mutation rate at multiple scales (Buisson et al., 2019). Here we focus on a small-scale feature, the trinucleotide context, which is known to strongly affect the prior probability of single-nucleotide mutation (Nik-Zainal et al., 2012a; Alexandrov et al., 2015; Lee-Six et al., 2018). The trinucleotide context of a genomic site consists of the identity of the reference base and the 3' and 5' flanking bases. Folding the central base to the pyrimidines, there are two possible bases at the focal site, and there are four possible bases 3' and 5' of the focal site, yielding $2 \cdot 4 \cdot 4$ possible tri-nucleotide contexts C. At the focal site, a mutation m can be to any of three alternate alleles. Indexing by the $c = \{1 \dots 32\}$ contexts and by the $m = \{1 \dots 3\}$ alternate bases, we have 96 possible substitution types $S_{m,c}$. Eq. 3.5 is then

$$\mathrm{p}(\mathrm{S}_{m,c}) = \mathrm{p}(m \mid C = c)\mathrm{p}(C = c \mid M)\frac{\mathrm{p}(M)}{\mathrm{p}(C = c)}. \tag{3.6}$$

The first two terms on the right-hand side can be estimated from the observed mutation profile (Fig. 3.1).

We model the observed mutation profile S as multinomial with parameter $\boldsymbol{\pi} = \{\pi_{m,c}\}$. Each element of $\boldsymbol{\pi}$ represents the expected proportion of mutations that are to allele m and in context c. In a tumor with many high-confidence observed mutations, $\boldsymbol{\pi}$ could be estimated directly from the observed mutation profile S. But in practice many entries in $\boldsymbol{\pi}$ would then have zero weight. We thus model the distribution of S

as Dirichlet-multinomial with pseudo-count hyper-parameter $\boldsymbol{\alpha}$,

$$\begin{aligned} \boldsymbol{\pi} \mid \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \text{S} \mid \boldsymbol{\pi} &\sim \text{Multinomial}(\boldsymbol{\pi}). \end{aligned} \tag{3.7}$$

In BATCAVE we use the symmetric non-informative hyper-parameter $\boldsymbol{\alpha} = \mathbf{1}$, so a priori mutation is equally likely to any allele and in any context.

To estimate $\boldsymbol{\pi}$, we identify a subset of high confidence variants, based on an initial calculation of their likelihood given the data. These are variants for which the evidence in the read data overwhelms any reasonable value of the site-specific prior probability of mutation. Let D be the set of high confidence variant calls, which we define as those having posterior odds greater than 10 to 1 without the site-specific prior, and $\text{s} \in \text{D}$ be the substitution type of each mutation in D. The posterior distribution of $\boldsymbol{\pi}$ is then $\text{p}(\boldsymbol{\pi} \mid \text{D}) \sim \text{Dirichlet}(\boldsymbol{\alpha'})$ where

$$\alpha'_{m,c} = \alpha_{m,c} + \sum_{\text{s}\in\text{D}} \text{I}\{\text{s} = \text{s}_{m,c}\}, \tag{3.8}$$

and I is the indicator function. Returning to Eq. 3.6, given that a mutation has occurred, the posterior probability it occurred in context c is

$$\text{p}(C = c \mid M, \text{D}) = \frac{\sum_m \alpha'_{m,C=c}}{\sum_{m,c} \alpha'_{m,c}}. \tag{3.9}$$

The posterior probability of mutation to allele m given that a mutation has occurred in context $C = c$ is then

$$\text{p}(m \mid C = c, \text{D}) = \frac{\alpha'_{m,C=c}}{\sum_m \alpha'_{m,C=c}}. \tag{3.10}$$

The prior probability of each particular trinucleotide context $p(C = c)$ is computed simply as the proportion of sequenced trinucleotide contexts that have context c. The R implementation of BATCAVE ships with pre-computed tables for both human whole exomes and whole genomes.

3.3.4 Estimation of the mutation rate

The final piece of Eq. 3.6 is $p(M)$, the prior probability of mutation, which we specify as the per-base per-division mutation rate μ. In an exponentially growing and neutrally evolving tumor, branching process calculations (Williams et al., 2018b) show that the expected total number of mutations $\mathrm{M_{tot}}$ between two allele frequencies (f_{min},f_{max}) is

$$\mathrm{M_{tot}}(f_{min}, f_{max}) = N\frac{\mu}{\beta}\left(\frac{1}{f_{min}} - \frac{1}{f_{max}}\right). \tag{3.11}$$

The number of bases N is $3{\cdot}10^9$ for a whole genome and $3{\cdot}10^7$ for a whole exome. The quantity μ/β is the effective mutation rate, where β is the fraction of cell divisions that lead to two surviving lineages. We make the simplifying assumption that there is no cell death ($\beta = 1$), so we somewhat over-estimate μ. We then estimate μ by counting observed high-confidence mutations between allele frequencies f_{min} and f_{max}. We set f_{max} to be the largest allele frequency in D, but we must choose f_{min} conservatively, depending on sequencing depth. In the R implementation of BATCAVE, f_{min} is a free parameter. For this paper, we set $f_{min} = 0.05$, because we are working at high depth.

3.3.5 Likelihood function

The current implementation of BATCAVE builds on MuTect, because MuTect reports the log ratio of the likelihood functions for the null and alternative hypotheses (Eq. 3.1) as TLOD (MuTect1) or t_lod_fstar (MuTect2). We used MuTect 1.1.7 for all analyses in this paper, so we have

$$\mathrm{TLOD} = \log_{10}\left(\frac{\prod_{i=1}^{\mathrm{d}} \mathrm{f}_{m,\nu=\hat{f}}(x_i)}{\prod_{i=1}^{\mathrm{d}} \mathrm{f}_{m,\nu=0}(x_i)}\right). \tag{3.12}$$

The log posterior odds is the log likelihood ratio (TLOD) plus the log prior odds, so the posterior odds in favor of the alternate hypothesis for a given substitution type is

$$\frac{\mathrm{P}(m, \nu = \hat{f})}{1 - \mathrm{P}(m, \nu = \hat{f})} = 10^{\mathrm{TLOD} + \mathrm{logit}_{10}(\mathrm{p}(\mathrm{S}_{m,c}))}. \tag{3.13}$$

Here $\mathrm{p}(\mathrm{S}_{m,c})$ is the prior probability of a substitution of type $\mathrm{S}_{m,c}$, as described in Eq. 3.6 and specified in Eq. 3.9-3.11. When comparing our posterior odds to those of MuTect, we assume a uniform per-base probability of mutation of $3 \cdot 10^{-6}$ (Cibulskis et al., 2013), so

$$\frac{\mathrm{P}_{\mathrm{MuTect}}(m, \nu = \hat{f})}{1 - \mathrm{P}_{\mathrm{MuTect}}(m, \nu = \hat{f})} = 10^{\mathrm{TLOD} - 6}. \tag{3.14}$$

3.3.6 Implementation

We have implemented the BATCAVE algorithm as an R package `batcaver`. The package leverages the Bioconductor packages BSgenome (Pagès, 2019), GenomicAlignments (Lawrence et al., 2013), VariantAnnotation (Obenchain et al., 2014), and SomaticSignatures (Gehring et al., 2015) for fast and memory-efficient variant annotation and genomic context identification. Reference sequences are specified as BSgenome objects, allowing efficient access to genomic context information.

