Impact of sample source on variant discovery: a saliva vs blood comparison

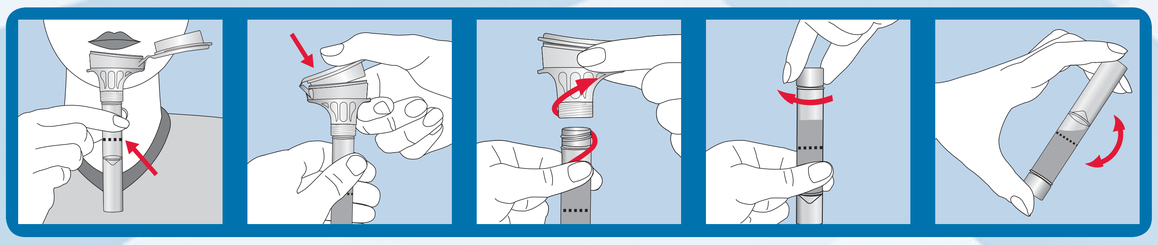
Mike Tayeb, Christina Dillane, Aaron Del Duca and Rafal M. Iwasiow
DNA Genotek Inc., Ottawa, Canada

Introduction

Recent advances in Next Generation Sequencing technologies, coupled with cost reductions, have made NGS a more practical tool for studies requiring larger cohorts. Saliva collected using Oragene® self-collection kits is a non-invasive alternative to blood for obtaining high quality genomic DNA.

Previously, we have shown that saliva DNA performs similarly to DNA from blood in targeted NGS applications such as exome sequencing and HLA typing. In this study, we investigate the performance of DNA extracted from saliva collected using Oragene in whole genome sequencing and the effect of sample type on variant calling.

Methods

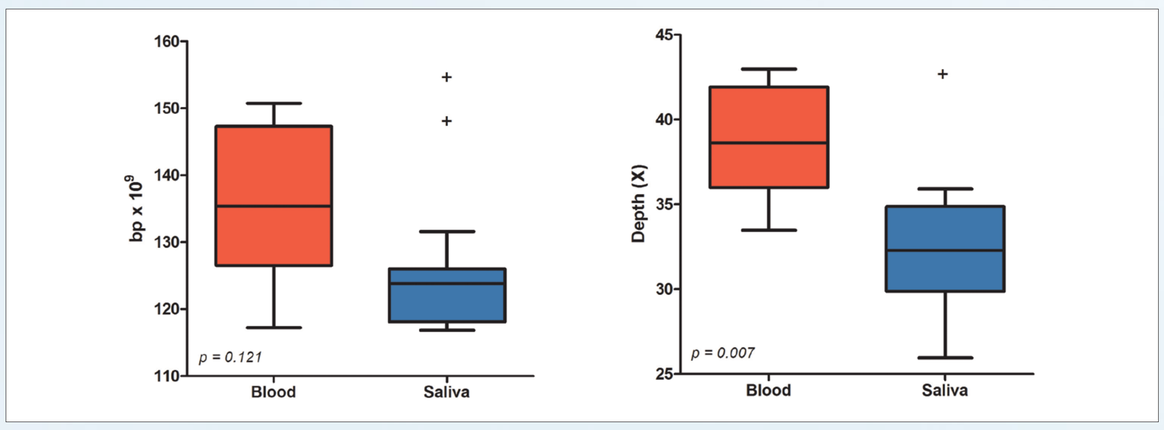


Saliva samples were collected from two multi-generational families (7 donors total) using the Oragene self-collection kit. **Donors were specifically selected for this study based on low, medium and high bacterial DNA content of their saliva in order to facilitate investigation of the effect of bacterial DNA on sequencing.** DNA was extracted from a 500 µL aliquot of sample using the prepIT®•L2P DNA extraction kit from DNA Genotek. DNA pellets were dissolved in 50 µL TE buffer. Bacterial DNA content was assessed using qPCR targeting the bacterial 16S ribosomal RNA gene.[1] Whole blood was collected into K-EDTA vacutainers and DNA was extracted using QIAGEN QIAamp DNA Mini kits.

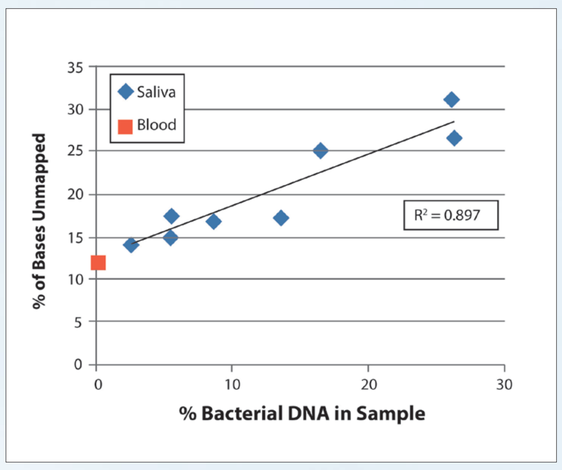
DNA was sequenced on the Illumina HiSeq 2000 using 100 bp paired-end reads (Illumina FastTrack Service). Alignment and variant calling were performed using Illumina's CASAVA software (version 1.8.2). Standard metrics (coverage, sequence depth, numbers and qualities of total and aligned reads) were extracted using Picard AlignmentSummaryMetrics and QualityScoreDistribution tools (version 1.66) and Bedtools GenomeCoverageBed (version 2.16.2). Concordances and breakdown of variants were done using multiple custom scripts written in Python.

