Targets

- Nucleic Acids Research published EBCall in 2013
- BioinformaticEmprical Bayes estimate of the posterior probability of mutation
- NAR has a brand new genomics and bioinformatics journal

Introduction

Cancer develops as the result of the accumulation of somatic mutations and clonal selection of cells with mutations that confer a selective advantage on the cell. Understanding the forces that shaped the evolutionary history of a tumor, the mutations that are responsible for its growth, the rate at which mutations are occurring, or how much genetic diversity is likely present in the tumor, requires accurate variant calling, particularly at low variant allele frequency (Williams et al. 2016; Bozic, Gerold, and Nowak 2016; Williams et al. 2018). Accurate variant identification is also critical in optimizing the treatment regime for an individual patients disease (J. Ding et al. 2012; E. R. Mardis 2012; X. Chen et al. 2013; Borad et al. 2014; Findlay et al. 2016). Low frequency mutations present a significant problem for current mutation calling methods because their signature in the data is difficult to distinguish from the noise introduced by Next Generation Sequencing (NGS), and this problem increases as sequencing depth increases.

Methods for identifying true somatic mutations - i.e. variant calling - from NGS data are an active area of research in bioinformatics. The earliest widely used somatic variant callers aimed specifically at tumors, Mutect1 and Varscan2, used a combination of heuristic filtering and a model of sequencing errors to identify and score potential variants, setting a threshold for that score designed to balance sensitivity and specificity (D. C. Koboldt et al. 2012; Cibulskis et al. 2013). Subsequent research gave rise to a number of alternate variant calling strategies including haplotype based callers (Garrison and Marth 2012), joint genotype analysis (SomaticSniper, JointSNVMix2, Seurat, and CaVEMan,MuClone)(D. E. Larson et al. 2012; Roth2012a;@Christoforides2013; D. Jones et al. 2016; Dorri et al. 2019), allele frequency based analysis (Strelka, MuTect, LoFreq, EBCall, deepSNV, LoLoPicker, and MuSE)(Saunders et al. 2012; Wilm et al. 2012; Shiraishi2013b;Gerstung2012;@Carrot-Zhang2017; Fan et al. 2016), and a mixture of ensemble and deep learning methods (MutationSeq, SomaticSeq, SNooPer, and BAYSIC). All of these methods have varying levels of complexity, and some are focused on specific types of data. The one thing they all have in common is that they either implicitly or explicitly assume that the probability of a mutation occuring at a give site is proportional to the overall mutation rate, and the same at every site in the genome.

Single nucleotide substitions, i.e. simple mutations, arise in tumors at a rate and at genomic locations driven by two main processes. The first is the spontaneous accumulation of mutations that occurs in all dividing tissues, and has a characteristic mutation signature that describes the probability of mutation in a given genomic context (Nik-Zainal et al. 2012; Ludmil B Alexandrov et al. 2015; Lee-Six et al. 2018). The second, and far more complex, process is the accumulation of mutations through exposure to mutagens or degradation - via mutation or deletion - of cellular machinery responsible for the identification and repair of damage or replication errors. Many mutagens and DNA repair mechanism defects also have highly specific mutation signatures, such that they can be identified by observing the mutations in the tumor (Alexandrov et al. 2013; Helleday, Eshtad, and Nik-Zainal 2014; Nik-Zainal et al. 2016; Kandoth et al. 2013; L. B. Alexandrov et al. 2016).

Here we present an empirical bayes method for estimating the prior probability of mutation at a given site using the observed mutation spectrum of the tumor, and show that the addition of this prior to the MuTect variant calling model produces a superior variant classifier in both simulated and real tumor data. We then extend the method with an application of the local false discovery rate by computing the probability that a site is non-null under an assumption of clonal expansion with either early or small selective differences between clones. We provide a simple implementation in R that takes MuTect caller output as input, and returns the posterior probability that a site is variant for every site observed by MuTect.

Results

Performance measurements

We are interested in measuring two different quantities here. First, we want to know whether the addition of a biological prior improves the statistical model. We use the area under the ROC curve to measure whether the biological prior provides a better ordering of true and false positive variants. Second, we want to know whether a particular decision threshold implemented on the biological prior provides an improved tradeoff of true to false positives than MuTect. Both the shape and the area under the precision-recall curve are informative here.

Implementation

We implement our model on top of the MuTect 1.1.7 output. MuTect1 and MuTect2 both report the log likelihood ratio of two models, one with the variant and one without, which we can directly convert to posterior odds in favor of a mutation. Other variant callers have probability models that could be converted to use the mutation signature prior, but MuTect's is most directly accessible. We use MuTect 1.1.7 rather than MuTect2 because MuTect2 also does haplotype calling and realignment, making it difficult to use with simulated data (i.e. MuTect2 does local realignment after mutations are spiked and sometimes loses true mutations as a result). We chose to run MuTect with an initial probability sufficiently low to ensure that nearly every potential variant was evaluated and assigned a log likelihood ratio in order to have the largest possible range of true and false positive/negative variants to evaluate the performance of our algorithm. However, no sensible analysis would include exceptionally low likelihood variants, so in our results we show result ony for those potential variants which have a log likelihood ratio (TLOD) > 4, which implies log posterior odds of -2.3, i.e. very small. This adjustment does not change the results, it just makes the analysis easier and more meaningful. The algorithm processes a whole genome simulation consisting of 53 million potential variants in 2400 seconds, of which 1600 seconds are consumed reading the data into R, and 800 seconds collecting genomic contexts from the reference genome. For a whole exome with 2.3M potential variants the run time is 142 seconds, with 56 seconds to read the data and 33 seconds to collect the contexts. The portion of the algorithm that actually computes the prior is a trivial fraction of the whole process. If integrated into an already existing variant caller which is already walking the reference genome it should add no significant processing time.

Sensitivity and specificity in simulated data

In order to describe the operating characteristics of our score as a classifier compared to MuTect, we simulated NGS reads and called variants six tumor-normal pairs as described in Methods. We made three 100X whole genomes and three 500X whole exomes, with three differnent mutation spectra. Differences in performance between our method and MuTect are driven by two main factors; the concentration of the data generating mutation signature, and the fraction of the total mutations in the tumor that are at low frequency and thus near the threshold for calling.

