Statistical Scores for Rare Variant Calls in Ultra-deep Sequencing

Wei-min Liu[1]
[1]Roche Molecular Systems, Inc., 4300 Hacienda Dr, Pleasanton, CA 94588

Abstract

NGS technology is changing bio-medical sciences from academic research to clinical diagnostics. For non-invasive blood-based tests, it is crucial to distinguish the rare variants from sequencing noise in ultra-deep sequencing. Several approaches have been developed to make the variant calls reliable. They include base call quality scores, unique molecular identifiers, etc. Most software packages only call variants of 1% or higher by their default setting to avoid false positives. I describe the new variant quality scores based on the distribution of false positives in sequencing, as well as the fact that the false positive rates are dependent on the sequence contexts and locations. With statistical tests based on these considerations, we can detect variants with percentages significantly below 1% (depending on variants, sample types and sequencing protocols) when sufficient number of DNA molecules is present in the sample. I also describe a non-standard usage of MiSeq Reporter (MSR) to verify the low-frequency variants we found using this approach.

Key Words: non-invasive test, sequencing, variant, ultra-deep sequencing

1. Introduction

This paper proposes to establish quality scores of variant calls based on statistical tests to distinguish low frequency true mutations from the false positive sequencing noise in ultra-deep sequencing (UDS) experiments. Next generation sequencing (NGS) can be a reference method for PCR-based diagnostic tests to detect mutations of cancer or other diseases, and can also be developed as IVD tests. A number of recent studies have shown that it may be valuable to develop NGS-based applications as non-invasive blood tests[1,2]. One challenge for blood-based detection of cell free circulating tumor mutations is that the frequency of biomarker mutation is often quite low. Therefore, UDS will almost certainly be required. Many software packages report only 1% or higher variants to avoid high false positive rates[4]. However, these thresholds are too high for blood-based tests where frequencies of 0.01-0.1% have been reported[2,3]. In order to distinguish low frequency true mutations from noise, we can use the observed distribution of false positive variants to establish the variant calling quality scores to determine how likely a variant is a true mutation and hence to establish more accurate thresholds for low frequency variants.

The false positive rates of variant calls depend on the sequence context and location. Therefore, we may compare mutation and wild type counts of a variant at the same location in different samples. This method is especially useful if there are wild type (usually normal) samples as negative control in the sequencing run. We use the chi-squared test of a 2 x 2 table to compare different samples. We use the symbol QS to denote this quality score (S stands for special variants at the same location).

We may also compare variants of the same type, such as T>C, at different locations in all samples. We use the model based on distribution of logarithmic frequencies for this type of comparison. We use the symbol QM to denote this quality score based on the model.

We also use QB to denote the quality score of a base call used in the FASTQ files

2. Variant calling quality score QS based on 2x2 table of observed counts

This method can be applied to very few, such as two, data points. We set a 2x2 table as follows.

Table 1: The 2x2 table of reference sample and sample in question

	Reference sample	*Sample in question*	*Row sum*
Variant count	a_1	a_2	a
Wild type count	w_1	w_2	w
Total count (depth)	n_1	n_2	n

The reference sample represents the data theoretically without true mutation (but sequencing noise can generate small variant count) and the column of sample in question represents the data we want to determine whether its variant frequency is significantly larger than the reference frequency. The value a_1 is the count of a particular variant in reference, n_1 is the depth of the reference, and $w_1 = n_1 – a_1$ represents the wild type count. The values a_2, w_2 and n_2 are defined similarly for the sample in question. We define the row sums $a = a_1 + a_2$, $w = w_1 + w_2$ and define $n = n_1 + n_2$ as the total counts of the 2x2 table.

