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

Milena Kovacevic[1], Ana Mijalkovic Lazic[1], Milos Popovic[1], Sebastian Wernicke[1],
Mike Tayeb[2], Rafal Iwasiow[2] and Aaron Del Duca[2]

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

## Introduction

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

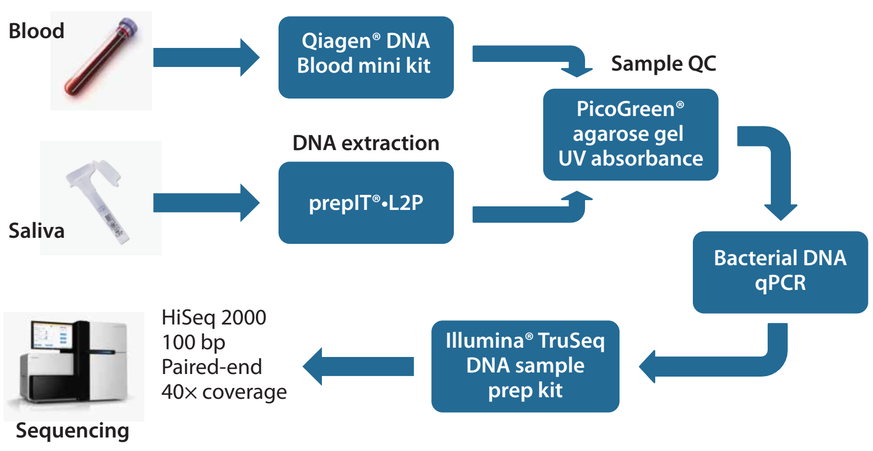
In this study, we investigate the source of unaligned reads in both the blood and saliva sample data. Additionally, we examine the effects of sample type on the detected variants (SNPs and INDELs) using paired blood and saliva WGS data from two family trios.

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.

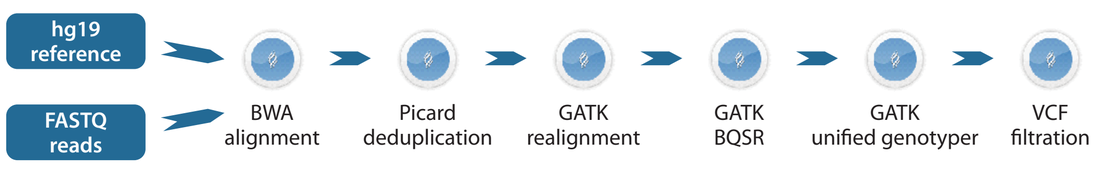
## Materials and methods

**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 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 platform for bioinformatics analysis in accordance with the Broad Institute's best-practices guidelines.




Reads were aligned to the hg19/b37 reference. 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:

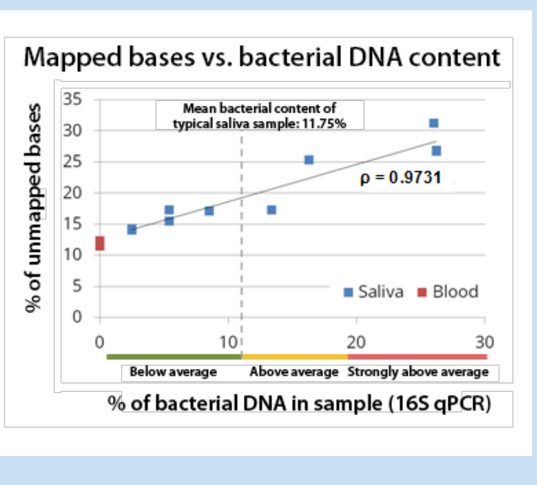
- **SNPs:** Qual by Depth 2.0, Fisher Strand 60.0, RMS Mapping Quality 40.0, Haplotype Score 13.0, Mapping Quality Rank Sum Test 12.5, Read Position Rank Sum Test 8.0
- **INDELs:** Qual by Depth (QD) 2.0, Fisher Strand (FS) 200.0, Read Position Rank Sum Test 20.0

The variants obtained from the blood- and saliva-derived DNA were compared in terms of their total number. In order to investigate whether the unaligned reads from saliva (and blood) samples have bacterial origin, they were aligned to human and bacterial and viral sequences obtained from "Human Microbiome Project". [1] Additionally, we compared the number of aligned reads with the percentage of bacterial DNA in the samples as measured by 16S qPCR.

