On Genomic Repeats and Reproducibility

Can Firtina[1] and Can Alkan[1]

[1]Dept. of Computer Engineering, Bilkent University, 06800 Ankara, Turkey

Abstract

The advancements in high throughput sequencing (HTS) technologies have increased the demand on producing genome sequence data for many research questions. However, analyzing and interpreting HTS data with high accuracy is still a challenge. Although many algorithms were developed for this purpose, a handful of computational pipelines from mapping to variant calling may be considered standard, as they are commonly used in large-scale genome projects.

Approximately half of the human genome consists of repeats, which cause ambiguity in read mapping when the read length is short. On the average, a 100-bp read generated by the Illumina platform may align to hundreds of genome locations with similar edit distance. The BWA [4] mapper's approach to handle such ambiguity is randomly selecting one location, and assigning the mapping quality to zero to inform the variant calling algorithms that the alignment may not be accurate.

In this study, we investigated whether some of the commonly used variant discovery algorithms make use of this mapping quality information, and how they react to genomic repeats. We aligned two whole genome shotgun (WGS) datasets with one low (5X) and one high (45X) coverage genome sequenced as part of the 1000 Genomes Project [1] to the human reference genome (GRCh37) **twice** using BWA with the same parameters. However, we shuffled the order of reads in the second mapping experiment to make sure that the same random number is not used for the same reads. As expected, BWA reported different map locations to repetitive regions (∼2.8% of reads). Interestingly, although BWA reported zero mapping qualities for most of the discrepant read mappings (∼94%), it also assigned high MAPQ values (≥30) for a fraction of them (∼1.75%). We then generated two single nucleotide polymorphism (SNP) and indel callsets each using GATK [2] HaplotypeCaller, GATK UnifiedGenotyper, Freebayes [3], Platypus [7], and SAMTools [5], and structural variation (SV) callsets using DELLY [6] from each genome.

We observed substantial differences in the callsets generated by all of the tools we tested. GATK's HaplotypeCaller showed a discrepancy of 0.4% to 1.1% , where UnifiedGenotyper showed the highest number of different calls between two alignments of the same data set (up to 12.76%). As expected, 72 to 80% of the discrepant calls were found within common repeats. However, we also observed 165 to 4,397 SNVs that were called within one alignment but not the other that map to coding exons. Furthermore, 691 of the 4397 (15.7%) discrepant exonic SNVs predicted by GATK UnifiedGenotyper, did not intersect with any common repeats or segmental duplications. Freebayes, Platypus, and SAMtools predictions were more reproducable, as >98.5% of the calls were identical. DELLY also predicted different calls: ∼3% of deletion, ∼4% of tandem duplication, ∼6% of inversion, and ∼3.6% of translocation calls were spefic to a single alignment (i.e. BAM file), and >91% of these differences intersected with common repeats. More interestingly, when we ran GATK's both HaplotypeCaller and UnifiedGenotyper on the **same** BAM file twice, we observed similar differences. Other tools produced no discrepancies.

Our results raise questions about reproducibility and accuracy of several commonly used genomic variation discovery tools. The differences in callsets we observed in this study may have similar sensitivity and specificity. It is expected to have differences between different algorithms and/or parameters, but obtaining different results should not be due to the order of *independently generated* reads in the input file. We argue that computational predictions should not change by "luck", and that one would prefer full reproducibility.

Keywords

genome analysis, reproducibility, repeats

References

[1] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov 2012.

[2] Mark A. DePristo, Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, Guillermo del Angel, Manuel A. Rivas, Matt Hanna, Aaron McKenna, Tim J. Fennell, Andrew M. Kernytsky, Andrey Y. Sivachenko, Kristian Cibulskis, Stacey B. Gabriel, David Altshuler, and Mark J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491–498, May 2011.

[3] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. arXiv, Jul 2012.

[4] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[5] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[6] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, Sep 2012.

[7] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F. Twigg, W. G. S500 Consortium , Andrew O M. Wilkie, Gil McVean, and Gerton Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet*, 46(8):912–918, Aug 2014.

No Publishing