Precision - Recall in simulated data

The fraction of positive calls that are false positives grows as the threshold used to call variants goes down. In such cases precision-recall curves give a better sense of the risk/reward tradeoff between the methods in an actual variant calling situation. We computed precision-recall curves for each our six tumor simulations (Figures 2 and 3). We find that for 100X depth whole genomes, MuTect has a slight advantage (1-2% area under the curve), driven by a sharp dropoff in precision which occurs at lower recall for our method than for MuTect, while our method has a slight advantage at 500X depth (1-3% area under the curve). The differences are driven by the allele frequency distributions we simulated, the concentration in the mutation

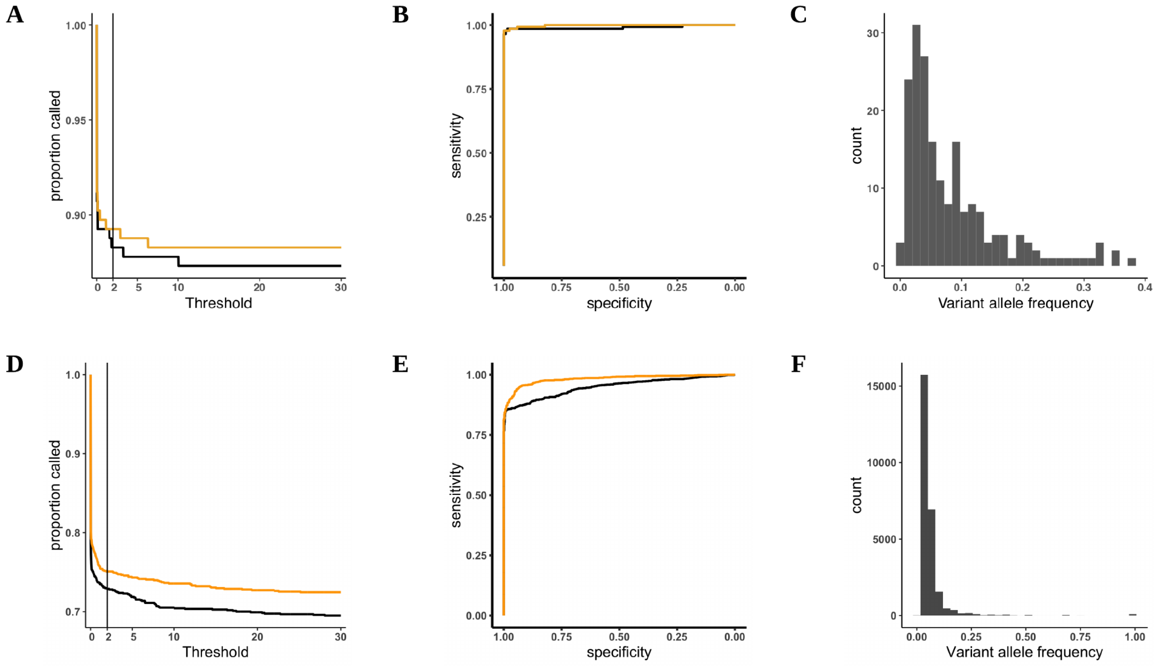

Figure 1: Sensitivity in simulated tumors. A-C) Whole exome simulation. D-F) Whole genome simulation

signature, and the discrete nature of the distribution of TLOD values. Figure 4 shows the distribution of TLOD scores for 1, 2, 3, 4, 5, and six alterate reads as a function of total sequencing depth. At 100X, the mutect threshold falls exactly between 2 and 3 alternate reads, representing variant allele frequencies of 2 and 3%. The allele frequency distribution of the whole genome simulations has less that 5% of variants at or below 2%, so that nearly all positive findings below 2% are false positives. When our algorithm elevates the probability of variants at below 2% they are nearly all false positives, and as a result precision suffers. At 500X sequencing depth, combined with the allele frequency distribution used in the simulations, there is a broader range of recall values for which the precision of our algorithm is better than MuTect, and as shown in figure 4 the alternate read count at which MuTect begins to perform better than our method is now between 4 and 5 reads, or variant allele frequencies of 1%. As depth increases the allele frequency at which our method performs better than MuTect continues to decrease. Thresholding on vaf remains necessary, but as as sequencing depth increases that advantage of our method increases.

Convergence of the prior to simulated target distributions.

In both whole genome whole exome simulations, the estimated mutation spectrum is very close to the simulated spectrum (Supplementary Figure1 and Figure 3). We ranked all mutations called by MuTect by their TLOD score from highest to lowest, and computed the Kullback-Leibler divergence between the prior and the target distribution as each new mutation was observed (Figure 2). In our simulations, which have high read depth, the prior converges to the target well before all mutations passed by MuTect are evaluated. The quality of the estimate increases with the number of mutations and will likely be suboptimal for low depth sequence with a small number of high confidence mutations. Convergence is faster and the prior moves closer to the target distribution the more concentrated the simulated signature is. - this is what we would expect.

