Exploring differences in sequencing quality and variant-detection concordance across DNA sources

Mariana Chichkova (s205694)[1], Mette Christoffersen (s192033)[2], Laura Sans-Comerma (s192437)[3], Natasia Thornval (s143493)[1], and Huijiao Yang (s202360)[3]

1 DTU Bioengineering, Technical University of Denmark, 2 DTU HealthTech, Technical University of Denmark, 3 DTU Bioinformatics, Technical University of Denmark

Introduction

The recent development of next-generation sequencing technologies has made whole-genome sequencing (WGS) attractive for clinical applications and large genetic studies with thousands of participants [1]. Obtaining this data requires sample collection methods that are convenient and suitable for both patients and practitioners [2].

Traditionally, blood has been the main source of DNA for genetic analysis, however DNA isolated from saliva has become an attractive alternative [2, 3]. Saliva collection is non-invasive, has lower overall costs and better stability for shipping and storage. Furthermore, in large scale genetic studies, saliva collection leads to significantly higher response rates [4].

Despite these advantages, saliva samples generate lower yields of DNA and are often contaminated with bacterial DNA [3, 4]. A study by *Trost et al.* assessed how the DNA source and bacterial DNA contamination affect the quality of sequencing data and the accuracy of single-nucleotide variant (SNV), indel, and copy-number variation (CNV) detection using industry-standard short-read WGS data [3].

The aim of the project is to compare blood and saliva as DNA sources in regards to sequencing quality and variant-detection concordance.

Data set

The study uses a subset of eight samples from *Trost et al*. DNA from one blood sample and one saliva sample was isolated from four individuals (table 1). DNA library preparation was PCR-free using NxSeq AmpFREE Low DNA kit (Lucigen). Sequencing WGS was performed using ILLUMINA HiSeq X, generating 151 bp paired-end reads. The raw data was downloaded from the European Nucleotide Archive (ENA) and belongs to the Personal Genome Project Canada (PGPC).

Table 1: **Overview of samples used.**

Individual	Sample ID	Source	Bases	Size
PGPC-0002	SRR8595490	Blood	174.5 G	72.1Gb
	SRR8595494	Saliva	127.3 G	52.3 Gb
PGPC-005	SRR8595491	Blood	111.6 G	46 Gb
	SRR8595495	Saliva	132.5 G	53.1 Gb
PGPC-006	SRR8595492	Blood	151.9 G	62.4 Gb
	SRR8595496	Saliva	138.8G	56.3 Gb
PGPC-0050	SRR8595493	Blood	129.3 G	52.9 Gb
	SRR8595497	Saliva	137.7 G	54.9 Gb

