Published online 6 February 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 1 **1**
doi: 10.1093/nargab/lqaa004

BATCAVE: calling somatic mutations with a tumor- and site-specific prior

Brian K. Mannakee[1] and Ryan N. Gutenkunst[2,*]

[1]Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ 85721, USA and [2]Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721, USA

Received October 08, 2019; Revised January 13, 2020; Editorial Decision January 19, 2020; Accepted January 23, 2020

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.

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 (1–4). Accurate variant calling is also critical for optimizing the treatment of individual patients' disease (5–9). 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 (10) and Varscan2 (11), 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 (12), joint genotype analysis (SomaticSniper (13), JointSNVMix2 (14), Seurat (15), CaVEMan (16) and MuClone (17)), allele frequency-based analysis (Strelka2 (18), LoFreq (19), EBCall (20), deepSNV (21), LoLoPicker (22), and MuSE (23)), and ensemble and deep learning methods (MutationSeq (5), BAYSIC (24), SomaticSeq (25) and SNooPer (26)). 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 (27–29). 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 (30–34). The specific mutational processes active in a particular tumor generate its unique mutation profile, and differences within and between tumor types are pronounced (35–39). For example, the mutation profiles differ substantially among the three breast tumors illustrated in Figure 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,

*To whom correspondence should be addressed. Email: rgutenk@email.arizona.edu

© The Author(s) 2020. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nargab/article-abstract/2/1/lqaa004/5728473 by University of Arizona Library user on 06 February 2020

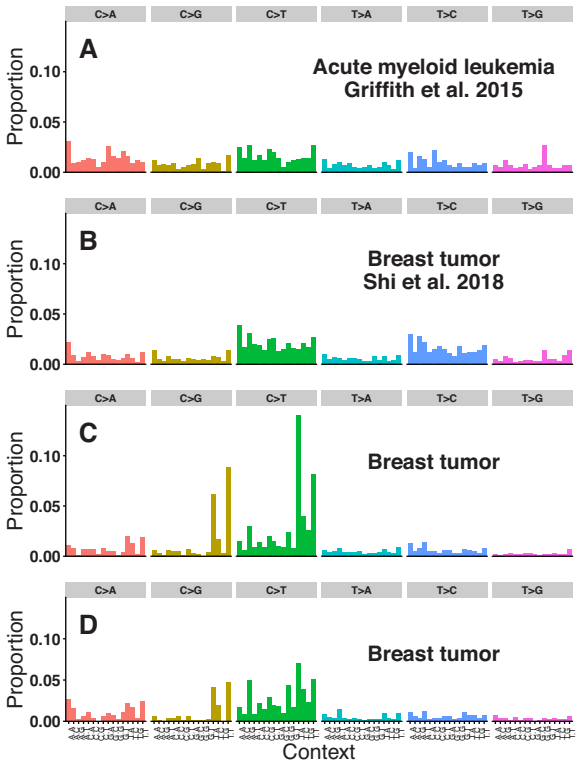

Figure 1. Real tumor mutation profiles. In each panel, the x-axis corresponds to each of the 96 possible mutation types, and the y-axis is the proportion of total mutations of each type. (**A**) The observed mutation profile of an acute myeloid leukemia used in our real data analysis (52). (**B**) The observed mutation profile of a breast tumor used in our real data analysis (4). (**C** and **D**) The observed mutation profiles of two additional breast tumors (53).

`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.

MATERIALS AND METHODS

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

$$\mathbf{H_0}: \quad \text{Alt allele} = m; \quad \nu = 0 \tag{1}$$

$$\mathbf{H_1}: \quad \text{Alt allele} = m; \quad \nu = \hat{f}. \tag{2}$$

Here m is any of the three 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

$$\mathrm{P}(m, \nu) = \mathrm{p}(m, \nu) \cdot \prod_{i=1}^{\mathrm{d}} \mathrm{f}_{m,\nu}(x_i), \tag{3}$$

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, equation (3) becomes

$$\mathrm{P}(m, \nu) = \mathrm{p}(m) \cdot \prod_{i=1}^{\mathrm{d}} \mathrm{f}_{m,\nu}(x_i). \tag{4}$$

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

Site-specific prior probability of mutation

The probability that we have denoted $\mathrm{p}(m)$ in equation (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 $\mathrm{p}(m, M|C)$ (40). Assuming that m and M are independent conditional on the genomic context, $\mathrm{p}(m, M|C) = \mathrm{p}(m|C)\mathrm{p}(M|C)$, which we can use Bayes' theorem to further decompose as

$$\mathrm{p}(m, M \mid C) = \mathrm{p}(m \mid C)\mathrm{p}(C \mid M)\frac{\mathrm{p}(M)}{\mathrm{p}(C)}. \tag{5}$$

We next show how to estimate the quantities in equation (5).

