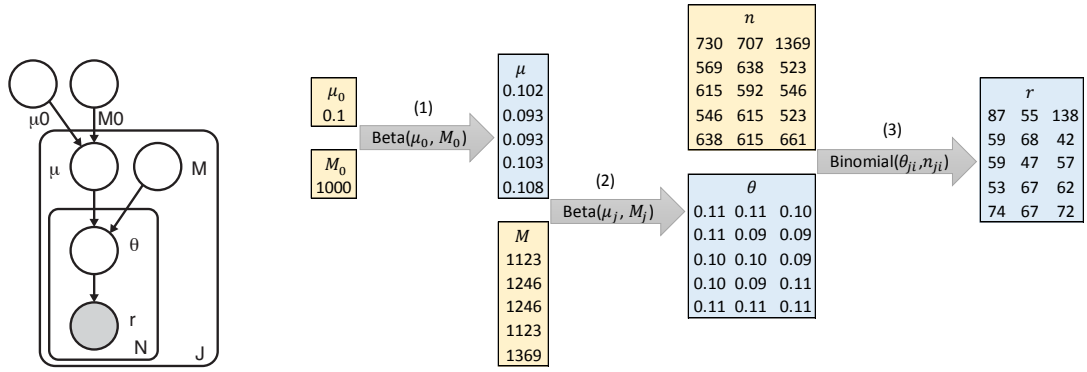
SUPPLEMENTARY INFORMATION FOR RVD2: AN ULTRA-SENSITIVE VARIANT DETECTION MODEL FOR LOW-DEPTH HETEROGENEOUS NEXT-GENERATION SEQUENCING DATA

1. An simulation example to illustrate the proposal graphical model and generative process.



(A) (B)

FIGURE 1. (A) RVD2 Graphical Model.(B) An simulation example to illustrate the proposal graphical model and generative process. A generative process is to generate randomly observable data following your generative model, which is the reverse process of doing inference from the actual data. Here we want to generate a simulation sample with 3 replicates, and each replicate is a 5bp long sequence. Therefore, we have $J = 5, N = 3$. The numbers in orange are known, including $\mu_0 \in \mathbb{R}^{1\times1}$, $M_0 \in \mathbb{R}^{1\times1}$, $M \in \mathbb{R}^{5\times1}$, $n \in \mathbb{R}^{5\times3}$. (1) We sample local error rate $\mu \in \mathbb{R}^{5\times1}$ by drawing random numbers $\mu_j \sim \text{Beta}(\mu_0, M_0)$; (2) and then use μ and M to sample $\theta \in \mathbb{R}^{5\times3}$ by drawing random samples $\theta_{ji} \sim \text{Beta}(\mu_j, M_j)$; (3) finally we use $\theta \in \mathbb{R}^{5\times3}$ and $n \in \mathbb{R}^{5\times3}$ to sample $r \in \mathbb{R}^{5\times3}$ by drawing random numbers $r_{ji} \sim \text{Binomial}(\theta_{ji}, n_{ji})$. In actual practice, non-reference read counts r and coverage n are the data we use to do inference.

2. 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

$$E[\theta_{ji}] = \quad \mu_j, \tag{2}$$

$$\text{Var}[\theta_{ji}] = \quad \tfrac{\mu_j(1-\mu_j)}{M_j+1}. \tag{3}$$

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 μ_j and M_j gives

$$\hat{\mu}_j = \quad \tfrac{1}{N}\sum_{i=1}^{N}\hat{\theta}_{ji}, \tag{4}$$

$$\hat{M}_j = \quad \tfrac{\hat{\mu}_j(1-\hat{\mu}_j)}{\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_{ji}^2} - 1. \tag{5}$$

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

$$\hat{\mu}_0 = \quad \tfrac{1}{J}\sum_{j=1}^{J}\hat{\mu}_j \tag{6}$$

$$\hat{M}_0 = \quad \tfrac{\hat{\mu}_0(1-\hat{\mu}_0)}{\frac{1}{J}\sum_{j=1}^{J}\hat{\mu}_j^2} - 1. \tag{7}$$

3. 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 2.

