BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–8

RVD2: An ultra-sensitive variant detection model for low-depth targeted next-generation sequencing data

Yuting He [1], Patrick Flaherty [1*]

[1]Department of Biomedical Engineering, Worcester Polytechnic Institute, Worcester, MA, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Next-generation sequencing technology is increasingly being used for clinical diagnostic tests. Unlike research cell lines, clinical samples are often genomically heterogeneous due to low sample purity or the presence of genetic subpopulations. However, many variant calling algorithms are optimized to call single nucleotide polymorphisms in homogeneous rather than heterogeneous samples.

Results: We present a novel variant calling algorithm that uses a hierarchical Bayesian model to estimate allele frequency and call variants in heterogeneous samples. We show that our algorithm improves upon current classifiers and has higher sensitivity and specificity over a wide range of median read depth and minor allele frequency. We identify six mutations in the PAXP1 gene in a matched clinical breast ductal carcinoma tumor sample; two of which are loss-of-heterozygosity events.

Availability: http://genomics.wpi.edu/rvd2/rvd2.html

Contact: pjflaherty@wpi.edu

1 INTRODUCTION

Next-generation sequencing (NGS) technology has enabled the systematic interrogation of the genome for a fraction of the cost of traditional assays (Koboldt *et al.*, 2013). Protocol and platform engineering improvements have enabled the generation of 1×10^9 bases of sequence data in 27 hours for approximately $1000 (Quail *et al.*, 2012). As a result, NGS is increasingly being used as a general platform for research assays for methylation state (Laird, 2010), DNA mutations (Consortium *et al.*, 2013), copy number variation (Alkan *et al.*, 2009), promoter occupancy (Ouyang *et al.*, 2009) and others (Rivera and Ren, 2013). NGS diagnostics are being translated to clinical applications including noninvasive fetal diagnostics (Kitzman *et al.*, 2012), infectious disease diagnostics (Capobianchi *et al.*, 2012), cancer diagnostics (Navin *et al.*, 2010), and human microbial analysis (Consortium, 2013).

Increasingly, NGS is being used to interrogate mutations in heterogeneous clinical samples. For example, NGS-based non-invasive fetal DNA testing uses maternal blood sample to sequence the minority fraction of cell-free fetal DNA (Fan *et al.*, 2008). Infectious diseases such as HIV and influenza may contain many genetically heterogeneous sub-populations (Flaherty *et al.*, 2011; Ghedin *et al.*, 2010). DNA sequencing of individual regions of a solid tumor has revealed genetic heterogeneous within an individual sample (Navin *et al.*, 2010).

However, the primary statistical tools for calling variants from NGS data are optimized for homogeneous samples. Samtools/bcftools and GATK uses naive Bayesian decision rule to call variants (Li, 2011; DePristo *et al.*, 2011). GATK involves more sophisticate pre- and post-processing steps wherein the genotype prior is fixed and constant across all loci and the likelihood of an allele at a locus is a function of the phred score (McKenna *et al.*, 2010).

Recently, researchers have developed algorithms to call low-frequency or rare variants in heterogeneous samples. VarScan2 combines algorithmic heuristics to call genotypes in the tumor and normal sample pileup data and then applies a Fisher's exact test on the read count data to detect a significant difference in the genotype calls (Koboldt *et al.*, 2012). Strelka uses a hierarchical Bayesian approach to model the joint distribution of the allele frequency in the tumor and normal samples at each locus (Saunders *et al.*, 2012). With the joint distribution available, one is able to identify locations with dissimilar allele frequencies. muTect uses a Bayesian posterior probability in its decision rule to evaluate the likelihood of a mutation (Cibulskis *et al.*, 2013). RVD uses a hierarchical Bayesian model to capture the error structure of the data and call variants (Flaherty *et al.*, 2011; Cushing *et al.*, 2013). However, that algorithm requires a very high read depth to estimate the sequencing error rate and call variants.

Several studies have compared the relative performance of these algorithms. Spencer *et al.* (2013) demonstrated that VarScan-somatic performed the best comparing with SAMtools, GATK and SPLINTER in detecting minor allele frequencies (MAFs) of 1% to 8%, with $>$500 coverage required for optimal performance. However, Spencer *et al.* (2013) also highlighted the fact that VarScan2 yielded more false positives at high read depth. Stead *et al.* (2013) showed that VarScan-somatic outperformed Strelka had performance on-par with muTect in detecting a 5% MAF for read depths between 100 and 1000.

The remainder of this article is organized as follows. In the next section we describe the statistical model structure of our new algorithm, RVD2. Then, we derive a sampling algorithm for computing the posterior distribution over latent variables in the model and use those samples in a Bayesian posterior distribution hypothesis test to call variants. We compare the performance of RVD2 to several other variant calling algorithms for a range of read depths and minor allele fractions. Finally, we show that RVD2 is able to call

*to whom correspondence should be addressed

variants on a heterogeneous clinical sample and identify two novel loss-of-heterozygosity events.

2 MODEL STRUCTURE

RVD2 uses a two-stage approach for detecting rare variants. First, it estimates the parameters of a hierarchical Bayesian model under two sequencing data sets: one from the sample of interest (case) and one from a known reference sample (control). Then, it tests for a significant difference between key model parameters in the case and control samples and returns called variant positions.

For a given sample, the observed data consists of two matrices $r \in I\!R^{J\times N}$ and $n \in I\!R^{J\times N}$, where r_{ji} is the number of reads with a non-reference base at location j in experimental replicate i and n_{ji} is the total number of reads at location j in replicate i. The model generative process is as follows:

1. For each location j:
   a. Draw an error rate $\mu_j \sim \text{Beta}(\mu_0, M_0)$
   b. For each replicate i:
      (1) Draw $\theta_{ji} \sim \text{Beta}(\mu_j, M_j)$
      (2) Draw $r_{ji}|n_{ji} \sim \text{Binomial}(\theta_{ji}, n_{ji})$

The generative process involves several hyperparameters: μ_0, a global error rate; M_0, a global precision; μ_j, a local error rate. M_j, a local precision. The global error rate, μ_0, estimates the expected error rate across all locations. The global precision, M_0, estimates the variation in the error rate across locations. The local error rate, μ_j, estimates the exepected error rate across replicates at location j. The local precision, M_j, estimates the variation in the error rate across replicates at location j.