Estimation of the mutation profile

Many aspects of genomic architecture can affect the somatic mutation rate at multiple scales (40). Here we focus on a small-scale feature, the trinucleotide context, which is known to strongly affect the prior probability of single-nucleotide mutation (27–29). The trinucleotide context of a genomic site consists of the identity of the reference base and the 3' and 5' flanking bases. There are six classes of base substitution: C>A, C>G, C>T, T>A, T>C, T>G (all referred to by the pyrimidine of the mutated Watson-Crick pair). This results in a total of 32 ($4 \times 2 \times 4$) reference contexts and 3 alternate bases per reference base. Indexing by the $c = \{1 \ldots 32\}$ contexts and by the $m = \{1 \ldots 3\}$ alternate bases, we have 96 possible substitution types $S_{m,c}$. equation (5) is then

$$\mathrm{p}(\mathrm{S}_{m,c}) = \mathrm{p}(m \mid C = c)\mathrm{p}(C = c \mid M)\frac{\mathrm{p}(M)}{\mathrm{p}(C = c)}. \tag{6}$$

The first two terms on the right-hand side can be estimated from the observed mutation profile (Figure 1).

Downloaded from https://academic.oup.com/nargab/article-abstract/2/1/lqaa004/5728473 by University of Arizona Library user on 06 February 2020

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 $\boldsymbol{\alpha}$,

$$\begin{aligned} \boldsymbol{\pi} \mid \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \text{S} \mid \boldsymbol{\pi} &\sim \text{Multinomial}(\boldsymbol{\pi}). \end{aligned} \tag{7}$$

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 >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

$$\alpha'_{m,c} = \alpha_{m,c} + \sum_{\text{s}\in\text{D}} \text{I}\{\text{s} = \text{s}_{m,c}\}, \tag{8}$$

and$I$ is the indicator function. Returning to Equation (6), given that a mutation has occurred, the posterior probability it occurred in context c is

$$\text{p}(C = c \mid M, \text{D}) = \frac{\sum_m \alpha'_{m,C=c}}{\sum_{m,c} \alpha'_{m,c}}. \tag{9}$$

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

$$\text{p}(m \mid C = c, \text{D}) = \frac{\alpha'_{m,C=c}}{\sum_m \alpha'_{m,C=c}}. \tag{10}$$

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.

Estimation of the mutation rate

The final piece of Equation (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 (3) show that the expected total number of mutations M_{tot} between two allele frequencies $(f_{\min}, f_{\max})$ is

$$\text{M}_{\text{tot}}(f_{\min}, f_{\max}) = N\frac{\mu}{\beta}\left(\frac{1}{f_{\min}} - \frac{1}{f_{\max}}\right). \tag{11}$$

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 $f_{\max}$. Sample contamination by normal cells can result in underestimation of tumor mutation allele frequencies. To account for this, our implementation of BATCAVE takes as input an estimated sample purity p_{ur}. All estimated allele frequencies are then multiplied by a factor of $1/p_{ur}$ before being used in equation (11). 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, $f_{\min}$ and p_{ur} are free parameters. For this paper, we set $f_{\min} = 0.05$, because we are working at high depth. For our simulated data analyses, we set $p_{ur} = 1$, and for our real data analyses, we set p_{ur} to the estimated purity.

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 (Equation 1) as TLOD (MuTect1) or t_lod_fstar (MuTect2). We used MuTect 1.1.7 for all analyses in this paper, so we have

$$\text{TLOD} = \log_{10}\left(\frac{\prod_{i=1}^{\text{d}} \text{f}_{m,\nu=\hat{f}}(x_i)}{\prod_{i=1}^{\text{d}} \text{f}_{m,\nu=0}(x_i)}\right). \tag{12}$$

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

$$\frac{\text{P}(m, \nu = \hat{f})}{1 - \text{P}(m, \nu = \hat{f})} = 10^{\text{TLOD}+\text{logit}_{10}(\text{p}(\text{S}_{m,c}))}. \tag{13}$$

Here $\text{p}(\text{S}_{m,c})$ is the prior probability of a substitution of type $\text{S}_{m,c}$, as described in equation (6) and specified in equations (9)–(11). When comparing our posterior odds to those of MuTect, we assume a uniform per-base probability of mutation of $3 \cdot 10^{-6}$ (10), so

$$\frac{\text{P}_{\text{MuTect}}(m, \nu = \hat{f})}{1 - \text{P}_{\text{MuTect}}(m, \nu = \hat{f})} = 10^{\text{TLOD}-6}. \tag{14}$$

Implementation

We have implemented the BATCAVE algorithm as an R package called `batcaver`. The package leverages the Bioconductor packages BSgenome (41), GenomicAlignments (42), VariantAnnotation (43), and SomaticSignatures (44) 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.

An assumption of the BATCAVE algorithm is that the mutation profile of low-confidence mutations is similar to that of the high-confidence mutations from which the prior is inferred. To test this assumption, `batcaver` performs a statistical hypothesis test comparing these two mutation profiles (45). If the two profiles differ at a significance

Downloaded from https://academic.oup.com/nargab/article-abstract/2/1/lqaa004/5728473 by University of Arizona Library user on 06 February 2020

