# SUPPLEMENTARY INFORMATION FOR RVD2: AN ULTRA-SENSITIVE VARIANT DETECTION MODEL FOR LOW-DEPTH TARGETED NEXT-GENERATION SEQUENCING DATA

## 1. Parameter Initialization

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

$$\hat{\theta}_{ji} = \frac{r_{ji}}{n_{ji}} \tag{1}$$

We take the MoM estimate, $\hat{\theta}_{ji}$, 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

$$\begin{aligned} E[\theta_{ji}] = &\quad \mu_j, & (2) \\ \text{Var}[\theta_{ji}] = &\quad \tfrac{\mu_j(1-\mu_j)}{M_j+1}. & (3) \end{aligned}$$

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

$$\begin{aligned} \hat{\mu}_j = &\quad \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \hat{\theta}_{ji}, & (4) \\ \hat{M}_j = &\quad \frac{\hat{\mu}_j(1-\hat{\mu}_j)}{\frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_{ji}^2} - 1. & (5) \end{aligned}$$

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

$$\begin{aligned} \hat{\mu}_0 = &\quad \tfrac{1}{J}\textstyle\sum_{j=1}^{J} \mu_j & (6) \\ \hat{M}_0 = &\quad \frac{\hat{\mu}_0(1-\hat{\mu}_0)}{\frac{1}{J}\sum_{j=1}^{J} \mu_j^2} - 1. & (7) \end{aligned}$$

## 2. RVD2 Estimated Parameters

The RVD2 algorithm provides estimates of model parameters and latent variables given the data. We show several of these parameters in Figure 1.

The left column of Figure 1 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

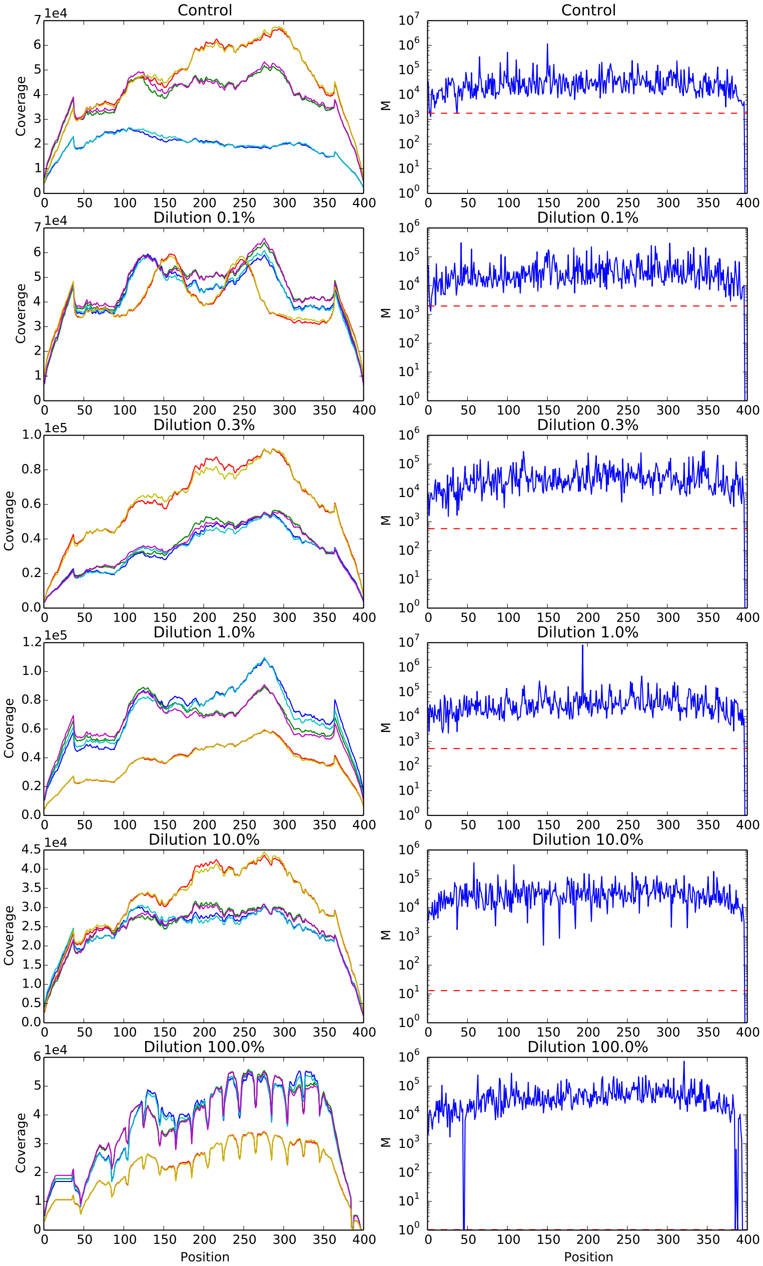

Figure 1. Key parameters for RVD2 model for synthetic DNA data sets.

