# VARIANT DETECTION MODEL WITH IMPROVED ROBUSTNESS AND ACCURACY FOR LOW-DEPTH TARGETED NEXT-GENERATION SEQUENCING DATA

Abstract. Massively parallel sequencing data generated by next-generation sequencing (NGS) technology is routinely used to detect single nucleotide variants (SNVs) in research samples. An emerging challenge for this technology is the identification of SNVs in heterogeneous cell populations with low read-depth data. We have developed a Bayesian statistical model is able to share information between correlated positions and call low-frequency variants in heterogeneous samples. We present a Bayesian sensitivity analysis of the model to variations in the prior function. Our model with different priors both performs a high accuracy, and a Jeffreys prior gives a lower false discovery rate (FDR) to detect a 0.1% minor allele frequency event within minor read depth compared with an improper prior. In an analysis of a directed evolution experiment, we are able to detect the emergence of a beneficial SNV earlier than was previously shown.

## 1. Introduction

Massively parallel sequencing data has been generated by Next-generation sequencing (NGS) technology to benefit clinical diagnostics and sequencing based phylogenetic analyses. One primary application of NGS is variation detection among related populations and separate novel single-nucleotide variants (SNVs) candidates. Somatic SNVs are detected by comparing the tumor and corresponding normal samples.

To address the detection of SNVs at low allele frequencies, a number of algorithms from NGS platform are being under-represented. Strelka (Saunders et al., 2012), VarScan2 (Koboldt et al., 2012), JointSNVMix (Roth et al., 2012) are highly used to differentiate somatic SNVs from germline cells. Also, SAMtools (Li et al., 2009), Genome Analysis Toolkit (GATK) (McKenna et al., 2010), and MuTect (Cibulskis et al., 2013) broadly concentrate on detecting low-frequency variants. Through the comparision of these somatic mutation callers (Wang et al., 2013), VarScan2 excelled at the detection of high coverage and allele frequecy, while MuTect outperformed the other methods in detecting the low allelic fraction SNVs. However identifying the true SNVs remains challenging

because of the high false positive rate or high false negative rate which are mainly caused by the clonal heterogeneity.

Recently empirical Bayesian approaches have been made use to identify SNVs, which can automatically adjust for multiple testing and selection bias (Liao et al., 2014). A empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data - EBCall, enables accurate mutations calling with low allele frequencies (less than 10%) in a minor tumour subpopulation (Shiraishi et al., 2013). Another method based on empirical Bayesian hierarchical model - RVD, was proposed for ultrasensitive rare SNV detection using beta-binomial model (Flaherty et al., 2011). RVD method is demonstrated to robustly detect mutations at 0.1% fractional representation, which means accurately call one mutant per every 1000 wild-type alleles. With the shortcoming of high read depth estimation in RVD, we originally built an improved robustness and accuracy model - RVD3 for low-depth SNVs detection. Variants calling ability is tested and analyzed both on the synthetic sequence data and the true yeast sequencing data with different read depths.

In this article, RVD3 model - a novel Bayesian structure to accurately identify SNVs with small false discovery rate is first described in detail. Secondly, Metropolis-within-Gibbs sampling is evolved for inference. And then to detect variants, a Bayesian posterior distribution is taken for hypothesis test. Furthermore, we analyze the sensitivity of the Bayesian model to different priors - Jeffreys prior and log-normal prior, and apply the different priors on the synthetic sequence data. Finally, we choose Jeffreys prior for the RVD3 model and demonstrate its performance on the yeast sequence data. Thus our Bayesian model achieves a enhanced robustness and accuracy when calling variants for the low read depth and minor allele frequencies.

## 2. Data Sets

2.1. **Synthetic DNA Sequence Data.** Two 400bp DNA sequences(control/case) were synthesized with only 14 different single nucleotide positions. Sample of the case and control DNA were mixed to yield 0.1%, 0.3%, 1%, 10%, and 100% defined minor allele frequencies (MAFs). The details of the experimental protocol are available from the original publication (Flaherty et al., 2011). We used BWA v0.7.5a to align the short sequencing reads to the reference sequence. The -C50 option of BWA was taken to remove the reads of low mapping quality. BAM files were sampled by

$10\times$, $100\times$, $1,000\times$, and $10,000\times$ using Picard v1.104 (http://picard.sourceforge.net). The final data set contains read pairs for $N = 6$ replicates for the control at different MAF levels.