The left column of Figure 2 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 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 2 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

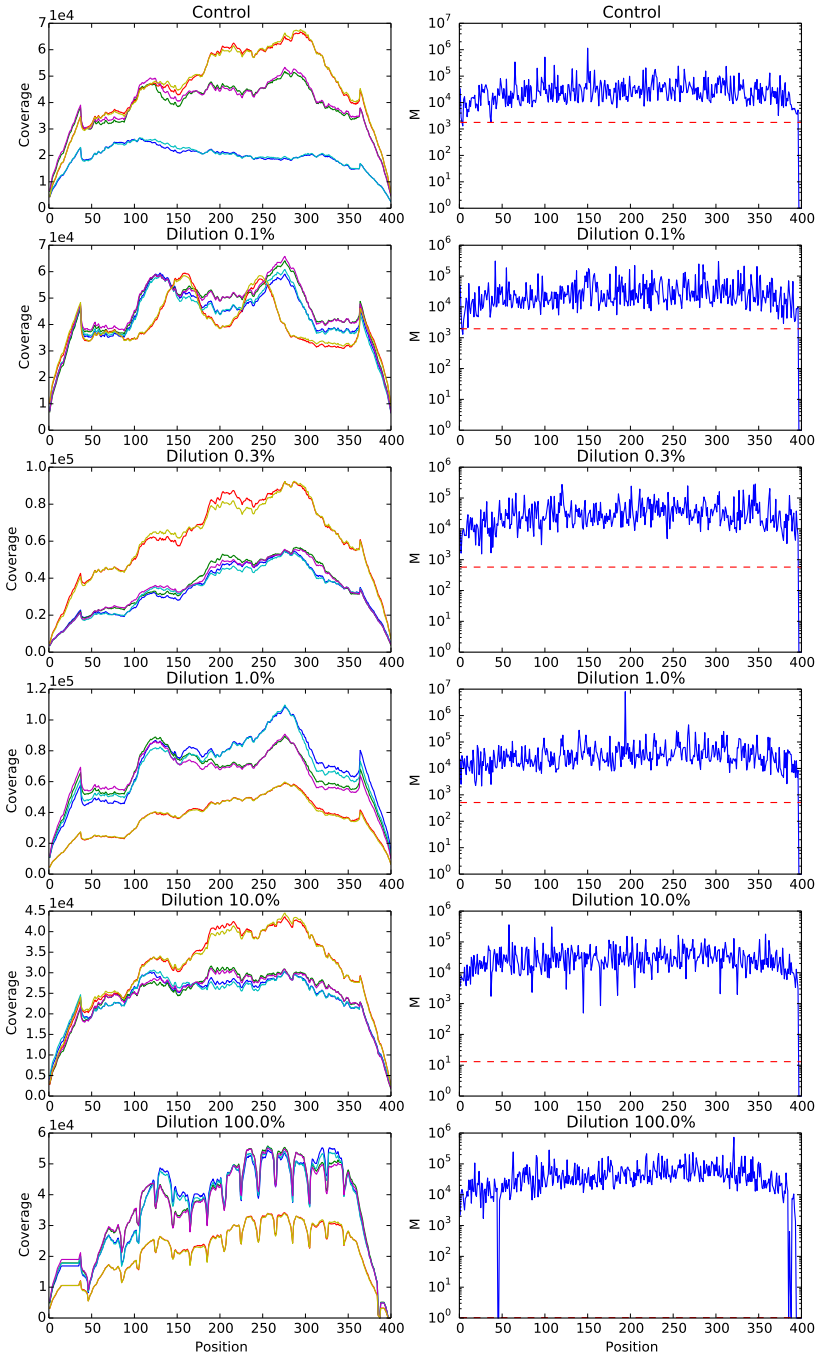


FIGURE 2. Key parameters for RVD2 model for synthetic DNA data sets.