RVD2 has three levels of sampling. First, a global error rate and global precision are chosen once for the entire data set. Then, at each location, a local precision is chosen and a local error rate is sampled from a Beta distribution. Finally, the error rate for replicate i at location j is drawn from a Beta distribution and the number of non-reference reads is drawn from a binomial.

RVD2 hierarchically partitions sources of variation in the data. $r_{ji}|n_{ji} \sim \text{Binomial}(\theta_{ji}, n_{ji})$ models the variation due to sampling the pool of DNA molecules on the sequencer. $\theta_{ji} \sim \text{Beta}(\mu_j, M_j)$ models the variation due to experimental reproducibility. The variation in error rate due to sequence context is modeled by $\mu_j \sim \text{Beta}(\mu_0, M_0)$. Importantly, increasing the read depth n_{ji} only reduces the sampling error, but does nothing to reduce experimental variation or variation due to sequence context.

Figure 1 shows a graphical representation of the RVD2 statistical model. In this graphical model framework a shaded node represents an observed random variable, an unshaded node represents an unobserved or latent random variable and a directed edge represents a functional dependency between the two connected nodes (Jordan, 2004). A rounded box or "plate" represents replication of the nodes within the plate. The graphical model framework connects graph theory and probability theory in a way that facilitates algorithmic methods for statistical inference.

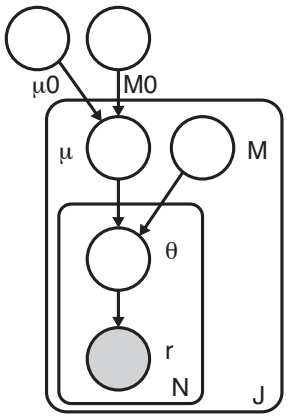

Fig. 1. RVD2 Graphical Model

The joint distribution over the latent and observed variables for data at location j in replicate i given the parameters can be factorized as

$$p\left(r_{ji}, \theta_{ji}, \mu_j | n_{ji}; \mu_0, M_0, M_j\right) = p\left(r_{ji}|\theta_{ji}, n_{ji}\right) p\left(\theta_{ji}|\mu_j; M_j\right) p\left(\mu_j; \mu_0, M_0\right), \quad (1)$$

where

$$\begin{aligned} p\left(\mu_j; \mu_0, M_0\right) &= \frac{\Gamma(M_0)}{\Gamma(\mu_0 M_0)\Gamma(M_0(1-\mu_0))} \cdot \\ &\quad \mu_j^{M_0\mu_0 - 1}(1-\mu_j)^{M_0(1-\mu_0)-1}, \\ p\left(\theta_{ji}|\mu_j; M_j\right) &= \frac{\Gamma(M_j)}{\Gamma(\mu_j M_j)\Gamma(M_j(1-\mu_j))} \cdot \\ &\quad \theta_{ji}^{M_j\mu_j - 1}(1-\theta_{ji})^{M_j(1-\mu_j)-1}, \\ p\left(r_{ji}|\theta_{ji}, n_{ji}\right) &= \frac{\Gamma(n_{ji}+1)}{\Gamma(r_{ji}+1)\Gamma(n_{ji}-r_{ji}+1)} \cdot \\ &\quad \theta_{ji}^{r_{ji}}(1-\theta_{ji})^{n_{ji}-r_{ji}}. \end{aligned} \quad (2)$$

Integrating over the latent variables θ_{ji} and μ_j yields the marginal distribution of the data,

$$p\left(r_{ji}|n_{ji}; \mu_0, M_0, M_j\right) = \int_{\mu_j}\int_{\theta_{ji}} p\left(r_{ji}|\theta_{ji}, n_{ji}\right) p\left(\theta_{ji}|\mu_j; M_j\right) p\left(\mu_j; \mu_0, M_0\right) d\theta_{ji} d\mu_j. \quad (3)$$

Finally, the log-likelihood of the data set is

$$\log p\left(r|n; \mu_0, M_0, M\right) = \sum_{j=1}^{J}\sum_{i=1}^{N} \log \int_{\mu_j}\int_{\theta_{ji}} p\left(r_{ji}|\theta_{ji}, n_{ji}\right) p\left(\theta_{ji}|\mu_j; M_j\right) p\left(\mu_j; \mu_0, M_0\right) d\theta_{ji} d\mu_j. \quad (4)$$

3 INFERENCE AND HYPOTHESIS TESTING

The primary object of inference in this model is the joint posterior distribution function over the latent variables,

$$p(\mu, \theta | r, n; \phi) = \frac{p(\mu, \theta, r | n; \phi)}{p(r | n; \phi)}, \quad (5)$$

where the parameters are $\phi \triangleq \{\mu_0, M_0, M\}$.

The Beta distribution over μ_j is conjugate to the Binomial distribution over θ_{ji}, so we can write the posterior distribution as a Beta distribution. However, there is not a closed form for the product of a Beta distribution with another Beta distribution, so exact inference is intractable.

Instead, we have developed a Metropolis-within-Gibbs approximate inference algorithm shown in Algorithm 1. First, the hyperparameters are initialized using method-of-moments (MoM). Given those hyperparameter estimates, we sample from the marginal posterior distribution for μ_j given its Markov blanket using a Metropolis-Hasting rejection sampling rule. Finally, we sample from the marginal posterior distribution for θ_{ji} given its Markov blanket. Samples from θ_{ji} can be drawn from the posterior distribution directly because the prior and likelihood form a conjugate pair. This sampling procedure is repeated until the chain converges to a stationary distribution then we draw samples from the posterior distribution over latent variables.