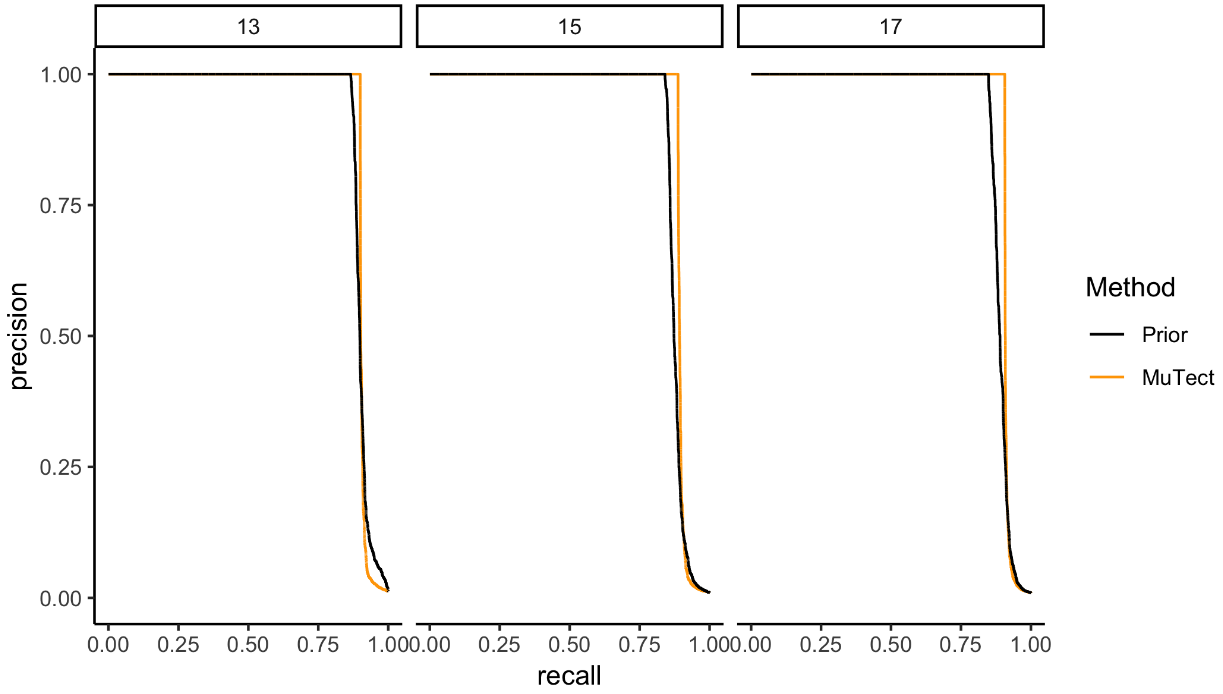

Figure 2: WGS precision-recall plot

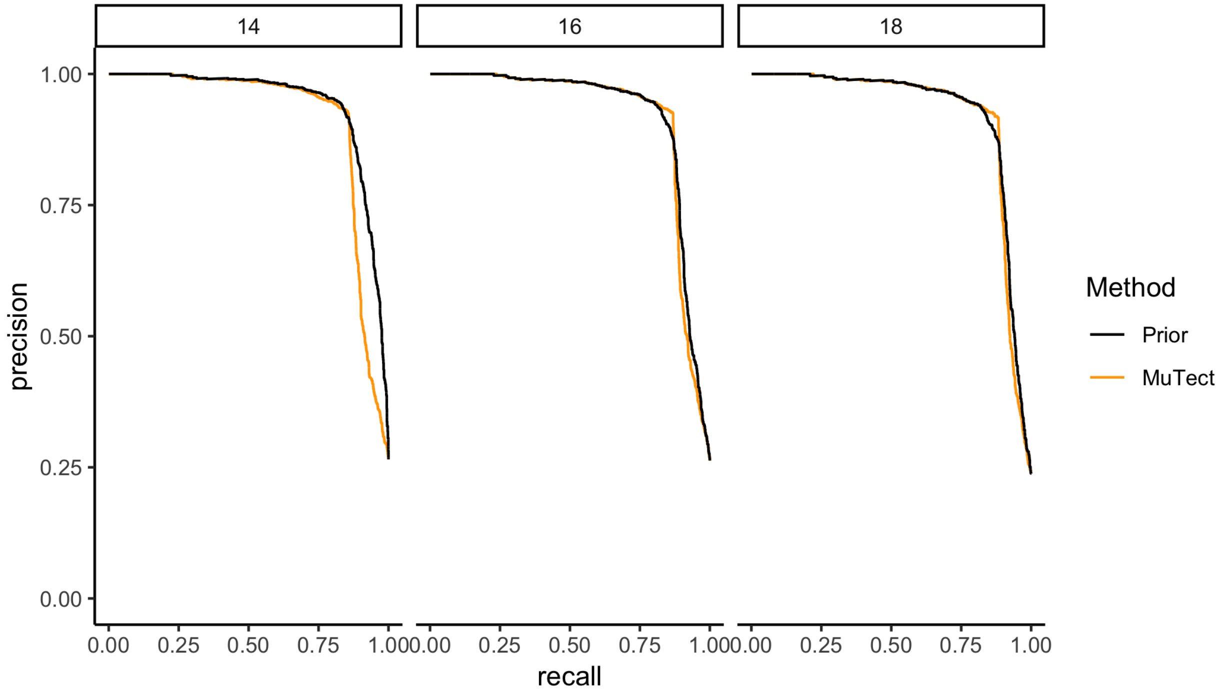

Figure 3: WES precision-recall plot

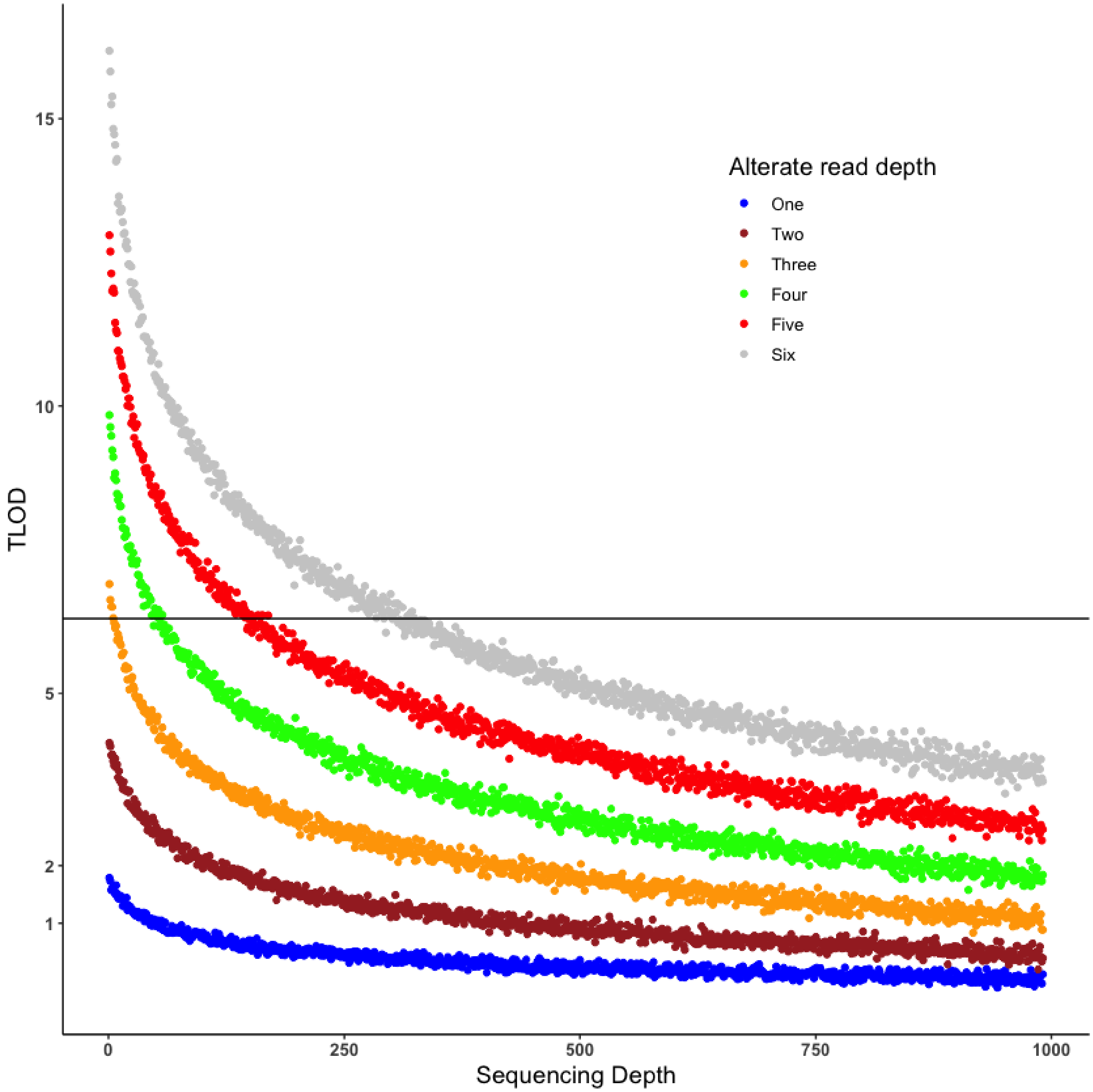

Figure 4: Plot of dependence of TLOD on sequencing depth

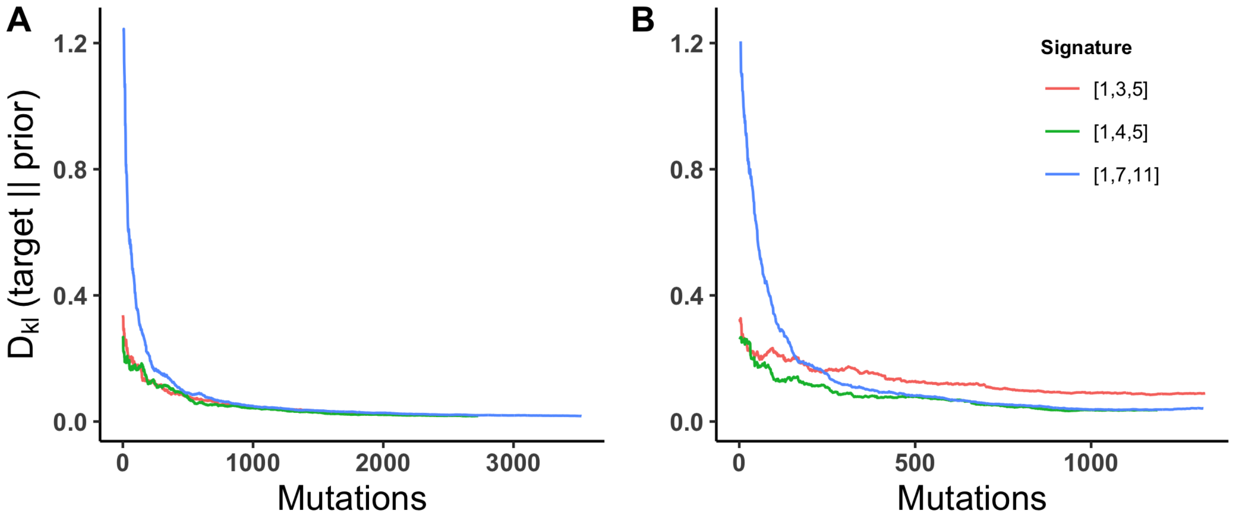

Figure 5: Convergence of the prior to simulated target mutation signatures. The prior distribution converges quickly to the target distribution, and after 200-300 mutations is as close as it will get in both A) WGS simulations and B) Whole exome simulations

Sensitivity in real data

We examined two real tumor datasets in which variants had been validated by deep targeted resequencing (M. Griffith et al. 2015; W. Shi et al. 2018). M. Griffith et al. (2015) performed whole genome sequencing of an acute myeloid leukemia to a depth of ~312X, called variants with seven different variant callers and validated over 200,000 variants by targeted re-sequencing to a depth of ~1000X. This led to a platinum set of variant calls containg 1,343 SNVs. We obtained BAM files from this experiment and called variants using MuTect 1.1.7, then compared the sensitivity of the calls between MuTect and our method (Figure 1A). At any relevant threshold our method is slightly more sensitive than MuTect. MuTect is unable to recover 100% of the calls due to hueristic filtering and other differences between MuTect and the other variant callers used.

W. Shi et al. (2018) performed multi-region sequencing of 6 breast tumors to evaluate the effects of variant calling and sequencing depth on estimates of tumor heterogeneity, validating 1,385 somatic SNVs. As with the leukemia we obtained BAM files for this experiment and compared our method to raw MuTect calls (Figure 1B). We again find that our method is more sensitive than MuTect across the full range of relevant thresholds.