Methods

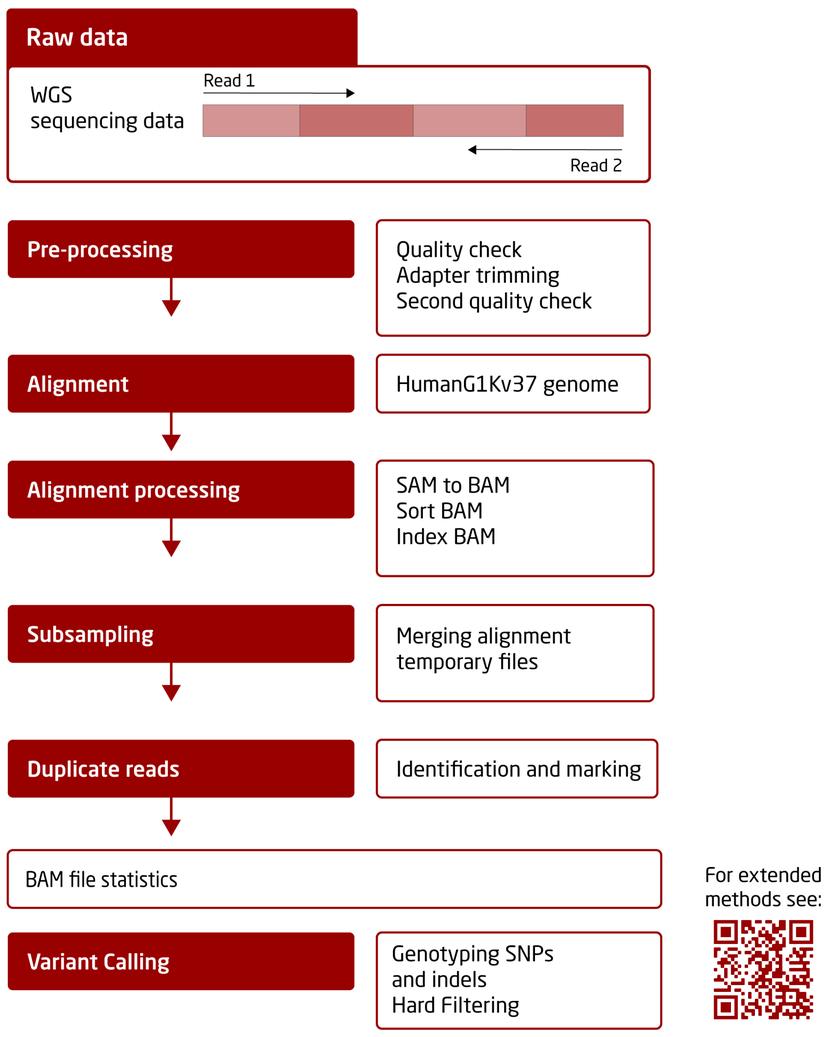

Figure 1: **Overview of the workflow.** (1) Pre-processing step, `FastQC` is used for both quality checks, and `fastp` for trimming. (2) The alignment against HumanG1KV37 genome is done using Burrows Wheeler Aligner command, `bwa mem`. (3) The processing post-alignment is done using several functions from `SAMTools`. (4) The duplicate reads are marked using `Picard`. (5) Production of alignment statistics. (6) The variant calling was done using `GATK`.

Sequencing quality and alignment metrics

Several differences can be observed between blood and saliva when comparing sequencing quality and alignment metrics. Overall, WGS of saliva samples generated a higher number of raw reads (Fig. 2A) and a lower number of aligned reads when compared with the blood samples (Fig. 2B). The average insert size was lower for the saliva samples (Fig. 2D), while average quality was similar for both DNA sources (Fig. 2C). The average coverage in the blood samples were similar (mean 18X), while average coverage in the saliva samples were lower and more variable (mean 14X) (Fig. 2E). A probable explanation for this is contamination from bacterial DNA in the saliva samples.

