Variant Call Differences between Paired Reads in Illumina Amplicon Sequencing: An Issue for Rare Mutation Detection

Wei-min Liu[1], Julie Chen[2]
[1] Roche Molecular Systems, Inc., 4300 Hacienda Drive, Dublin, CA 94588
[2] Department of Statistics, University of California Berkeley, Berkeley, CA 94720

Abstract
Deep amplicon sequencing is crucial for finding rare somatic mutations from liquid biopsy or formalin-fixed paraffin-embedded tissues (FFPET). It can be used as a reference method for developing PCR assay or used directly as diagnosis assay for individualized health care based on mutation detection. It is important to extract actual mutation signals from sequencing noise. Using the RMS proprietary software SWISSA, we studied a special type of sequencing noise appearing in one of the paired reads but not both in Illumina amplicon sequencing. We found that the variants occurring in read 2 but not read 1 are often higher than those occurring in read 1 but not read 2. We also calculated the proportion of this type of noise in all observed variants. We compared the results of using paired-end reads and using only single-end reads for mutation frequencies in a dataset of dilution experiments. We found that the two methods generated similar results and using paired-end reads with removal of this type of sequencing noise may be slightly better than using single-end reads. Our proposed metrics R1Not2 and R2Not1 in VPKH could be used to optimize the experimental conditions in assay development.

Key Words: FFPET, liquid biopsy, mutation detection, deep amplicon sequencing

1. Introduction

Next generation sequencing (NGS) has become a useful tool in biological research, medical diagnosis, oncology, virology and many other areas. Deep amplicon sequencing is important for finding somatic mutations on oncogenes from liquid biopsy or formalin-fixed paraffin-embedded tissues (FFPET). Most software packages only call variants of 1% or higher by their default setting to avoid false positives[1]. The mutation frequencies in blood-based tests can be as low as 0.01-0.1%[2,3]. Thus, distinguishing actual rare mutations from sequencing noise remains a challenge to researchers and sequencing device makers. We studied a special type of sequencing noise of Illumina amplicon sequencing. Our work shows that we can reduce this type of noise using proper algorithms proposed[4-6] and implemented in the SWISSA (SWIft and Succinct Sequencing Analyzer) software.

2. Method

We worked with Illumina MiSeq paired-end amplicon sequencing data, specifically analyzing the sequencing noise of variant call differences between read 1 and read 2 for the EGFR, ESR1, BRAF, KRAS, NRAS, and PIK3CA genes. We examined total 128

Illumina runs of 5 assays: 64 EGFR runs, 5 ESR1 runs, 32 KRAS runs, 10 runs of the 2-gene panel (BRAF and NRAS), 17 runs of the 5-gene panel (BRAF, EGFR, KRAS, NRAS and PIK3CA). Sequencing noise can appear in both reads. If one read calls differently from the other in a pair, there must be noise in at least one of them.

Figure 1 shows 5 cases of variants in a read pair (cluster). The top horizontal bar (in light brown color) represents read 1, and the second bar (in light green color) represents read 2. Case (a) indicates that read 2 is too short to cover the mutation due to trimming bad quality bases and the mutation is only observed in read 1. Case (b) indicates that read 1 is too short to cover the mutation and the mutation is only observed in read 2. Case (c) shows that the mutation is detected in both read 1 and read 2. Case (d) shows that the mutation is detected in read 1 but not in read 2 (R1NotR2, or simply R1Not2). Case (e) shows that the mutation is detected in read 2 but not in read 1 (R2NotR1, or R2Not1). SWISSA uses read pair as a basic unit and only tally mutations in the first three cases (R1 only, R2 only, or R1AndR2). R1Not2 and R2Not1 are reported, but not used for mutation tally.