2.2. **Yeast Data.** We first mapped the wild-tpye strain GSY1135 (Kvitek and Sherlock, 2011) to Chromosome 10 in S288c reference genome (SGD; http://www.yeastgenome.org/) by BWA v0.7.5a (Li and Durbin, 2009). Then called SNPs by GATK v2.5 UnifiedGenotyper (McKenna et al., 2010; DePristo et al., 2011) and created a FASTA GSY1135 reference using GATK FastaAlternative. Secondly, we downloaded generation 7 as control and generation 133 as case in experiment 1 from (Kvitek and Sherlock, 2013), and removed WT population using FASTX Barcode Splitter and cut down the pair ends accordingly. The FASTQ files of case and control were mapped to the corresponding reference genome created before. Then wen used SAMtools v0.1.19 (Li et al., 2009) to convert the alignment files to the binary alignment map (BAM) format. Next, pileup files were generated by SAMtools and depth chart file were derived for further SNVs detection.

## 3. RVD3 Model

3.1. **Model Structure.** RVD is based on a two-stage hierarchical Bayesian model for variant detection (Flaherty et al., 2011). Through hypothesis test on case and control samples by RVD, we can call the variants successfully. Now RVD3 has three-stage model including priors built on the former RVD model. The definitions for sample data are given: $r_{ji}$ is the number of reads with a non-reference base at position $j$ in replicate $i$, and $n_{ji}$ is the total number of reads at position $j$ in replicate $i$. Three parameters of the model are: $\mu_0$, a global error rate; $M_0$, the global position which estimates the variation in the error rate across the positions; to choose a priori distribution for $M_j$, log-normal prior and Jeffreys prior (Jeffreys, 1946) are employed, which enhances the former RVD model (Flaherty et al., 2011). Here $M_j$ is the local precision measures the alteration in the error rate across replicates at position $j$. The graphical chart for RVD3 is shown in Figure 1.

RVD3 hierarchically includes three levels of samplings: $r_{ji}|n_{ji} \sim \text{Binomial}(\theta_{ji}, n_{ji})$ 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 caused by experimental repeatability. The variation in error rate due to sequence context is modeled by $\mu_j \sim \text{Beta}(\mu_0, M_0)$. And the local precision is modeled by $M_j \sim \text{log-normal}(\mu, \sigma)$ (log-normal prior), and Jeffreys prior for $M_j$.

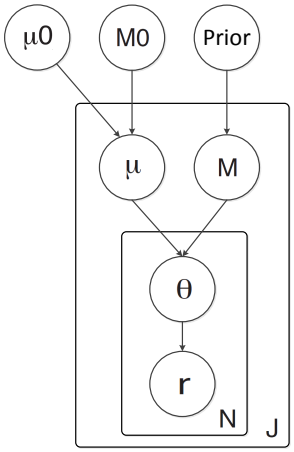


FIGURE 1. RVD3 Graphical Model

3.2. **Inference and Hypothesis Testing.** Metropolis-within-Gibbs sampling is evolved for inference. Algorithm 1 shows the inference process and the detail are also illustrated.

**Algorithm 1** Inference process for Metropolis-within-Gibbs

Initialize $\theta$, $\mu$, $M_j$, $\mu_0$, $M_0$
**repeat**
  **for** each location j **do**
    Samples from $p(\mu_j|\theta_{ij}, \mu_0, M_0)$ ▷ Sample $\mu_j$
    Set $\mu_j$ to the sample median for the samples
    Samples from $p(M_j|\mu, \sigma, \theta_{ji}, \mu_j)$ ▷ Sample $M_j$
    **for** each replicate i **do**
      Sample from $p(\theta_{ij}|r_{ij}, n_{ij}, \mu_j, M)$ ▷ Sample $\theta_{ji}$
    **end for**
  **end for**
**until** sample size sufficient

3.2.1. *Initialization.* [This paragrah COPIED]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.

The initial parameter estimates and derivations are provided in Appendix A. Below is the MoM estimate for replicate-level parameters $\tilde{\theta}_{ji} = \frac{r_{ji}}{n_{ji}}$. The estimates for the position-level parameters

are $\tilde{\mu}_j = \frac{1}{N}\sum_{i=1}^{N}\theta_{ji}$ and $\tilde{M}_j = \frac{\tilde{\mu}_j(1-\tilde{\mu}_j)}{\frac{1}{N}\sum_{i=1}^{N}\theta_{ji}^2} - 1$. The estimates for the genome-level parameters are $\tilde{\mu}_0 = \frac{1}{J}\sum_{j=1}^{J}\mu_j$ and $\tilde{M}_0 = \frac{\tilde{\mu}_0(1-\tilde{\mu}_0)}{\frac{1}{J}\sum_{j=1}^{J}\mu_j^2} - 1$.