3.3.7 Tumor simulations

We used a neutral branching process with no death and $\mu = 3 \cdot 10^{-6}$ to simulate realistic distributions of mutation frequencies. Tumors were simulated with three different mutation profiles composed of COSMIC mutation signatures (version 2) (COSMIC Consortium, 2019). Each simulated profile includes COSMIC signature 1, which is found in nearly all tumors and is associated with spontaneous cytosine deamination. The "Concentrated" profile (Fig. 3.2A) is an equal combination of COSMIC signatures 1, 7, and 11, which has a large percentage of C > T substitutions such as are

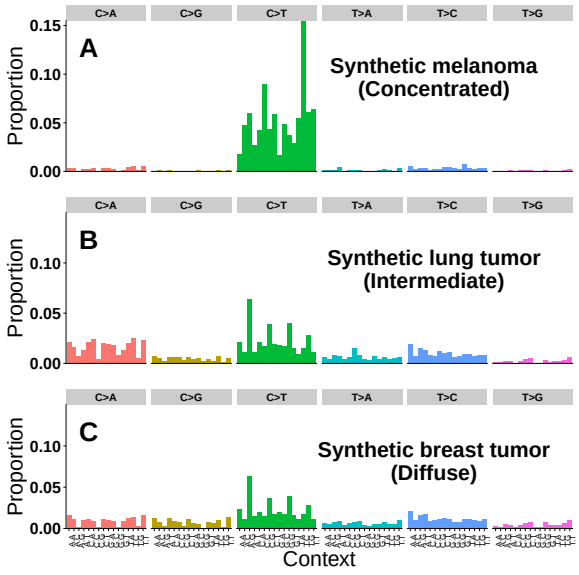

FIGURE 3.2. Simulated tumor mutation profiles. As in Fig. 3.1, in each panel the x-axis corresponds to each of the 96 possible mutation types, and the y-axis is the proportion of total mutations of each type. (A) A mutation profile used for simulating tumors, made up of equal proportions of COSMIC mutation signatures 1, 7, & 11. (B) Equal proportions of signatures 1, 4, & 5. (C) Equal proportions of signatures 1, 3, & 5.

often seen in cancers caused by UV exposure (Alexandrov et al., 2013a). The "Intermediate" profile (Fig. 3.2B) is an equal combination of COSMIC signatures 1, 4, and 5, which has been associated with tobacco carcinogens and is representative of some lung cancers (Alexandrov et al., 2013a). The "Diffuse" profile (Fig. 3.2C) is an equal combination of COSMIC signatures 1, 3, and 5, which has been associated with inactivating germline mutations in the BRCA1/2 genes leading to a deficiency in DNA double strand break repair (Nik-Zainal et al., 2016). Simulated variants were sampled from a combination of the Cancer Genome Atlas (TCGA) and Pan-Cancer Analysis of Whole Genomes (PCAWG) databases, which include mutations found in all types of cancer. Whole genome (100X depth) and whole exome (500X depth) reads were simulated from the GRCh38 reference genome using VarSim (Mu et al., 2015) and aligned with BWA (Li & Durbin, 2009), both with default parameters. Variants were inserted to create tumors with BAMSurgeon with default parameters (Ewing et al., 2015) and called with MuTect 1.1.7 (Cibulskis et al., 2013) with the following parameters:

```
java -Xmx24g -jar $MUTECT_JAR --analysis_type MuTect --reference_sequence $ref_path
--dbsnp $db_snp
--enable_extended_output --fraction_contamination 0.00 --tumor_f_pretest 0.00
--initial_tumor_lod -10.00 --required_maximum_alt_allele_mapping_quality_score
1 --input_file:normal $tmp_normal --input_file:tumor $tmp_tumor --out $out_path/$chr.txt
--coverage_file $out_path/$chr.cov.
```

Variants identified by MuTect are labelled as to whether they pass all filters, fail to pass only the the evidence threshold tlod_f_star filter, or fail to pass any other filter. Variants that passed all filters or failed only tlod_f_star were then passed to BATCAVE for prior estimation and rescoring.

3.3.8 Calibration metric

To quantify the difference in calibration between MuTect and BATCAVE, we used the Integrated Calibration Index (Austin & Steyerberg, 2019). Briefly, a loess-smoothed regression was fit by regressing the binary (True=1, False=0) true variant classification against the reported posterior probability for both MuTect and BATCAVE. For a perfectly calibrated caller, the regression fit would be the diagonal line $y = x$. The Integrated Calibration Index is a weighted average of the absolute distance between the calibration curve and the diagonal line of perfect calibration.

3.3.9 Real data

We analyzed two real data sets, one from an acute myeloid leukemia (AML) (Griffith et al., 2015) and one from a multi-region sequencing experiment in breast cancer (Shi et al., 2018). We downloaded the normal and primary whole-genome AML tumor bam files from dbGaP accession number phs000159.v8.p4. Griffith et al. generated a platinum set of variant calls for this tumor (Griffith et al., 2015), which we used for our true positive dataset. We downloaded the normal and tumor whole-exome breast cancer bam files from NCBI Sequence Read Archive accession SRP070662. Shi et al. generated a gold set of variant calls for each tumor region sequenced (Shi et al., 2018), which we used for our true positive dataset. For these multi-region data, we ran BATCAVE separately on each sequenced region and combined results to generate precision-recall curves. We called variants using Mutect 1.1.7 as in our simulations, except that both these data sets were originally aligned to GRCr37, so we used that reference.

3.4 Results

We implemented BATCAVE as a post-call variant evaluation algorithm to be used with MuTect (Versions 1.1.7 or >2.0) (Cibulskis et al., 2013). BATCAVE extracts the log-likelihood ratio for each potential variant site from the MuTect output, and then it uses that ratio to separate the potential sites into high and low confidence groups. The mutation profile and mutation rate are estimated from the high confidence sites, and the posterior probability of mutation is then recomputed for all sites. The BATCAVE algorithm is inexpensive, processing 22,000 variants per second on a typical desktop computer, which corresponds to roughly 100 seconds to process a 500X exome and 2,000 seconds for a 100X whole genome.

To test the performance of BATCAVE, we generated six different tumor/normal pairs, corresponding to 100X whole genomes and 500X whole exomes for three different mutation profiles. The three mutation profiles were chosen to resemble a melanoma (concentrated), a lung cancer (intermediate), and a BRCA-driven breast cancer (diffuse) (Fig. 3.2). We also tested BATCAVE using two real cancer data sets, a whole-genome Acute Myeloid Leukemia (AML) (Griffith et al., 2015) and a whole-exome multi-region breast cancer (Shi et al., 2018). In both, deep sequencing and variant validation were performed with the specific purpose of evaluating tumor variant calling pipelines. Because our focus is on evaluating the statistical calling model, we computed all test metrics using only those potential variants that passed MuTect's heuristic filters and entered the statistical model.

