BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–8

RVD2: An ultra-sensitive variant detection model for low-depth heterogeneous 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. Therefore, a variant calling algorithm for calling low-frequency polymorphisms in heterogeneous samples is needed.

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 apply our model and identify twelve 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 (**?**). Protocol and platform engineering improvements have enabled the generation of 1×10^9 bases of sequence data in 27 hours for approximately $1000 (**?**). As a result, NGS is increasingly being used as a general platform for research assays for methylation state (**?**), DNA mutations (**?**), copy number variation (**?**), promoter occupancy (**?**) and others (**?**). NGS diagnostics are being translated to clinical applications including noninvasive fetal diagnostics (**?**), infectious disease diagnostics (**?**), cancer diagnostics (**?**), and human microbial analysis (**?**).

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 (**?**). Infectious diseases such as HIV and influenza may contain many genetically heterogeneous sub-populations (**??**). DNA sequencing of individual regions of a solid tumor has revealed genetic heterogeneous within an individual sample (**?**).

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 (**??**). 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 (**?**).

Recently, some have developed algorithms to call low-frequency or rare variants in heterogeneous samples. **?** developed a Bayesian framework which can model the normal DNA contamination and intra-tumor heterogeneity by parameterizing the normal genotype cell proportion at each SNP. 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 (**?**). Strelka uses a hierarchical Bayesian approach to model the joint distribution of the allele frequency in the tumor and normal samples at each locus (**?**). 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 (**?**). RVD uses a hierarchical Bayesian model to capture the error structure of the data and call variants (**??**). 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. **?** 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, **?** also highlighted the fact that VarScan2 yielded more false positives at high read depth. **?** showed that VarScan-somatic outperformed Strelka and 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

*to whom correspondence should be addressed

© Oxford University Press 2014.

able to call 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 $r \in \mathbb{R}^{J\times N}$ and $n \in \mathbb{R}^{J\times N}$, 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. J is the region of interest length and N is the number of technical replicates in the sample. Technical replicates are used to establish experimental variability in next-generation sequencing procedure (**??**), though multiple replicates are not necessary for RVD2.

The model generative process given hyperparameters μ_0, M_0 and $M$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; and 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. The distribution $r_{ji}|n_{ji} \sim \text{Binomial}(\theta_{ji}, n_{ji})$ models the variation due to sampling the pool of DNA molecules on the sequencer. The distribution $\theta_{ji} \sim \text{Beta}(\mu_j, M_j)$ 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 (**?**). 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.

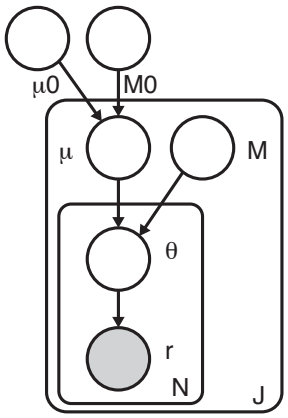


Fig. 1: RVD2 Graphical Model.

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

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

RVD2 improves on RVD in three ways. First, RVD2 has a Beta(μ_0, M_0) prior on local error rate μ_j, which captures the global across-position error rate. The prior distribution allows μ_j to borrow information from adjacent positions and allows RVD2 to handle low read depths. Second, RVD2 handles multiple replicates in case samples. Third, RVD2 has a more accurate Bayesian hypothesis testing method compared to a frequentist normal z-test in RVD. We show a performance comparison between RVD and RVD2 in Section 5.2.

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 are 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

$$p(\theta_{ji}|r_{ji},n_{ji},\mu_j,M_j)$$
$$\sim \text{Beta}\left(r_{ji}+M_j\mu_j, n_{ji}-r_{ji}+M_j(1-\mu_j)\right). \quad (6)$$

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

Posterior Difference 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,

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

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),

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

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 $\tau = 0$.

Posterior Somatic Test. We use a two-sided posterior difference test using control and case samples to test for somatic mutations. We consider 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 $\Pr(\mu_j^{\text{case}} - \mu_j^{\text{control}} > \tau) > 1-\alpha$ and $\Pr(\mu_j^{\text{case}} - \mu_j^{\text{control}} < \tau) > 1-\alpha$. We typically set the hreshold τ to zero.

