Systematic multi-sample analysis of the effect of sample type (blood vs. saliva) on variant calling confidence for WGS

Goran Rakocevic[1], Sebastian Wernicke[1], Mike Tayeb[2], Rafal Iwasiow[2], Aaron Del Duca[2]

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

Introduction

The Oragene®•DNA collection kit enables scalable donor access and large-scale population studies. While the general quality and utility of DNA collected from saliva with Oragene is supported by over 1000 peer-reviewed publications, data on **Whole Genome Sequencing** (WGS) is more limited in previous studies.

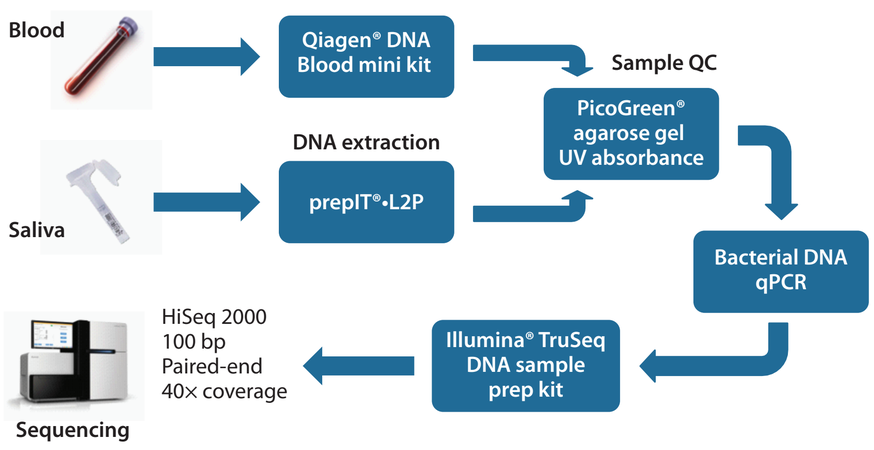
This study presents the first systematic multi-sample analysis of the effect of sample type (blood vs. saliva) on variant calling confidence for WGS. Using paired blood and saliva WGS data from two family trios, we investigate the effects of sample type on the detected variants (SNPs and INDELs) and systematically investigate the causes of any differences.

Overall, our analysis shows that variants that differ between saliva and blood are caused by lower sequencing coverage in certain saliva samples, which is a direct result of bacterial contamination. Fortunately, this bacterial contamination is simple to account for prior to sequencing, thus all relevant sources of variation can be minimized.

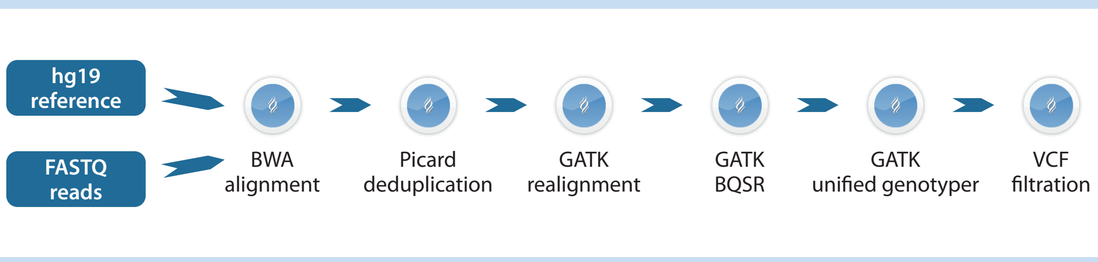
Materials and methods

Biological samples: Two families were selected for sequencing. Both blood and saliva were collected from each donor (14 biological samples). The bacterial content in each saliva sample was assessed using 16S qPCR.

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 30×. Samples from Family 2 were prepared and sequenced twice to obtain technical replicates, yielding a total of 20 sequenced samples (4 blood/saliva pairs from Family 1, 2 × 3 blood/saliva pairs from Family 2).



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 Genomics' cloud computing platform in accordance with the Broad Institute's best-practices guidelines.



Reads were aligned to the hg19/b37 reference using BWA v0.6; while duplicate reads were marked with Picard Tools v1.66. Indel Realignment and Base Quality Recalibration were performed with GATK v2.39Lite; SNP and Indel calls were done using the Unified Genotyper from the GATK v2.39Lite. This pipeline is generally recognized for its sensitivity. To limit the number of false positives and low-confidence variants, all called variants were filtered using hard filters set according to Broad Institute's hard filtering recommendations:

- **SNPs:** Qual by Depth 2.0, Fisher Strand 60.0, 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, as well as in terms of genotype-based concordance. A variant call was considered concordant only if there is an exact genotype match.