For the calling of the *de novo* mutations in trios, the data were re-processed using the BWA aligner (version 0.6) and GATK-Lite (version 2.3-9) according to best practice recommendations by the Broad Institute. Alignments were then processed with samtools mpileup and the DeNovoGear (version 0.5.3) was used to call the *de novo* mutations, which were then annotated with 1000 Genomes allele frequency and dbSNP membership using custom Python scripts.

Results

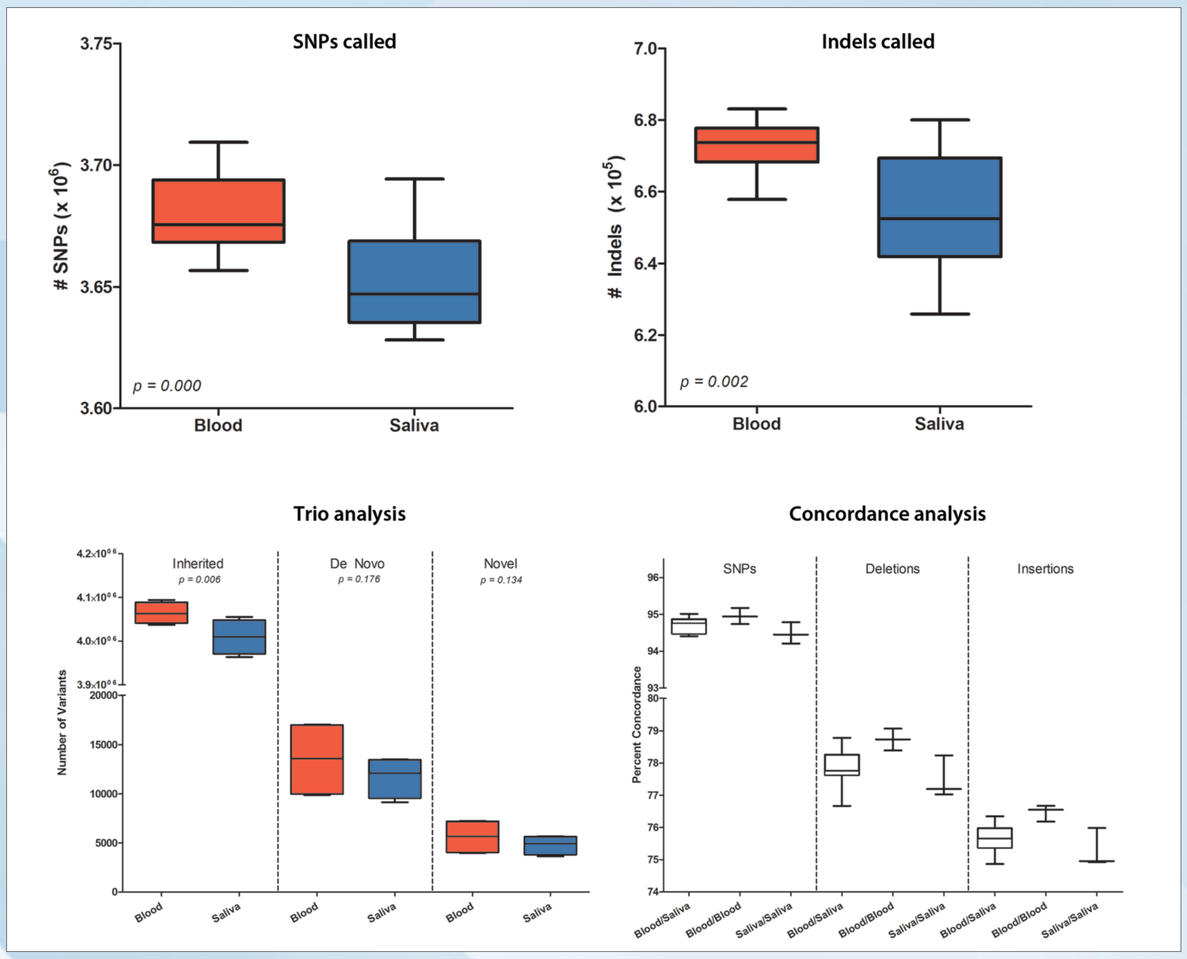


There was no difference in sequence yield between blood and saliva samples, however, the mean sequence depth of the saliva sample group was, on average, 15% lower than the blood sample group. In order to determine the cause of this difference we examined the relationship between mean depth and bacterial DNA content of each sample.