Posterior Germline Test. We use a one-sided posterior distribution test using a single control sample to identify germline mutations. We call a germline mutation if $\mu_j^{control} \geq \tau$ with high confidence,

$$\Pr(\mu_j^{\text{control}} \geq \tau) \approx \frac{1}{N_{\text{MH}}} \sum_{k=1}^{N_{\text{MH}}} \mathbb{1}_{\tilde{\mu}_{jk}^{\text{control}} \geq \tau} > 1-\alpha. \quad (11)$$

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

$$2nI^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{k=1}^{3} r_{ji}^{(k)} \left[\left(\frac{r_{ji}^{(k)}}{E_{ji}^{(k)}} \right)^{\lambda} - 1 \right]; \lambda \in R, \quad (12)$$

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. **?** recommended $\lambda = 2/3$ when no knowledge of the alternative distribution is available and we choose that value.

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

$$X_j^2 = -2 \sum_{i=1}^{N} \ln(p_{ji}). \quad (13)$$

Equation (13) gives a test statistic that follows a χ^2 distribution with $2N$ degrees of freedom when the null hypothesis is true. If the sample average depth is higher than 500, we use the Benjamini-Hochberg method to control the family-wise error rate (FWER) over positions that have been called by the posterior distribution test (**??**). The average depth threshold is set because Benjamini-Hochberg method is a highly conservative method and will reject many true calls when the read depth is not high enough.

4 DATA SETS

We used two independent data sets to evaluate the performance of RVD2 and compare it with other variant calling algorithms. The synthetic DNA sequence data provides true positive and true negative positions as well as define minor allele fractions. The 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

Experimental process. 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 minor allele fractons (MAFs) of 0.1%, 0.3%, 1%, 10%, and 100%. More details of the experimental protocol are available from the original publication (**?**).

Computational process. 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

Experimental process. The HCC1187 dataset is a well-recognized baseline dataset from Illumina for evaluating sequence analysis algorithms (**???**). 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.

Computational process. Sequencing libraries were prepared according to the protocol described in the original technical report (**?**). The raw FASTQ read files were aligned to hg19 using the Isaac aligner to generate BAM files (**?**). 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 (**?**).

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 $\tilde{\mu}^{\Delta} = \tilde{\mu}_j^{\text{case}} - \tilde{\mu}_j^{\text{control}}$ to estimate the posterior probability of a variant.

We performed the posterior distribution test for somatic mutations on the synthetic data set to identify mutations in the haploid DNA sequence. We set the threshold $\tau = 0$ and the size of the test $\alpha = 0.05$.

For the HCC1187 data set, we identified both somatic and germline mutations. In the posterior somatic test, we set the threshold $\tau = 0$ and the size of the test $\alpha = 0.05$. In the posterior germline test, We set the threshold $\tau = 0.05$ considering the low average coverage (40x). The size of the test is set at $\alpha = 0.15$. We performed χ^2 non-uniformity test after the posterior density tests.

5.1 Performance by read depth

We generated receiver-operating characteristic curves (ROCs) for a range of median read depth and a range of minor allele frequencies (MAFs). For these ROC curves, we used the posterior density test without the χ^2 test to evaluate the performance of posterior density test individually. 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 set by the false discovery rate and sensitivity/specificity. Among these algorithms,