3.4.1 Tests using simulated data

To improve variant identification, the context-dependent prior probability of mutation must converge to an accurate representation of the data generating distribution within the set of high-confidence mutations. When applied to simulated data, the prior converged within a few hundred mutations (Fig. 3.3). For comparison, in our

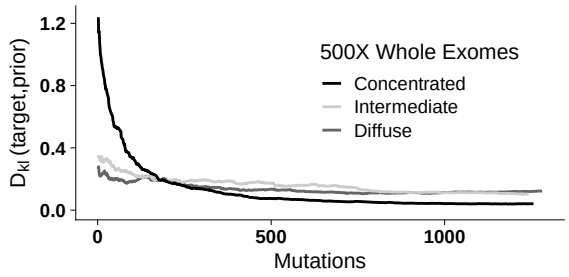

FIGURE 3.3. Convergence of the mutational prior to the data generating distribution. Plotted is the Kullback-Leibler divergence between the simulated and estimated profiles versus number of incorporated mutations for whole exomes. Convergence for whole genomes is similar.

simulated data sets the number of high-confidence mutations ranged between 1,500 and 5,000, and in the real AML we test on it is over 17,000 (Griffith et al., 2015).

We assessed classification performance using the areas under both the receiver operating characteristic and the precision-recall curves, because the classes are unbalanced (approximately 5 to 1 ratio of false to true variants in our simulated data). By both metrics BATCAVE outperforms MuTect (Fig. 3.4A&B, Fig. 3.6A&B, and Table 3.1). The extent of the performance difference is dependent on both the sequencing depth and the concentration of the mutation profile. Deeper sequencing and more concentrated mutation profiles increase the performance advantage of BATCAVE.

For all simulated tumors, the estimated mutation rate was approximately $3 \cdot 10^{-7}$

TABLE 3.1. Variant calling metrics for all data sets.

Scenario	Mutation profile	μ	AUROC		AUPRC		ICI	
		(estimated)	MuTect	BATCAVE	MuTect	BATCAVE	MuTect	BATCAVE
100X whole genome	Concentrated	3.6e-7	.987	.993	.972	.975	.117	.287
100X whole genome	Intermediate	3.2e-7	.987	.989	.972	.973	.118	.214
100X whole genome	Diffuse	3.2e-7	.988	.989	.971	.973	.120	.219
500X whole exome	Concentrated	3.6e-7	.848	.929	.674	.758	.138	.109
500X whole exome	Intermediate	3.6e-7	.847	.881	.677	.706	.108	.112
500X whole exome	Diffuse	3.6e-7	.850	.873	.676	.698	.105	.116
real AML (Griffith et al., 2015)	Actual	3.6e-8	–	–	.995	.996	–	–
real breast (Shi et al., 2018)	Actual	3.6e-8	–	–	.972	.972	–	–

μ = per-base mutation rate, AUROC/AUPRC = Area Under Receiver Operating Characteristic / Precision-Recall Curve, ICI = Integrated Calibration Index. Smaller values of ICI are superior.

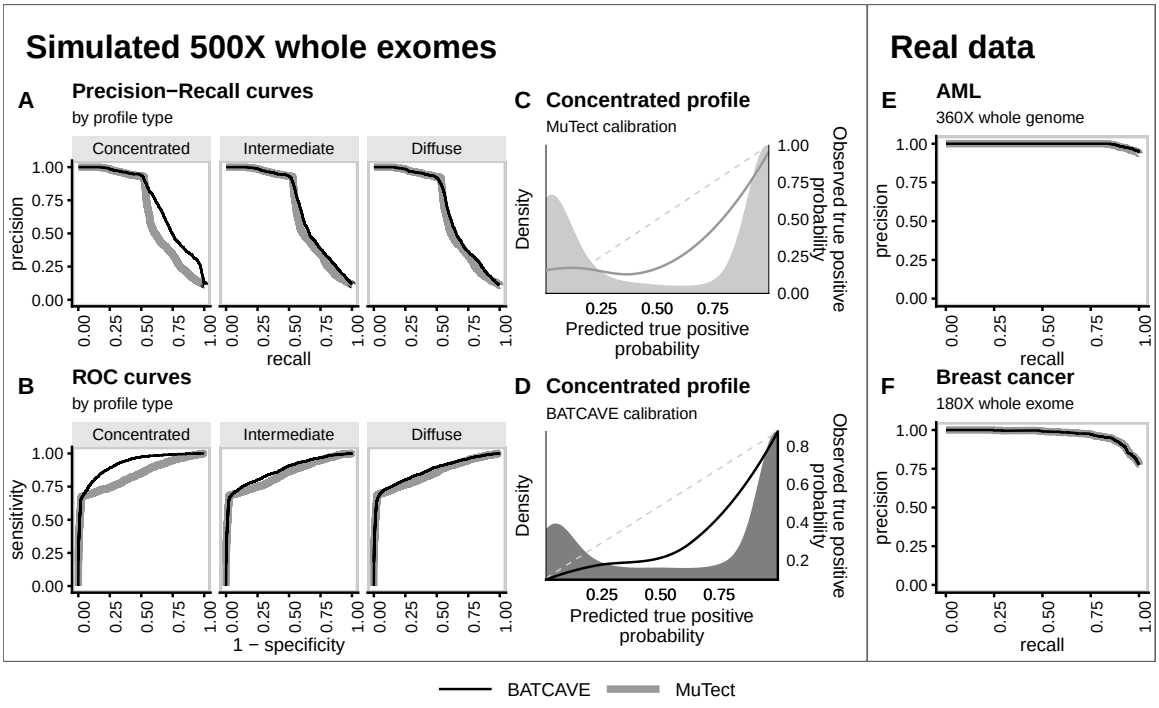

FIGURE 3.4. Variant-calling performance on simulated and real data. Throughout, MuTect results are plotted with gray lines and BATCAVE results with black lines. (A) Precision-recall curves and (B) receiver operating characteristic curves for different mutation profiles. (C) and (D) Calibration plots. Shaded regions show distributions of posterior probabilities for true positive variants, and smooth lines show loess-smoothed relationships, from which the Integrated Calibration Index is calculated. For a perfectly calibrated caller, those curves would match the dashed y=x line. (E) and (F) Precision-recall curves for real data in which substantial mutation validation was performed (Griffith et al., 2015; Shi et al., 2018).

(Table 3.1), which is lower than the simulated rate of $3 \cdot 10^{-6}$. This is likely due to restrictions within BAMSurgeon, such as sequencing depth and quality, that prevent 100% of simulated variants from being inserted into the reads.

We also assessed calibration, the likelihood that a potential variant with a given posterior probability is actually a true variant. We measured overall calibration performance using the Integrated Calibration Index (ICI) (Austin & Steyerberg, 2019), which integrates the difference between predicted and observed probabilities, weighted by the density of the predicted probabilities. This metric is particularly useful in our case, because the density of posterior probabilities is bi-modal (Fig. 3.4C&D and 3.6C&D). A large fraction of true negative variants have posterior probabilities less than 10^{-4}, far below any meaningful threshold, so we evaluated calibration only on potential variants with posterior probability greater than 0.01. For these potential variants, BATCAVE tends to increase posterior probabilities of low probability but true variants (density curves in Fig. 3.4C&D and 3.6C&D) while decreasing probabilities of low probability but false variants. For 500X exomes, the calibration of BATCAVE is better than MuTect across the full spectrum of posterior probabilities (Fig. 3.4 and Table 3.1). For 100X whole genomes, the calibration of BATCAVE is slightly worse (Fig. 3.6 and Table 3.1), likely because there are few low probability true positive variants in tumors sequenced to 100X depth. As with the other metrics, the advantage of BATCAVE increases with the concentration of the mutation profile and the sequencing depth.

