Unmapped reads of bacterial and viral origin from blood and saliva do not affect variant calling

Milena Kovacevic 1, Ana Mijalkovic Lazic 2, Milos Popovic 1, Sebastian Wernicke 1, Mike Teyeb 1, Rafael Iwasowi 2 and Aaron Del Duc 3

1 Seven Bridges Genomics Inc., Cambridge, Massachusetts
2 DNA Genotek Inc., Ottawa, Ontario

Introduction

The Orogen®-gDNA collection kit facilitates access to donor and enables large-scale population studies. While the general quality and utility of DNA collected from saliva with Orogen® is supported by over one thousand peer-reviewed publications, data on Whole-Genome Sequencing (WGS) is more limited.

In this study, we investigate the source of unmapped reads in both the blood and saliva sample data. Additionally, we examine the effects of sample type on the detected variants (SNPs and INDELs) using paired blood and saliva WGS data from two family trios.

We show that many of the reads failing to map to the human reference either align directly to species contained in the human microbiome database or bear similarities to other known bacterial and viral species. Overall, our analysis shows that there is no significant difference in variants detected between saliva and blood when samples are sequenced to the same coverage.

Materials and methods

Biological samples: Under IRB consent two families volunteered for the sequencing study. Both blood and saliva were collected from each donor (14 biological samples). The bacterial content in each saliva sample was assessed by testing for the presence of Enterobacteria phage PhiX174.

Sample preparation and sequencing: A standard sample preparation protocol was used to prepare all samples for sequencing on an Illumina® HiSeq 2000 sequencer to a target coverage of 30x. Samples from Family 1 were processed as paired-end sequencing and sequencing reads from Family 2 were obtained on an Illumina® MiSeq with 100 bp paired-end reads.

Reads were aligned to the hg19/b37 reference. To limit the number of false positives and low-confidence bases we called variants only when found using hard filters set according to Broad Institute’s hard filtering recommendations:

- SNPs: Qual by Depth 2.0, Fisher Strand 3.0, RMS Mapping Quality 40.0, Haplotype Score 13.0, Mapping Quality Rank Sum Test 12.5, Read Position Rank Sum Test 8.0
- INDELs: Qual by Depth (QD) 2.0, Fisher Strand (FS) 200.0, Read Position Rank Sum Test 20.0

The variants obtained from the blood- and saliva-derived DNA were compared in terms of their total number. In order to investigate whether the unmapped reads from saliva (and blood) samples have bacterial origin, they were mapped to human and bacterial and viral sequences obtained from the Human Microbiome Project (HMP) 4. Additionally, we compared the number of aligned reads with the percentage of bacterial DNA in the samples as measured by 16S rDNA.

All bioinformatics analyses were performed through reproducible pipelines on the Seven Bridges Genomics platform for bioinformatics analyses.

Results

Bacterial content in the samples correlates very closely with the number of bases/reads that cannot be aligned to the human reference genome by the BWA aligner in the bioinformatics pipeline, with a Pearson correlation factor of 0.9731 between the percentage of bacterial DNA in a sample and the number of bases/reads that cannot be mapped.

Conclusions

There is a close correlation between the amount of bacterial DNA in a sample and the number of reads that do not align to the human reference. Most of the non-mapped reads in blood (72%) and saliva were not aligned to the human reference genome by the BWA aligner in the bioinformatics pipeline. In saliva, the average difference in SNP count was 0.06%, the average difference in INDEL count 0.30%.

By far the most significant portion of reads from saliva (and blood) samples have bacterial/viral origin. This is confirmed by the presence of PhiX174 in both blood and saliva samples. The most abundant bacterial sequence in blood samples, with more than double the number of reads that do not align to the human reference, is, at the same time, most present bacterial sequence in the samples.

To quantify the amount of different bacteria relative to the sample we counted the reads aligning to each bacterial sequence and expressed the number of reads as a percentage of the total number of reads in the sample. Here, we show the number of reads originating from the top 13 different bacteria and viruses found in the sample.

References


Data analysis:

To call variants from the original FASTQ reads, all 20 samples were processed with a BWA-GATK pipeline, set up and implemented on the Seven Bridges platform for bioinformatics analysis in accordance with the Broad Institute’s best-practice guidelines.

Reads were aligned to the hg19/b37 reference. To limit the number of false positives and low-confidence bases we called variants only when found using hard filters set according to Broad Institute’s hard filtering recommendations:

- SNPs: Qual by Depth 2.0, Fisher Strand 3.0, RMS Mapping Quality 40.0, Haplotype Score 13.0, Mapping Quality Rank Sum Test 12.5, Read Position Rank Sum Test 8.0
- INDELs: Qual by Depth (QD) 2.0, Fisher Strand (FS) 200.0, Read Position Rank Sum Test 20.0