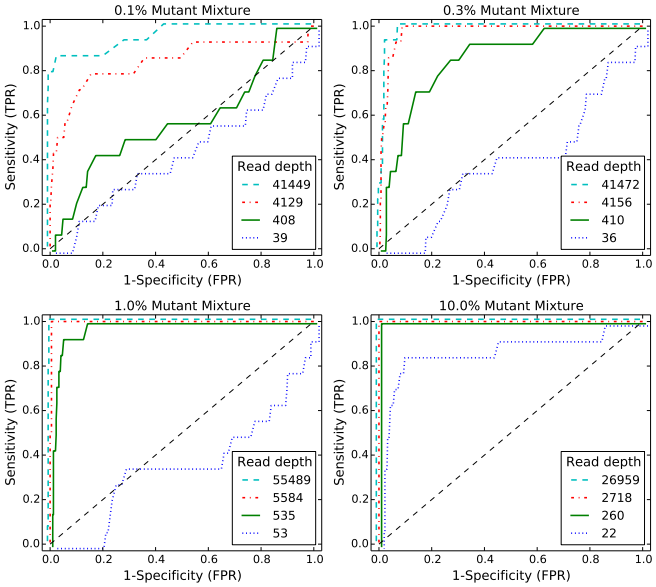


Fig. 2: ROC curve varying read depth showing detection performance. Each subfigure shows ROC curves across four different read depths for one MAF level. The performance improves monotonically with read depth. The performance improves for all read depths with increasing MAF level.

Samtools and GATK are optimized for homogeneous samples, while RVD, VarScan2 Somatic, Strelka and muTect are have good performance to call variants in heterogeneous samples. In a research applications, the false discovery rate is a more relevant performance metric 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 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. 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 performance statistics for RVD are an average three sets of pair-end case replicates. RVD performed the best among all algorithms when the read depth is near 40,000. RVD called all the mutated positions across all MAF levels with no false positives when MAF level is 0.3% or lower. However, RVD can not call any mutations when the depth is too low to measure the baseline error rate and therefore is not useful for low-depth data.

RVD2 has a high sensitivity and specificity for a broad range of read depths and minor-allele frequencies. The sensitivity increases

MAF	Median Depth	SAMtools	GATK	VarScan2 mpileup	VarScan2 somatic	Strelka	MuTect	RVD	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	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	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	0.00/1.00	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	1.00/1.00	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	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	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	0.00/1.00	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	1.00/1.00	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	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	0.00/1.00	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	0.00/1.00	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	1.00/0.99	1.00/0.95	1.00/0.99	1.00/0.87	1.00/0.99
10.0%	22	0.21/1.00	0.00/1.00	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	0.00/1.00
	260	0.00/1.00	0.00/1.00	0.00/1.00	0.86/1.00	1.00/1.00	1.00/0.99	0.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
	2718	0.00/1.00	0.00/1.00	0.00/1.00	0.57/0.78	1.00/1.00	1.00/0.98	0.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
	26959	0.00/1.00	0.00/1.00	0.00/1.00	0.64/0.53	1.00/0.99	1.00/0.98	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	0.00/1.00	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	0.00/1.00	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	0.00/1.00	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	1.00/0.98	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 using synthetic DNA data.

MAF	Median Depth	SAMtools	GATK	VarScan2 mpileup	VarScan2 somatic	Strelka	MuTect	RVD	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	0.04	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	0.08	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	0.19	0.59	0.22	0.78	0.12
10.0%	22	0.00			0.00	0.00	0.25					
	260				0.08	0.00	0.18		0.00	0.00	0.00	0.00
	2718				0.91	0.07	0.36		0.00	0.00	0.00	0.00
	26959				0.95	0.18	0.33	0.31	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	0.3	0.00	0.00	0.00	0.00

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

considerably with read depth at a slight expense to specificity. For the most difficult test with a low read depth and low MAF, performs on-par with muTect. 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