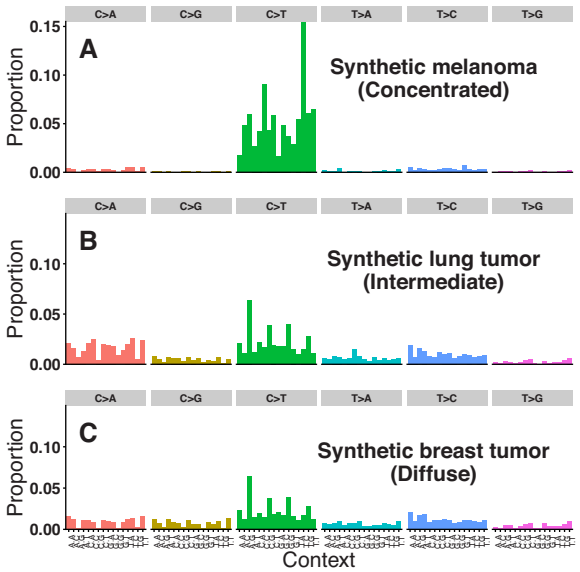

Figure 2. Simulated tumor mutation profiles. As in Figure 1, in each panel the *x*-axis corresponds to each of the 96 possible mutation types, and the *y*-axis is the proportion of total mutations of each type. (**A**) A mutation profile used for simulating tumors, made up of equal proportions of COSMIC mutation signatures 1, 7 and 11. (**B**) Equal proportions of signatures 1, 4 and 5. (**C**) Equal proportions of signatures 1, 3 and 5.

threshold of $\alpha = 0.05$, `batcaver` issues a warning and outputs the profiles.

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) (46). Each simulated profile includes COSMIC signature 1, which is found in nearly all tumors and is associated with spontaneous cytosine deamination. The 'Concentrated' profile (Figure 2A) is an equal combination of COSMIC signatures 1, 7,and 11, which has a large percentage of C > T substitutions such as are often seen in cancers caused by UV exposure (47). The 'Intermediate' profile (Figure 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 (47). The 'Diffuse' profile (Figure 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 (32). 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 (100× depth) and whole exome (500X depth) reads were simulated from the GRCh38 reference genome using VarSim (48) and aligned with BWA (49), both with default parameters. Variants were inserted to create tumors with BAMSurgeon with default parameters (50) and called with MuTect 1.1.7 (10) 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 labeled 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.

Calibration metric

To quantify the difference in calibration between MuTect and BATCAVE, we used the Integrated Calibration Index (51). 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.

Real data

We analyzed two real datasets, one from an acute myeloid leukemia (AML) (52) and one from a multi-region sequencing experiment in breast cancer (4). 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 (52), 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. Griffith *et al.* estimated the purity of their primary sample to be 90.7%, and we used this value for tumor purity in `batcaver`. Shi *et al.* generated a gold set of variant calls for each tumor region sequenced (4), which we used for our true positive dataset. For these multi-region data, we analyzed the three biological replicates from each of six tumors (Table 1), running BATCAVE separately on each of the eighteen samples, and we combined results to generate precision-recall curves. Shi *et al.* estimated sample purities of between 26 and 80%, and we used these estimates in `batcaver`. We called variants using Mutect 1.1.7 as in our simulations, except that both these datasets were originally aligned to GRCh37, so we used that reference.

RESULTS

We implemented BATCAVE as a post-call variant evaluation algorithm to be used with MuTect (Versions 1.1.7 or

Downloaded from https://academic.oup.com/nargab/article-abstract/2/1/lqaa004/5728473 by University of Arizona Library user on 06 February 2020

Table 1. Variant calling metrics for all datasets

			AUROC		AUPRC		ICI	
Scenario	Mutation profile	μ (estimated)	MuTect	BATCAVE	MuTect	BATCAVE	MuTect	BATCAVE
100× whole genome	Concentrated	3.6e-7	0.987	0.993	0.972	0.975	0.117	0.287
100× whole genome	Intermediate	3.2e-7	0.987	0.989	0.972	0.973	0.118	0.214
100× whole genome	Diffuse	3.2e-7	0.988	0.989	0.971	0.973	0.120	0.219
500× whole exome	Concentrated	3.6e-7	0.848	0.929	0.674	0.758	0.138	0.109
500× whole exome	Intermediate	3.6e-7	0.847	0.881	0.677	0.706	0.108	0.112
500× whole exome	Diffuse	3.6e-7	0.850	0.873	0.676	0.698	0.105	0.116
real AML (52)	Actual	3.6e-8			0.996	0.988		
real breast (4)	Actual	6.6e-7			0.972	0.968		

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.

>2.0) (10). 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 s to process a 500× exome and 2000 s for a 100× 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) (Figure 2). We also tested BATCAVE using two real cancer datasets, a whole-genome Acute Myeloid Leukemia (AML) (52) and a whole-exome multi-region breast cancer (4). 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.

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 (Figure 3), and the `batcaver` package emits a warning if there are fewer than 500 high-confidence mutations. For comparison, in our simulated datasets the number of high-confidence mutations ranged between 1500 and 5000, and in the real AML we test on it is over 17 000 (52).

We assessed classification performance using the areas under both the receiver operating characteristic and the precision-recall curves, because the classes are unbalanced (∼5 to 1 ratio of false to true variants in our simulated data). By both metrics BATCAVE outperforms MuTect (Figure 4A-B, Supplementary Figure S1A-B and Table 1).