Discussion

- Relevance to germline mutations (Rahbari et al. 2016), and somatic mutation in healthy tissue (Lee-Six et al. 2018)
- Relevance to deep learning, should at least be a feature.
- Standalone package, but approach really should be integrated into callers
- Computational efficiency if integrated
- Applicability to other algorithms for somatic variant calling
- Why are false negative rates important?
- heterogeneity
- selection inference
- rare but druggable variant identification
- Caveat: Need for better real tumor validation sets. Focus on false negatives as well as false positives.

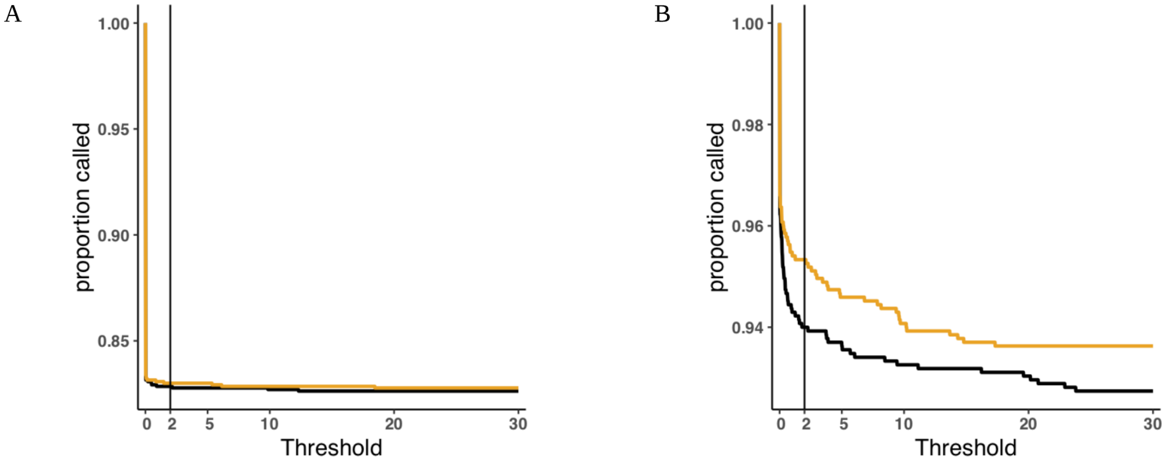

Figure 6: Sensitivity in real tumors. A) AML31 platinum SNV calls (M. Griffith et al. 2015). B) Validated SNV in 6 breast cancers(W. Shi et al. 2018).

The aml31 paper gets alot of them, but if they had for instance just used mutect to identify any potential variant that passed all other heuristic filters they would have a better sense of false negative rates.

- Caveat: evolution of mutational spectrum [Rubanova2018a]

Methods

Algorithm

At every site in the genome with non-zero coverage, Next Generation Sequencing (NGS) produces a vector $\mathbf{x} = (\{b_i\}, \{q_i\}), i = 1 \dots D$ of base calls and their associated quality scores, where D is total read depth. The goal is to use $\mathbf{x}$ to select between competing hypotheses;

$$\begin{aligned} \mathbf{H_0}: &\quad \text{Alt allele} = m; \quad \nu = 0 \\ \mathbf{H_1}: &\quad \text{Alt allele} = m; \quad \nu = \hat{f}, \end{aligned}$$

where ν is the variant allele frequency, $\hat{f}$ is the maximum likelihood estimate of ν given data $\mathbf{x}$, i.e. the ratio of the count of variant reads and total read depth, and m is any of the 3 possible alternative non-reference bases. For a given read with base b_i and q-score q_i, the density function under a particular hypothesis is defined as

$$\mathrm{f}_{\nu,m}(b_i, q_i) = \begin{cases} \nu \frac{10^{-q_i/10}}{3} + (1-\nu)(1 - 10^{-q_i/10}) & b_i = \text{reference} \\ \nu(1 - 10^{-q_i/10}) + (1-\nu)\frac{10^{-q_i/10}}{3} & b_i = m \\ \frac{10^{-q_i/10}}{3} & \mathit{otherwise}. \end{cases}$$

The likelihood under the hypothesis is then $\mathcal{L}_{\nu,m}(\mathbf{x}) = \prod_{i=1}^{D} \mathrm{f}_{\nu,m}(x_i)$. MuTect reports the log likelihood ratio $\log(\mathcal{L}_{\nu=\hat{f},m}(\mathbf{x})/\mathcal{L}_{\nu=0,m}(\mathbf{x}))$ as either TLOD or t_lod_fstar depending on the version. By fixing the threshold posterior odds at two, the site-specific mutation probability a constant $\mathrm{p}(M) = \mu = 3\mathrm{e}{-6}$, and $\mathrm{p}(m \mid M)$ the prior probability of mutation to specific allele m constant $\mathrm{p}(m \mid M) = \mu/3 = 1\mathrm{e}{-6}$, they derive a TLOD threshold of 6.3 for classifying a site as a somatic variant. Here we examine the effect of the assumption of a constant prior probability of mutation.

Site-specific prior probability of mutation

While variant calling algorithms typically assume a constant probability of mutation at every site in the genome, work by Alexandrov and others show that the random mutation generating process actually varies from site to site in a nucleotide context specific manner. We develop a model of the prior probability of mutation to allele m conditional on the observed genomic context $\mathrm{p}(m, M \mid C)$, and demonstrate an empirical Bayes method for computing this probability from MuTect output. The prior probability $\mathrm{p}(m, M \mid C)$ can be decomposed as

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

since the probability of a mutation at a site and the probability that it is to allele m are independent conditional on the context. Here $\mathrm{p}(M) = \mu$ as above, and the empirical distribution of contexts $\mathrm{p}(C)$ is the fraction of the genome made up of each context. We model $\mathrm{p}(C \mid M)$ as a multinomial distribution with parameter $\boldsymbol{\pi} = \{\pi_i\}, i = 1 \ldots 96$. Mutations are drawn from this multinomial distribution such that $\mathrm{p}(C = i \mid M) = \pi_i$. The final quantity $\mathrm{p}(m \mid C)$, the probability of mutation to m given a particular three letter context, is a function of $\boldsymbol{\pi}$. We are left to estimate only the vector of probabilities $\boldsymbol{\pi}$.

$$\begin{aligned} C \mid M, \boldsymbol{\pi} &\sim \text{Multinomial}(\boldsymbol{\pi}) \\ \boldsymbol{\pi} \mid \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}). \end{aligned}$$

The posterior distribution of $\boldsymbol{\pi}$ is $\boldsymbol{\pi} \mid C, \boldsymbol{\alpha} \sim \text{Dirichlet}(\mathbf{C} + \boldsymbol{\alpha})$, where $\mathbf{C} = (C_1, \ldots, C_{96})$ are the counts of mutations present in the tumor for each of the 96 contexts. We compute an empirical bayes estimate of $\boldsymbol{\pi}$ by choosing $\mathbf{C}$ as the set of mutations assigned a TLOD by MuTect above some threshold, which we choose as 10. We show through extensive simulation that our estimate of $\boldsymbol{\pi}$ converges quickly its true simulated value.