In practice, variant callers are typically used with a threshold score above which a variant is called. The user's choice of threshold ideally meets their need to balance precision and recall; accurate posterior probability estimates enable an informed choice. For posterior probability thresholds between 60 and 90%, the precision of BATCAVE calls is similar to the chosen threshold (Fig. 3.5&3.7). For this range of thresholds, however, the posterior probabilities from MuTect poorly predict precision (Fig. 3.5&3.7). For any posterior probability threshold above 70%, MuTect

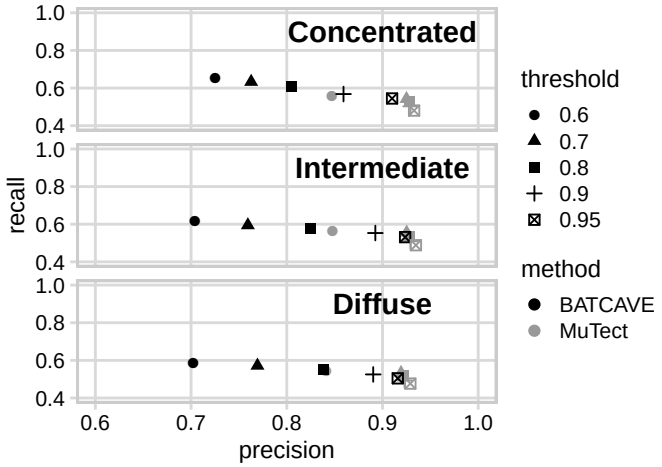

FIGURE 3.5. Posterior probability calibration for realistic calling thresholds, for 500X exomes. Plotted is precision and recall for variants identified using various realistic posterior probability thresholds. At these thresholds, the precision of BATCAVE is much closer to the given threshold than MuTect, no matter the concentration of the mutation profile.

has a false positive rate of roughly 8%, whereas BATCAVE has a false positive rate that decreases as the threshold increases. The cost of MuTect's compressed range of posterior probabilities is recall; at any posterior probability threshold BATCAVE has recall better than MuTect. Consequently, BATCAVE posterior probabilities are more informative than MuTect's with regard to choosing a calling threshold.

3.4.2 Tests using real tumor data

We tested BATCAVE using two data sets for which deep sequencing and variant validation were performed with the express purpose of evaluating tumor variant calling pipelines, yielding high quality true and false positive data (Griffith et al., 2015; Shi et al., 2018). However, only variants called by at least one variant caller were validated. As a result, there are no validated true or false negative calls, so we considered only precision-recall comparisons for these data.

Griffith et al. sequenced the whole genome of an acute myeloid leukemia (AML)

primary tumor to a depth of >360X and used targeted sequencing to validate nearly 200,000 mutations (Griffith et al., 2015). We estimated a per-base mutation rate for this tumor of $4 \cdot 10^{-8}$, which is consistent with previous estimates of AML mutation rates (Griffith et al., 2015; Williams et al., 2018b). For both MuTect and BATCAVE, the precision-recall curve is almost perfect for the validated variants (.995 & .996 area under the curve) (Fig. 3.4E and Table 3.1).

Shi et al. performed multi-region whole exome sequencing on six individual breast tumors to a mean target sequencing depth of 160X and validated all variants identified by three different variant calling pipelines (Shi et al., 2018). We estimated an average per-base mutation rate for these tumor regions of $4 \cdot 10^{-8}$, which is consistent with observed mutation rates for breast cancers (Alexandrov et al., 2019) and with the low number of validated somatic mutations. For the validated variants, MuTect and BATCAVE yielded almost identical precision-recall curves (Fig. 3.4F and Table 3.1)

3.5 Discussion

BATCAVE is an algorithm that leverages the biology of individual tumor mutation profiles to improve identification of low allelic frequency somatic variants. Our implementation is built on MuTect, one of the most widely used somatic variant callers. BATCAVE improves on the classification accuracy of MuTect in synthetic data (Fig. 3.4A-D, 3.6, and Table 3.1) across the entire range of recall and specificity. Moreover, BATCAVE is better calibrated than MuTect at relevant posterior probability thresholds (Fig. 3.5 and 3.7), allowing researchers and clinicians to make informed choices about the trade-off between precision and recall. For real data, testing on validated calls shows that BATCAVE does not degrade performance for variants that are relatively easy to identify (Fig. 3.4E&F and Table 3.1). The BATCAVE algorithm can thus be included in a wide variety of sequencing pipelines.

We evaluated BATCAVE with simulated tumors with three different mutation pro-

files and two real tumors. The simulated diffuse and intermediate profiles (Fig. 3.2A&B) represent baseline profiles of lung and breast tumors, respectively. And the concentrated profile (Fig. 3.2C) represents a tumor driven by a particular mutational process, such as UV exposure. But mutational profiles are highly heterogeneous, so concentrated profiles can be found in any tumor type (e.g., Fig. 3.1C). The two real data sets we considered are among the few for which extensive validation of variant calls has been performed (Griffith et al., 2015; Shi et al., 2018). They happen, however, to have diffuse mutation profiles (Fig. 3.1A&B), which reduces the expected advantage of BATCAVE over MuTect (Table 3.1). A more fundamental challenge of using these real data for testing callers is that only a subset of potential variants are validated. This subset tends to be relatively easy to call, so both MuTect and BATCAVE have almost perfect precision and recall for variants that pass heuristic filters (Fig. 3.4 and Table 3.1). Moreover, few true negative sites are validated, so specificity and calibration are impossible to calculate. Deep sequencing experiments that validate random samples of uncalled potential variants would give much-needed insight into the differences among statistical models in variant calling.

The improved calibration of BATCAVE posterior probabilities compared to MuTect provides several advantages. In practice, called variants are often manually reviewed to further reduce false positives (Barnell et al., 2019). Improved calibration enables users to focus review on the most questionable variants. In the clinic, identified variants act as biomarkers for susceptibility to targeted drugs (Boutros, 2015). Well-calibrated posterior probabilities facilitate the use of probabilistic risk models in the choice of treatment (Holmberg & Vickers, 2013), rather than an all or nothing approach. For research purposes, the International Cancer Genome Consortium recommends that catalogs of somatic mutations target a precision of 95% and a recall of 80% (International Cancer Genome Consortium, 2019). Achieving this goal while minimizing cost demands well-calibrated posterior probabilities.

Our current implementation of BATCAVE is as a post-calling algorithm for Mu-

Tect, but the algorithm is broadly applicable. We chose to build BATCAVE off MuTect because MuTect is widely used, has state-of-the-art sensitivity and specificity, and includes numerous heuristic filters and alignment adjustments that reduce the prevalence of sequencing errors in results (Cibulskis et al., 2013; Griffith et al., 2015). But the mutational prior can be incorporated into almost any caller with an underlying probabilistic model. For example, Strelka2 computes a joint posterior probability over tumor and normal genotypes, assuming a constant somatic mutation probability at each genomic site (Kim et al., 2018). Replacing that constant probability with a mutational prior would require a more complicated manipulation of the quality scores output by Strelka than for MuTect, but it is conceptually straightforward.

