Blood vs. saliva: analysis of the effect of sample type on variant calling confidence for human Whole Genome Sequencing

Mike Tayeb1, Ana Mijalkovic Lazic2, Milena Kovacevic2, Milos Popovic2, Sebastian Wernicke2, Christina Dillane1, Aaron Del Duca1 and Rafal M. Iwasiow1

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

Introduction
Saliva collected using the Oragene® self-collection kit is a non-invasive alternative to blood as a source of large amounts of high quality genomic DNA. Oragene enables large-scale population studies by improving donor access and compliance, and its utility has been well-documented in over one thousand peer-reviewed publications. However, data on the performance of DNA from saliva in Whole Genome Sequencing is scarce in the existing literature. In this study, we present a systematic, multi-sample analysis of the effect of sample type (blood vs. saliva) on variant calling confidence and the effect of bacterial DNA in saliva on sequence alignment.

Materials and methods
Sample collection: Blood and saliva samples were collected from each member of two families using K-EDTA tubes and Oragene self-collection kits, respectively. These particular study participants were selected because the bacterial DNA content in the saliva samples (determined by 16S qPCR) ranged from below average to significantly above average. Four and three blood/saliva pairs were obtained from family 1 and 2, respectively.

Sample preparation and sequencing: Standard sample preparation protocols were used to extract and quantify DNA, and to prepare TruSeq libraries for sequencing on the Illumina HiSeq 2000. Samples from Family 2 were prepared and sequenced in duplicate to provide technical replicates. All 20 prepared libraries were sequenced to a target coverage of 30x.

Data analysis: Variants were called from the sequencing reads on the Seven Bridges platform for bioinformatics analysis using a BWA-GATK pipeline configured according to the Broad Institute’s best-practices recommendations. Reads were aligned to the hg19/B37 reference and all called variants were filtered using hard filters set according to Broad Institute’s recommendations.

To determine if unaligned reads in blood and saliva samples were of bacterial origin, they were aligned to sequences contained in the Human Microbiome Project (HMP) database using BWA MEM 0.7.4.

Results
Bacterial DNA content in the samples correlates very closely with the number of bases (reads) that align to the hg19 reference, having a Pearson correlation coefficient of 0.973. This indicates that the bacterial DNA content of a sample has a linear effect on sequencing coverage. Reads that did not map to the hg19 reference were aligned to the HMP database. An average of 37% of the unaligned reads in saliva are of bacterial or viral origin while blood reads were higher at 72%, on average. Assembly of the reads that did not align to either hg19 or HMP revealed contigs that resemble bacteria and some unknown organisms previously identified in gut and soil samples. The presence of sequences from such organisms suggests that the HMP database is incomplete.

Concordance of SNPs and indels between blood replicates, saliva replicates and between blood-saliva pairs is generally very high, however a small, systematic difference between blood and saliva can be observed. In order to determine if the concordance difference was due to coverage differences, the blood samples were downsampled to a coverage equal to that of the saliva samples. Once differences in coverage were accounted for, the average SNP and indel concordances for replicates are within 0.05% and 0.29%, of each other respectively.

In order to check if there are any regions of the human genome which were enriched with bacterial reads, human reference-aligned reads were also aligned to the HMP reference. All reads not aligning to both hg19 and HMP were discarded and a moving average coverage was calculated per base with a 100 bp window. A region was classified as enriched if a 20x average coverage was observed. These regions were inspected for the following things to identify potential bacterial contamination:

• Unusually high mismatch ratio (# mismatches in reads/total bases in region)
• Existence of HMP-enriched regions detected only in saliva samples
• High ratio of alignments with map quality zero
• Unusually low concordance between blood and saliva

Manual inspection of regions falling into one or more of the above categories revealed that none of them showed contamination with bacterial reads. Although this is not conclusive proof that there is no bacterial read contamination, it nonetheless provides confidence that bacterial reads do not accumulate enough to affect the overall mutation calling quality.

Conclusions
The amount of bacterial DNA in a saliva sample and the number of reads that do not align to the human reference are closely correlated. However, the coverage loss due to bacterial DNA is relatively small, with coverage dropping approximately 3% for every 5% bacterial DNA in the sample. The majority (72%) of the unaligned reads in blood aligned to the HMP database indicating that the source of these sequences is indeed bacterial or viral. In saliva, this metric was lower (37%), however many of the remaining unaligned reads showed similarity to other bacterial/viral species not found in the HMP, suggesting that other likely environmentally derived, species are present in the oral cavity.

In spite of the reduced coverage due to the presence of bacterial DNA in the saliva samples, there was no significant difference in the number of SNPs and indels called. The differences in concordance between replicates and saliva/blood pairs was virtually eliminated when blood data was downsampled to a coverage equal to that of saliva, suggesting that coverage differences are, by far, the most significant reason for differences in concordance between sample types.

Finally, a close inspection of HMP-enriched regions of the genome revealed that it is likely that bacterial reads do not accumulate enough to affect mutation calling.

References
1. NHG Human Microbiome Project. http://hmpdacc.org/HMREFG