Returning to the model above, we can calculate the log posterior odds in favor of $\mathbf{H_1}$ as

$$\log_{10}\left(\frac{(\mathcal{L}_{\nu=\hat{f},m}(\mathbf{x})\mathrm{p}(m, M \mid C)}{(\mathcal{L}_{\nu=0,m}(\mathbf{x})(1 - \mathrm{p}(m, M \mid C))}\right) = \text{TLOD} + \text{log prior odds},$$

and the posterior odds ratio in favor of $\mathbf{H_1}$ as

$$10^{(\text{TLOD+log prior odds})}.$$

False positive rate control.

We develop a method, following Efron(2008), for controlling the false positive rate. Every site with sufficient coverage and at least 1 alternate read falls into one of two classes, they are either *null* (non-variant with $\nu = 0$) or *nonnull* (variant with $\nu = \hat{f}$) with prior probabilities p_0 and $p_1 = 1 - p_0$,

$$\begin{array}{lll} p_0 = \mathrm{P}\{\text{null}\} & \mathrm{f}_0(\mathbf{x}) & \text{density if null} \\ p_1 = \mathrm{P}\{\text{nonnull}\} & \mathrm{f}_1(\mathbf{x}) & \text{density if nonnull.} \end{array}$$

Controlling false positive rate implies $\mathrm{P}\{\text{nonnull} \mid \mathbf{x}\}/\mathrm{P}\{\text{null} \mid \mathbf{x}\}$ to some ratio of true positives to false positives. Defining $\mathrm{f}(\mathbf{x}) = \mathrm{f}_0(\mathbf{x}) + \mathrm{f}_1(\mathbf{x})$, then $\mathrm{fdr}(\mathbf{x}) = p_0\mathrm{f}_0(\mathbf{x})/\mathrm{f}(\mathbf{x})$, and

$$\frac{\mathrm{P}\{\text{nonnull} \mid \mathbf{x}\}}{\mathrm{P}\{\text{null} \mid \mathbf{x}\}} = \frac{1 - \mathrm{frd}(\mathbf{x})}{\mathrm{frd}(\mathbf{x})} = \frac{p_1\mathrm{f}_1((x))}{p_0\mathrm{f}_0((x))}.$$

The posterior odds ratio $\mathrm{f}_1(\mathbf{x})/\mathrm{f}_0(\mathbf{x})$ is the one computed by the algorithm above.

Prior odds site is non-null

The local, or site-specific, true positive probability p_1 can be estimated as the fraction of all sequenced sites that are expected to be positive. In a neutrally evolving tumor, or any tumor without very strong late selective sweeps, the count of variants with a given allele frequency $N(f)$ is(bozic et al)

$$N(f) = \frac{N\mu}{f},$$

Where N is the total number of sites sequenced and μ is the per-site mutation probability(Bozic, Gerold, and Nowak 2016).(*Ryan, this seems like a wierd citation for this although it is where I first understood it. Is it the kind of thing that needs a citation?*). The estimated local nonnull probability is then

$$\hat{p}_1 = \frac{N(f)}{N} = \frac{\mu}{f}$$

Variant allele frequency distribution

We simulated tumors with three different allele frequency distributions. For whole genome simulations with depth 100X we generated variant allele frequencies from a Beta(1,6) distribution, where 20% of variants have frequency between .017 and .057 and 50% are less than .1. Whole exome simulations at 500X depth were generated from a Beta(2,40) distribution where 20% of variants have frequency between .01 and .025 and 50% are less than .05.

Simulated tumors spectra

We simulated tumors with three different mutation spectra. Each is an equal mixture of three COSMIC signatures as described in Ludmil B Alexandrov et al. (2015) and downloaded from https://cancer.sanger.ac.uk/cosmic/signatures. We used mutation signatures 1, 7, and 11 to represent a highly concentrated mutation signature, signatures 1, 4, and 5 to represent intermediate concentration, and 1,3, and 5 to represent a diffuse mutation signature. - We selected mutations according to these signatures from a set of previously reported cancer mutations derived from the combined TCGA and PCAWG databases.

Simulated bam files

We simulated 100X whole genome and 500X exome normal reads from the GRCH38 reference genome with VarSim/art (Mu et al. 2015), and aligned them to GRch38 with BWA (H. Li and Durbin 2009), both with default parameters. Variants were spiked 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 MuTect filters, pass all filters *other* than the evidence threshold `tlod_f_star`, or fail to pass any filter other than `tlod_f_star`. Variants that pass all filters or fail only `tlod_f_star` are then passed to {method} for prior estimation and rescoring.

Supplementary Figures

References

Alexandrov, L. B., Y. S. Ju, K. Haase, P. Van Loo, I. Martincorena, S. Nik-Zainal, Y. Totoki, et al. 2016. "Mutational signatures associated with tobacco smoking in human cancer." *Science* 354 (6312): 618–22. doi:10.1126/science.aag0299.

Alexandrov, Ludmil B, Philip H Jones, David C Wedge, Julian E Sale, Peter J Campbell, Serena Nik-Zainal, and Michael R Stratton. 2015. "Clock-like mutational processes in human somatic cells." *Nature Genetics* 47 (12). Nature Publishing Group: 1402–7. doi:10.1038/ng.3441.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Peter J. Campbell, and Michael R. Stratton. 2013. "Deciphering Signatures of Mutational Processes Operative in Human Cancer." *Cell Reports* 3 (1). Cell Press: 246–59. doi:10.1016/j.celrep.2012.12.008.