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 reads are 36bp long and we required no more than 2 mismatches. Therefore, reads that overlapped two mutations (spaced 20bp apart by design) and included one additional mutation would not align.

The right column of Figure 1 shows 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. Furthermore, we see that $M_j$ does not change with read depth (except where the depth goes to zero) indicating that $M_j$ because $M_j$ is capturing a different process than the read depth.

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. We see that the variation between replicates is smaller than the variation between location. $M_j$ and $M_0$ are precision parameters, they are inversely proportional to the variance. Where $M_j$ is greater than $M_0$ the precision between replicates is higher than the precision across positions.

## 3. Algorithm Comparison Statistics

Figure 2 compares RVD2 with samtools, GATK, varscan, strelka and muTect using Matthews Correlation Coefficient (MCC) (Matthews *et al.*, 1985).

| | | | | | | | | N = 1 | | N = 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | | | **VarScan2** | **VarScan2** | | | **RVD2** | **RVD2** | **RVD2** | **RVD2** |
| **MAF** | **Depth** | **samtools** | **GATK** | **mpileup** | **somatic** | **strelka** | **MuTect** | **(T=0)** | **(T*)** | **(T=0)** | **(T*)** |
| 0.1% | 39 | | | | | | -0.02 | | | | |
| | 408 | | | | -0.00 | | 0.12 | | | | |
| | 4129 | | | | 0.03 | | 0.25 | | | 0.37 | 0.53 |
| | 41449 | | | | 0.19 | | 0.05 | 0.60 | 0.70 | 0.64 | 0.84 |
| 0.3% | 36 | | | | | | 0.60 | | | | |
| | 410 | | | | 0.14 | | 0.31 | | | | |
| | 4156 | | | | 0.04 | | 0.17 | 0.37 | 0.53 | 0.85 | 0.85 |
| | 41472 | | | | 0.16 | | 0.19 | 0.71 | 0.81 | 0.41 | 0.71 |
| 1.0% | 53 | | | | -0.02 | | 0.29 | | | | |
| | 535 | | | | 0.18 | | 0.36 | | | 0.46 | 0.46 |
| | 5584 | | | | 0.01 | | 0.41 | 0.86 | 0.90 | 0.83 | 0.96 |
| | 55489 | | | | 0.13 | -0.01 | 0.43 | 0.62 | 0.88 | 0.43 | 0.93 |
| 10.0% | 22 | 0.46 | 0.65 | | 0.59 | 0.53 | 0.79 | | | | |
| | 260 | | 0.75 | | 0.89 | 1.00 | 0.90 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 2718 | | 0.88 | | 0.16 | 0.96 | 0.79 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 26959 | | 0.75 | | 0.06 | 0.90 | 0.81 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0% | 27 | 0.93 | 0.96 | 0.96 | 1.00 | 0.96 | 0.79 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 298 | 0.93 | 0.96 | 1.00 | 0.93 | 0.90 | 0.77 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 3089 | 0.92 | 0.96 | 1.00 | 0.25 | 0.90 | 0.81 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 30590 | 0.84 | 0.96 | 1.00 | 0.15 | 1.00 | 0.85 | 1.00 | 1.00 | 1.00 | 1.00 |

Figure 2. Matthews correlation coefficient (MCC) comparison with other variant calling algorithms.

Samtools and VarScan2-mpileup achieved MCC value generally higher than 0.90 on 100% MAF sample across all read depthes, with 1.0 represents for a perfect prediction. However, both of them detected no variant when MAF is 10.0% or lower, with only one exception for samtools when MAF is 10.0% and read depth 22. GATK, Varscan2-somatic, Strelka and GATK outperformed Samtools and VarScan2-mpile on the 10.0% MAF sample, while approximately tied in other cases. Strelka achieved best MCC on 10% MAF sample comparing to Varscan2-somatic and GATK, more specifically around 1.00 when read depth is 260 or higher. There is a very obvious but unconventional decreasing trend in VarScan2-somatic MCC value across different read depth and MAF level, a phenomenon also observed by Stead *et al.*, 2013. It is because VarScan2-somatic tends to call more false positives as read depth gets higher. Mutect seems to performs the best among all the algorithms expect RVD2 when MAF is 1.0% or lower. It achieves MCC values varying from -0.02 to 0.43, though too low to be practically meaningful. However, muTect achieved relatively lower MCC values when the MAF level is 10% and 100%, as a counteractive of being oversensitive.