The BATCAVE algorithm is computationally inexpensive; our current implementation adds 1 second per 22,000 variants evaluated to a standard GATK best-practices variant calling pipeline. The majority of the computational cost is associated with extracting the trinucleotide context for each potential variant site from the reference genome. Since most callers are already walking the reference genome during the calling process, extracting the trinucleotide context simultaneously would virtually eliminate the computational cost of implementing a mutational prior.

The BATCAVE algorithm incorporates genomic context into the probabilistic model for variant calling. Our current implementation focuses on trinucleotide context, which is known to have a large effect on local mutation rates (Martincorena & Campbell, 2015; Hollstein et al., 2017). There are, however, many other aspects of genomic context that can affect local mutation rates (Buisson et al., 2019), including replication timing (Stamatoyannopoulos et al., 2009), expression level (Pleasance et al., 2010), and chromatin organization (Schuster-Böckler & Lehner, 2012). Some of these, such as replication timing and chromatin organization, could be incorporated into the BATCAVE mutational prior using the empirical distribution of mutations in the human germline (Hodgkinson & Eyre-Walker, 2011). Others, such as expression level, could be tumor-specific, but would require information not available in the

variant calls to compute. In the long run, we believe that incorporating more tumor biology into variant calling models will continue to improve performance.

BATCAVE divides the data into two classes: high- and low-confidence variants. The high-confidence variants are used to estimate the mutational prior and mutation rate, which are then used to improve the calling of low-confidence variants. Statistically, this is an empirical Bayesian approach (Robbins, 1954), in which the high and low-confidence variants are treated as parallel experiments (Morris, 1983; Efron, 2014). In general, high-confidence variants tend to have relatively high allelic frequencies, and consequently tend to have arisen early in tumor development. An implicit assumption of our approach is that the mutational process does not change between high- and low-confidence variants, implying that the mutational profile of the tumor is temporally constant. Recent studies have found differences in mutational profiles among variants of different allelic frequencies (Rubanova et al., 2018), although those differences are relatively small. A potential extension of the BATCAVE algorithm is to process potential variants in order of descending allelic frequency and to update the estimated mutational prior as the algorithm proceeds. This approach might increase sensitivity to low-frequency variants generated by recently-arisen mutational processes, at the cost of potentially increasing sensitivity to patterns of sequencing error.

Our results show that adding a mutational prior substantially improves probabilistic variant calling, particularly for tumors with concentrated profiles. Improved variant calling increases the benefit-to-cost ratio of deep sequencing in both research and clinical applications. Moreover, BATCAVE proves to be a better calibrated caller than vanilla MuTect (Fig. 3.5). Different users will prefer different tradeoffs in terms of precision and recall, which can be more accurately made with BATCAVE. Our R implementation, `batcaver`, can be easily incorporated into any MuTect-based pipeline, and the mutational profile algorithm can be incorporated into many other callers.

3.6 Software availability

The `batcaver` R package can be downloaded or installed from

`http://github.com/bmannakee/batcaver`

The version of `batcaver` used to generate results and all analysis code have been preserved on Zenodo

`https://doi.org/10.5281/zenodo.3471715`

Python code used to generate simulated tumors has been preserved on Zenodo

`https://doi.org/10.5281/zenodo.3471741`

3.7 Acknowledgments

This work was supported by the National Science Foundation via Graduate Research Fellowship award number DGE-1143953 to BKM and by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM127348 to RNG. We thank Prof. Edward J. Bedrick for fruitful discussions about the statistical model. This material is based upon High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and RDI and maintained by the UA Research Technologies department.

3.8 Supplementary figures

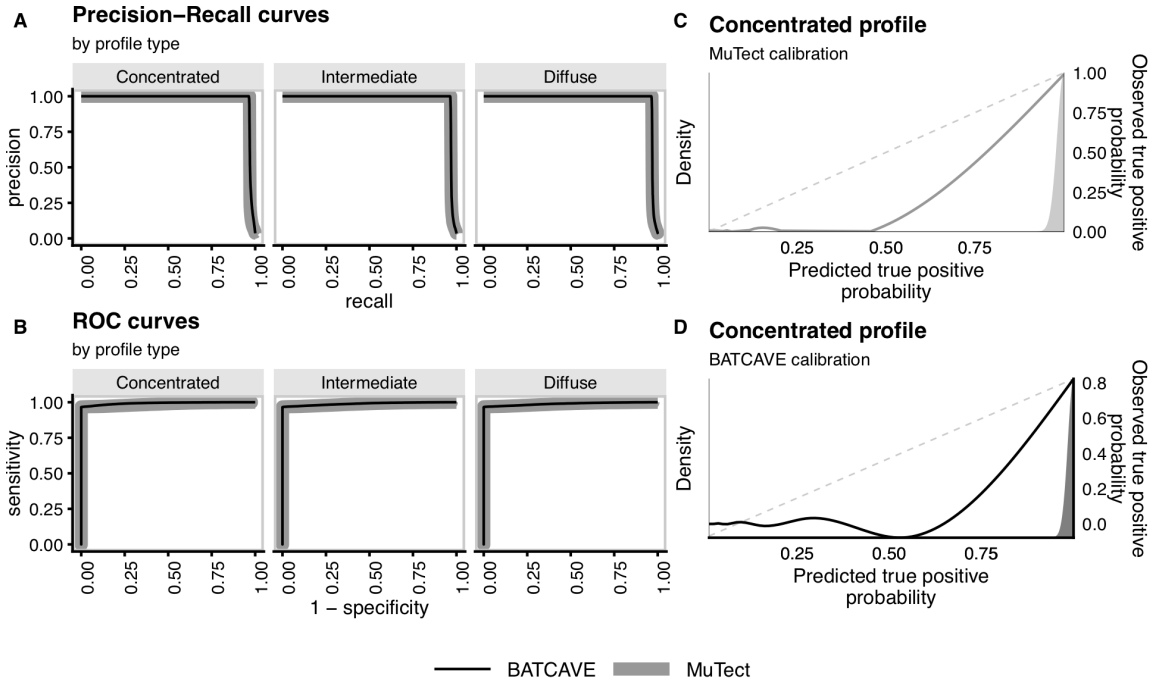

FIGURE 3.6. Variant-calling performance on simulated 100X whole genomes. As in Fig. 3.4A-D, but for 100X whole genomes.