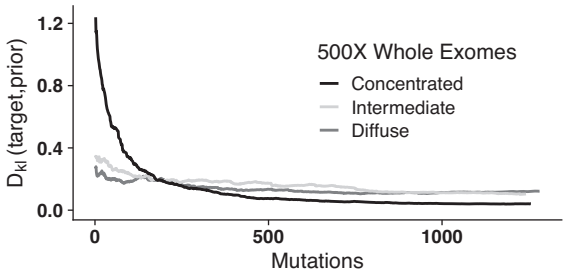

Figure 3. Convergence of the mutational prior to the data generating distribution. Plotted is the Kullback–Leibler divergence between the simulated and estimated profiles versus number of incorporated mutations for whole exomes. Convergence for whole genomes is similar.

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 $\sim 3 \cdot 10^{-7}$ (Table 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) (51), 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 (Figure 4C-D and Supplementary Figure S1C-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 >0.01. For these potential variants, BATCAVE tends to increase posterior probabilities of low probability but true variants (density curves in Figures 4C-D and Supplementary Figure S1C-D) while decreasing probabilities of low probability but false variants. For 500× exomes, the calibration of BATCAVE is better than MuTect across

Downloaded from https://academic.oup.com/nargab/article-abstract/2/1/lqaa004/5728473 by University of Arizona Library user on 06 February 2020

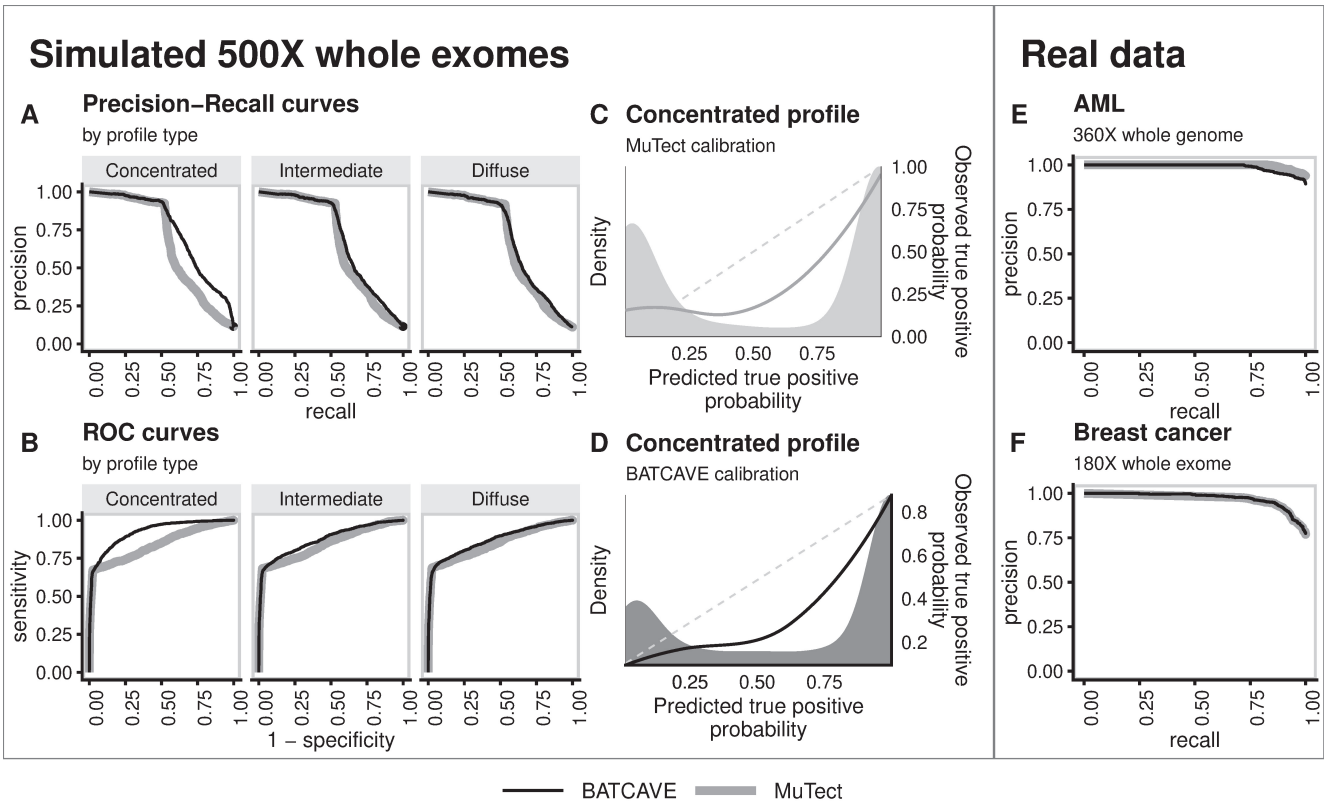

Figure 4. Variant-calling performance on simulated and real data. Throughout, MuTect results are plotted with gray lines and BATCAVE results with black lines. (**A**) Precision-recall curves and (**B**) receiver operating characteristic curves for different mutation profiles. (**C** and **D**) Calibration 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 (4,52).

the full spectrum of posterior probabilities (Figure 4 and Table 1). For 100× whole genomes, the calibration of BATCAVE is slightly worse (Supplementary Figure S1 and Table 1), likely because there are few low probability true positive variants in tumors sequenced to 100× 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 (Figure 5 and Supplementary Figure S2). For this range of thresholds, however, the posterior probabilities from MuTect poorly predict precision (Figure 5 and Supplementary Figure S2). For any posterior probability threshold above 70%, MuTect 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.