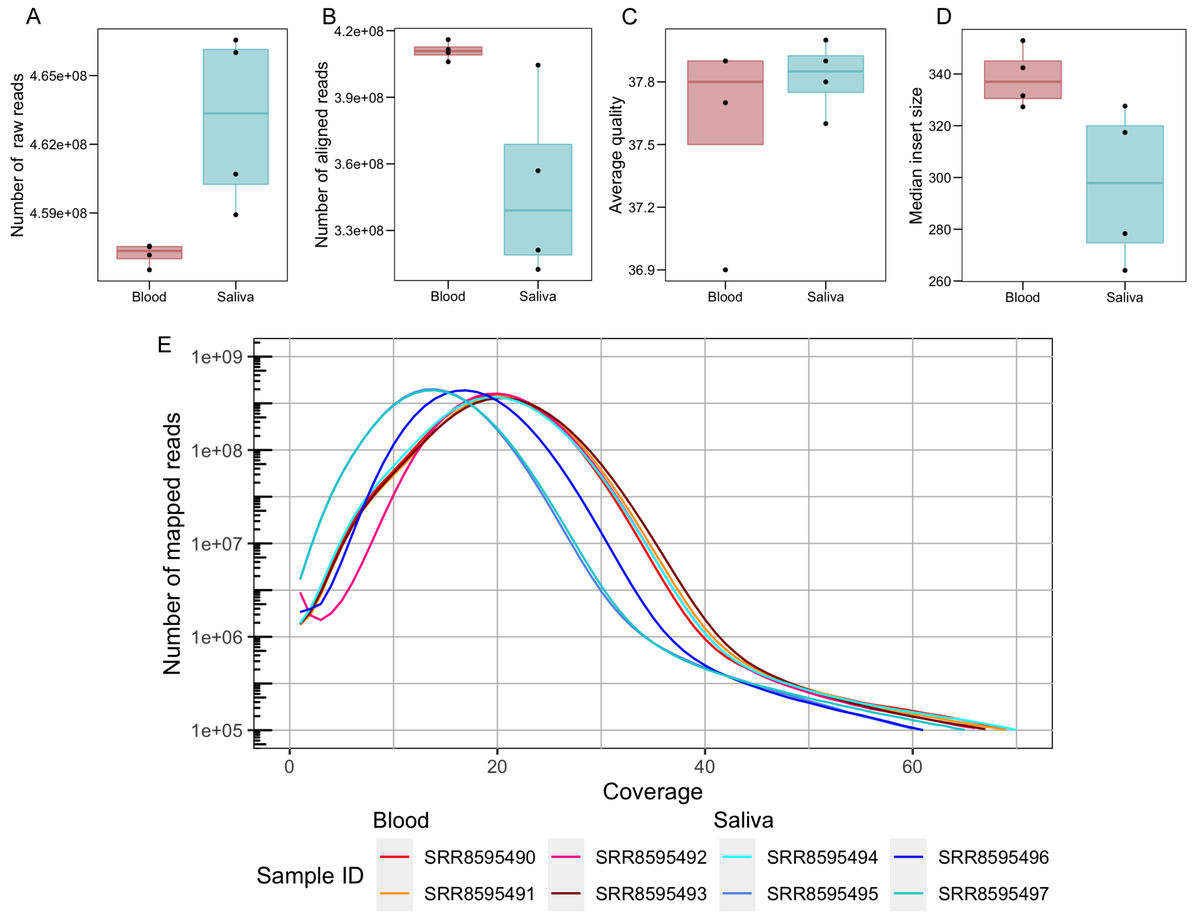

Figure 2: Comparison between different sequencing and alignment metrics for blood (Red) and saliva (Blue) samples. A) Number of raw reads. B) Number of aligned reads. C) Average Quality per base. D) Median insert size. E) Coverage.

Variant detection: Novel SNPs and indels

Variant-detection concordance between blood and saliva samples was compared before and after hard filtering. Hard filtering increased the concordance between blood and saliva samples from 89% to 93%.

Variant-detection concordance was further compared for different types of variants. Higher concordances were found for SNPs compared with indels and for known variants compared with novel variants. The highest concordance was observed for known SNPs (95%) followed by known indels (86%). A higher inherent error rate is usually observed for indels compared to SNPs. The concordance may therefore improve if the indels were filtered with a more stringent set of thresholds compared to the SNPs. For novel variants the concordance was poor for both SNPs (27%) and indels (20%). A likely explanation for this finding is a high probability of false positive variants among the novel variants.

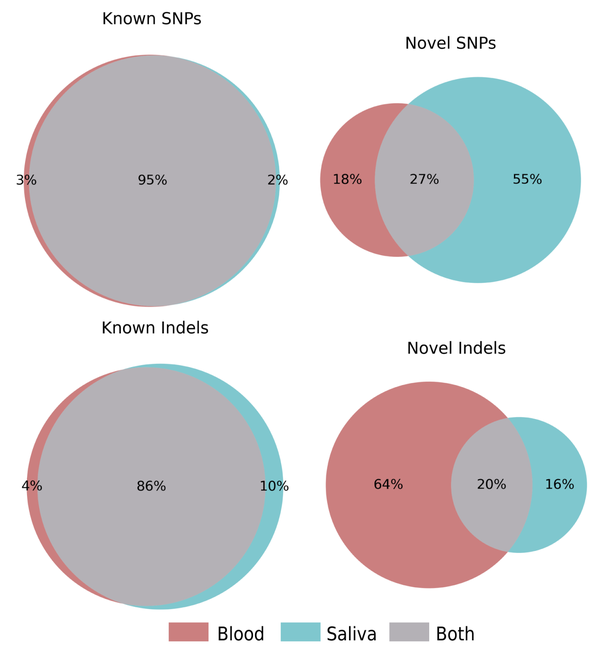

Figure 3: Venn diagrams of the variant detection concordance between blood and saliva samples for four different types of variants.

Conclusions

DNA extracted from blood and saliva was evaluated for sequencing quality, alignment characteristics and variant-detection concordance, in summary:

- Higher number of raw reads and unmapped reads in the saliva samples.
- Lower and less uniform coverage in the saliva samples.
- Smaller insert sizes in the saliva samples.
- Average quality was similar for blood and saliva samples.
- Higher concordance of variant detection across samples after hard filtering.
- Higher concordance of variant detection of known variants across samples.

All authors contributed equally.

[1] J. D. Wall, ‘‘Estimating genotype error rates from high-coverage next-generation sequence data,’’ *Genome Research*, vol. 24, no. 11, pp. 1734--1739, 2014.
[2] J. E. Abraham, ‘‘Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping,’’ *BMC Medical Genomics*, vol. 5, pp. 1--6, 2012.
[3] B. Trost, ‘‘Impact of dna source on genetic variant detection from human whole-genome sequencing data,’’ *Journal of Medical Genetics*, vol. 56, no. 12, pp. 809--817, 2019.
[4] H. V. Gudiseva, ‘‘Saliva DNA quality and genotyping efficiency in a predominantly elderly population,’’ *BMC Medical Genomics*, vol. 9, no. 1, pp. 1--8, 2016.

matach@student.dtu.dk, s192033@student.dtu.dk, s192437@student.dtu.dk, s143493@student.dtu.dk, s202360@student.dtu.dk