Algorithm 1 Metropolis-within-Gibbs Algorithm

```
Initialize θ, μ, M, μ_0, M_0
repeat
    for each location j do
        Draw T samples from p(μ_j|θ_ij, μ_0, M_0) using M–H
        Set μ_j to the sample median for the T samples
        for each replicate i do
            Sample from p(θ_ij|r_ij, n_ij, μ_j, M)
        end for
    end for
until sample size sufficient
```

3.1 Initialization

The initial values for the model parameters and latent variables is obtained by a method-of-moments (MoM) procedure. MoM works by setting the population moment equal to the sample moment. A system of equations is formed such that the number of moment equations is equal to the number of unknown parameters and the equations are solved simultaneously to give the parameter estimates. We simply start with the data matrices r and n and work up the hierarchy of the graphical model solving for the parameters of each conditional distribution in turn.

We present the initial parameter estimates here and provide the derivations in Supplementary Information. The MoM estimate for replicate-level parameters are $\hat{\theta}_{ji} = \frac{r_{ji}}{n_{ji}}$. The estimates for the position-level parameters are $\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_{ji}$ and $\hat{M}_j = \frac{\hat{\mu}_j(1-\hat{\mu}_j)}{\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_{ji}^2} - 1$. The estimates for the genome-level parameters are $\hat{\mu}_0 = \frac{1}{J}\sum_{j=1}^{J}\hat{\mu}_j$ and $\hat{M}_0 = \frac{\hat{\mu}_0(1-\hat{\mu}_0)}{\frac{1}{J}\sum_{j=1}^{J}\hat{\mu}_j^2} - 1$.

3.2 Sampling from $p\left(\theta_{ij} | r_{ij}, n_{ij}, \mu_j, M\right)$

Samples from the posterior distribution $p(\theta_{ji}|r_{ji}, n_{ji}, \mu_j, M_j)$ are drawn analytically because of the Bayesian conjugacy between the prior $p(\theta_{ji}|\mu_j, M_j) \sim \text{Beta}(\mu_j, M_j)$ and the likelihood $p(r_{ji}|n_{ji}, \theta_{ji}) \sim \text{Binomial}(\theta_{ji}, n_{ji})$. The posterior distribution is

$$p(\theta_{ji}|r_{ji}, n_{ji}, \mu_j, M_j) \sim \text{Beta}\left(\frac{r_{ji} + M_j\mu_j}{n_{ji} + M_j}, n_{ji} + M_j\right). \quad (6)$$

3.3 Sampling from $p\left(\mu_j | \theta_{ji}, M_j, \mu_0, M_0\right)$

The posterior distribution over μ_j given its Markov blanket is

$$p(\mu_j|\theta_{ji}, M_j, \mu_0, M_0) \propto p(\mu_j|\mu_0, M_0)p(\theta_{ji}|\mu_j, M_j). \quad (7)$$

Since the prior, $p(\mu_j|\mu_0, M_0)$, is not conjugate to the likelihood, $p(\theta_{ji}|\mu_j, M_j)$, we cannot write an analytical form for the posterior distribution. Instead, we sample from the posterior distribution using the Metropolis-Hastings algorithm.

A candidate sample is generated from the symmetric proposal distribution $Q(\mu_j^*|\mu_j^{(p)}) \sim \mathcal{N}(\mu_j^{(p)}, \sigma_j^2)$, where $\mu_j^{(p)}$ is the pth from the posterior distribution. The acceptance probability is then

$$a = \frac{p(\mu_j^*|\mu_0, M_0)p(\theta_{ji}^{(p+1)}|\mu_j^*, M_j)}{p(\mu_j^{(p)}|\mu_0, M_0)p(\theta_{ji}^{(p+1)}|\mu_j^{(p)}, M_j)} \quad (8)$$

We fixed the proposal distribution variance for all the Metropolis-Hastings steps within a Gibbs iteration to $\sigma_j = 0.1 \cdot \hat{\mu}_j \cdot (1 - \hat{\mu}_j)$ if $\hat{\mu}_j \in (10^{-3}, 1 - 10^{-3})$ and $\sigma_j = 10^{-4}$ otherwise, where $\hat{\mu}_j$ is the MoM estimator of μ_j. Though it is not theoretically necessary, we have found that the algorithm performance improves when we take the median of five or more M-H samples as a single Gibbs step for each position.

We resample from the proposal if the sample is outside of the support of the posterior distribution. We typically discard 20% of the sample for burn-in and thin the chain by a factor of 2 to reduce autocorrelation among samples. Since, each position j is exchangeable given the global hyperparameters μ_0 and M_0 this sampling step can be distributed across up to J processors.

3.4 Posterior Distribution Test

Metropolis-within-Gibbs provides samples from the posterior distribution of μ_j given the case or control data. For notational simplicity, we define the random variables associated with these two distributions μ_j^{case} and μ_j^{control} and the associated samples as $\tilde{\mu}_j^{\text{case}}$ and $\tilde{\mu}_j^{\text{control}}$.

A variant is called if $\mu_j^{\text{case}} > \mu_j^{\text{control}}$ with high confidence,

$$\Pr(\mu_j^{\text{case}} - \mu_j^{\text{control}} \geq \tau) \approx \frac{1}{N_{\text{MH}}}\sum_{k=1}^{N_{\text{MH}}} \mathbb{1}_{\tilde{\mu}_{jk}^{\text{case}} - \tilde{\mu}_{jk}^{\text{control}} \geq \tau} > 1 - \alpha, \quad (9)$$

where τ is a detection threshold and $1 - \alpha$ is a confidence level. We draw a sample from the posterior distribution $\tilde{\mu}_j^{\Delta} \triangleq \tilde{\mu}_j^{\text{case}} - \tilde{\mu}_j^{\text{control}}$ by simple random sampling with replacement from $\tilde{\mu}_j^{\text{case}}$ and $\tilde{\mu}_j^{\text{control}}$.

The threshold, τ, may be set to zero or optimized for a given median depth and desired MAF detection limit. The optimal τ

maximizes the Matthews Correlation Coefficient (MCC) to perfect classification in the ROC curve plot,

$$\tau^* = \arg\max_{\tau} \{\text{MCC}(\tau)\}. \quad (10)$$

While we are able to compute the optimal τ threshold for a test data set, in general we would not have access to τ^*. With sufficient training data, one would be able to develop a lookup table or calibration curve to set τ based on read depth and MAF level of interest. Absent this information we set $\tau = 0$.