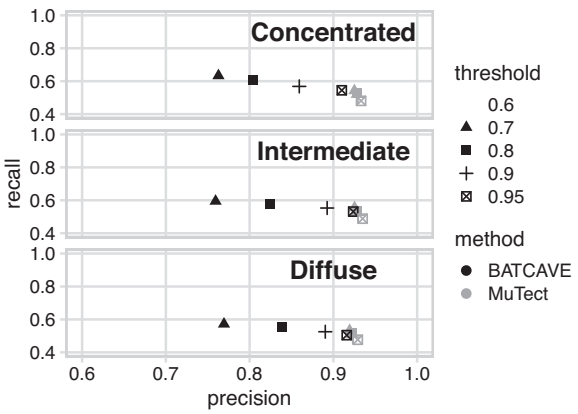

Figure 5. Posterior probability calibration for realistic calling thresholds, for 500× exomes. Plotted is precision and recall for variants identified using various realistic posterior probability thresholds. At these thresholds, the precision of BATCAVE is much closer to the given threshold than MuTect, no matter the concentration of the mutation profile.

Tests using real tumor data

We tested BATCAVE using two datasets for which deep sequencing and variant validation were performed with the express purpose of evaluating tumor variant calling

Downloaded from https://academic.oup.com/nargab/article-abstract/2/1/lqaa004/5728473 by University of Arizona Library user on 06 February 2020

pipelines, yielding high quality true and false positive data (4,52). 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 >360× and used targeted sequencing to validate nearly 200 000 mutations (52). We estimated a per-base mutation rate for this tumor of 3.6×10^{-8}, which is consistent with previous estimates of AML mutation rates (3,52). For this tumor, the mutation profiles of high- and low-confidence mutations differ ($p \sim 10^{-12}$). The low-confidence mutation profile contains additional spikes of mutations corresponding to five substitution types (Supplementary Figure S3). Nevertheless, for both MuTect and BATCAVE, the precision-recall curve is almost perfect for the validated variants (0.988 and 0.996 area under the curve) (Figure 4E and Table 1).

Shi *et al*. performed multi-region whole exome sequencing on six individual breast tumors to a mean target sequencing depth of 160× and validated all variants identified by three different variant calling pipelines (4). We estimated an average per-base mutation rate for these tumor regions of 6.6×10^{-7}, which is consistent with observed mutation rates for breast cancers (53) and with the low number of validated somatic mutations. For the validated variants, MuTect and BATCAVE yielded almost identical precision-recall curves (Figure 4F and Table 1)

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 (Figure 4A–D, Supplementary Figure S1 and Table 1) across the entire range of recall and specificity. Moreover, BATCAVE is better calibrated than MuTect at relevant posterior probability thresholds (Figure 5 and Supplementary Figure S2), 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 (Figure 4E-F and Table 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 profiles and two real tumors. The simulated diffuse and intermediate profiles (Figure 2A and B) represent baseline profiles of lung and breast tumors, respectively. And the concentrated profile (Figure 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. Figure 1C). The two real datasets we considered are among the few for which extensive validation of variant calls has been performed (4,52). They happen, however, to have diffuse mutation profiles (Figure 1A and B), which reduces the expected advantage of BATCAVE over MuTect (Table 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 (Figure 4 and Table 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 (54). 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 (55). Well-calibrated posterior probabilities facilitate the use of probabilistic risk models in the choice of treatment (56), 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% (57). Achieving this goal while minimizing cost demands well-calibrated posterior probabilities.

Our current implementation of BATCAVE is as a post-calling algorithm for MuTect, 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 (10,52). 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 (18). 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.

Following MuTect (10), BATCAVE assumes a uniform prior on the variant allele frequency. In general, low-frequency variants are expected to be more common than high-frequency variants. Nevertheless, the uniform frequency prior is appropriate when the goal is to make the best call possible at each potential variant site. By favoring low-frequency and thus typically low-evidence variants, a nonuniform prior would increase the false positive rate. Note however, that if the goal were not to best identify individual variants, but rather to best estimate the distribution of variant allele frequencies within the tumor, then a nonuniform frequency prior would be appropriate (58).

To estimate the tumor mutation rate, BATCAVE uses a model of the distribution of variant allele frequencies generated during exponential growth (equation 11). Mutations that change copy number might distort the empirical allele frequency distribution, biasing the estimate. But substantially biasing the estimate would require copy number changes in a large number of genomic regions. Nevertheless, a potential extension of the BATCAVE algorithm would be to incorporate external copy number calls in order

Downloaded from https://academic.oup.com/nargab/article-abstract/2/1/lqaa004/5728473 by University of Arizona Library user on 06 February 2020

to adjust allele frequencies before estimating the mutation rate.

Sequencing multiple regions of the same tumor is of growing interest for understanding tumor heterogeneity and evolution (59). In fact, the breast cancer data we analyze came from a multi-region study (4). We called variants independently across samples, because MuTect only considers a single sample at a time. But calling variants jointly across samples could substantially increase sensitivity. In the future, the BATCAVE algorithm could be integrated into a multi-sample caller such as MultiSNV (60). In that case, an optimal implementation would likely exploit the relationship between mutational priors in related samples.