3.2.2. *Sampling from* $p\left(\theta_{ij}|r_{ij}, n_{ij}, \mu_j, M\right)$. Because the Bayesian conjugacy between the prior $p(\theta_{ji}|\mu_j, M_j) \sim \text{Beta}(\mu_j, M_j)$ and the likelihood $p(r_{ji}|n_{ji}, \theta_{ji}) \sim \text{Binomial}(\theta_{ji}, n_{ji})$, we draw the samples from the posterior distribution $p(\theta_{ji}|r_{ji}, n_{ji}, \mu_j, M_j)$ The posterior distribution is

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

3.2.3. *Sampling from* $p\left(\mu_j|\theta_{ji}, M_j, \mu_0, M_0\right)$. Based on Markov blanket, the posterior distribution over $\mu_j$ is

$$p(\mu_j|\theta_{ji}, M_j, \mu_0, M_0) \propto p(\mu_j|\mu_0, M_0)p(\theta_{ji}|\mu_j, M_j). \tag{2}$$

[This paragrah COPIED]Since the prior, $p(\mu_j|\mu_0, M_0)$, is not conjugate to the likelihood, $p(\theta_{ji}|\mu_j, M_j)$, we sample from the posterior distribution using the Metropolis-Hastings algorithm. By experience when $\mu_j^{(p)} \in (10^{-3}, 1 - 10^{-3})$, the proposal distribution variance for all the Metropolis-Hastings steps within a Gibbs iteration is set to $\sigma_j = 0.1 \cdot \mu_j^{(p)}$ ; otherwise, we set $\sigma_j = 10^{-4}$ if $\mu_j^{(p)} < 10^{-3}$ and $\sigma_j = 10^{-1} - 10^{-4}$ if $\mu_j^{(p)} > 1 - 10^{-3}$. 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 resampled from the proposal if the sample is outside of the support of the posterior distribution. We throw away a burn-in period - 20% of the samples, and thin the chain by a factor 2 to reduce autocorrelation among samples, resulting in a sample with size 1600 from the posterior distribution.

3.2.4. *Sampling from* $p\left(M_j|\mu, \sigma, \theta_{ji}, \mu_j\right)$. Since Jeffreys prior is from the Fisher information and in RVD3 model $\theta_{ji} \sim Beta\left(\mu_j, M_j\right)$,

$$I\left(M_j\right) = E_{M_j}\left[-\frac{\delta^2 \log p\left(\theta_j|\mu_j, M_j\right)}{\delta M_j^2}\right] \tag{3}$$

We calculated the equations Appendix B and obtained the Jeffreys' prior for $M_j$:

$$[-\left(\Psi_1(M_j) - \Psi_1(\mu_j M_j)\mu_j^2 - \Psi_1((1-\mu_j)M_j)(1-\mu_j)^2\right)]^{\frac{1}{2}} \tag{4}$$

For log-normal prior, the posterior distribution over $M_j$ given its Markov blanket is

$$p(M_j|\mu,\sigma,\theta_{ji},\mu_j) \propto p(\theta_{ji}|\mu_j,M_j)p(M_j|\mu,\sigma) \tag{5}$$

We have $\theta_{ji} \sim \text{Beta}(\mu_j, M_j)$, and $M_j \sim \text{log-normal}(\mu,\sigma)$. Instead of computing the posterior distribution directly, Metropolis-Hastings algorithm was taken to sample from the posterior distribution.

3.2.5. *Posterior Density Test.* Posterior distributions of $\mu_j$ for the control and case are achieved - $\tilde{\mu}_j^{\text{case}}$ and $\tilde{\mu}_j^{\text{control}}$, by Metropolis-within-Gibbs. So we called a variant when $\tilde{\mu}_j^{\text{case}} > \tilde{\mu}_j^{\text{control}}$ with $1-\alpha$ confidence,

$$\Pr(\tilde{\mu}_j^{\text{case}} - \tilde{\mu}_j^{\text{control}} \geq \tau) > 1-\alpha, \tag{6}$$

where $\tau$ is a detection threshold and $1-\alpha$ is the confidence level. We set $\tau = 0$ in our experiment [XXX].

3.2.6. $\chi^2$ *test for non-uniform base distribution.* [This part is COPIED]

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 $\chi^2$ 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 $\lambda$, 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

$$2nI^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{k=1}^{3} r_{ji}^{(k)} \left[ \left( \frac{r_{ji}^{(k)}}{E_{ji}^{(k)}} \right)^{\lambda} - 1 \right]; \lambda \in R, \tag{7}$$

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 and 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 $j$,

$$X_j^2 = -2\sum_{i=1}^{N} \ln(p_{ji}). \tag{8}$$

Equation (8) gives a test statistic that follows a $\chi^2$ distribution with $2N$ degrees of freedom when the null hypothesis is true. Finally, we use the Bejamini-Hochberg method to control the family-wise error rate (FWER) over positions that have been called by the Bayesian hypothesis test (6) (Benjamini and Hochberg, 1995; Efron, 2010).