3.5 χ^2 test for non-uniform base distribution

An abundance of non-reference bases at a position called by the posterior density test may be due to a true mutation or due to a random sequencing error; we would like to differentiate these two scenarios. We assume non-reference read counts caused by a non-biological mechanism results in a uniform distribution over three non-reference bases. In contrast, the distribution of counts among three non-reference bases caused by biological mutation would not be uniform.

We use a χ^2 goodness-of-fit test on a multinomial distribution over the non-reference bases to distinguish these two possible scenarios. The null hypothesis is $H_0 : p = (p_1, p_2, p_3)$ where $p_1 = p_2 = p_3 = 1/3$. Cressie and Read (1984) identified a power-divergence family of statistics, indexed by λ, that includes as special cases Pearson's $\chi^2(\lambda = 1)$ statistic, the log likelihood ratio statistic ($\lambda = 0$), the Freeman-Tukey statistic ($\lambda = -1/2$), and the Neyman modified statistic $X^2(\lambda = -2)$. The test statistic is

$$2nI^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{k=1}^{3} r_{ji}^{(k)} \left[\left(\frac{r_{ji}^{(k)}}{E_{ji}^{(k)}} \right)^{\lambda} - 1 \right]; \lambda \in R, \quad (11)$$

where $r_{ji}^{(k)}$ is the observed frequency for non-reference base k at position j in replicate i and $E_{ji}^{(k)}$ is the corresponding expected frequency under the null hypothesis. Cressie and Read (1984) recommended $\lambda = 2/3$ when no knowledge of the alternative distribution is available; we choose that value.

We control for multiple hypothesis testing in two ways. We use Fisher's combined probability test (Fisher *et al.*, 1970) to combine the p-values for N replicates into a single p-value at position j,

$$X_j^2 = -2 \sum_{i=1}^{N} \ln(p_{ji}). \quad (12)$$

Equation (12) gives a test statistic that follows a χ^2 distribution with $2N$ degrees of freedom when the null hypothesis is true. We use the Bejamini-Hochberg method to control the family-wise error rate (FWER) over positions that have been called by the posterior distribution test (Benjamini and Hochberg, 1995; Efron, 2010).

4 DATA SETS

We used two independent data sets to evaluate the performance of RVD2 and compare it with other variant calling algorithms. Synthetic DNA sequence data provides true positive and true negative positions as well as define minor allele fractions. HCC1187 data is used to test the performance on a sequenced cancer genome with less than 100% tumor purity.

4.1 Synthetic DNA Sequence Data

Two 400bp DNA sequences that are identical except at 14 loci with variant bases were synthesized and clonally isolated and labeled case and control. Sample of the case and control DNA were mixed at defined fractions to yield defined MAFs of 0.1%, 0.3%, 1%, 10%, and 100%. More details of the experimental protocol are available from the original publication (Flaherty *et al.*, 2011). We aligned the reads to the reference sequence using BWA v0.7.4 with the -C50 option to filter for high mapping quality reads.

To simulate lower coverage data while retaining the error structure of real NGS data, BAM files for the synthetic DNA data were downsampled $10\times$, $100\times$, $1,000\times$, and $10,000\times$ using Picard v1.96. The final data set contains read pairs for three replicates of each case and pairs of reads three replicates for the control sample giving $N = 6$ replicates for the control and each MAF level.

4.2 HCC1187 Sequence Data

The HCC1187 dataset is a well-recognized baseline dataset from Illumina for evaluating sequence analysis algorithms (Newman *et al.*, 2013; Howarth *et al.*, 2011, 2007). The HCC1187 cell line was derived from epithelial cells from primary breast tissue from a 41 y/o adult with TNM stage IIA primary ductal carcinoma. The estimated tumor purity was reported to be 0.8. Matched normal cells were derived from lymphoblastoid cells from peripheral blood. Sequencing libraries were prepared according to the protocol described in the original technical report (Allen, 2013). The raw FASTQ read files were aligned to hg19 using the Isaac aligner to generate BAM files (Raczy *et al.*, 2013). The aligned data had an average read depth of 40x for the normal sample and 90x for the tumor sample with about 96% coverage with 10 or more reads. We used samtools mpileup to generate pileup files using hg19 as reference sequence (Navin *et al.*, 2010).

5 RESULTS

We tested RVD2 using synthetic DNA and data from a primary ductal carcinoma sample. The inference algorithm parameters were set to yield $4,000$ Gibbs samples with a 20% burn-in and $2\times$ tinning rate for a final total of $1,600$ samples. We drew $1,000$ samples from μ^{Δ} to estimate the posterior probability of a variant.

We used RVD2 to identify germline and somatic mutations in the diploid HCC1187 sample. To identify germline mutations, we used the control data in the following hypothesis test $\Pr(\tilde{\mu}_j^{\text{control}} \in [\tau_{\text{l}}, \tau_{\text{u}}]) > 1 - \alpha$, where the intervals are: homozygous reference $[0, 0.2]$, heterozygous mutant $[0.1, 0.9]$, and homozygous mutant $[0.8, 1.0]$ and the size of the test is $\alpha = 0.05$.

To identify somatic mutations, we considered scenarios when the case(tumor) error rate is lower than the control(germline) error rate (e.g. loss-of-heterozygosity) as well as scenarios when the case(tumor) error rate is higher than the control(germline) error rate (e.g. homozygous somatic mutation). The two hypothesis tests are then $\Pr(\mu_j^{\Delta} \geq \tau) > 1 - \alpha$ and $\Pr(\mu_j^{\Delta} \leq \tau) > 1 - \alpha$. The size of the test is $\alpha = 0.05$.