Concordance was also compared for sets of high-confidence *de novo* mutation calls. These sets were determined by grouped variant calls, making use of the pedigree structure of the samples.

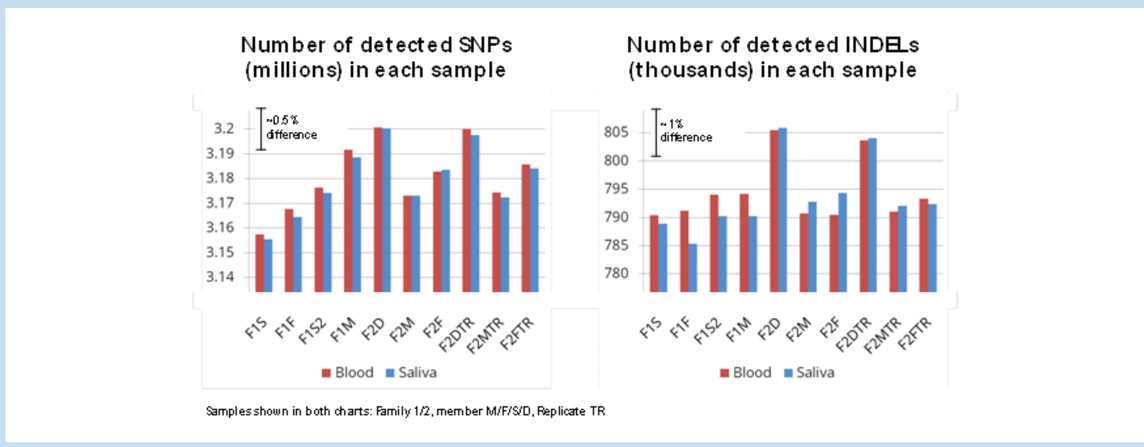
To assess the effect of bacterial DNA on coverage, we compared the number of aligned reads with the percentage of bacterial DNA in the samples as measured by 16S qPCR.

Bacterial DNA introduces systematic differences in sequencing coverage between blood and saliva samples. To assess the effect of these differences, we subsequently eliminated coverage differences between samples through *in silico* "downsampling" (randomly removing reads from samples with higher coverage until their numbers approximate the lower coverage samples). The concordance analysis was repeated after down-sampling.

All bioinformatic analyses were performed through reproducible pipelines on the Seven Bridges Genomics' cloud platform.

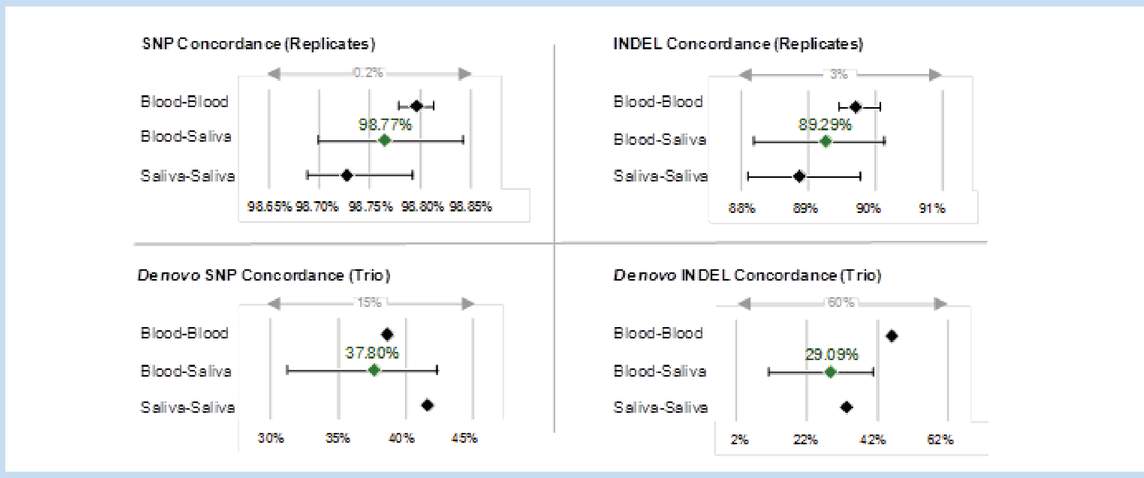
Results

A direct comparison of the variants called from the blood and saliva samples shows no significant systematic differences in their total number. The average difference in SNP count was 0.06%, the average difference in INDEL count 0.30%.