The BATCAVE algorithm is computationally inexpensive; our current implementation adds 1 s 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 (61,62). There are, however, many other aspects of genomic context that can affect local mutation rates (40), including replication timing (63), expression level (64) and chromatin organization (65). 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 (66). 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 (67), in which the high and low-confidence variants are treated as parallel experiments (68,69). 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 BATCAVE 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 (70), and although in general those differences are relatively small, they can be significant in some cases. Our `batcaver` software warns the user if the high- and low-confidence mutational profiles are statistically different. It also outputs those profiles, so the user can assess whether the difference will substantially affect their results. 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 (Figure 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.

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

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

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.

FUNDING

National Science Foundation via Graduate Research Fellowship Award [DGE-1143953 to B.K.M.]; National Institute of General Medical Sciences of the National Institutes of Health [R01GM127348 to R.N.G.].

Conflict of interest statement. None declared.

REFERENCES

1. Williams,M.J., Werner,B., Barnes,C.P., Graham,T.A. and Sottoriva,A. (2016) Identification of neutral tumor evolution across cancer types. *Nat. Genet.*, **48**, 238–244.
2. Bozic,I., Gerold,J.M. and Nowak,M.A. (2016) Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS Comput. Biol.*, **12**, e1004731.
3. Williams,M.J., Werner,B., Heide,T., Curtis,C., Barnes,C.P., Sottoriva,A. and Graham,T.A. (2018) Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.*, **50**, 895–903.
4. Shi,W., Ng,C.K.Y., Lim,R.S., Pusztai,L., Reis-Filho,J.S., Hatzis,C., Jiang,T., Kumar,S., Li,X., Wali,V.B. *et al.* (2018) Reliability of Whole-Exome sequencing for assessing intratumor genetic heterogeneity. *Cell Rep.*, **25**, 1446–1457.

Downloaded from https://academic.oup.com/nargab/article-abstract/2/1/lqaa004/5728473 by University of Arizona Library user on 06 February 2020