5.1 Performance with read depth

We generated receiver-operating characteristic curves (ROCs) for a range of median read depth and a range of minor allele frequencies

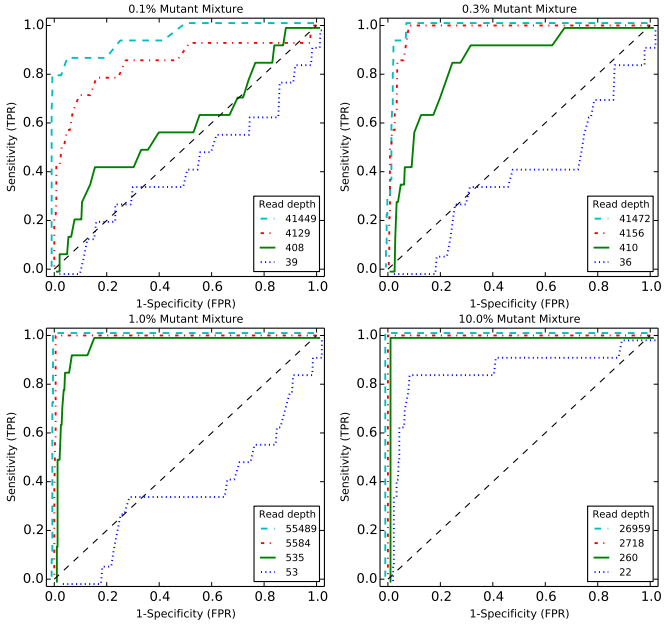

Fig. 2. ROC curve varying read depth showing detection performance.

(MAFs). For these ROC curves, we used the posterior distribution test without the χ^2 test. Adding the χ^2 improves specificity at the expense of sensitivity. Figure 2 shows ROC curves generated by varying the threshold τ with a fixed $\alpha = 0.05$. Figure 2A shows ROC curves for a true 0.1% MAF for a range of median coverage depths. At the lowest depth the sensitivity and specificity is no better than random. However, we would not expect to be able to call a 1 in 1000 variant base with a coverage of only 43. The performance improves monotonically with read depth. Figures 2B-C show a similar relationship between coverage depth and accuracy for higher MAFs.

5.2 Empirical performance compared with other algorithms

We compare the empirical performance of RVD2 to other variant calling algorithms using the synthetic DNA data sets using the false discovery rate as well as sensitivity/specificity. In a research applications, the false discovery rate is a more relevant performance metrics because the aim is generally to identify interesting variants. The sensitivity/specificity metric is more relevant in clinical applications where one is more interested in correctly calling all of the positive variants and none of the negatives. GATK, Varscan2, strelka and muTect are only able to make use of one case and cone control sample, so we provide results of RVD2 with the same data ($N = 1$) for a fair comparison.

Sensitivity/Specificity Comparison Figure 3 shows that samtools, GATK and VarScan2-mpileup all have similar performance. They call the 100% MAF experiment well even at low depth, but are unable to identify true variants in mixed samples with much success. GATK, samtools and VarScan2-mpileup are optimized to call genotypes on pure samples. Therefore, those algorithms are expected to perform well on the 100% dilution (pure mutant) sample and poorly on heterogeneous samples. VarScan2-somatic is able to call more mixed samples. However, as the read depth increases the specificity degrades. Strelka is able to call 10% MAF variants with good performance, but is limited at 1% MAF and below. muTect has good performance across a wide range of MAF levels. But even at the highest depth only has around 0.5 sensitivity for low MAF levels.

The sensitivity for RVD2 with $\tau = 0$ is low for low read depths and MAF levels and $N = 1$ case and control sample. The sensitivity increases considerably with read depth at a slight expense to specificity. With τ^* the performance is much better with high sensitivity and specificity across a wide range of read depths and MAFs. However, in practice one may not know the optimal τ^* a-priori. With $N = 6$ replicates, the sensitivity increases considerably for low MAF variants with a slight degradation in specificity due to false positives. When the median read depth is at least $10\times$ the MAF, RVD2 has higher specificity than all of the other algorithms tested and has a lower sensitivity in only three cases.

False Discovery Rate Comparison Figure 4 shows the false discovery rate for RVD2 compared to samtools, GATK, varscan, strelka and muTect. Blank cells indicate no positive calls were made.

Samtools performs well on 100% MAF sample and performance improves for read depths 3,089 and 30,590. GATK performs well on both the 10% and 100% variants, but makes a false positive call at the 100% MAF level for all read depth levels. VarScan2-pileup performs perfectly for all but the lowest depth for the 100% MAF.

VarScan2-somatic is able to make calls for all but the lowest MAF and coverage level. However, the FDR is high due to many false positives. Interestingly, at a MAF of 100% the FDR is zero for lowest read depth and over 0.9 for the highest read depth. Strelka has a better FDR than the samtools, GATK or Varscan2-somatic algorithms for almost all read depths at the 10% and 100% MAF. However, it does not call any variants at or below 1% MAF. MuTect has the best FDR performance of the other algorithms we tested over a wide range of MAF and depths. But the FDR level is relatively high at around 0.7 for 0.1% – 1% MAF and 0.3 for 10% – 100% MAF.

RVD2 has a lower FDR than other algorithms when the read depth is greater than $10\times$ the MAF with $N = 1$ and τ set to the default value of zero or to the optimal value. The FDR is higher when $N = 6$ because the variance of the control error rate distribution $P(\mu_j^{\text{control}}|r^{\text{control}})$ is smaller. The smaller variance yields improvements in sensitivity at the expense of more false positives. Since the FDR only considers positive calls, the performance degrades.

5.3 HCC1187 primary ductal carcinoma sample

RVD2 identified six variants in the 44kbp PAXIP1 gene from chr7:154738059 to chr7:154782774. Of the six called variants, three were homozygous germline variants and two were heterozygous germline variants. The two heterozygous variants had a loss-of-heterozygosity in the tumor sample. The remaining one is a homozygous somatic variant. Because the median read depth for the case sample is 60, we set the threshold of the test for μ^{Δ} to $\tau = 16\%$ to minimize the false discovery rate (Figure 4).