There are many ways to test whether (a_1, w_1) and (a_2, w_2) are significantly different in their ratios (odds). Because the values of n_1 and n_2 can be very large for UDS, we propose to use the one-sided chi-squared test. First, we calculate the proportions: $f_1 = a_1 / n_1$ and $f_2 = a_2 / n_2$. If $f_2 <= f_1$, i.e., the proportion of sample in question is not higher than the proportion of reference (false positive), we can set the quality score to a very small number such as 2 (corresponding to error rate $p = 0.63$), and there is no need for further computation. If $f_2 > f_1$, we calculate the chi squared statistic:

$$\chi^2 = n * (a_1 * w_2 – a_2 * w_1)^2 / (n_1 * n_2 *a * w).$$

The one-sided p-value is $p = 0.5 * (1 – \text{pchisq}(\chi^2, d))$, where pchisq is the chi-squared cumulative distribution function with degrees of freedom $d = 1$. The corresponding quality score can be defined as QS = $-10 * \log_{10}(\max(p, minP))$. Note that p is in the range of (0, 0.5). To avoid the difficulty of numerical computation when p is close to 0, we use $minP = 10^{-13}$, which is equivalent to set $maxQ = 130$. We use two different methods to set the reference counts for a particular variant in a sequencing run of multiple samples. One method is to use the sum of variant counts and the sum of the depths of the two samples with the lowest frequencies and enough depth (depth >= *minD*, and we may set *minD* = 3000) of the variant. To avoid the rare possibility that all samples have high frequency of the particular variant, when the reference proportion > f_0 (e.g., f_0 = 0.01 = 1% or smaller), we set $a_1 = f_0 * n_1$, i.e., the used a_1 value is either the true a_1 value or $f_0 * n_1$ whichever is smaller.

We may also use known wild type sample as a reference. A problem of this approach is that if wild type samples are contaminated with the variant, then the quality scores QS of other samples will be low. With the approach we described above, the wild type sample with variant contamination will show high QS (hence not likely wild type for the particular variant) and the QS of other samples are usually not influenced.

It is time consuming to directly calculate *p* and then quality score for every variant. Since we only report quality scores as an integer, we can discretize the value of QS, e.g., when $f_2 <= f_1$, we set QS = 2, when $f_2 > f_1$, we allow QS to be 3, 4, …, *maxQ* (we set *maxQ* = 130). Therefore, we calculate the χ^2 values for Q = 3.5, 4.5, …, 129.5 (Table 2 lists a part of the whole table) and use binary search to determine the best approximate integer value of Q in 3, 4, …, 130.

Table 2: Partial List of Q, qnorm and qchisq at half integers

Q	**qnorm**	**qchisq**	**Q**	**qnorm**	**qchisq**	**Q**	**qnorm**	**qchisq**
3.5	0.1340	0.0180	16.5	2.0068	4.0271	29.5	3.0559	9.3384
4.5	0.3724	0.1387	17.5	2.1019	4.4178	30.5	3.1243	9.7610
5.5	0.5774	0.3334	18.5	2.1938	4.8127	31.5	3.1914	10.1850
6.5	0.7592	0.5764	19.5	2.2828	5.2113	32.5	3.2573	10.6102
7.5	0.9237	0.8532	20.5	2.3692	5.6133	33.5	3.3221	11.0365
8.5	1.0747	1.1550	21.5	2.4532	6.0182	34.5	3.3858	11.4639
9.5	1.2149	1.4760	22.5	2.5349	6.4259	35.5	3.4485	11.8923
10.5	1.3462	1.8122	23.5	2.6146	6.8360	36.5	3.5102	12.3216
11.5	1.4699	2.1606	24.5	2.6923	7.2484	37.5	3.5710	12.7519
12.5	1.5872	2.5192	25.5	2.7682	7.6629	38.5	3.6308	13.1830
13.5	1.6989	2.8863	26.5	2.8424	8.0793	39.5	3.6898	13.6148
14.5	1.8057	3.2606	27.5	2.9150	8.4974	40.5	3.7480	14.0475
15.5	1.9082	3.6412	28.5	2.9862	8.9171	41.5	3.8054	14.4809

3. Variant calling quality score QM based on frequency distribution

In case we have multiple data points, we can use the normal approximation to make statistical inference. We first make a logarithmic transformation of frequency. Let the original frequency be $f = a / n$, where *a* is the mutant read count and *n* is the total read count. The logarithmic transformation we use is

$$x = \log_{10}(f + e),$$

where e=10^{-6} is the constant to avoid the negative infinity value when $f = 0$.