3.3. **Priors for precision parameter.** The prior distribution characterizes the knowledge of the parameters in the statistical model. Including prior information in the Bayesian approach is difficult but meaningful. An improper prior is the prior distribution integrates to infinity, and may cause an improper posterior which results in an invalid inferences (Lesaffre and Lawson, 2012). Furthermore when the Markov chain Monte Carlo method is taken to derive the posterior, it is possibly hard to sniff out the improper posterior. Even though no problems happened in estimation by improper prior, other troubles could be caused in the Bayesian inference and analysis by it (Stein, 1965). Playing no prior on $M_j$ is exactly an implicit improper prior. Therefore we considered about non-information prior and information prior for sensitivity analysis.

Non-information prior seems to be more unbiased and objective. Various non-informative prior distributions have been suggested for parameters in hierarchical models. Jeffreys prior, as a typical and influential one, is proposed to establish a least informative prior that is automatically invariant to transformations by Harold Jeffreys (Jeffreys, 1946). It is defined in terms of the Fisher information and works well with a single parameter. In our research Jeffreys prior for $M_j$ is the square root of Fisher information of $M_j$.

Informative prior distribution is the most specific type of prior. A good informative prior is imperative to promote accurate posterior estimates. In our study, log-normal prior is a typical prior for the beta density's parameters, $\theta_{ji} \sim \text{Beta}(\mu_j, M_j)$. The parameters of it denoted $\mu$ (mean) and $\sigma$ (standard deviation) respectively.

## 4. Results

4.1. **Sensitivity analysis of priors.** To examing the robustness of our Bayesian model, different plausible prior distributions can be utilized (Gelman, 2006). For this purpose, we performed information prior (log-normal) and non-information prior (Jeffreys). We found that RVD3 is robust to changing of the priors. The RVD3 algorithm provides estimates of model parameters and latent variables given the data.

We show several of these parameters of the model with Jeffreys prior in Figure 2. The left column of is shows the read depth for each of the six BAM files (three replicates each with two read pairs) for each data set. Because the DNA was not sheared and ligated prior to sequencing, the read depth drops to zero at the boundaries. For the 100% mutant data set, the read depth drops at the mutant locations. This is due to the parameters imposed at the alignment stage. The right column of Figure 2 reveals the parameter estimates $\hat{M}_j$ and $\hat{M}_0$ for each data set. $M_j$ measures the variance between replicates at location $j$. There is little variability across positions indicating that the replication variance does not change greatly across position. Interestingly, $M_j$ from Jeffreys prior [Figure 2] across the M value drops at the mutant positions when the data is 10% and 100% mutant, which is not shown in log-normal prior [Figure 8] in Appendix C. Additionally, the error rate across positions is captured by the $M_0$ parameter shown as a horizontal dotted line in the plots in the right column. $M_j$ is greater than $M_0$ shows that the precision between replicates is higher than the precision across positions.

The Bayesian posterior predictive distributions and the priors distributions are shown in Figure 3. These plots indicate the probability distribution over different M values when dilution is 10% at 100× read depth rate. The positions are chosen in the middle of the base length- position 104 and 244 are mutant, and position 100 and 300 are non-mutant. The posterior probability distributions are estimated by Gaussian Kernel Density (Silverman, 1986). Generally the non-mutant positions display normal and stable without a peak nor a strange shape. The two distributions show different