5. Ding,J., Bashashati,A., Roth,A., Oloumi,A., Tse,K., Zeng,T., Haffari,G., Hirst,M., Marra,M.A., Condon,A. *et al.* (2012) Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics*, **28**, 167–175.
6. Mardis,E.R. (2012) Applying next-generation sequencing to pancreatic cancer treatment. *Nat. Rev. Gastroenterol. Hepatol.*, **9**, 477–486.
7. Chen,X., Stewart,E., Shelat,A.A., Qu,C., Bahrami,A., Hatley,M., Wu,G., Bradley,C., McEvoy,J., Pappo,A. *et al.* (2013) Targeting oxidative stress in embryonal rhabdomyosarcoma. *Cancer Cell*, **24**, 710–724.
8. Borad,M.J., Champion,M.D., Egan,J.B., Liang,W.S., Fonseca,R., Bryce,A.H., McCullough,A.E., Barrett,M.T., Hunt,K., Patel,M.D. *et al.* (2014) Integrated genomic characterization reveals novel, therapeutically relevant drug targets in FGFR and EGFR pathways in sporadic intrahepatic cholangiocarcinoma. *PLoS Genet.*, **10**, e1004135.
9. Findlay,J.M., Castro-Giner,F., Makino,S., Rayner,E., Kartsonaki,C., Cross,W., Kovac,M., Ulahannan,D., Palles,C., Gillies,R.S. *et al.* (2016) Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. *Nat. Commun.*, **7**, 11111.
10. Cibulskis,K., Lawrence,M.S., Carter,S.L., Sivachenko,A., Jaffe,D., Sougnez,C., Gabriel,S., Meyerson,M., Lander,E.S. and Getz,G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.
11. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
12. Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. arXiv doi: https://aps.arxiv.org/abs/1207.3907v2, 20 July 2012, preprint: not peer reviewed.
13. Larson,D.E., Harris,C.C., Chen,K., Koboldt,D.C., Abbott,T.E., Dooling,D.J., Ley,T.J., Mardis,E.R., Wilson,R.K. and Ding,L. (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics (Oxford, England)*, **28**, 311–317.
14. Roth,A., Ding,J., Morin,R., Crisan,A., Ha,G., Giuliany,R., Bashashati,A., Hirst,M., Turashvili,G., Oloumi,A. *et al.* (2012) JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics (Oxford, England)*, **28**, 907–913.
15. Christoforides,A., Carpten,J.D., Weiss,G.J., Demeure,M.J., Von Hoff,D.D. and Craig,D.W. (2013) Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics*, **14**, 302.
16. Jones,D., Raine,K.M., Davies,H., Tarpey,P.S., Butler,A.P., Teague,J.W., Nik-Zainal,S. and Campbell,P.J. (2016) cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinform.*, **56**, 15.10.1–15.10.18.
17. Dorri,F., Jewell,S., Bouchard-Côté,A. and Shah,S.P. (2019) Somatic mutation detection and classification through probabilistic integration of clonal population information. *Commun. Biol.*, **2**, 44.
18. Kim,S., Scheffler,K., Halpern,A.L., Bekritsky,M.A., Noh,E., Källberg,M., Chen,X., Kim,Y., Beyter,D., Krusche,P. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.
19. Wilm,A., Aw,P.P.K., Bertrand,D., Yeo,G.H.T., Ong,S.H., Wong,C.H., Khor,C.C., Petric,R., Hibberd,M.L. and Nagarajan,N. (2012) LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.*, **40**, 11189–11201.
20. Shiraishi,Y., Sato,Y., Chiba,K., Okuno,Y., Nagata,Y., Yoshida,K., Shiba,N., Hayashi,Y., Kume,H., Homma,Y. *et al.* (2013) An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.*, **41**, e89.
21. Gerstung,M., Beisel,C., Rechsteiner,M., Wild,P., Schraml,P., Moch,H. and Beerenwinkel,N. (2012) Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.*, **3**, 811–818.
22. Carrot-Zhang,J. and Majewski,J. (2017) LoLoPicker: detecting low allelic-fraction variants from low-quality cancer samples. *Oncotarget*, **8**, 37032–37040.
23. Fan,Y., Xi,L., Hughes,D.S., Zhang,J., Zhang,J., Futreal,P.A., Wheeler,D.A. and Wang,W. (2016) MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.*, **17**, 178.
24. Cantarel,B.L., Weaver,D., McNeill,N., Zhang,J., Mackey,A.J. and Reese,J. (2014) BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics*, **15**, 104.
25. Fang,L.T., Afshar,P.T., Chhibber,A., Mohiyuddin,M., Fan,Y., Mu,J.C., Gibeling,G., Barr,S., Asadi,N.B., Gerstein,M.B. *et al.* (2015) An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol.*, **16**, 197.
26. Spinella,J.F., Mehanna,P., Vidal,R., Saillour,V., Cassart,P., Richer,C., Ouimet,M., Healy,J. and Sinnett,D. (2016) SNooPer: A machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics*, **17**, 912.
27. Nik-Zainal,S., Alexandrov,L.B., Wedge,D.C., Van Loo,P., Greenman,C.D., Raine,K., Jones,D., Hinton,J., Marshall,J., Stebbings,L.A. *et al.* (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell*, **149**, 979–993.
28. Alexandrov,L.B., Jones,P.H., Wedge,D.C., Sale,J.E., Campbell,P.J., Nik-Zainal,S. and Stratton,M.R. (2015) Clock-like mutational processes in human somatic cells. *Nat. Genet.*, **47**, 1402–1407.
29. Lee-Six,H., Øbro,N.F., Shepherd,M.S., Grossmann,S., Dawson,K., Belmonte,M., Osborne,R.J., Huntly,B. J.P., Martincorena,I., Anderson,E. *et al.* (2018) Population dynamics of normal human blood inferred from somatic mutations. *Nature*, **561**, 473–478.
30. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Campbell,P.J. and Stratton,M.R. (2013) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, **3**, 246–259.
31. Helleday,T., Eshtad,S. and Nik-Zainal,S. (2014) Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, **15**, 585–598.
32. Nik-Zainal,S., Davies,H., Staaf,J., Ramakrishna,M., Glodzik,D., Zou,X., Martincorena,I., Alexandrov,L.B., Martin,S., Wedge,D.C. *et al.* (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, **534**, 47–54.
33. Kandoth,C., McLellan,M.D., Vandin,F., Ye,K., Niu,B., Lu,C., Xie,M., Zhang,Q., McMichael,J.F., Wyczalkowski,M.A. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.
34. Alexandrov,L.B., Ju,Y.S., Haase,K., Van Loo,P., Martincorena,I., Nik-Zainal,S., Totoki,Y., Fujimoto,A., Nakagawa,H., Shibata,T. *et al.* (2016) Mutational signatures associated with tobacco smoking in human cancer. *Science*, **354**, 618–622.
35. Stephens,P., Edkins,S., Davies,H., Greenman,C., Cox,C., Hunter,C., Bignell,G., Teague,J., Smith,R., Stevens,C. *et al.* (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat. Genet.*, **37**, 590–592.
36. Burrell,R.A., McGranahan,N., Bartek,J. and Swanton,C. (2013) The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501**, 338–345.
37. Nakamura,H., Arai,Y., Totoki,Y., Shirota,T., Elzawahry,A., Kato,M., Hama,N., Hosoda,F., Urushidate,T., Ohashi,S. *et al.* (2015) Genomic spectra of biliary tract cancer. *Nat. Genet.*, **47**, 1003–1010.
38. Witkiewicz,A.K., McMillan,E.A., Balaji,U., Baek,G., Lin,W.-C., Mansour,J., Mollaee,M., Wagner,K.-U., Koduru,P., Yopp,A. *et al.* (2015) Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat. Commun.*, **6**, 6744.
39. Kumar,A., Coleman,I., Morrissey,C., Zhang,X., True,L.D., Gulati,R., Etzioni,R., Bolouri,H., Montgomery,B., White,T. *et al.* (2016) Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat. Med.*, **22**, 369–378.
40. Buisson,R., Langenbucher,A., Bowen,D., Kwan,E.E., Benes,C.H., Zou,L. and Lawrence,M.S. (2019) Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science (New York, N.Y.)*, **364**, eaaw2872.