Figure 5 shows the estimated minor allele frequencies for the normal and tumor samples at the called locations. Positions

MAF	Median Depth	samtools	GATK	VarScan2 mpileup	VarScan2 somatic	strelka	MuTect	N = 1 RVD2 (T=0)	N = 1 RVD2 (T*)	N = 6 RVD2 (T=0)	N = 6 RVD2 (T*)
0.1%	39	0.00/1.00	0.00/1.00	0.00/1.00	0.00/1.00	0.00/1.00	0.00/0.99	0.00/1.00	0.00/1.00	0.00/1.00	0.00/1.00
	408	0.00/1.00	0.00/1.00	0.00/1.00	0.07/0.92	0.00/1.00	0.29/0.91	0.00/1.00	0.00/1.00	0.00/1.00	0.00/1.00
	4129	0.00/1.00	0.00/1.00	0.00/1.00	0.57/0.52	0.00/1.00	0.64/0.86	0.00/1.00	0.00/1.00	0.14/1.00	0.29/1.00
	41449	0.00/1.00	0.00/1.00	0.00/1.00	0.64/0.79	0.00/1.00	0.14/0.93	0.43/1.00	0.57/1.00	0.86/0.97	0.79/1.00
0.3%	36	0.00/1.00	0.00/1.00	0.00/1.00	0.00/1.00	0.00/1.00	0.43/1.00	0.00/1.00	0.00/1.00	0.00/1.00	0.00/1.00
	410	0.00/1.00	0.00/1.00	0.00/1.00	0.21/0.95	0.00/1.00	0.50/0.94	0.00/1.00	0.00/1.00	0.00/1.00	0.00/1.00
	4156	0.00/1.00	0.00/1.00	0.00/1.00	0.57/0.53	0.00/1.00	0.36/0.91	0.14/1.00	0.29/1.00	1.00/0.99	1.00/0.99
	41472	0.00/1.00	0.00/1.00	0.00/1.00	0.64/0.75	0.00/1.00	0.43/0.90	0.93/0.97	0.93/0.99	1.00/0.85	0.93/0.97
1.0%	53	0.00/1.00	0.00/1.00	0.00/1.00	0.00/0.99	0.00/1.00	0.29/0.98	0.00/1.00	0.00/1.00	0.00/1.00	0.00/1.00
	535	0.00/1.00	0.00/1.00	0.00/1.00	0.43/0.89	0.00/1.00	0.71/0.91	0.00/1.00	0.00/1.00	0.21/1.00	0.21/1.00
	5584	0.00/1.00	0.00/1.00	0.00/1.00	0.57/0.47	0.00/1.00	0.64/0.95	0.93/0.99	1.00/0.99	1.00/0.98	1.00/1.00
	55489	0.00/1.00	0.00/1.00	0.00/1.00	0.64/0.69	0.00/1.00	0.86/0.90	1.00/0.95	1.00/0.99	1.00/0.87	1.00/0.99
10.0%	22	0.21/1.00	0.43/1.00	0.00/1.00	0.36/1.00	0.29/1.00	0.86/0.99	0.00/1.00	0.00/1.00	0.00/1.00	0.00/1.00
	260	0.00/1.00	0.57/1.00	0.00/1.00	0.86/1.00	1.00/1.00	1.00/0.99	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
	2718	0.00/1.00	0.79/1.00	0.00/1.00	0.57/0.78	1.00/1.00	1.00/0.98	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
	26959	0.00/1.00	0.57/1.00	0.00/1.00	0.64/0.53	1.00/0.99	1.00/0.98	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
100.0%	27	1.00/0.99	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/0.98	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
	298	1.00/0.99	1.00/1.00	1.00/1.00	1.00/0.99	1.00/0.99	1.00/0.98	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
	3089	0.86/1.00	1.00/1.00	1.00/1.00	1.00/0.65	1.00/0.99	1.00/0.98	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
	30590	0.71/1.00	1.00/1.00	1.00/1.00	1.00/0.39	1.00/1.00	1.00/0.99	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00

Fig. 3. Sensitivity/Specificity comparison of RVD2 with other variant calling algorithms.

MAF	Median Depth	samtools	GATK	VarScan2 mpileup	VarScan2 somatic	strelka	MuTect	N = 1 RVD2 (T=0)	N = 1 RVD2 (T*)	N = 6 RVD2 (T=0)	N = 6 RVD2 (T*)
0.1%	39						1.00				
	408				0.97		0.89				
	4129				0.96		0.86			0.00	0.00
	41449				0.90		0.93	0.14	0.11	0.50	0.08
0.3%	36						0.14				
	410				0.86		0.76				
	4156				0.96		0.87	0.00	0.00	0.26	0.26
	41472				0.92		0.87	0.43	0.28	0.80	0.43
1.0%	53				1.00		0.67				
	535				0.87		0.78			0.00	0.00
	5584				0.96		0.70	0.19	0.18	0.30	0.07
	55489				0.93	1.00	0.76	0.59	0.22	0.78	0.12
10.0%	22	0.00	0.00		0.00	0.00	0.25				
	260		0.00		0.08	0.00	0.18	0.00	0.00	0.00	0.00
	2718		0.00		0.91	0.07	0.36	0.00	0.00	0.00	0.00
	26959		0.00		0.95	0.18	0.33	0.00	0.00	0.00	0.00
100.0%	27	0.12	0.07	0.07	0.00	0.07	0.36	0.00	0.00	0.00	0.00
	298	0.12	0.07	0.00	0.12	0.18	0.39	0.00	0.00	0.00	0.00
	3089	0.00	0.07	0.00	0.91	0.18	0.33	0.00	0.00	0.00	0.00
	30590	0.00	0.07	0.00	0.94	0.00	0.26	0.00	0.00	0.00	0.00

Fig. 4. False discovery rate comparison of RVD2 with other variant calling algorithms. Blank cells indicate no locations were called variant.

chr7:154754371T>C, chr7:154758813G>A, chr7:154760439A>C are called significantly different in tumor and normal sample MAF. Positions chr7:154743899C>T, chr7:154753635T>C and chr7:154780960C>T are called germline homozygous mutations. Positions chr7:154754371 and chr7:154758813 are called heterozygous in the normal sample. In the tumor sample we identified six mutations chr7:154743899C>T, chr7:154753635T>C, chr7:154754371T>C, chr7:154758813G>A, chr7:154760439A>C and chr7:154780960C>T. Positions chr7:154754371T>C and chr7:154758813G>A are called loss-of-heterozygosity events. Some of these mutations are also found to be common population SNPs according to dbSNPv138. The corresponding identities are shown in the figure.