RVD2 achieved MCC value 1.00 when the MAF is 100.0% at all read depth and 10% when read depth is not lower than 260. This indicates that RVD2($\tau = 0, N = 1$) is more accurate than the other algorithms when the median read depth is at least 10× the MAF.

## 4. Estimated MAF for called variants in synthetic gene.

Figure 3 shows the posterior mean and 95% credible intervals for $\mu_j$ for called variant positions with $\bar{n} = 5584$ and MAF = 1.0%. All of the called positions show a clear difference between the case and control error rates. The posterior mean estimates are all shrunken towards the global error rate parameter $\mu_0 = 0.0023$ due to the hierarchical structure of the model.

## 5. Parameter settings for other variant calling algorithms

**Samtools**. We used samtools (v0.1.19) function mpileup to call variants and bcftools to save the result in standard VCF files. In mpileup, we set the -d option sufficiently high at $10^6$ to avoid truncating read deapth. Option -u was enabled to make sure the output bcf files were uncompressed.

**GATK**. We used GATK (v2.1-8) UnifiedGenotyper function to detect mutations on our synthetic data following the recommended workflow. Due to some format incompatibility, we applied Picard to format read group and GATK for realignment. In UnifiedGenotyper, -ploid (Number of samples in each pool × Sample Ploidy) was set at 1 because our synthetic data is haploid; -dcov was set at $10^6$ to avoid downsampling coverage within GATK.

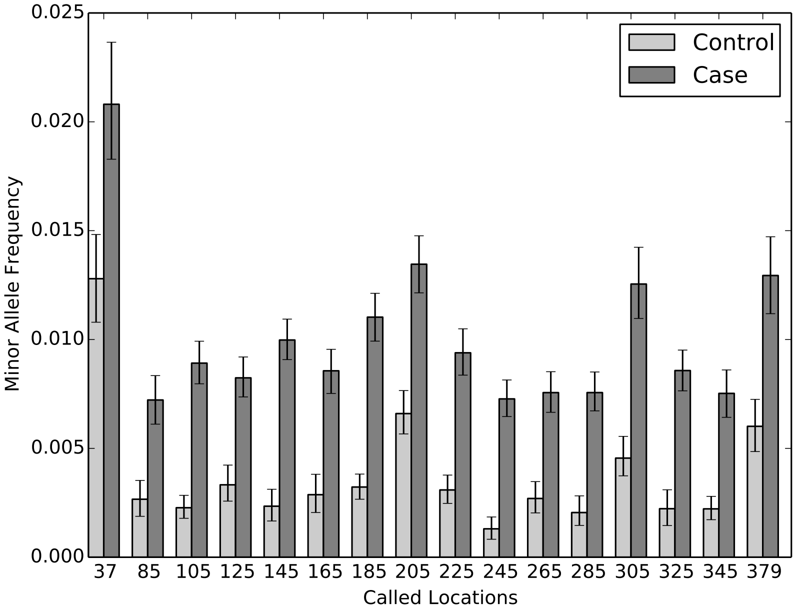

Figure 3. Estimated minor allele fraction for called variants in 1.0% dilution.

**VarScan2-mpileup**. VarScan2 (v2.3.4) mpileup2snp is a SNP calling program which takes multi-samples from samtools mpileup pipeline.We assigned parameter -C value 50 as the synthetic data was aligned using BWA and set -d at $10^6$. In mpileup2snp, –min-var-freq, the only non-default parameter, was set low enough at $10^{-5}$ because the variant frequency can be as low as $10^{-3}$.

**VarScan2-somatic**. We tested VarScan2 somatic on our synthetic dataset. The parameter –normal-purity set was at 1.00, –tumor-purity at the dilution rate. The parameter –min-var-freq was set at $10^{-5}$. We combined all the positions VarScan2-somatic called regardless the somatic status (Germline/LOH/Somatic/Unknown) to compare with performance of RVD2.

**Strelka and muTect**. Since configuration and Analysis for Strelka and muTect is standardized and no parameter needs to be specified, we installed these two programs and ran them on our data set separately.

Samtools mpileup and GATK can accept multiple ”tumor” replicates for variant calling, so we fed six bam files from each case replicate group to mpileup. VarScan2-mpileup takes multiple “tumor-normal” pair replicates so we passed six pair replicates to each algorithm. Varscan2-somatic, strelka and muTect do not accept replicate data for the “normal” or "tumor" bam files so we used a single bam file from each replicate group with a read depth that most closely matched the overall median depth of the replicates.

## References

Matthews, D., Hosker, J., Rudenski, A., Naylor, B., Treacher, D., and Turner, R. (1985). Homeostasis model assessment: insulin resistance and $\beta$-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*, **28**(7), 412–419.

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.