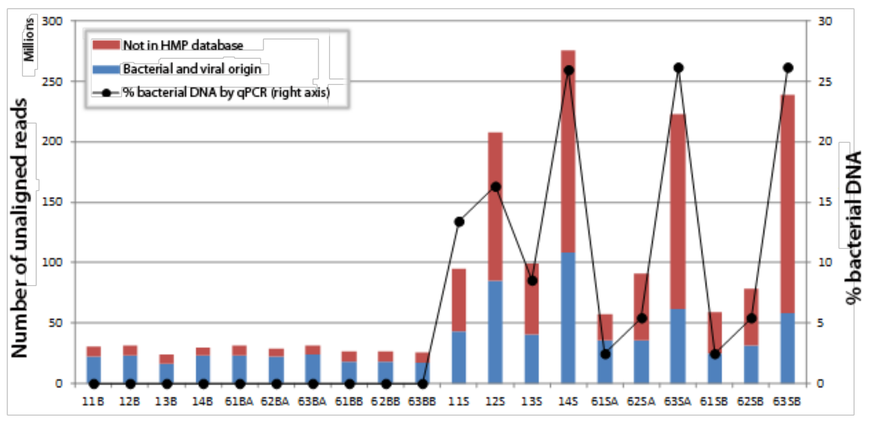
All bioinformatics analyses were performed through reproducible pipelines on the Seven Bridges Genomics platform for bioinformatics analysis.

## Results

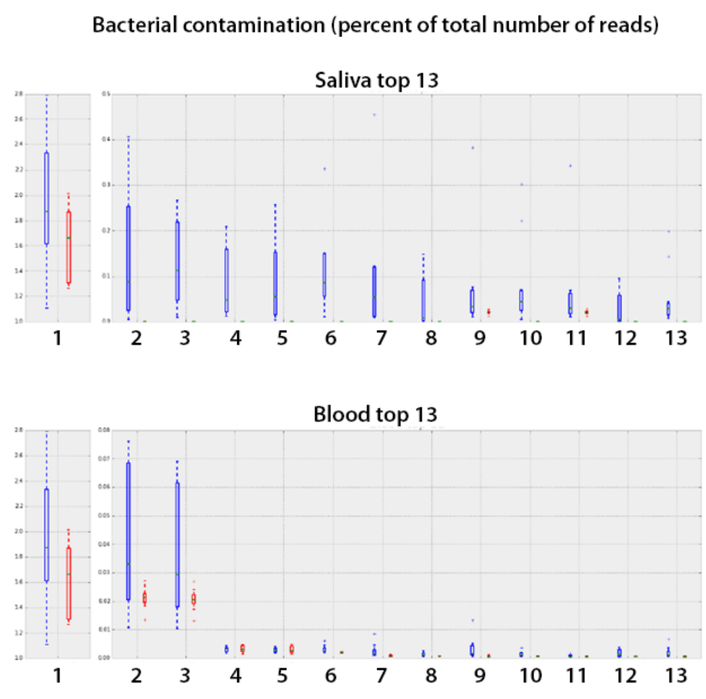
Bacterial content in the samples correlates very closely with the number of bases/reads that cannot be aligned to the human reference genome by the BWA aligner in the bioinformatics pipeline, with a Pearson correlation factor of 0.9731 between the percentage of bacterial DNA in a sample and the percentage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




Reads that did not map to the human reference were aligned to sequences from the Human Microbiome Project database. On average, 72% of the unaligned reads in blood are of bacterial/viral origin while saliva reads look less so with 37% on average. Upon closer examination of the assembly of the reads not aligning to either human or bacteria we found that assembly formed contigs which were then run through BLAST. The results showed that these contigs bare similarities with some bacteria (though many parts were missing or highly variant) and to some unknown organisms which were previously identified in gut and soil samples. The presence of sequences from these organisms suggests that the oral microbiome database is incomplete.




To quantify the amount of different bacteria relative to the sample we counted the reads aligning to each bacterial sequence and expressed the numbers 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.