Borad, Mitesh J., Mia D. Champion, Jan B. Egan, Winnie S. Liang, Rafael Fonseca, Alan H. Bryce, Ann E. McCullough, et al. 2014. "Integrated Genomic Characterization Reveals Novel, Therapeutically Relevant Drug Targets in FGFR and EGFR Pathways in Sporadic Intrahepatic Cholangiocarcinoma." *PLoS Genetics* 10 (2). doi:10.1371/journal.pgen.1004135.

Bozic, Ivana, Jeffrey M. Gerold, and Martin A. Nowak. 2016. "Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution." *PLoS Computational Biology* 12 (2): e1004731. doi:10.1371/journal.pcbi.1004731.

Chen, Xiang, Elizabeth Stewart, Anang A. Shelat, Chunxu Qu, Armita Bahrami, Mark Hatley, Gang Wu, et al. 2013. "Targeting Oxidative Stress in Embryonal Rhabdomyosarcoma." *Cancer Cell* 24 (6). Elsevier Inc.: 710–24. doi:10.1016/j.ccr.2013.11.002.

Cibulskis, Kristian, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. 2013. "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples." *Nature Biotechnology* 31 (3). Nature Publishing Group: 213–19. doi:10.1038/nbt.2514.

Ding, J., A. Bashashati, A. Roth, A. Oloumi, K. Tse, T. Zeng, G. Haffari, et al. 2012. "Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data." *Bioinformatics* 28 (2): 167–75. doi:10.1093/bioinformatics/btr629.

Dorri, Fatemeh, Sean Jewell, Alexandre Bouchard-Côté, and Sohrab P. Shah. 2019. "Somatic mutation detection and classification through probabilistic integration of clonal population information." *Communications Biology* 2 (1). Nature Publishing Group: 44. doi:10.1038/s42003-019-0291-z.

Ewing, Adam D, Kathleen E Houlahan, Yin Hu, Kyle Ellrott, Cristian Caloian, Takafumi N Yamaguchi, J Christopher Bare, et al. 2015. "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection." *Nature Methods* 12 (7). Nature Publishing Group: 623–30. doi:10.1038/nmeth.3407.

Fan, Yu, Liu Xi, Daniel S.T. Hughes, Jianjun Zhang, Jianhua Zhang, P. Andrew Futreal, David A. Wheeler, and Wenyi Wang. 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 (1). Genome Biology: 178. doi:10.1186/s13059-016-1029-6.

Findlay, John M, Francesc Castro-Giner, Seiko Makino, Emily Rayner, Christiana Kartsonaki, William Cross,