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.

4. Algorithm Comparison Statistics

Figure 3 compares RVD2 with samtools, GATK, varscan, strelka and muTect using Matthews Correlation Coefficient (MCC) (Matthews *et al.*, 1985).

									N=1		N=6	
MAF	Median Depth	SAMtools	GATK	VarScan2 mpileup	VarScan2 somatic	Strelka	MuTect	RVD	RVD2 (T=0)	RVD2 (T*)	RVD2 (T=0)	RVD2 (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	0.98	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	0.95	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	0.9	0.62	0.88	0.43	0.93
10.0%	22	0.46			0.59	0.53	0.79					
	260				0.89	1.00	0.90		1.00	1.00	1.00	1.00
	2718				0.16	0.96	0.79		1.00	1.00	1.00	1.00
	26959				0.06	0.90	0.81	0.82	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	0.83	1.00	1.00	1.00	1.00

Figure 3. 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 $10\times$ the MAF.

5. Estimated MAF for called variants in synthetic gene.

Figure 4 shows the posterior mean and 95% credible intervals for μ_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.

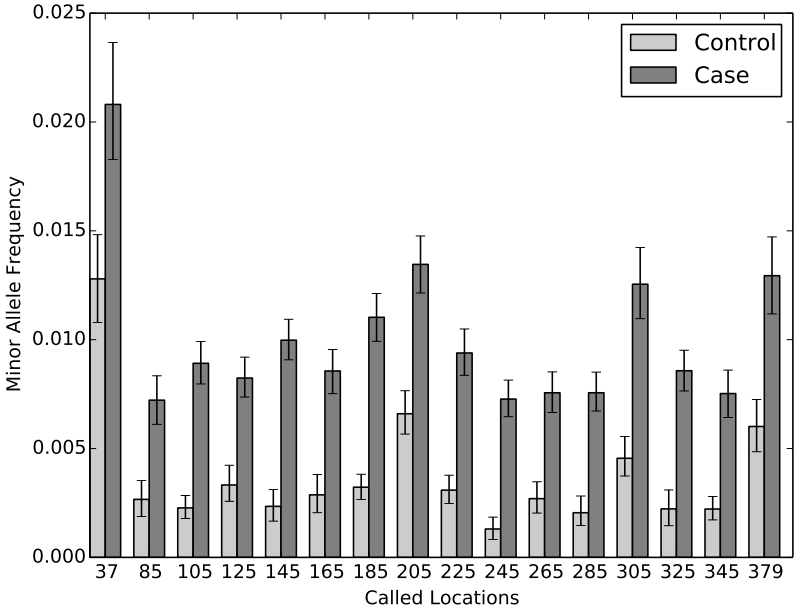


Figure 4. Estimated minor allele fraction for called variants in 1.0% dilution.

6. 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 $\times$ Sample Ploidy) was set at 1 because our synthetic data is haploid; -dcov was set at 10^6 to avoid downsampling coverage within GATK.

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.

7. HISTOGRAM OF $\hat{\mu}_j$ FOR SOMATIC MUTATIONS CALLED BY RVD2 IN THE 44KBP PAXIP1 GENE FROM CHR7:154738059 TO CHR7:154782774 IN HCC1187 DATASET