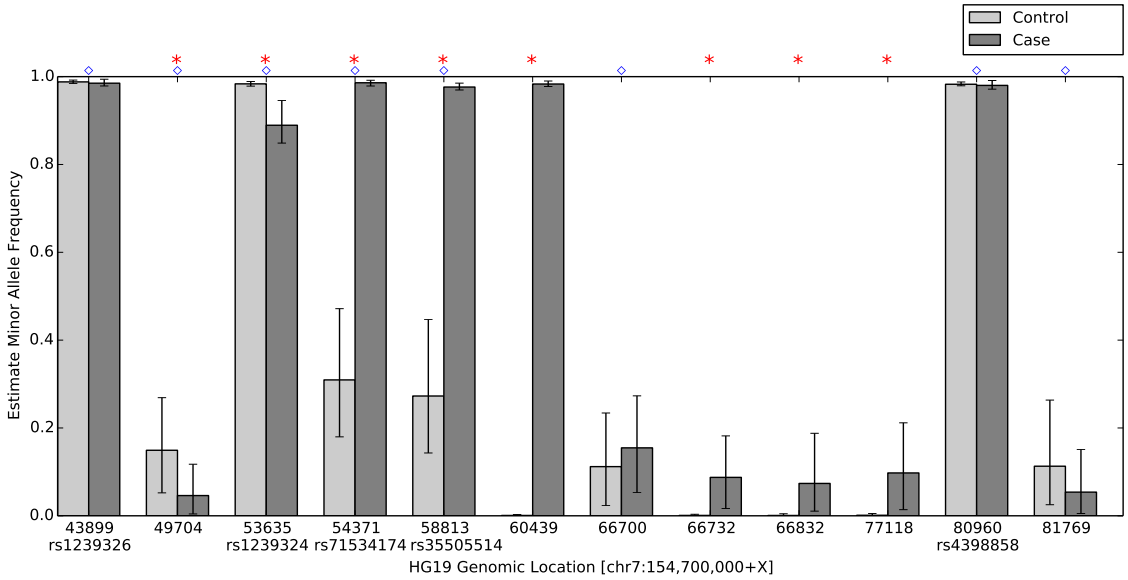


Fig. 5: Estimated minor allele fraction for germline and somatic mutations called by RVD2 in the 44kbp PAXIP1 gene from chr7:154738059 to chr7:154782774. Blue diamonds (◇) indicates germline mutations, where μ^{control} is significantly different from the reference sequence. Red stars (*) indicates somatic mutations, where μ^{case} is significantly different from μ^{control}. The vertical lines represent 95% credible interval around posterior mean MAF. Five positions are common population SNPs according to dbSNPv138, and the identities are shown below the positions.

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. RVD has best FDR performance in the high read depth for 0.1% – 1% MAF levels. The FDR increases to around 0.3 for 10% – 100% MAF in the high read depth.

RVD2 has a lower FDR than other algorithms when the read depth is greater than $10\times$ the inverse 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 by that measure degrades.

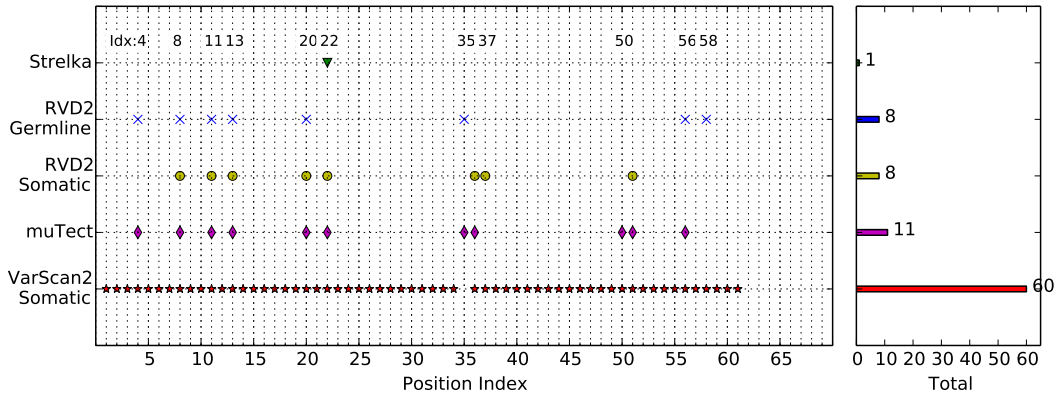
5.3 HCC1187 primary ductal carcinoma sample

RVD2 identified twelve variants in the 44kbp PAXIP1 gene from chr7:154738059 to chr7:154782774. There were eight germline variants and eight somatic mutations. Figure **??** shows the estimated minor allele frequencies for the normal and tumor samples at the called locations. Positions chr7:154743899C>T, chr7:154749704G>A, chr7:154753635T>C, chr7:154754371T>C, chr7:154758813G>A, chr7:154766700C>A, chr7:154780960C>T, and chr7:154781769G >T were called germline mutations. Positions chr7:154749704G>G, chr7:154753635T>C, chr7:154754371T>C, chr7:154758813G>A, chr7:154760439A>C, chr7:154766732T>G, chr7:154766832A>C, chr7:154777118A>C were identified as significantly different in tumor and normal sample MAF and called somatic mutation. Positions chr7:154754371 and chr7:154758813 were considered 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 **??**. The read depth distribution for positions called by RVD2 are provided in Supplementary Table 1.