We divide the simple variants in a sequencing run at different locations into 19 classes. In every class, the majority of the calls are false positive, and we can calculate the parameters of the distribution of every class. The 19 variant classes are as follows
(a) 12 single-base substitutions (A>C, A>G, A>T, C>A, C>G, C>T, G>A, G>C, G>T, T>A, T>C and T>G),
(b) multiple-base substitutions,
(c) deletions of 1-2 bases, 3 bases, 4-5 bases, 6 or more bases,
(d) insertions
(e) other simple variants such as a substitution followed by a deletion.

We can use the normal approximation to calculate the quality score QM. For an observed variant in a class with frequency f_1, $x_1 = \log_{10}(f_1 + e)$, with sufficient depth (total count n_1), we can calculate a statistic similar to z-score

$z = (x_1 - m) / (s / \mathrm{sqrt}(n))$,

where m and s are the sample mean and sample standard deviation. and n is the number of reference data points used in estimation of s and m. Our calculation indicates that the standard z score is large for large n, and it can generate very small p-value and hence too large quality score. Therefore, we calculate the z-like statistic, i.e, use min(n, N) to replace n in the above formula. We set $N = 36$. We also set a lower bound s_2 (with default value 0.01), for s / sqrt(min(n, N), to handle the situation where s is too small. Our z-like statistic can be expressed as

$z' = (x_1 - m) / \max(s_2,\ s / \mathrm{sqrt}(\min(n, N)))$.

We calculate $p = 1 - F(z')$, where F is the cumulative distribution function of the standard normal. The variant calling quality score, QM, is defined as the Phred-like score

$\mathrm{QM} = -10 \log_{10}(\max(p, minP))$

as we set before $minP = 10^{-13}$, which implies that $maxQ = 130$.

We may also use robust estimation for central position and variation of data instead of sample mean and sample deviation to calculate the quality scores.

4. Application and Verification

The two types of quality scores defined above help us to determine the threshold of variant frequencies as the limit of detection. We can successfully detect substitutions with 0.1-0.7% frequency given sufficient input DNA amount for the Illumina MiSeq system. Since the false positive rate depends on mutation context and location, for particular substitutions at particular location, we may even detect variant at 0.03%.

For moderate size insertions, deletions and complex mutations, such as a 15-base deletion, it is almost impossible to generate this type of mutations randomly in sequencing, and the main source of false positive is the carry-over contamination from other samples. With well-established washing protocol between runs, we may detect as low as 0.0025% of such variants.

We also used MiSeq Reporter in a non-standard way to verify the low frequency variants we report. MiSeq Reporter (MSR) uses a somatic variant caller with a built-in Poisson model to report low-frequency variants[4]. The lowest frequency that MSR reports is depth-dependent and with default settings it is above 1%.

We ran MSR with the sequence containing known variants as the reference and make MSR to report the wild type as a "variant" of this reference. MSR also reports the actual variant as "wild type". In this way, we can verify our own variant calls. This non-standard usage of MSR has the following disadvantages. (1) It can only be used to verify known variants. (2) The variant calling quality score that MSR reports is for the wild type and is not for actual variant. (3) When there are multiple overlapping known variants, it becomes tedious or difficult to use this method. However, it can be used as a verification method for known variants after the above concerns are considered.

Acknowledgements

The author thanks Yan Li for helpful discussion, usage of the proposed approaches, and suggestions on revision of the paper. The author also thanks Mari Christensen, Janet Jin, Julie Tsai, and Wei Wen for providing data. He also thanks Eugene Spier for revision suggestions.

References

1. Kidess, E. and S.S. Jeffrey (2013), Circulating tumor cells versus tumor-derived cellfree DNA: rivals or partners in cancer care in the era of single-cell analysis? Genome. Med., 5:70, doi:10.1186/gm474.

2. L.A. Diaz Jr and A. Bardelli (2014), Liquid biopsies: genotyping circulating tumor DNA, J. Clin. Oncol., 32:579-586.

3. F. Diehl1, K. Schmidt, M.A. Choti, K. Romans, S. Goodman, M. Li, K. Thornton, N. Agrawal, L. Sokol, S.A. Szabo, K.W. Kinzler, B.Vogelstein, and L. A. Diaz Jr (2008), Circulating mutant DNA to assess tumor dynamics, Nat Med., 14:985–990.

4. Illumina Inc., Somatic Variant Caller, Pub No. 970-2012-014.