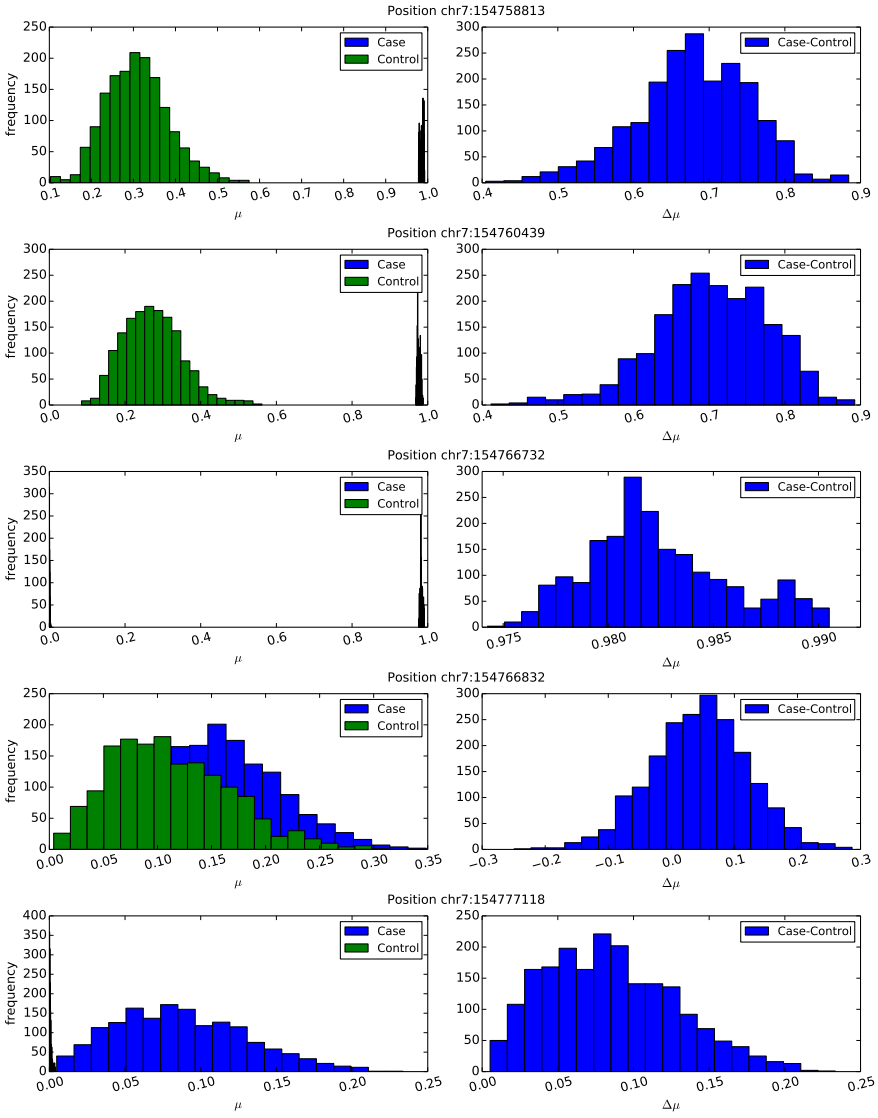


FIGURE 5. Histogram of $\hat{\mu}_j$ for positions where μ^{case} is significantly higher than μ^{control}, namely $\Pr(\Delta\hat{\mu}_j > \tau) > 1 - \alpha$, where $\tau = 0, \alpha = 0.05$.