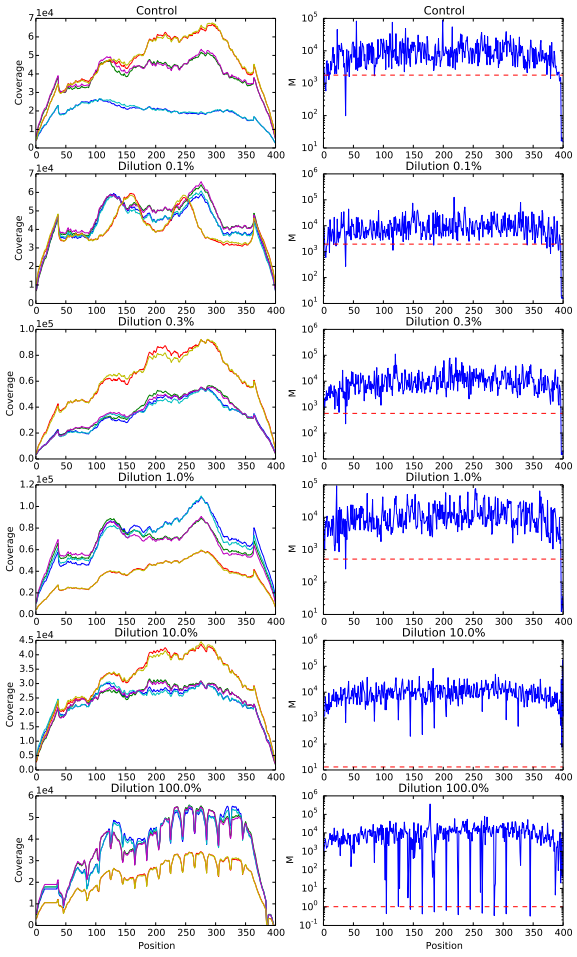


FIGURE 2. Key parameters for RVD3 model with Jeffreys prior on synthetic DNA data sets.

ways of the prior and posterior. The log-normal prior attributes a higher prior probability to M values between 0 and 2000. The Jeffreys prior assigns more information than the log-normal prior (flat) from the posterior curve. They both want to search for a small value for M from the prior curve.

4.2. **Results of priors on synthetic data.** The RVD3 model is analyzed by the selecting two reasonable prior distributions, and the corresponding results are compared from the aspects of performance with different read depths, sensitivity and specificity, and FDR.