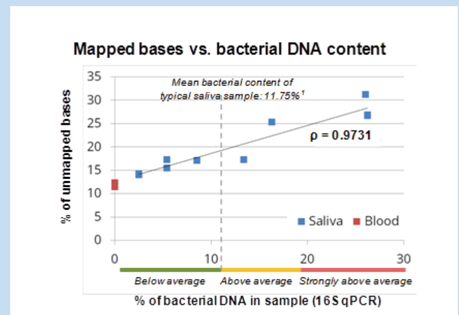


SNP and INDEL concordance between technical replicates was high overall with only slight variation between sample types (less than 0.15% for SNPs and less than 1% for INDELs). In the plots, a small but systematic difference in concordance between blood and saliva can be observed.

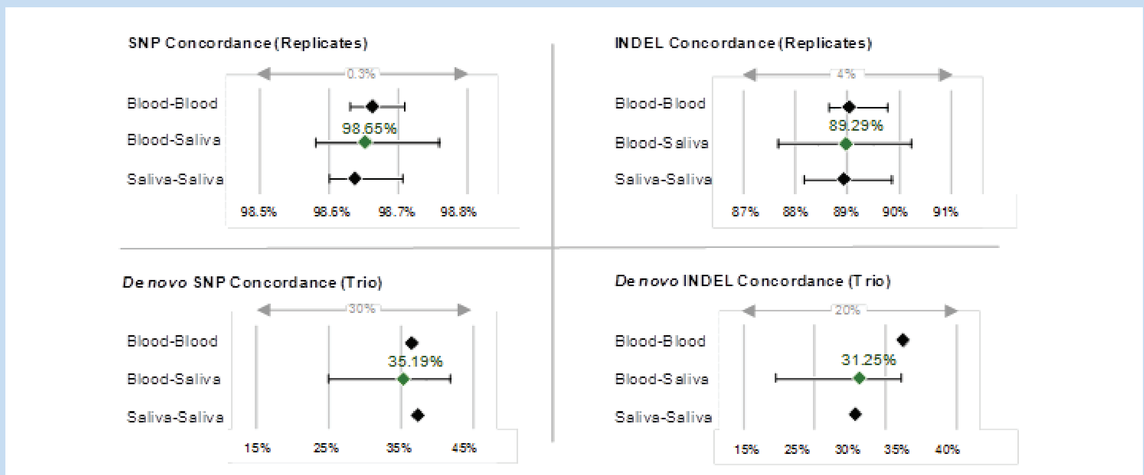
For *de novo* SNPs and INDELs, only a single trio (Family 2) was available for analysis. The data seems to give no indication for systematic differences of any kind, but concordance is generally low as has been previously reported in the literature.[1]



Bacterial content in the samples correlates very closely with the number of bases/reads that can 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 percentage of unmapped bases. The bacterial DNA content has a linear effect on the sequencing coverage, increasing the percentage of unmapped bases by about 3 percentage points for every 5 percentage points of bacterial DNA in the sample.



Once differences in coverage are accounted for by downsampling to the same coverage before comparison, differences in concordance virtually disappear between sample types: The average concordances for the replicates are within 0.05% of each other for SNPs and within 0.25% of each other for INDELs. *De novo* mutations also become significantly more concordant once coverage differences are accounted for.



Discussion and outlook

The main difference between sequencing blood and saliva samples is a difference in coverage depth which can result from bacterial contamination. These coverage differences appear, by far, to be the most significant reason for differences in concordance between sample types. Eliminating coverage effects results in very high blood-saliva concordance, such that only a few mutations are observed over an entire genome. Furthermore, these differences do not appear to be systematic.

Coverage loss due to bacterial DNA is comparatively small; approximately 3% of coverage is lost for every 5% of bacterial DNA in the sample. Given the highly linear relationship between bacterial contamination and coverage, we therefore recommend assessing the bacterial contamination of saliva samples prior to sequencing and to target 3% higher coverage for each 5% of bacterial content measured. This will allow samples to be sequenced to the same depth of coverage, thus minimizing the most relevant source of variation.

Seven Bridges and DNA Genotek are currently collaborating to advance this study by systematically investigating the impact of bacterial DNA on the sequencing and alignment of the human DNA in saliva samples.

Reference

1. O'Rawe et al., Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing, *Genome Medicine, Vol. 5, Issue 3, 2013*.

SevenBridges genomics

DNA genotek

Superior samples
Proven performance

www.dnagenotek.com • support@dnagenotek.com