5.3.1 Performance comparison with other algorithms. We ran muTect and VarScan2-somatic to call mutations in the PAXIP1 gene in HCC1187 sample. We also compared to the result shown in original research report where Strelka was used to identify mutations in the same sample (**?**). Figure 6a shows mutation detection result from Strelka, RVD2, muTect, and VarScan2-somatic, the state-of-art algorithms able to call mutation from heterogeneous samples. For notation simplicity, we use position index to present actual positions in Figure 6 (the correspondence to genomic positions is provided in Supplementary Table 1).

The mutations called by RVD2 and muTect are the most consistent among all the techniques. RVD2 detected twelve germline and somatic mutations, while muTect reported eleven; ten were called by both. In the disagreements, RVD2 did not call position 50 while muTect did not call position 37 and 58. Referring to the depth distribution shown in Figure 6b, it can be seen that position 37 and 58 are more likely mutated while position 50 is less likely mutated.

Strelka was the least sensitive algorithm among all the algorithms. According to the technical report, Strelka identified position 22 (chr7:154760439) as variant, but did not call any other variants. In particular Strelka missed the two LOH events called by RVD2. VarScan2-somatic called most positions among all algorithms, sixty



(a)

Index	REF	Normal	Case
		A:C:G:T	A:C:G:T
4	C	0:0:0:45	0:0:0:71
8	G	8:0:36:0	2:0:42:0
11	T	0:38:0:0	0:63:0:1
13	T	0:19:0:31	0:61:0:0
20	G	14:0:28:0	54:0:0:0
22	A	37:0:0:0	0:38:0:0
35	C	4:21:0:0	10:42:0:0
36	T	0:0:0:35	0:0:5:44
37	A	34:0:0:0	42:4:0:0
50	C	0:49:0:0	0:64:2:0
51	A	46:0:0:0	31:4:0:0
56	C	0:0:0:42	0:0:0:56
58	G	0:0:21:4	0:0:31:2

(b)

Fig. 6: (a) Positions called by VarScan2-somatic, muTect, RVD2 and Strelka in the 44kbp PAXIP1 gene from chr7:154738059 to chr7:154782774. VarScan2-somatic reported 60 positions, muTect called 11 positions, RVD2 called 12 positions and Strelka called 1 position. The unique called positions are sorted by index (correspondance to genomic positions shown in Supplementary Table 1) (b) Read counts for each base for positions called by RVD2 and muTect from raw pileup data.

positions as shown in Figure 6a. VarScan2-somatic detected all the positions called by RVD2 except position 35, which turns out to be a very likely mutation given the depth distribution in Figure 6b. VarScan2-somatic reported fifty positions which were not called by any other three algorithms. The read depth in Supplementary Table 1 suggestions that these positions are very likely to be false positives. The fact that VarScan2-somatic can be over-sensitive has appeared in synthetic dataset analysis. As shown in Figure 4, the False Discovery Rate for VarScan2-somatic at read depth 53 MAF level 1.0% is as high as 1.00. **?** also mentioned that VarScan2 has tendency to call many false positives at high read depth.

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. Another factor that can affect the speed of RVD2 is the number of Metropolis-Hasting sample within one Gibbs sampling run. However, RVD2 is able to use multiple cores in parallel, which can significantly improve time efficiency. On a computational node with 64 2.34GHz processor and 256Gb of RAM, RVD2 takes around 10 minutes to analyze the one 400bp long synthetic data set. The memory requirement is no higher than the size of the gene sequence times the number of Gibbs samples. We are currently developing a variational method to estimate RVD2, aiming at decreasing time cost.

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.