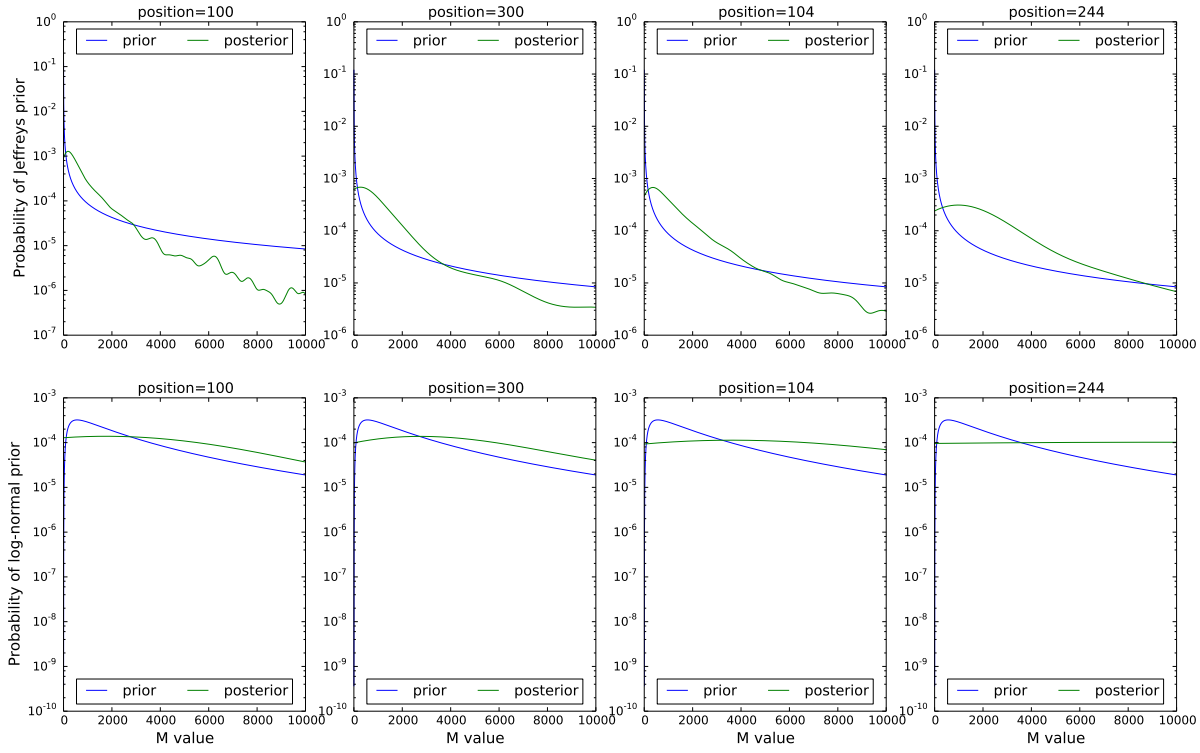


FIGURE 3. Distribution of priors and posteriors when dilution is 10%

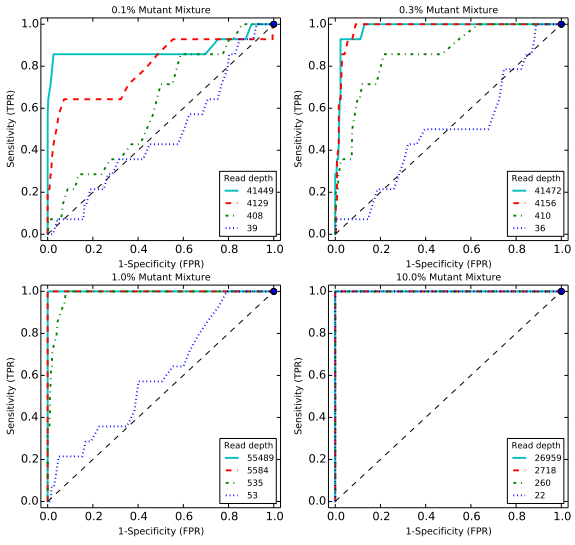


FIGURE 4. ROC curve for variants detection performance by Jeffreys prior.

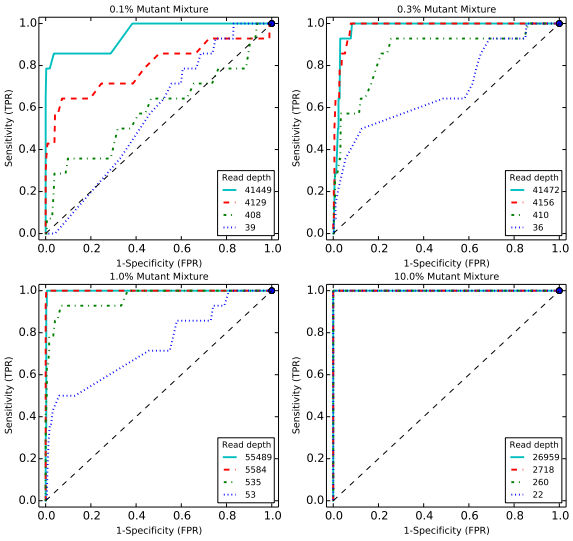


FIGURE 5. ROC curve for variants detection performance by log-normal prior.

4.2.1. *Performance with read depth.* To evaluate the performance of RVD3 model with priors, we generated receiver-operating characteristic curves (ROCs) for median read depth and minor allele frequencies (MAFs). Here the Bayesian test is used without the $\chi^2$ test. Figure 4 and Figure 5 shows ROC curves with a fixed $\alpha = 0.05$. The performance improves when the read depth goes up. (ROC shows that the model with priors performs better than improper prior situation especially on the small read depth). Noticed at the lowest depth (22) with 10.0% mutant mixture, the sensitivity and specificity value are 1 and much better than the model with improper priors for $M_j$, which definitely demonstrates the advantage of priors.

4.2.2. *Sensitivity/Specificity/FDR.* Figure 6 shows that the sensitivity and specificity of the RVD3 of different priors compared with the model with improper priors. Log-normal prior shows a higher sensitivity and specificity value than the Jeffreys prior. Figure **??** shows the false discovery rate of the RVD3 with different priors. Jeffreys prior shows a smaller false discovery rate than others. It is obvious no matter Jeffreys or log-normal, the variant detection performance acquires lower FDR to a known 0.1% minor allele frequency event, campared with improper priors. Based on the various advantages for Jeffreys and log-normal prior, RVD3 can afford a more appropriate choice for the precision parameter. Here we chose Jeffreys prior model because it's a non-information prior, and more attention should be paid to false discovery rate and accuracy for variants calling research,

compared with the clinical experiment or diagnosis which cares more on true positive rate and true negative rate.

| | | RVD3 (T=0) | |
|---|---|---|---|
| MAF | Median Depth | Log-normal Prior | Jeffreys Prior |
| 0.10% | 39 | 0.00/1.00 | 0.00/1.00 |
| | 408 | 0.00/1.00 | 0.00/1.00 |
| | 4129 | 0.07/1.00 | 0.00/1.00 |
| | 41449 | 0.79/0.98 | 0.36/1.00 |
| 0.30% | 36 | 0.00/1.00 | 0.00/1.00 |
| | 410 | 0.00/1.00 | 0.00/1.00 |
| | 4156 | 1.00/0.99 | 0.86/0.99 |
| | 41472 | 1.00/0.88 | 0.93/0.92 |
| 1.00% | 53 | 0.00/1.00 | 0.00/1.00 |
| | 535 | 0.21/1.00 | 0.14/1.00 |
| | 5584 | 1.00/0.99 | 1.00/0.99 |
| | 55489 | 1.00/0.88 | 1.00/0.91 |
| 10.00% | 22 | 0.00/1.00 | 0.00/1.00 |
| | 260 | 1.00/1.00 | 1.00/1.00 |
| | 2718 | 1.00/1.00 | 1.00/1.00 |
| | 26959 | 1.00/1.00 | 1.00/1.00 |
| 100.00% | 27 | 1.00/1.00 | 1.00/1.00 |
| | 298 | 1.00/1.00 | 1.00/1.00 |
| | 3089 | 1.00/1.00 | 1.00/1.00 |
| | 30590 | 1.00/1.00 | 1.00/1.00 |