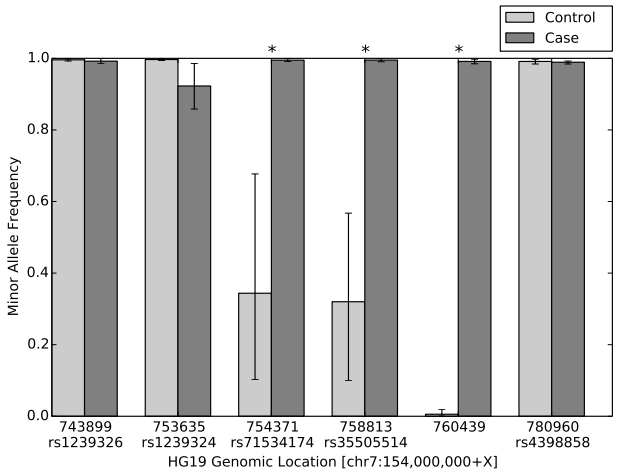

Fig. 5. Estimated minor allele fraction for called variants in PAXIP1 gene. Star (*) indicates $\mu^{\text{case}} - \mu^{\text{control}} \geq \tau$.

The original research describing this sample used Strelka to identify mutations in the same sample. They identified chr7:154760439 as variant, but did not call the other five variants. In particular strelka missed the two LOH events.

6 DISCUSSION

We describe here a novel algorithm for model estimation and hypothesis testing for identifying single-nucleotide variants in heterogeneous samples using next-generation sequencing data. Our algorithm has a higher sensitivity and specificity than many other approaches for a range of read depths and minor allele frequencies.

Our inference algorithm uses Gibbs sampling to estimate the RVD2 hierarchical empirical Bayes model. This sampling procedure provides a guarantee to identify the global optimal parameter settings asymptotically. However, it may require many samples to achieve that guarantee causing the algorithm to be slower than other deterministic approaches. We opted for this balance of speed and accuracy because computational time is often not limiting and the cost of a false positive or false negative greatly outweighs the cost of more computation.

We have focused on the statistical model and hypothesis test in this study and our results do not include any pre-filtration of erroneous reads or post-filtration of mutation calls beyond a simple quality score threshold. Incorporation of such data-cleaning steps will likely improve the accuracy of the algorithm.

Our approach does not address identification of indels, structural variants or copy number variants. Those mutations typically require specific data analysis models and tests that are different than those for single-nucleotide variants. Furthermore, analysis of RNA-seq data or other data generated on the NGS platform may require different models that are more appropriately tuned to the particular noise feature of that data.

The availability of clinical sequence data is increasing as the technical capability to sequence clinical samples at low cost improves. Consequently, we require statistically accurate algorithms that are able to call germline and somatic point mutations in heterogeneous samples with low purity. Such accurate algorithms are a step towards greater access to genomics for clinical diagnostics.

ACKNOWLEDGEMENT

We would like to thank Fan Zhang for careful review of this manuscript and helpful comments.

Funding: P.F. and Y.H. were supported by seed funding from Worcester Polytechnic Institute.

REFERENCES

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O., Sahinalp, S. C., Gibbs, R. A., and Eichler, E. E. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics*, **41**(10), 1061–1067.

Allen, E. (2013). Molecular characterization of tumors using next-generation sequencing. Technical Report 770-2013-011, 2013 Illumina, Inc.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Capobianchi, M. R., Giombini, E., and Rozera, G. (2012). Next-generation sequencing technology in clinical virology. *Clinical Microbiology and Infection*, **19**(1), 15–22.

Cibulskis, K., Lawrence, M. S., and Carter, S. L. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature*.

Consortium, T. . G. P., The 1000 Genomes Consortium Participants are arranged by project role, t. b. i. a., finally alphabetically within institutions except for Principal Investigators, as indicated, P. L., author, C., committee, S., Medicine, P. g. B. C. o., BGI-Shenzhen, Broad Institute of MIT and Harvard, European Bioinformatics Institute, Illumina, Max Planck Institute for Molecular Genetics, US National Institutes of Health, University of Oxford, Washington University in St Louis, Wellcome Trust Sanger Institute, Affymetrix, A. g., Medicine, A. E. C. o., Medicine, B. C. o., BGI-Shenzhen, College, B., Hospital, B., Women's, Broad Institute of MIT and Harvard, Laboratory, C. S. H., Dankook University, Laboratory, E. M. B., European Bioinformatics Institute, Cornell University, Harvard University, Database, H. G. M., Illumina, Leiden University Medical Center, Louisiana State University, Hospital, M. G., Max Planck Institute for Molecular Genetics, Pennsylvania State University, Stanford University, Tel-Aviv University, Translational Genomics Research Institute, US National Institutes of Health, University of California, San Diego, University of California, San Francisco, University of California, Santa Cruz, University of Chicago, University College London, University of Geneva, University of Maryland School of Medicine, University of Medicine and Dentistry of New Jersey, University of Michigan, University of Montréal, University of Oxford, University of Puerto Rico, University of Texas Health Sciences Center at Houston, University of Utah, University of Washington, Washington University in St Louis, Wellcome Trust Sanger Institute, Yale University, BGI-Shenzhen, S. v. g., Hospital, B., Women's, College, B., Broad Institute of MIT and Harvard, Laboratory, C. S. H., Cornell University, European Bioinformatics Institute, Laboratory, E. M. B., Illumina, Leiden University Medical Center, Louisiana State University, Stanford University, Translational Genomics Research Institute, US National Institutes of Health, University of California, San Diego, University of Maryland School of Medicine, University of Oxford, University of Utah, University of Washington, Washington University in St Louis, Wellcome Trust Sanger Institute, Yale University, Medicine, E. g. B. C. o., BGI-Shenzhen, College, B., Broad Institute of MIT and Harvard, Cornell University, European Bioinformatics Institute, Hospital, M. G., Stanford University, Translational Genomics Research Institute, US National Institutes of Health, Geneva, i. o., University of Michigan, University of Oxford, Washington University in St Louis, Wellcome Trust Sanger Institute, Yale University, College, F. i. g. B., Medicine, B. C. o., Broad Institute of MIT and Harvard, Laboratory, C. S. H., Cornell University, Dankook University, European Bioinformatics Institute, Harvard University, Hospital, M. G., Stanford University, Translational Genomics Research Institute, University of Geneva, University of Medicine and Dentistry of New Jersey, University of Montréal, University of Oxford, Washington University in St Louis, Wellcome Trust Sanger Institute, Yale University, Medicine, D. c. c. g. B.