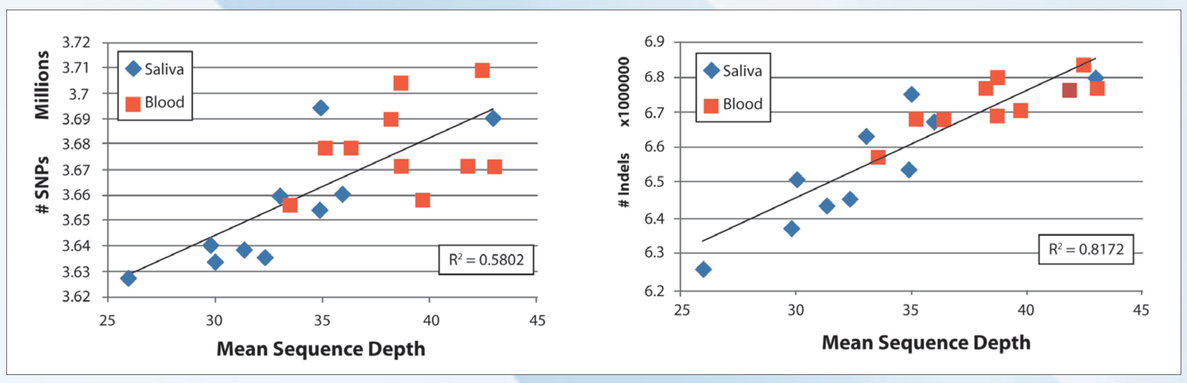


A direct, linear correlation was observed between the bacterial DNA content of a saliva sample and the number of unmapped bases suggesting that reads from non-human DNA did not map to the human reference. For blood samples 12% of reads were unmapped. When corrected for this baseline non-alignment, between 2 and 19% of the reads from the saliva samples do not map to the human reference.

The samples used represent a boundary condition and are not indicative of the typical bacterial DNA content of human saliva. For an average saliva sample containing 11% bacterial DNA[2], approximately 6% of the reads are unmapped.



Very small differences between blood and saliva samples were observed with respect to number of SNPs (0.7%) and indels (2.7%) called. When a trio analysis was performed on the two families included, a small difference was detected in the number of inherited variants called (1.3%). There was no statistically significant difference in the number of *de novo* and novel mutations in the paired blood and saliva samples. No differences were observed in the SNP concordance between blood and saliva samples compared to blood replicates.



Correlations between mean sequence depth and number of variants called (SNPs and Indels) were observed.

Discussion

- Saliva and blood produce similar sequencing yield.
- Percent alignment correlates with bacterial DNA content of sample indicating that bacterial DNA likely does not align significantly to human reference.
- An average saliva sample with 11% bacterial DNA will have only 6% of bases unaligned.
- Despite reduced depth and alignment, changes to numbers of variants called and concordance, if any, are minimal.
- Correlation between mean sequence depth and number of variants called indicates that the differences in variant detection could be overcome with additional sequencing. In addition, contribution to variant calls and concordance by analytical pipeline is also relevant and will be explored in future analysis.
- Additional work will be done to investigate the cause of any persistent differences between blood and saliva samples, including the effect of sequencing depth, as well as to validate a subset of the variants called to ensure sequencing accuracy with saliva samples.

References

1 Bacterial DNA Assay Protocol. DNA Genotek. PD-PR-065.
2 Human genomic DNA content of saliva samples collected with the Oragene® self-collection kit. DNA Genotek. PD-WP-011.

Acknowledgements

The authors would like to thank Miloš Popović, Ana Mijalković and Goran Rakočević of Seven Bridges Genomics, Inc. for their kind assistance with data analysis and results interpretation.

Software used:

DeNovoGear *(http://www.nature.com/nmeth/journal/v10/n10/full/nmeth.2611.html)*
samtools *(http://bioinformatics.oxfordjournals.org/content/25/16/2078.full)*
BWA *(http://www.ncbi.nlm.nih.gov/pubmed/19451168)*
GATK *(http://www.ncbi.nlm.nih.gov/pubmed?term=20644199, http://www.ncbi.nlm.nih.gov/pubmed?term=21478889)*
bedtools *(http://bioinformatics.oxfordjournals.org/content/26/6/841.short)*
Picard *(no paper, but the site can be cited http://picard.sourceforge.net)*
dbSNP *(http://www.ncbi.nlm.nih.gov/pubmed/11125122)*
1000Genomes *(http://www.nature.com/nature/journal/v491/n7422/full/nature11632.html)*

DNA genotek
Superior samples
Proven performance

SevenBridges genomics

www.dnagenotek.com • support@dnagenotek.com