Downloaded from https://academic.oup.com/nargab/article-abstract/2/1/lqaa004/5728473 by University of Arizona Library user on 06 February 2020

41. Pagès,H. (2019) BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. R package version 1.52.0.
42. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M. and Carey,V. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
43. Obenchain,V., Lawrence,M., Carey,V., Gogarten,S., Shannon,P. and Morgan,M. (2014) VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, **30**, 2076–2078.
44. Gehring,J.S., Fischer,B., Lawrence,M. and Huber,W. (2015) SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*, **31**, 3673–3675.
45. Plunkett,A. and Park,J. (2019) Two-sample test for sparse high-dimensional multinomial distributions. *TEST*, **28**, 804–826.
46. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E. *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **47**, D941–D947.
47. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S.A.J.R., Behjati,S., Biankin,A.V., Bignell,G.R., Bolli,N., Borg,A., Børresen-Dale,A.-L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.
48. Mu,J.C., Mohiyuddin,M., Li,J., Bani Asadi,N., Gerstein,M.B., Abyzov,A., Wong,W.H. and Lam,H.Y.K. (2015) VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*, **31**, 1469–1471.
49. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
50. Ewing,A.D., Houlahan,K.E., Hu,Y., Ellrott,K., Caloian,C., Yamaguchi,T.N., Bare,J.C., P'ng,C., Waggott,D., Sabelnykova,V.Y. *et al.* (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods*, **12**, 623–630.
51. Austin,P.C. and Steyerberg,E.W. (2019) The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stati. Med.*, **38**, 4051–4065.
52. Griffith,M., Miller,C.A., Griffith,O.L., Krysiak,K., Skidmore,Z.L., Ramu,A., Walker,J.R., Dang,H.X., Trani,L., Larson,D.E. *et al.* (2015) Optimizing cancer genome sequencing and analysis. *Cell Syst.*, **1**, 210–223.
53. Alexandrov,L.B., Kim,J., Haradhvala,N.J., Huang,M.N., Ng,A.W., Wu,Y., Boot,A., Covington,K.R., Gordenin,D.A., Bergstrom,E.N. *et al.* (2019) The Repertoire of Mutational Signatures in Human Cancer. bioRxiv doi: https://doi.org/10.1101/322859, 03 July 2019, preprint: not peer reviewed.
54. Barnell,E.K., Ronning,P., Campbell,K.M., Krysiak,K., Ainscough,B.J., Sheta,L.M., Pema,S.P., Schmidt,A.D., Richters,M., Cotto,K.C. *et al.* (2019) Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. *Genet. Med.*, **21**, 972–981.
55. Boutros,P.C. (2015) The path to routine use of genomic biomarkers in the cancer clinic. *Genome Res.*, **25**, 1508–1513.
56. Holmberg,L. and Vickers,A. (2013) Evaluation of prediction models for decision-making: beyond calibration and discrimination. *PLoS Med.*, **10**, e1001491.
57. International Cancer Genome Consortium (2008) *Goals, Structure, Policy & Guidelines*. https://icgc.org/icgc/goals-structure-policies-guidelines/e8-genome-analyses.
58. Nielsen,R., Korneliussen,T., Albrechtsen,A., Li,Y. and Wang,J. (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS One*, **7**, e37558.
59. McGranahan,N. and Swanton,C. (2017) Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, **168**, 613–628.
60. Josephidou,M., Lynch,A.G. and Tavaré,S. (2015) multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. *Nucleic Acids Res.*, **43**, e61.
61. Martincorena,I. and Campbell,P.J. (2015) Somatic mutation in cancer and normal cells. *Science (New York, N.Y.)*, **349**, 1483–1489.
62. Hollstein,M., Alexandrov,L.B., Wild,C.P., Ardin,M. and Zavadil,J. (2017) Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. *Oncogene*, **36**, 158–167.
63. Stamatoyannopoulos,J.A., Adzhubei,I., Thurman,R.E., Kryukov,G.V., Mirkin,S.M. and Sunyaev,S.R. (2009) Human mutation rate associated with DNA replication timing. *Nat. Genet.*, **41**, 393–395.
64. Pleasance,E.D., Cheetham,R.K., Stephens,P.J., McBride,D.J., Humphray,S.J., Greenman,C.D., Varela,I., Lin,M.-L., Ordóñez,G.R., Bignell,G.R. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.
65. Schuster-Böckler,B. and Lehner,B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, **488**, 504–507.
66. Hodgkinson,A. and Eyre-Walker,A. (2011) Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.*, **12**, 756–766.
67. Robbins,H. (1956) An empirical Bayes approach to statistics. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley. Vol. **1**, pp. 157–163.
68. Morris,C.N. (1983) Parametric empirical bayes Inference: theory and applications. *J. Am. Stat. Assoc.*, **78**, 47–55.
69. Efron,B. (2014) Two modeling strategies for empirical Bayes estimation. *Stat. Sci.*, **29**, 285–301.
70. Rubanova,Y., Shi,R., Harrigan,C.F., Li,R., Wintersinger,J., Sahin,N., Deshwar,A., Morris,Q. and PCAWG Evolution and Heterogeneity Working Group, PCAWG networkPCAWG Evolution and Heterogeneity Working Group, PCAWG network (2018) TrackSig: reconstructing evolutionary trajectories of mutations in cancer. bioRxiv doi: https://doi.org/10.1101/260471, 30 September 2019, preprint: not peer reviewed.

Downloaded from https://academic.oup.com/nargab/article-abstract/2/1/lqaa004/5728473 by University of Arizona Library user on 06 February 2020