C. o., BGI-Shenzhen, Broad Institute of MIT and Harvard, European Bioinformatics Institute, Illumina, Max Planck Institute for Molecular Genetics, Translational Genomics Research Institute, US National Institutes of Health, University of California, Santa Cruz, University of Michigan, University of Oxford, University of Washington, Washington University in St Louis, Wellcome Trust Sanger Institute, group, S., ELSI, GBR, S. c. B. f. E., Scotland, Colombians in Medellín, C. C., CHS, H. C. S., FIN, F. i. F., and IBS, I. (2013). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **490**(7422), 56–65.

Consortium, T. H. M. P. (2013). A framework for human microbiome research. *Nature*, **486**(7402), 215–221.

Cressie, N. and Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 440–464.

Cushing, A., Flaherty, P., Hopmans, E., Bell, J. M., and Ji, H. P. (2013). Rvd: a command-line program for ultrasensitive rare single nucleotide variant detection using targeted next-generation dna resequencing. *BMC research notes*, **6**(1), 206.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., *et al.* (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, **43**(5), 491–498.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L., and Quake, S. R. (2008). Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *PNAS*, **105**(42), 16266–16271.

Fisher, S. R. A., Genetiker, S., Fisher, R. A., Genetician, S., Britain, G., Fisher, R. A., and Généticien, S. (1970). *Statistical methods for research workers*, volume 14. Oliver and Boyd Edinburgh.

Flaherty, P., Natsoulis, G., Muralidharan, O., Winters, M., Buenrostro, J., Bell, J., Brown, S., Holodniy, M., Zhang, N., and Ji, H. P. (2011). Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Research*.

Ghedin, E., Laplante, J., DePasse, J., Wentworth, D. E., Santos, R. P., Lepow, M. L., Porter, J., Stellrecht, K., Lin, X., Operario, D., Griesemer, S., Fitch, A., Halpin, R. A., Stockwell, T. B., Spiro, D. J., Holmes, E. C., and George, K. S. (2010). Deep Sequencing Reveals Mixed Infection with 2009 Pandemic Influenza A (H1N1) Virus Strains and the Emergence of Oseltamivir Resistance. *Journal of Infectious Diseases*, **203**(2), 168–174.

Howarth, K., Blood, K., Ng, B., Beavis, J., Chua, Y., Cooke, S., Raby, S., Ichimura, K., Collins, V., Carter, N., *et al.* (2007). Array painting reveals a high frequency of balanced translocations in breast cancer cell lines that break in cancer-relevant genes. *Oncogene*, **27**(23), 3345–3359.

Howarth, K. D., Pole, J. C., Beavis, J. C., Batty, E. M., Newman, S., Bignell, G. R., and Edwards, P. A. (2011). Large duplications at reciprocal translocation breakpoints that might be the counterpart of large deletions and could arise from stalled replication bubbles. *Genome Research*, **21**(4), 525–534.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, pages 140–155.

Kitzman, J. O., Snyder, M. W., Ventura, M., Lewis, A. P., Qiu, R., Simmons, L. E., Gammill, H. S., Rubens, C. E., Santillan, D. A., Murray, J. C., Tabor, H. K., Bamshad, M. J., Eichler, E. E., and Shendure, J. (2012). Noninvasive Whole-Genome Sequencing of a Human Fetus. *Science Translational Medicine*, **4**(137), 137ra76–137ra76.

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. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, **22**(3), 568–576.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, **155**(1), 27–38.

Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nature Reviews Genetics*, **11**(3), 191–203.

Li, H. (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**(21), 2987–2993.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9), 1297–1303.

Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., Levy, D., Lundin, P., Maner, S., Zetterberg, A., Hicks, J., and Wigler, M. (2010). Inferring tumor progression from genomic heterogeneity. *Genome Research*, **20**(1), 68–80.

Newman, S., Howarth, K. D., Greenman, C. D., Bignell, G. R., Tavaré, S., and Edwards, P. A. (2013). The relative timing of mutations in a breast cancer genome. *PloS one*, **8**(6), e64991.

Ouyang, Z., Zhou, Q., and Wong, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *PNAS*, **106**(51), 21521–21526.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencingplatforms: comparison of Ion Torrent, PacificBiosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**(1), 1–1.

Raczy, C., Petrovski, R., Saunders, C. T., Chorny, I., Kruglyak, S., Margulies, E. H., Chuang, H.-Y., Källberg, M., Kumar, S. A., Liao, A., *et al.* (2013). Isaac: Ultra-fast whole genome secondary analysis on illumina sequencing platforms. *Bioinformatics*.

Rivera, C. M. and Ren, B. (2013). Mapping Human Epigenomes. *Cell*, **155**(1), 39–55.

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, **28**(14), 1811–1817.

Spencer, D. H., Tyagi, M., Vallania, F., Bredemeyer, A., Pfeifer, J. D., Mitra, R. D., and Duncavage, E. J. (2013). Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. *The Journal of Molecular Diagnostics*.

Stead, L. F., Sutton, K. M., Taylor, G. R., Quirke, P., and Rabbitts, P. (2013). Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing: Applications in Tumor Subclone Resolution. *Human Mutation*, **34**(10), 1432–1438.