FIGURE 6. Sensitivity/Specificity comparison of RVD3 with different priors.

4.3. **Results of Jeffreys prior on yeast data.** We demonstrated our RVD3 model with Jeffreys prior on yeast data to identify the variants (Kvitek and Sherlock, 2013).

## 5. Discussion

## 6. Conclusion

## Appendix A. Parameter Initialization

[This part is COPIED]

Since $r_{ji} \sim \text{Binomial}(n_{ji}, \theta_{ji})$, the first population moment is $E[r_{ji}] = \theta_{ji} n_{ji}$ and the first sample moment is simply $m_1 = r_{ji}$. Therefore the MoM estimator is

$$\tilde{\theta}_{ji} = \frac{r_{ji}}{n_{ji}} \tag{9}$$

| | | RVD3 (T=0) | | |
|---|---|---|---|---|
| | Median | Log-normal Prior | Jeffreys Prior | Improper Prior |
| MAF | Depth | FDR | FDR | FDR |
| 0.10% | 39 | | | |
| | 408 | | | |
| | 4129 | 0 | | 0 |
| | 41449 | 0.39 | 0 | 0.54 |
| 0.30% | 36 | | | |
| | 410 | | | |
| | 4156 | 0.26 | 0.2 | 0.26 |
| | 41472 | 0.77 | 0.7 | 0.8 |
| 1.00% | 53 | | | |
| | 535 | 0 | 0 | 0 |
| | 5584 | 0.26 | 0.18 | 0.33 |
| | 55489 | 0.77 | 0.71 | 0.78 |
| 10.00% | 22 | | | |
| | 260 | 0 | 0 | 0 |
| | 2718 | 0 | 0 | 0 |
| | 26959 | 0 | 0 | 0 |
| 100.00% | 27 | 0 | 0 | 0 |
| | 298 | 0 | 0 | 0 |
| | 3089 | 0 | 0 | 0 |
| | 30590 | 0 | 0 | 0 |

FIGURE 7. False Discovery Rate comparison with different priors on RVD3.

We take the MoM estimate, $\tilde{\theta}_{ji}$, as data for the next conditional distribution in the hierarchical model. The distribution is $\theta_{ji} \sim \text{Beta}(\mu_j M_j, (1-\mu_j)M_j)$. The first and second population moments are

$$E[\theta_{ji}] = \quad \mu_j, \tag{10}$$

$$\text{Var}[\theta_{ji}] = \quad \frac{\mu_j(1-\mu_j)}{M_j+1}. \tag{11}$$

The first and second sample moments are $m_1 = \frac{1}{N}\sum_{i=1}^{N}\theta_{ji}$ and $m_2 = \frac{1}{N}\sum_{i=1}^{N}\theta_{ji}^2$. Setting the population moments equal to the sample moments and solving for $\mu_j$ and $M_j$ gives

$$\tilde{\mu}_j = \quad \frac{1}{N}\sum_{i=1}^{N}\theta_{ji}, \tag{12}$$

$$\tilde{M}_j = \quad \frac{\tilde{\mu}_j(1-\tilde{\mu}_j)}{\frac{1}{N}\sum_{i=1}^{N}\theta_{ji}^2} - 1. \tag{13}$$

Following the same procedure for the parameters of $\mu_j \sim \text{Beta}(\mu_0, M_0)$ gives the following MoM estimates

$$\tilde{\mu}_0 = \quad \tfrac{1}{J}\sum_{j=1}^{J}\mu_j \tag{14}$$

$$\tilde{M}_0 = \quad \tfrac{\tilde{\mu}_0(1-\tilde{\mu}_0)}{\frac{1}{J}\sum_{j=1}^{J}\mu_j^2} - 1. \tag{15}$$