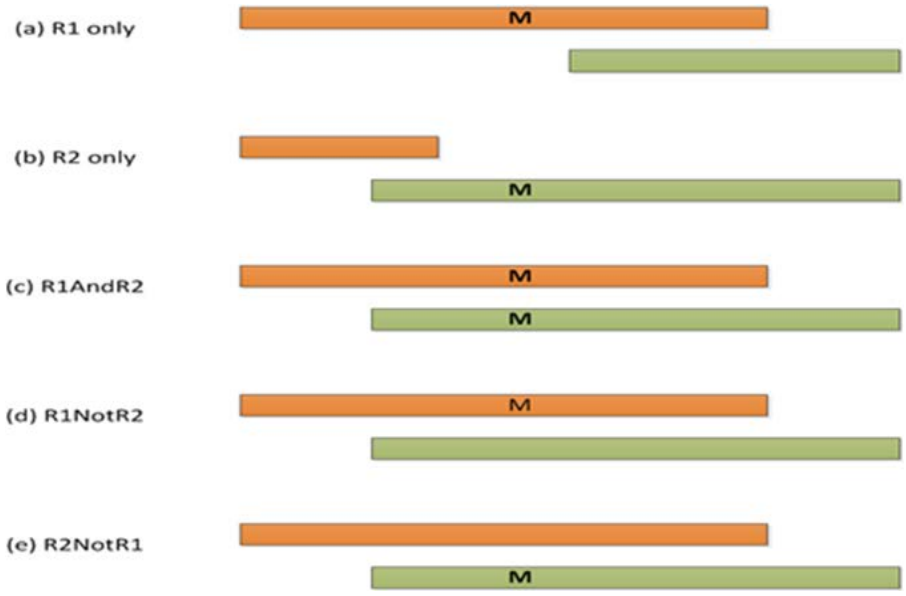

Figure 1: Five cases of a mutation in read pairs (clusters): (a) R1 only (R2 is too short to cover the mutation); (b) R2 only (R1 is too short to cover the mutation); (c) Mutation is observed in both R1 and R2; (d) Mutation appears in R1 but not in R2; (e) Mutation appears in R2 but not in R1

In this study, we counted the variants in read 1 but not in read 2 (R1Not2) and the variants in read 2 but not in read 1 (R2Not1) and normalized them in the unit of number of variants per kilobase per hundred mapped read pairs (VPKH). Figure 2 shows our results by 5 different assays. We see that the normalized statistic R2Not1 is consistently higher than R1Not2. This observation could be explained as read 1 is often more accurate than read 2 and the wild type reads are usually more than the mutant reads.

Furthermore, we summarized the differences by variant type. We considered 16 total variant types including 12 single substitutions, multiple substitutions (ms), deletions (del), insertions (ins), and complex mutations (x) such as a substitution followed

immediately by an insertion or a deletion. We also calculated the normalized counts in VPKH. Figures 3a and 3b show the results for substitutions and other mutations (because they are of different scale). Since the values of VPKH for insertions, deletions and complex mutations are significantly smaller than in those for substitutions, we split the figure to 3a and 3b so that we can use different ranges of VPKH. We see that for most mutations R1Not2 is smaller than R2Not1, but for the substitution C>T R1Not2 is slightly larger than R2Not1.

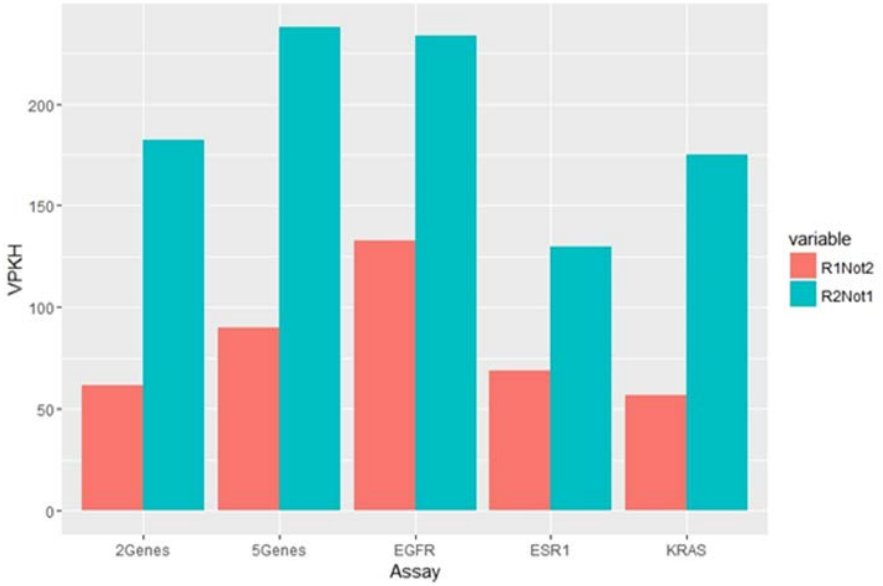

Figure 2: Numbers of variants in read 1 but not in read 2 (R1Not2) and in read 2 but not in read 1 (R2Not1) per kilobase per hundred mapped read pairs (VPKH) by five assays