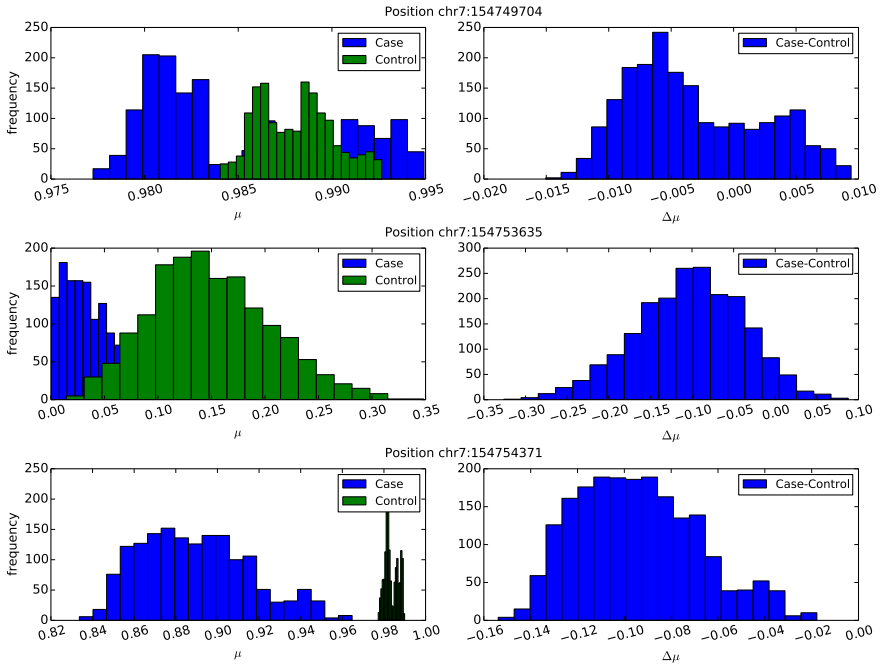


FIGURE 6. Histogram of $\hat{\mu}_j$ for positions where μ^{case} is significantly lower than μ^{control}, namely $\Pr(\Delta\hat{\mu}_j < \tau) > 1 - \alpha$, where $\tau = 0, \alpha = 0.05$.

Supplemetary Table 1. This table shows the sum of positions called by VarScan Somatic, RVD2, muTect and Strelka in 44kbp PAXIP1 gene from chr7:154738059 to chr7:154782774. The first column is the position index corresponding to Fig 6 in the paper. The second column is the actual position in PAXIP1 gene with respect to index. The Third column is the reference base. The forth and fifth columns are the depth matrix for Normal and Case sample repectively. The last four columns are the mutation detection status from the four algorithms. Blank cell indicates this position is not called by the algorithm in the header.

Index	Actual Position	REF	Control	Case	VarScan2 Somatic	RVD2	muTect	Strelka
Format			**A:C:G:T**	**A:C:G:T**				
1	chr7:154739981	T	0:0:0:50	0:0:2:58	Somatic			
2	chr7:154740723	T	0:0:0:60	0:0:2:69	Somatic			
3	chr7:154740848	A	37:0:0:0	56:2:0:0	Somatic			
4	chr7:154743899	C	0:0:0:45	0:0:0:71	Germline	Germline	Call	
5	chr7:154745094	G	0:0:47:0	2:0:68:0	Somatic			
6	chr7:154747295	T	0:0:0:41	0:0:2:81	Somatic			
7	chr7:154749452	T	0:0:3:29	0:0:2:37	Germline			
8	chr7:154749704	G	8:0:36:0	2:0:42:0	Germline	Germline/Somatic	Call	
9	chr7:154751731	A	53:0:0:0	73:2:0:0	Somatic			
10	chr7:154751818	T	0:0:0:57	0:1:4:75	Somatic			
11	chr7:154753635	T	0:38:0:0	0:63:0:1	Germline	Germline	Call	
12	chr7:154754218	G	0:0:49:0	0:0:65:2	Somatic			
13	chr7:154754371	T	0:19:0:31	0:61:0:0	LOH	LOH	Call	
14	chr7:154754860	T	0:0:0:49	0:0:2:61	Somatic			
15	chr7:154755810	T	0:0:0:49	0:2:0:58	Somatic			
16	chr7:154755836	A	57:0:0:0	64:3:0:0	Somatic			
17	chr7:154757232	A	40:0:0:0	54:2:0:0	Somatic			
18	chr7:154757503	T	0:0:0:50	0:0:2:69	Somatic			
19	chr7:154757988	A	57:0:0:0	69:3:0:0	Somatic			
20	chr7:154758813	G	14:0:28:0	54:0:0:0	LOH	LOH	Call	
21	chr7:154759833	T	0:0:0:48	0:0:2:62	Somatic			
22	chr7:154760439	A	37:0:0:0	0:38:0:0	Somatic	Somatic	Call	Somatic
23	chr7:154760538	G	0:0:62:0	0:0:42:2	Somatic			
24	chr7:154760592	T	0:0:0:47	0:0:2:38	Somatic			
25	chr7:154760658	T	0:0:0:52	0:0:2:47	Somatic			
26	chr7:154760790	A	55:0:0:0	66:2:0:0	Somatic			
27	chr7:154762296	T	0:0:0:51	0:0:2:74	Somatic			
28	chr7:154762409	T	0:0:0:32	0:0:2:51	Somatic			
29	chr7:154762415	A	27:2:0:0	47:3:0:0	Germline			
30	chr7:154762957	T	0:0:0:37	2:0:0:54	Somatic			
31	chr7:154763136	A	61:0:0:0	62:0:2:0	Somatic			
32	chr7:154763337	T	0:0:0:43	0:0:2:80	Somatic			
33	chr7:154766365	A	53:0:0:0	57:2:0:0	Somatic			
34	chr7:154766388	A	47:0:0:0	66:2:0:0	Somatic			
35	chr7:154766700	C	4:21:0:0	10:42:0:0		Germline	Call	
36	chr7:154766732	T	0:0:0:35	0:0:5:44	Somatic	Somatic	Call	
37	chr7:154766832	A	34:0:0:0	42:4:0:0	Somatic	Somatic		