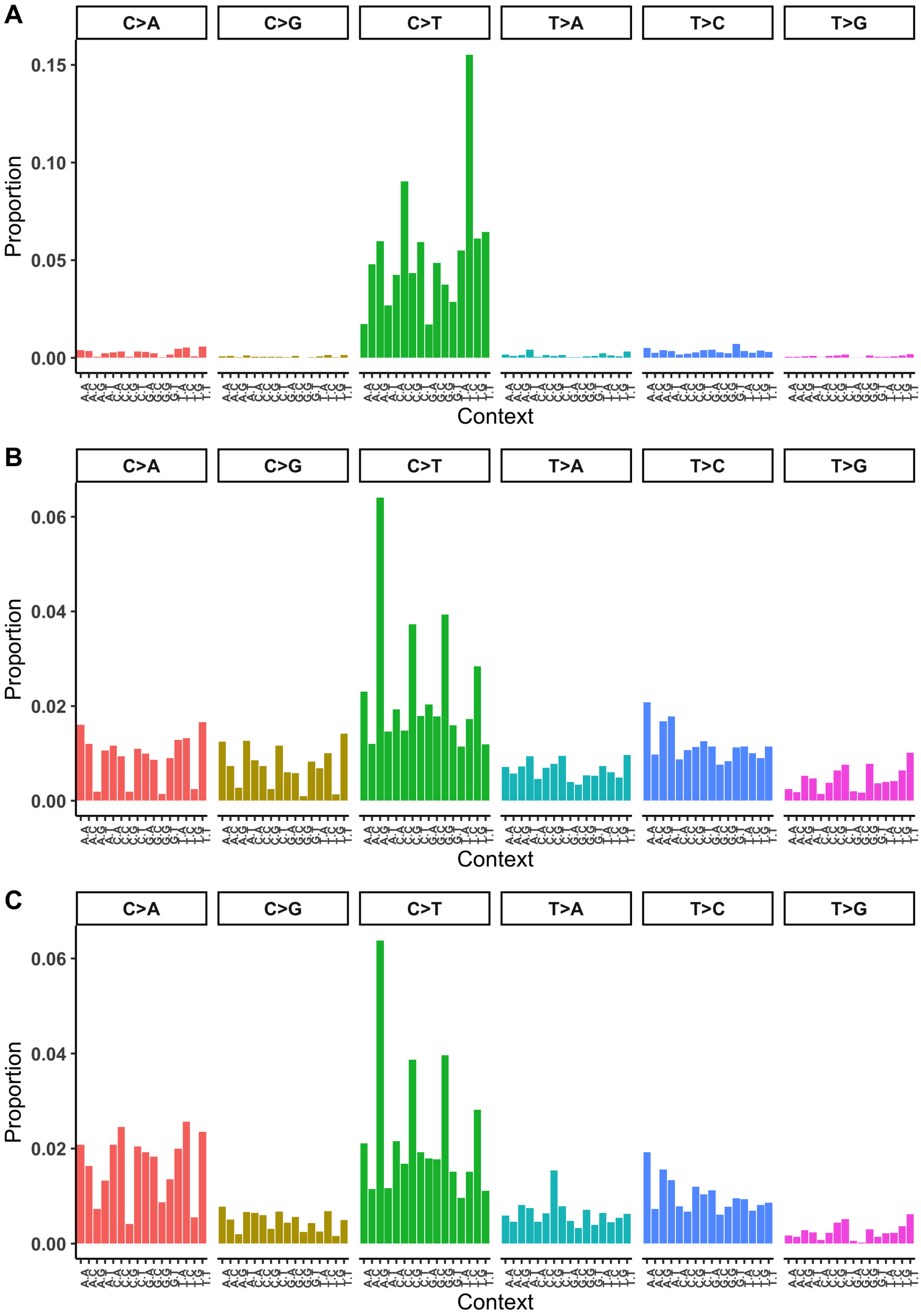

Figure 7: The signatures used to simulate both whole genome and whole exomes A) Equal combination of COSMIC signatures 1, 7, and 11 representing a highly concentrated signature of the type that might be observed in a melanoma. B) Equal combination of COSMIC signatures 1, 3, and 5 representing an intermediate level of concentration typical of a breast tumor. C) Equal combination of COSMIC signatures 1, 4, and 5 representing a diffuse signature typical of a lung tumor.

Michal Kovac, et al. 2016. “Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy.” *Nature Communications* 7. doi:10.1038/ncomms11111.

Garrison, Erik, and Gabor Marth. 2012. “Haplotype-based variant detection from short-read sequencing,” July. http://arxiv.org/abs/1207.3907.

Griffith, Malachi, Christopher A. Miller, Obi L. Griffith, Kilannin Krysiak, Zachary L. Skidmore, Avinash Ramu, Jason R. Walker, et al. 2015. “Optimizing Cancer Genome Sequencing and Analysis.” *Cell Systems* 1 (3). Elsevier Inc.: 210–23. doi:10.1016/j.cels.2015.08.015.

Helleday, Thomas, Saeed Eshtad, and Serena Nik-Zainal. 2014. “Mechanisms underlying mutational signatures in human cancers.” *Nature Reviews Genetics* 15 (9). Nature Publishing Group: 585–98. doi:10.1038/nrg3729.

Jones, David, Keiran M. Raine, Helen Davies, Patrick S. Tarpey, Adam P. Butler, Jon W. Teague, Serena Nik-Zainal, and Peter J. Campbell. 2016. “cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data.” *Current Protocols in Bioinformatics* 56 (1). John Wiley & Sons, Ltd: 15.10.1–15.10.18. doi:10.1002/cpbi.20.

Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, et al. 2013. “Mutational landscape and significance across 12 major cancer types.” *Nature* 502 (7471). Nature Research: 333–39. doi:10.1038/nature12634.

Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. 2012. “VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing.” *Genome Research* 22 (3): 568–76. doi:10.1101/gr.129684.111.

Larson, David E, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. 2012. “SomaticSniper: identification of somatic point mutations in whole genome sequencing data.” *Bioinformatics (Oxford, England)* 28 (3). Oxford University Press: 311–7. doi:10.1093/bioinformatics/btr665.

Lee-Six, Henry, Nina Friesgaard Øbro, Mairi S. Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J. Osborne, et al. 2018. “Population dynamics of normal human blood inferred from somatic mutations.” *Nature* 561 (7724). Nature Publishing Group: 473–78. doi:10.1038/s41586-018-0497-0.

Li, H., and R. Durbin. 2009. “Fast and accurate short read alignment with Burrows-Wheeler transform.” *Bioinformatics* 25 (14). Oxford University Press: 1754–60. doi:10.1093/bioinformatics/btp324.

Mardis, Elaine R. 2012. “Applying next-generation sequencing to pancreatic cancer treatment.” *Nature Reviews Gastroenterology & Hepatology* 9 (8). Nature Publishing Group: 477–86. doi:10.1038/nrgastro.2012.126.

Mu, J. C., M. Mohiyuddin, J. Li, N. Bani Asadi, M. B. Gerstein, A. Abyzov, W. H. Wong, and H. Y. K. Lam. 2015. “VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications.” *Bioinformatics* 31 (9). Oxford University Press: 1469–71. doi:10.1093/bioinformatics/btu828.

Nik-Zainal, Serena, Ludmil B. Alexandrov, David C. Wedge, Peter Van Loo, Christopher D. Greenman, Keiran Raine, David Jones, et al. 2012. “Mutational Processes Molding the Genomes of 21 Breast Cancers.” *Cell* 149 (5): 979–93. doi:10.1016/j.cell.2012.04.024.

Nik-Zainal, Serena, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, et al. 2016. “Landscape of somatic mutations in 560 breast cancer whole-genome sequences.” *Nature* 534 (7605). Nature Publishing Group: 47–54. doi:10.1038/nature17676.

Rahbari, Raheleh, Arthur Wuster, Sarah J Lindsay, Robert J Hardwick, Ludmil B Alexandrov, Saeed Al Turki, Anna Dominiczak, et al. 2016. “Timing, rates and spectra of human germline mutation.” *Nature Genetics* 48 (2). Nature Publishing Group: 126–33. doi:10.1038/ng.3469.

Saunders, Christopher T, Wendy S W Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. 2012. “Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs." *Bioinformatics (Oxford, England)* 28 (14). Oxford University Press: 1811–7. doi:10.1093/bioinformatics/bts271.

Shi, Weiwei, Charlotte K Y Ng, Raymond S Lim, Lajos Pusztai, Jorge S Reis-Filho, Christos Hatzis, Tingting Jiang, 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: 1446–57. doi:10.1016/j.celrep.2018.10.046.

Williams, Marc J, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. 2016. "Identification of neutral tumor evolution across cancer types." *Nature Genetics* 48 (3). Nature Research: 238–44. doi:10.1038/ng.3489.

Williams, Marc J, Benjamin Werner, Timon Heide, Christina Curtis, Chris P Barnes, Andrea Sottoriva, and Trevor A Graham. 2018. "Quantification of subclonal selection in cancer from bulk sequencing data." *Nature Genetics* 50 (June). Springer US: 895–903. doi:10.1038/s41588-018-0128-6.

Wilm, Andreas, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. 2012. "LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets." *Nucleic Acids Research* 40 (22): 11189–11201. doi:10.1093/nar/gks918.