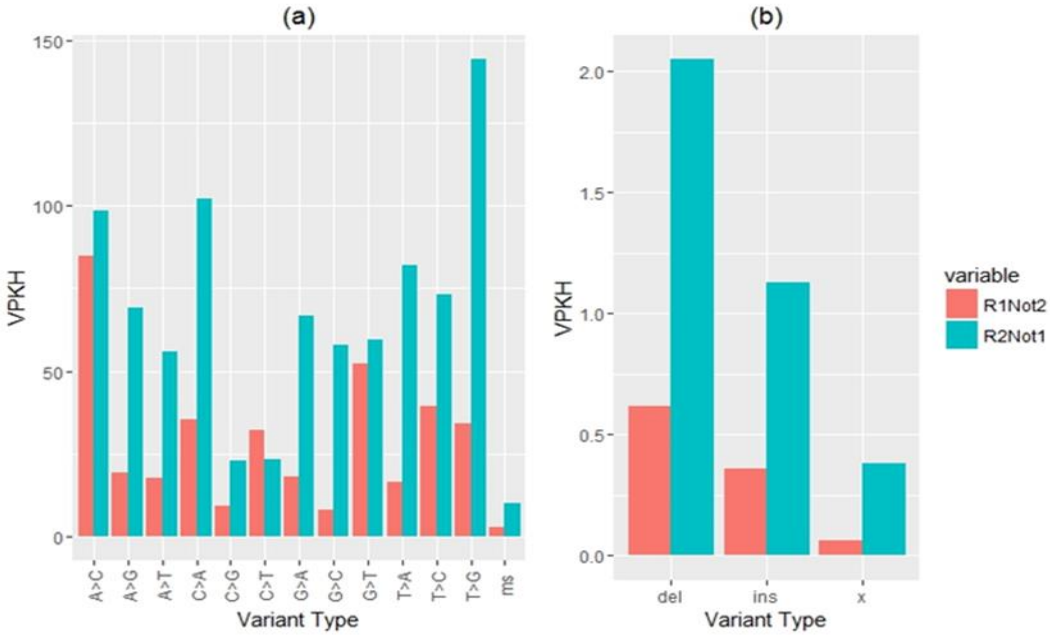

Figure 3: (a) VPKH of R1Not2 and R2Not1 in substitutions; (b) VPKH of R1Not2 and R2Not1 in deletions (del), insertions (ins) and complex mutations (x)

Please note that when we report mutations, we used the CDS (Coding Sequence) direction. In our experiments, most amplicons were designed using genomic direction for read 1 and anti-genomic direction for read 2. Thus, for positive genes ESR1, PIK3CA, and EGFR, read 1 is in the CDS (coding sequence) direction and read 2 is in the anti-CDS direction. For negative genes KRAS, NRAS and BRAF, the situation is opposite, with the exception of two amplicons of exon 4 of the NRAS gene where read 1 is in the anti-genomic direction (CDS direction) and read 2 is in the genomic direction (anti-CDS direction).

We also calculated the proportions of R1Not2 and R2Not1 in all mutations. The results are shown in Figure 4. The proportions indicate that the noise in the form of R1Not2 and R2Not1 is relatively severe, and they should be removed.