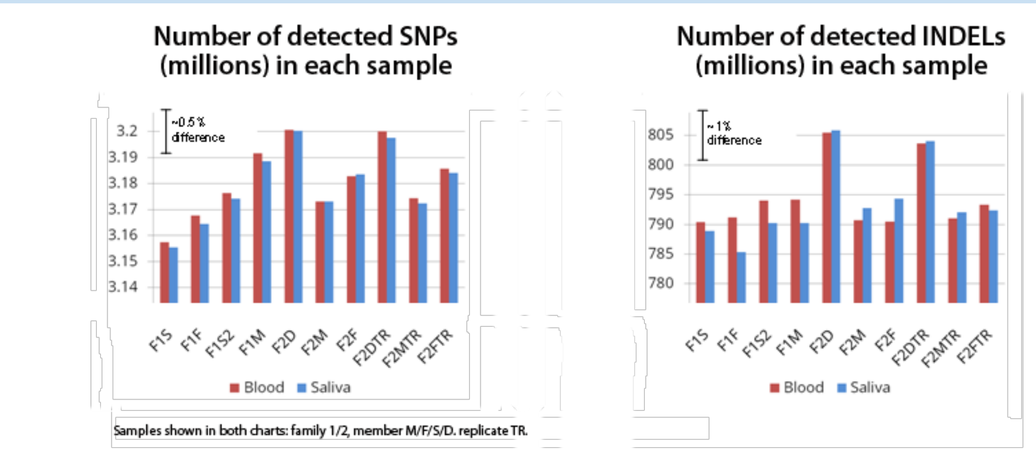

| Saliva | |
|---|---|
| 1 | Enterobacteria phage phiX174 |
| 2 | Prevotella melaninogenica ATCC 25845 chromosome chromosome I |
| 3 | Campylobacter concisus 13826 |
| 4 | Rothia mucilaginosa DY-18 |
| 5 | Prevotella melaninogenica ATCC 25845 chromosome chromosome II |
| 6 | Streptococcus parasanguinis ATCC 15912 genomic scaffold SCAFFOLD1 |
| 7 | Streptococcus mitis B6 |
| 8 | Gemella haemolysans ATCC 10379 ctg1119035638547 |
| 9 | Escherichia coli MS 45-1 genomic scaffold Scfld683 |
| 10 | Veillonella dispar ATCC 17748 V_dispar-1.0.1_Cont0.2 |
| 11 | Escherichia coli MS 21-1 genomic scaffold Scfld856 |
| 12 | Gemella haemolysans ATCC 10379 ctg1119035638550 |
| 13 | Streptococcus sanguinis SK36 |

| Blood | |
|---|---|
| 1 | Enterobacteria phage phiX174 |
| 2 | Escherichia coli MS 45-1 genomic scaffold Scfld683 |
| 3 | Escherichia coli MS 21-1 genomic scaffold Scfld856 |
| 4 | Fusobacterium ulcerans ATCC 49185 NZ_ACDH01000101 |
| 5 | Mollicutes bacterium D7 cont1.210 |
| 6 | Cyanothece sp. CCY0110 1101676644430 |
| 7 | Candidate division TM7 single-cell isolate TM7a NZ_ABBV01002500 |
| 8 | Trypanosoma brucei gambiense DAL972 chromosome 3 |
| 9 | Escherichia coli MS 198-1 genomic scaffold Scfld389 |
| 10 | Beggiatoa sp. PS contig24454 |
| 11 | Plasmodium falciparum 3D7 chromosome 9 |
| 12 | Chthoniobacter flavus Ellin428 ctg66 |
| 13 | Cyanothece sp. CCY0110 1101676644431 |

By far the most significant portion of reads (up to 2.0% and 2.8% in blood and saliva respectively) is contributed by *Enterobacteria phage PhiX174*. Interestingly, this virus is used by Illumina as a low concentration spike-in to improve calibration and quality control.[2]

The rest of the bacteria/viruses come in much lower concentrations of 0.5% and less for saliva and 0.08% and less for blood. Most of the species found in saliva are known inhabitants of the human mouth as expected. We also detected the presence of *Escherichia coli* in both blood and saliva samples. *E. coli* is, at the same time, most present bacterial sequence in blood samples, with more than double amount of reads aligned on its sequence than any other detected bacteria. It is important to note that the presence of bacterial sequences in blood is likely due to contamination during sample or library preparation and that similar contamination is also present in the saliva samples.

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%.




## 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 a lower proportion in saliva (37%) aligned to sequences in the human microbiome project database showing that the source of these sequences is indeed bacterial or viral.

In saliva, the remainder of the unmapped reads showed similarity to other known bacterial/viral species. Failure of these reads to map to HMP indicates that the HMP database is likely incomplete and that other, uncharacterized species are present in the oral cavity.

Although the presence of bacterial DNA results in a slight loss in coverage depth, there is no significant difference in the total number of variants detected between paired blood and saliva samples. Previously we have shown that, when sequenced to equal depth, blood-saliva concordance is very high, such that only a few differences are observed across an entire genome.[3]

Seven Bridges and DNA Genotek are currently collaborating to advance this study. Further investigation of the impact of bacterial DNA on the sequencing and alignment of the human DNA in saliva samples is currently underway.

### References

1. NIH Human Microbiome Project. http://hmpdacc.org/HMREFG
2. Using a PhiX Control for HiSeq Sequencing Runs. Illumina Inc. March 2013. http://res.illumina.com/documents/products/technotes/technote_phixcontrolv3.pdf
3. Systematic multi-sample analysis of the effect of sample type (blood vs. saliva) on variant calling confidence for WGS. Rakocevic et al. http://www.dnagenotek.com/ROW/pdf/MK-00360.pdf

Seven Bridges genomics

DNA genotek

*Superior samples*
*Proven performance*

© 2014 DNA Genotek Inc., a subsidiary of OraSure Technologies, Inc., all rights reserved.
All brands and names contained herein are the property of their respective owners.
Patent (www.dnagenotek.com/legalnotices) MK-00373 Issue 1/2014-05

**www.dnagenotek.com • support@dnagenotek.com**