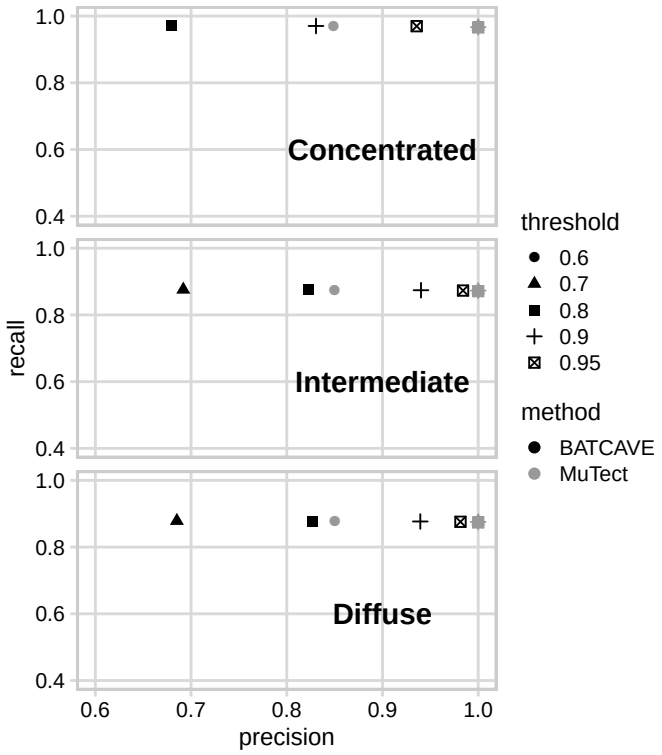

FIGURE 3.7. Posterior probability calibration for realistic calling thresholds, for 100X whole genomes. As in Fig. 3.5, but for 100X whole genomes.

Chapter 4

Conclusion

4.1 Summary of my work

In Chapter 2 I described MAPEX, an algorithm designed to facilitate improved variant calling from tumors grown in mice (patient-derived xenografts). When human tumors are grown in mice their blood supply, connective tissue, and immune infiltration are all performed by mouse cells. As a result, when these tumors are removed and dissected for sequencing a significant fraction of DNA extracted is from the mouse. With careful dissection this fraction can be low, leading to the problem of determining whether low frequency variants identified by a variant calling algorithm are somatic mutations in the tumor, or derived from wild-type mouse DNA. The MAPEX algorithm and associated R package `mapexr` takes as input the NGS reads derived from the xenograft along with a set of somatic variant calls, and uses the NCBI tool BLASTN to separate mouse reads from human. MAPEX then outputs a new set of variant calls classified as either tumor somatic or mouse wild type. I demonstrated that the algorithm performs as well as competing methods, while being easy to use.

In Chapter 3 I described BATCAVE, an extension of the allele-frequency based statistical variant calling model, along with an R implementation of the method. BATCAVE leverages information about the biological processes under way in every tumor to provide a tumor-and-site-specific prior probability of mutation for every site in the genome. I show that the new method improves measures of variant classification ability across a range of sequencing depths and mutation profiles, while adding little computation to existing variant calling pipelines. The algorithm is general, and can be easily extended to incorporate a wide variety of functional forms for both the prior and the likelihood.

4.2 Future directions

While ensemble and deep learning methods provide promise for improved accuracy in variant calling, they suffer from the need for external training data and a lack of biological interpretability. Meanwhile, there remain important improvements to be made in the older, more interpretable, class of variant callers focused on modeling posterior probabilities of variants. A major hurdle in the evaluation of variant calling methods is the scarcity of high quality validated datasets for testing. All major variant calling validation studies to date use the union of an ensemble of variant callers to select mutations to validate (Griffith et al., 2015; Shi et al., 2018). As a result, every validated mutation is by definition callable, and the set of very low frequency mutations for which the read evidence is ambiguous are not fully characterized in terms of true positives and true negatives (Griffith et al., 2015; Shi et al., 2018). What is required for a true test of statistical models for variant classification is to also validate at least a large random sample of variants for which the heuristic filters all pass, but the posterior probability for the variant is below the classification threshold.

The paucity of data validated in such a way has led the cancer sequencing methods field to move away from the underlying statistical model for variant calling and toward improvements to heuristic filters and alignment. The advantage of focusing on heuristic filters is that limited sequencing can be done to find common failure modes, and heuristic rules created to deal with those modes. However, in a world where every failure mode resulting from the complexity of the genome and the sequencing process there will remain a need for a statistical model to deal with the underlying variant generating process. The work in Chapter 3 of this dissertation is a demonstration of the potential benefits that can be derived from better models of the mutation generating process. My hope is that this demonstration will shift the cost-benefit characteristics of deep tumor sequencing and variant validation experiments, resulting in data sets amenable to the complete investigation of the properties

of detailed models of the mutation generating process.

Genomic contexts that effect the probability of mutation occur at multiple scales (Buisson et al., 2019). The tri-nucleotide context is a small scale feature comprised of three adjacent bases. Other important mutational processes, such as mutational hotspots resulting from DNA minor groove orientation around nucleosomes (Pich et al., 2018), non-canonical DNA secondary structures (Georgakopoulos-Soares et al., 2018), and DNA hairpins (Buisson et al., 2019), are characterized by mesoscale features comprised of approximately 30 nucleotides. Finally, large scale processes operating at the chromosome or chromosome region level (Hodgkinson & Eyre-Walker, 2011) and replication timing (Stamatoyannopoulos et al., 2009) are known to effect mutation rates. The method in Chapter 3 could be extended to account for many of these meso- and large-scale processes through the incorporation of annotated genome maps containing relative mutation rates at multiple scales.

An inherent assumption of my work is that while every tumor has its own characteristic mutation rate, that mutation rate is the same at every region of the genome. In other words, a particular genomic context has the same prior probability of mutation wherever it occurs. While this assumption is better that assuming than every tumor has the same underlying mutation rate, there is potentially a great deal of benefit from incorporating large scale features of the genome and their demonstrated effect on the local mutation rate. For instance, There is a strong 10-bp periodicity in mutation rate associated with DNA minor groove orientation around nucleosomes, and the magnitude of this effect is dependent on the same mutational processes underlying the tri-nucleotide context mutation profile (Pich et al., 2018). I believe that modeling minor groove orientation mapping and tri-nucleotide context mutation profiles my provide significant improvement to the statistical model developed in Chapter 3.

The accumulation of mutations in a tumor is an evolutionary process, with important implications for tumor biology. In Chapter 3 I make the explicit assumption that

the mutation processes operating in a tumor are constant throughout tumor evolution. One important addition to the work presented here will be to use large datasets of validated mutations to investigate the evolution of mutational processes. Better validated data sets and very deep sequencing will be required to do this because most variants present in a tumor are present at low frequency, and a large number of variants are required to identify change-points in mutation profiles (Rubanova et al., 2018). It is my sincere hope that the contributions in this dissertation help to facilitate the evolution of the field along these lines.

References

Ainscough BJ, Barnell EK, Ronning P, et al. (2018) A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. Nature Genetics 50:1735.

Alexandrov LB, Jones PH, Wedge DC, et al. (2015) Clock-like mutational processes in human somatic cells. Nature Genetics 47:1402.

Alexandrov LB, Ju YS, Haase K, et al. (2016) Mutational signatures associated with tobacco smoking in human cancer. Science 354:618.

Alexandrov LB, Kim J, Haradhvala NJ, et al. (2019) The Repertoire of Mutational Signatures in Human Cancer. bioRxiv page 322859.

Alexandrov LB, Nik-Zainal S, Wedge DC, et al. (2013a) Signatures of mutational processes in human cancer. Nature 500:415.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, and Stratton MR (2013b) Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Reports 3:246.

Allaway RJ, Fischer DA, de Abreu FB, et al. (2016) Genomic characterization of patient-derived xenograft models established from fine needle aspirate biopsies of a primary pancreatic ductal adenocarcinoma and from patient-matched metastatic sites. Oncotarget 7:17087.

Austin PC and Steyerberg EW (2019) The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine page sim.8281.