Index	Actual Position	REF	Control	Case	VarScan2 Somatic	RVD2	muTect	Strelka
38	chr7:154766834	T	0:0:0:34	0:2:0:43	Somatic			
39	chr7:154766857	A	29:0:0:0	36:2:0:0	Somatic			
40	chr7:154767108	T	0:0:0:55	0:0:1:64	Somatic			
41	chr7:154767456	T	0:0:0:40	0:0:2:74	Somatic			
42	chr7:154767801	A	47:0:0:0	67:2:0:0	Somatic			
43	chr7:154768640	T	0:0:0:66	0:0:2:97	Somatic			
44	chr7:154769094	G	0:0:46:0	0:0:63:2	Somatic			
45	chr7:154771488	T	0:0:0:50	0:0:2:60	Somatic			
46	chr7:154772261	A	47:0:0:0	71:2:0:0	Somatic			
47	chr7:154775220	A	47:0:0:0	84:2:1:0	Somatic			
48	chr7:154775236	T	0:0:0:47	0:0:2:88	Somatic			
49	chr7:154775872	A	55:0:0:0	71:2:0:1	Somatic			
50	chr7:154777014	C	0:49:0:0	0:64:2:0	Somatic		Call	
51	chr7:154777118	A	46:0:0:0	31:4:0:0	Somatic	Somatic	Call	
52	chr7:154777687	G	0:0:45:0	0:0:50:2	Somatic			
53	chr7:154777955	T	0:0:0:52	0:0:2:47	Somatic			
54	chr7:154778892	A	55:0:0:0	44:2:0:0	Somatic			
55	chr7:154779690	A	45:0:0:0	52:2:0:0	Somatic			
56	chr7:154780960	C	0:0:0:42	0:0:0:56	Germline	Germline	Call	
57	chr7:154781700	A	55:3:0:0	82:2:0:0	Germline			
58	chr7:154781769	G	0:0:21:4	0:0:31:2	LOH	Germline		
59	chr7:154781944	T	0:0:0:55	0:0:2:47	Somatic			
60	chr7:154782120	A	47:0:0:0	42:2:0:0	Somatic			
61	chr7:154782770	A	58:0:0:0	91:2:0:0	Somatic			

References

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

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics*, **32**, 496–501.

Robasky, K., Lewis, N. E., and Church, G. M. (2013). The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*.

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.