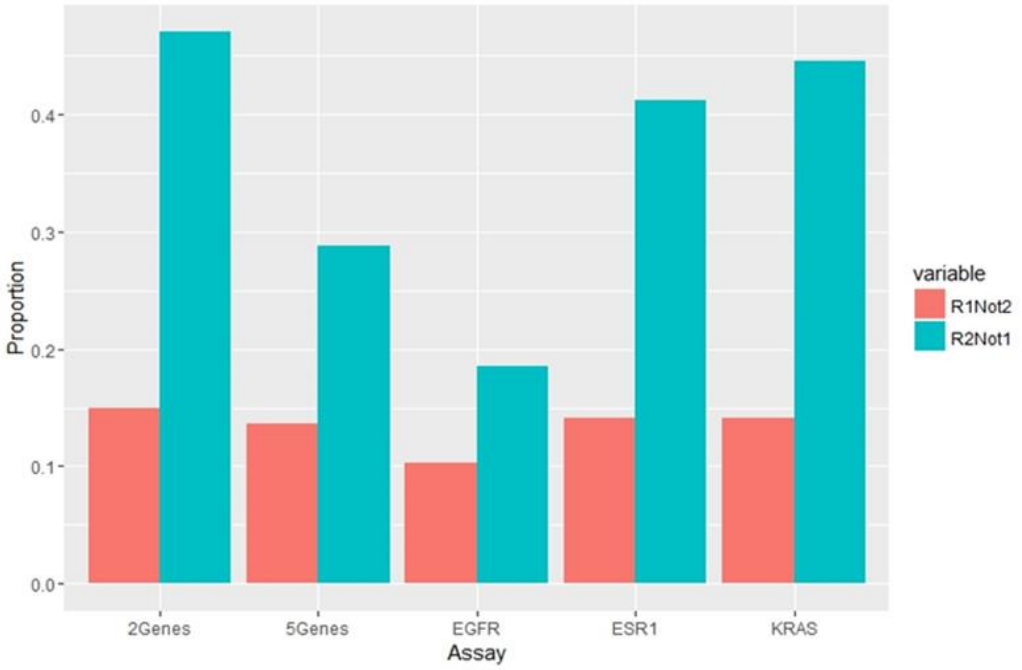

Figure 4: Proportions of R1Not2 and R2Not1 in all variants

To compare the difference between the protocol of paired-end reads (pe) and simulated single-end reads (se) by using read 1 only, we examined the dilution experiments of five EGFR spiked-in target mutations (G719S, Exon19Del, T790M, L858R and L861Q) at 1%, 0.5%, 0.25%, 0.125%, 0.0625%, 0.03125% and 0% with 300 data points for each target mutations, we did linear regression to compare the results of mutation percentages using paired-end reads (pe) and read 1 only (se). In Table 1, we can see that the intercept (in percentage, the closer to 0, the better), slope (the closer to 1, the better), and root mean squared error (RMSE, the smaller, the better), pe is better than se, except for Ex19Del intercept and slope. For R2 (the larger the better), pe is not worse than se, except for G719S. Therefore, it makes sense to use paired-end reads with noise reduction implemented in SWISSA, while using single-end reads is also feasible because the differences are small.

Table 1: Regression results of 5 EGFR target genes
for paired-end reads (pe) and simulated single-end reads (se)

Mutation	*Intercept*		*Slope*		*RMSE*		*R2*	
	pe	*se*	*pe*	*se*	*pe*	*se*	*pe*	*se*
G719S	0.0516	0.0606	1.3554	1.3763	0.1528	0.1543	0.896445	0.897444
Ex19Del	-0.0007	-0.0006	1.3013	1.2856	0.1841	0.1845	0.845994	0.842193
T790M	0.0616	0.0690	0.9045	0.9123	0.1416	0.1428	0.817746	0.817730
L858R	0.0029	0.0069	1.0072	1.0170	0.1454	0.1474	0.840598	0.839552
L861Q	0.0097	0.0122	1.0288	1.0303	0.1672	0.1675	0.806272	0.806267

3. Conclusions

The overall trend revealed from the MiSeq data shows that the variants appearing in read 1 but not in read 2 are often lower than those appearing in read 2 but not in read 1. It is proper to remove the noise R1Not2 and R2Not1 when reporting mutation percentages in samples. Assay developers may consider use the metrics R1Not2 and R2Not1 in VPKH proposed here to optimize the experimental conditions.

Acknowledgements

We thank the EGFR, ESR1, KRAS, and Multiple Gene Mutation Teams in RMS Research and Development for generating the MiSeq data. We also thank Yan Li, Yu Chuan Tai, Kevin Lee and many others for verification and validation of the SWISSA software.

References

1. Illumina Inc., Somatic Variant Caller, Pub No. 970-2012-014.
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. Liu, W-m., Li, Y., Tai, Y. C., Tsai, J., Christensen, M. and Wen, W. (2011) Notes on algorithms for detection of amplicon variants in next-generation sequencing, Proceedings of Joint Statistical Meetings 2011, 4268-4273, American Statistical Association, Alexandria, VA.
5. Liu, W.-m. and Li, Y. (2012) Development of fast, slim and accurate amplicon variant detection algorithm for next generation sequencing, Proceedings of Joint Statistical Meetings 2012, 373-376, American Statistical Association, Alexandria, VA.
6. Liu, W-m. (2014), Statistical scores for rare variant calls in ultra-deep sequencing, Proceedings of the Joint Statistical Meetings 2014, 1863-1867, American Statistical Association, Alexandria, VA.