## Appendix B. Inference of Jeffreys prior

We assume there is only one replicate (i=1),

$$p\left(\theta_j\right) = \frac{\Gamma\left(M_j\right)}{\Gamma\left(\mu_j M_j\right)\Gamma\left((1-\mu_j)M_j\right)}{\theta_j}^{\mu_j M_j - 1}(1-\theta)_j^{(1-\mu_j)M_j-1} \tag{16}$$

$$\begin{aligned}\log p\left(\theta_j|\mu_j, M_j\right) &= \log\Gamma\left(M_j\right) - \log\Gamma\left(\mu_j, M_j\right)\\ &- \log\Gamma\left(1-\mu_j, M_j\right) + \left(\mu_j M_j - 1\right)\log\theta_j\\ &+ \left((1-\ u_j)M_j - 1\right)\log(1-\theta_j)\end{aligned} \tag{17}$$

$$\begin{aligned}\frac{\delta \log p(\theta_j)}{\delta M_j} &= \Psi(M_j) - \Psi(\mu_j M_j)\mu_j\\ &- \Psi((1-\mu_j)M_j)(1-\mu_j) + \mu_j\log\theta_j + (1-\mu_j)\log(1-\theta_j)\end{aligned} \tag{18}$$

$$\frac{\delta^2 \log p(\theta_j)}{\delta M_j^2} = \Psi_1(M_j) - \Psi_1(\mu_j M_j)\mu_j^2 - \Psi_1((1-\mu_j)M_j)(1-\mu_j)^2 \tag{19}$$

Now we have the Jeffreys' prior $\pi\left(M_j\right)$ for $M_j$:

$$[-\left(\Psi_1(M_j) - \Psi_1(\mu_j M_j)\mu_j^2 - \Psi_1((1-\mu_j)M_j)(1-\mu_j)^2\right)]^{\frac{1}{2}} \tag{20}$$

## Appendix C. Key parameters for RVD3 model with log-normal prior

Key parameters of the RVD3 with log-normal prior is ploted in Figure 8.

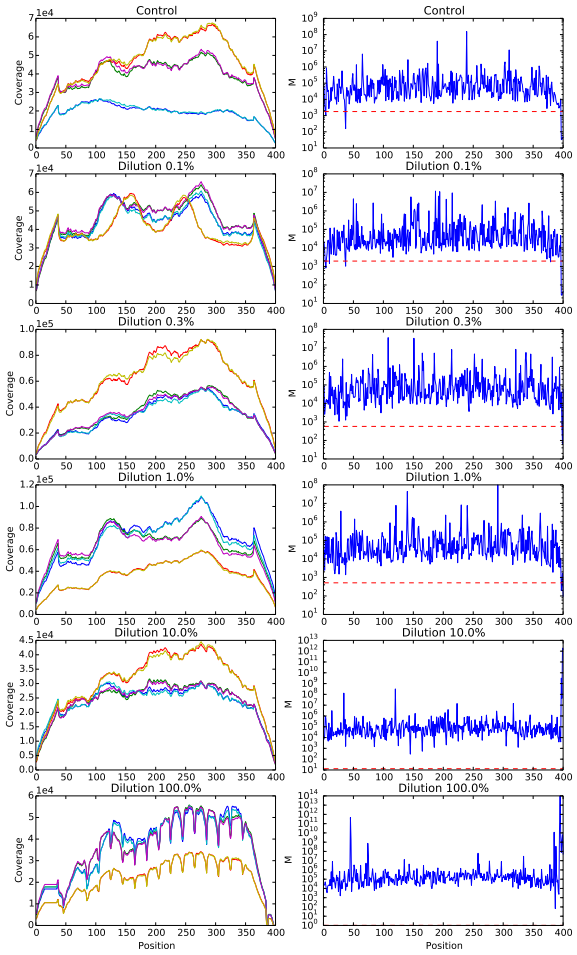


FIGURE 8. Key parameters for RVD3 model with log-normal prior for synthetic DNA data sets.

## REFERENCES

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.

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

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

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.

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*.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

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 Research*, 22(3):568–576.

Kvitek, D. J. and Sherlock, G. (2011). Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS genetics*, 7(4):e1002056.

Kvitek, D. J. and Sherlock, G. (2013). Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS genetics*, 9(11):e1003972.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian biostatistics*. Wiley. com.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

Liao, J., Mcmurry, T., and Berg, A. (2014). Prior robust empirical bayes inference for large-scale data by conditioning on rank with application to microarray data. *Biostatistics*, 15(1):60–73.

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.

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*, 28(7):907–913.

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.

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 research*, 41(7):e89–e89.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Stein, C. (1965). *Approximation of improper prior measures by prior probability measures*. Springer.

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K. B., Pao, W., and Zhao, Z. (2013). Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome*, 5(10):91.