Bailey MH, Tokheim C, Porta-Pardo E, et al. (2018) Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 173:371.

Barnell EK, Ronning P, Campbell KM, et al. (2019) Standard operating procedure for somatic variant refinement of sequencing data with paired tumor andnormal samples. Genetics in Medicine 21:972.

Borad MJ, Champion MD, Egan JB, et al. (2014) Integrated Genomic Characterization Reveals Novel, Therapeutically Relevant Drug Targets in FGFR and EGFR Pathways in Sporadic Intrahepatic Cholangiocarcinoma. PLoS Genetics 10.

Boutros PC (2015) The path to routine use of genomic biomarkers in the cancer clinic. Genome research 25:1508.

Bowler TG, Bartenstein M, Morrone KA, et al. (2014) Exome Sequencing of Familial MDS Reveals Novel Mutations and High Rates of False Positive Mutations in MLL3 Due to Pseudogene Effects. Blood 124:4591.

Bozic I, Gerold JM, and Nowak MA (2016) Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. PLoS Computational Biology 12:e1004731.

Bruna A, Rueda OM, Greenwood W, et al. (2016) A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds. Cell 167:260.

Buisson R, Langenbucher A, Bowen D, et al. (2019) Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science (New York, NY) 364:eaaw2872.

Burrell RA, McGranahan N, Bartek J, and Swanton C (2013) The causes and consequences of genetic heterogeneity in cancer evolution. Nature 501:338.

Cameron DL, Di Stefano L, and Papenfuss AT (2019) Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nature Communications 10:3240.

Cantarel BL, Weaver D, McNeill N, et al. (2014) BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinformatics 15:104.

Carrot-Zhang J and Majewski J (2017) LoLoPicker: detecting low allelic-fraction variants from low-quality cancer samples. Oncotarget 8:37032.

Carter SL, Cibulskis K, Helman E, et al. (2012) Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology 30:413.

Carvalho CMB and Lupski JR (2016) Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics 17:224.

Chen X, Stewart E, Shelat AA, et al. (2013) Targeting Oxidative Stress in Embryonal Rhabdomyosarcoma. Cancer Cell 24:710.

Christoforides A, Carpten JD, Weiss GJ, et al. (2013) Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. BMC Genomics 14:302.

Cibulskis K, Lawrence MS, Carter SL, et al. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology 31:213.

Conway T, Wazny J, Bromage A, et al. (2012) Xenome–a tool for classifying reads from xenograft samples. Bioinformatics 28:i172.

COSMIC Consortium (2019) COSMIC Mutational Signatures (Version 2) https://cancer.sanger.ac.uk/cosmic/signatures_v2.

Dawson CW, Port RJ, and Young LS (2012) The role of the EBV-encoded latent membrane proteins LMP1 and LMP2 in the pathogenesis of nasopharyngeal carcinoma (NPC). Seminars in Cancer Biology 22:144.

Day CP, Merlino G, and VanDyke T (2015) Preclinical Mouse Cancer Models: A Maze of Opportunities and Challenges. Cell 163:39.

Ding L, Ley TJ, Larson DE, et al. (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 481:506.

Dorri F, Jewell S, Bouchard-Côté A, and Shah SP (2019) Somatic mutation detection and classification through probabilistic integration of clonal population information. Communications Biology 2:44.

Efron B (2014) Two modeling strategies for empirical Bayes estimation. Statistical science : a review journal of the Institute of Mathematical Statistics 29:285.

Ewing AD, Houlahan KE, Hu Y, et al. (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nature Methods 12:623.

Fan Y, Xi L, Hughes DS, et al. (2016) MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome biology 17:178.

Fang LT, Afshar PT, Chhibber A, et al. (2015) An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biology 16:197.

Fearon EF and Vogelstein B (1989) A Genetic Model for Colorectal Tumorigenesis. Cell 61:759.

Findlay JM, Castro-Giner F, Makino S, et al. (2016) Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. Nature Communications 7.

Garrison E and Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv .

Gehring JS, Fischer B, Lawrence M, and Huber W (2015) SomaticSignatures: inferring mutational signatures from single-nucleotide variants: Fig. 1. Bioinformatics 31:btv408.

Georgakopoulos-Soares I, Morganella S, Jain N, Hemberg M, and Nik-Zainal S (2018) Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. Genome research 28:1264.

Gerstung M, Beisel C, Rechsteiner M, et al. (2012) Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nature Communications 3:811.

Griffith M, Miller CA, Griffith OL, et al. (2015) Optimizing Cancer Genome Sequencing and Analysis. Cell Systems 1:210.

Helleday T, Eshtad S, and Nik-Zainal S (2014a) Mechanisms underlying mutational signatures in human cancers. Nature Reviews Genetics 15:585.

Helleday T, Eshtad S, and Nik-Zainal S (2014b) Mechanisms underlying mutational signatures in human cancers. Nature Reviews Genetics 15:585.

Hodgkinson A and Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. Nature Reviews Genetics 12:756.

Hollstein M, Alexandrov LB, Wild CP, Ardin M, and Zavadil J (2017) Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. Oncogene 36:158.

Holmberg L and Vickers A (2013) Evaluation of Prediction Models for Decision-Making: Beyond Calibration and Discrimination. PLoS Medicine 10:e1001491.

International Cancer Genome Consortium (2019) International Cancer Genome Consortium Goals, Structure, Policies, and Guidelines https://icgc.org/icgc/goals-structure-policies-guidelines/e8-genome-analyses.

Jacobs MT, Mohindra NA, Shantzer L, et al. (2018) Use of Low-Frequency Driver Mutations Detected by Cell-Free Circulating Tumor DNA to Guide Targeted Therapy in NonSmall-Cell Lung Cancer: A Multicenter Case Series. JCO Precision Oncology pages 1–10.

Jia P, Li F, Xia J, et al. (2012) Consensus rules in variant detection from next-generation sequencing data. PLoS ONE 7.

Jones D, Raine KM, Davies H, et al. (2016) cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. Current Protocols in Bioinformatics 56:15.10.1.

Kandoth C, McLellan MD, Vandin F, et al. (2013) Mutational landscape and significance across 12 major cancer types. Nature 502:333.

Kelley MJ, Li S, and Harpole DH (2001) Genetic Analysis of the Beta-Tubulin Gene, TUBB, in Non-Small-Cell Lung Cancer. Journal of the National Cancer Institute 93:1886.

Khandelwal G, Girotti MR, Smowton C, et al. (2017) Next-Gen Sequencing Analysis and Algorithms for PDX and CDX Models. Molecular Cancer Research .

Kim S, Scheffler K, Halpern AL, et al. (2018) Strelka2: fast and accurate calling of germline and somatic variants. Nature Methods 15:591.

Knudsen ES, Balaji U, Mannakee B, et al. (2017) Pancreatic cancer cell lines as patient-derived avatars: genetic characterisation and functional utility. Gut pages gutjnl–2016–313133.

Koboldt DC, Zhang Q, Larson DE, et al. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research 22:568.

Kumar A, Coleman I, Morrissey C, et al. (2016) Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. Nature Medicine 22:369.

Larson DE, Harris CC, Chen K, et al. (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics (Oxford, England) 28:311.

Lawrence M, Huber W, Pagès H, et al. (2013) Software for Computing and Annotating Genomic Ranges. PLoS Computational Biology 9:1.

Lee-Six H, Øbro NF, Shepherd MS, et al. (2018) Population dynamics of normal human blood inferred from somatic mutations. Nature 561:473.

Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754.

Liu J, Lichtenberg T, Hoadley KA, et al. (2018) An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell 173:400.

Mandelker D, Schmidt RJ, Ankala A, et al. (2016) Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. Genetics in Medicine 18:1.

Mannakee BK, Balaji U, Witkiewicz AK, Gutenkunst RN, and Knudsen ES (2018) Sensitive and specific post-call filtering of genetic variants in xenograft and primary tumors. Bioinformatics 34:1713.

Mardis ER (2012) Applying next-generation sequencing to pancreatic cancer treatment. Nature Reviews Gastroenterology & Hepatology 9:477.

Mardis ER (2018) Insights from Large-Scale Cancer Genome Sequencing. The Annual Review of Cancer Biology is online at Annu Rev Cancer Biol 2:429.

Martincorena I and Campbell PJ (2015) Somatic mutation in cancer and normal cells. Science (New York, NY) 349:1483.

McKenna A, Hanna M, Banks E, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20:1297.

Morgan M, Pagès H, Obenchain V, and Hayden N (2017) Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import..

Morris CN (1983) Parametric Empirical Bayes Inference: Theory and Applications. Journal of the American Statistical Association 78:47.

Mu JC, Mohiyuddin M, Li J, et al. (2015) VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. Bioinformatics 31:1469.

Müller CI, Miller CW, Hofmann WK, et al. (2007) Rare mutations of the PIK3CA gene in malignancies of the hematopoietic system as well as endometrium, ovary, prostate and osteosarcomas, and discovery of a PIK3CA pseudogene. Leukemia Research 31:27.

Nakamura H, Arai Y, Totoki Y, et al. (2015) Genomic spectra of biliary tract cancer. Nature Genetics 47:1003.

Ng SB, Nickerson DA, Bamshad MJ, and Shendure J (2010) Massively parallel sequencing and rare disease. Human Molecular Genetics 19:119.

Nik-Zainal S, Alexandrov L, Wedge D, et al. (2012a) Mutational Processes Molding the Genomes of 21 Breast Cancers. Cell 149:979.

Nik-Zainal S, Davies H, Staaf J, et al. (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534:47.

Nik-Zainal S, Van Loo P, Wedge DC, et al. (2012b) The life history of 21 breast cancers. Cell 149:994.

Nowell PC (1976) The clonal evolution of tumor cell populations. Science 194:23.

Obenchain V, Lawrence M, Carey V, et al. (2014) VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. Bioinformatics 30:2076.

Pagès H (2019) BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs.

Park SY, Gönen M, Kim HJ, Michor F, and Polyak K (2010) Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. Journal of Clinical Investigation 120:636.

Pich O, Muiños F, Sabarinathan R, et al. (2018) Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. Cell 175:1074.

Pleasance ED, Cheetham RK, Stephens PJ, et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463:191.

Ramos AH, Lichtenstein L, Gupta M, et al. (2015) Oncotator: Cancer variant annotation tool. Human Mutation 36:E2423.

Robbins H (1954) An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163.

Rossello FJ, Tothill RW, Britt K, et al. (2013) Next-Generation Sequence Analysis of Cancer Xenograft Models. PLoS ONE 8.

Roth A, Ding J, Morin R, et al. (2012) JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics 28:907.

Rubanova Y, Shi R, Li R, et al. (2018) Reconstructing Evolutionary Trajectories of Mutations in Cancer. bioRxiv .

Sahraeian SME, Liu R, Lau B, et al. (2019) Deep convolutional neural networks for accurate somatic mutation detection. Nature Communications 10:1041.

Saunders CT, Wong WSW, Swamy S, et al. (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics (Oxford, England) 28:1811.

Schneeberger VE, Allaj V, Gardner EE, et al. (2016) Quantitation of Murine Stroma and Selective Purification of the Human Tumor Component of Patient-Derived Xenografts for Genomic Analysis. PLOS ONE 11:e0160587.

Schuster-Böckler B and Lehner B (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 488:504.

Shah SP, Morin RD, Khattra J, et al. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature 461:809.

Shi W, Ng CKY, Lim RS, et al. (2018) Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity In Brief Article Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity. Cell Reports 25:1446.

Shiraishi Y, Sato Y, Chiba K, et al. (2013) An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. Nucleic Acids Research 41:e89.

Spinella JF, Mehanna P, Vidal R, et al. (2016) SNooPer: A machine learning-based method for somatic variant identification from low-pass next-generation sequencing. BMC Genomics 17:912.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, et al. (2009) Human mutation rate associated with DNA replication timing. Nature Genetics 41:393.

Stephens P, Edkins S, Davies H, et al. (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nature Genetics 37:590.

Stratton MR (2011) Exploring the Genomes of Cancer Cells: Progress and Promise. Science 331:1553.

Tanaka Y, Kanai F, Tada M, et al. (2006) Absence of PIK3CA hotspot mutations in hepatocellular carcinoma in Japanese patients. Oncogene 25:2950.

Thorvaldsdottir H, Robinson JT, and Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14:178.

Tso KY, Lee S, Lo KW, and Yip KY (2014) Are special read alignment strategies necessary and cost-effective when handling sequencing reads from patient-derived tumor xenografts? BMC Genomics 15:1172.

Way GP, Sanchez-Vega F, La K, et al. (2018) Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell reports 23:172.

Williams MJ, Werner B, Barnes CP, Graham TA, and Sottoriva A (2016) Identification of neutral tumor evolution across cancer types. Nature Genetics 48:238.

Williams MJ, Werner B, Heide T, et al. (2018a) Quantification of subclonal selection in cancer from bulk sequencing data. Nature Genetics 50:895.

Williams MJ, Werner B, Heide T, et al. (2018b) Quantification of subclonal selection in cancer from bulk sequencing data. Nature Genetics 50:895.

Wilm A, Aw PPK, Bertrand D, et al. (2012) LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Research 40:11189.

Witkiewicz A, Balaji U, Eslinger C, et al. (2016) Integrated Patient-Derived Models Delineate Individualized Therapeutic Vulnerabilities of Pancreatic Cancer. Cell Reports 16:2017.

Witkiewicz AK, McMillan EA, Balaji U, et al. (2015) Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nature communications 6:6744.

Woo XY, Srivastava A, Graber JH, et al. (2019) Genomic data analysis workflows for tumors from patient-derived xenografts (PDXs): challenges and guidelines. BMC Medical Genomics 12:92.

Wood DE, White JR, Georgiadis A, et al. (2018) A machine learning approach for somatic mutation discovery. Science translational medicine 10:eaar7939.

Xu C (2018) A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Computational and Structural Biotechnology Journal 16:15.

Yachida S, Jones S, Bozic I, et al. (2010) Distant metas.tasis occurs late during the genetic evolution of pancreatic cancer. Nature 467:1114.

Zhang Y, Yang L, Kucherlapati M, et al. (2018) A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. CellReports 24:515.

Zhou W, Zhao H, Chong Z, et al. (2015) ClinSeK: a targeted variant characterization framework for clinical sequencing. Genome Medicine 7:34.