Validated WGS and WES protocols proved saliva-derived gDNA as an equivalent to blood-derived gDNA for clinical and population genomic analyses

Katerina Kvapilova[1,2], Pavol Misenko[3], Jan Radvanszky[3,4,5,6], Ondrej Brzon[2], Jaroslav Budis[3,6,7], Juraj Gazdarica[3,6,7], Ondrej Pos[3,6], Marie Korabecna[8], *Martin Kasny[2,9], Tomas Szemes[3,5,6], Petr Kvapil[2], Jan Paces[1,10] and Zbynek Kozmik[11]

2 Institute of Applied Biotechnologies, Sluzeb 3056/4, 108 00 Prague, Czech Republic
* Correspondence: kasny@iabio.eu; Tel.: +420 739 394 364
(All affiliations in detail are presented in supplement)

Whole exome sequencing (WES) and whole genome sequencing (WGS) have become standard methods in human clinical diagnostics as well as in population genomics (POPGEN). Blood-derived genomic DNA (gDNA) is routinely used in the clinical environment. Conversely, many POPGEN studies and commercial tests benefit from easy saliva sampling. Here, we evaluated the quality of variant call sets, the level of genotype concordance of single nucleotide variants (SNVs), and small insertions and deletions (indels) for WES and WGS using paired blood- and saliva-derived gDNA isolates employing genomic reference-based validated protocols.

Material and methods

Paired blood–saliva samples were processed utilizing the same protocol in accordance with the study design (Figure 1).
Three technical replicates were sequenced of the same DNA isolate of RS NA12878, each using both WGS and WES protocols (Figure 1A).
The accuracy parameters of the benchmark-derived DNA sequencing data were compared with those of blood-derived gDNA and saliva-derived gDNA from sequencing data outputs of 10 individuals, including the determination of protocol accuracy using F1 score calculations of blood–saliva paired samples (blood–saliva-based comparison); these were performed individually for both WGS and WES data calls (Figure 1B).

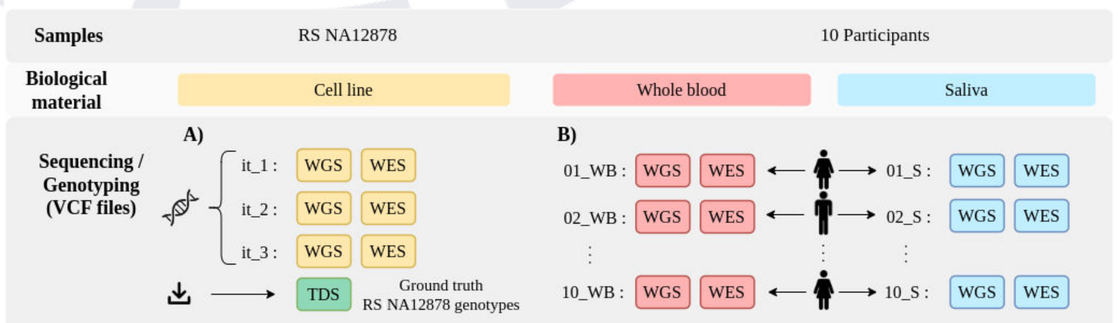

Figure 1. Graphical overview of the study design. A) Biological material used in the study; B) Sequencing analyses performed during the study. RS NA12878 – reference standard (Coriell Institute).

Isolation of gDNA and RNA

- within 24 hours after sample collection
- QIAamp® DNA Blood Mini Kit (QIAGEN, Germany) with a sample origin--dependent protocol
  Blood input: 1 mL Saliva input: 2 mL (saliva-media mixture)

Quality control of DNA

- Purity – NanoPhotometer P300 (Implen, Germany)
- Quantity – Qubit 1x dsDNA High Sensitivity (HS) Kit (Thermo Fisher Scientific, USA)
- Integrity – 0.8% agarose gel electrophoresis

NGS library preparation

1. WGS
   - TruSeq DNA PCR-Free kit (Illumina, USA)
   - gDNA input: 1 µg
2. WES
   - Illumina DNA Prep with Enrichment kit (Illumina, USA) with the Alliance VCGS Exome panel and Mitochondrial DNA panels (both Twist Bioscience, USA)
   - gDNA input: 100 ng

Sequencing

- NovaSeq 6000 (Illumina, USA), S4 chemistry with XP 4-Lane kit

Bioinformatic analysis

- FASTQ QC: FastQC (v2.20.0.422)
- FASTQ adapters and quality trimming: fastp (v. 0.20.1)
- Reference mapping and variant calling: DRAGEN v3.10, human GRCh38 reference genome

Results

First, the technical error rate of the sequencing process itself was determined for WGS and WES protocols, using two complementary approaches: pairwise-triplicate-based and TDS-based. For the WGS protocol the determined F1 scores ranged between 0.9469–0.9998 and for the WES protocol between 0.8880–1.000.

In the second comparison approach, genotype concordance rates were determined between the defined genotypes in the intersection of the truth dataset region (TDS) and the WGS and WES results of RS NA12878 iterations, all restricted to autosomes (to eliminate sex chromosome-related genetic male–female differences) and to the HCR (Table 1). Comparable numbers were identified when considering the WGS and WES results alone.

Defined Genotypes	Identified Number with TDS	Identified Number without TDS	Identified Number in Blood	Identified Number in Saliva
SNVs	17 360	17 368	17 548	17 537
Small-indels	377	378	394	394

Table 1. Comparison of identified number of SNVs and small-indels in the TDS and the WGS and WES results.

The third blood–saliva comparison was based on independent comparisons of sequencing metrics among blood–saliva pairs for the WGS and WES protocols. The majority of metrics were concordant between blood- and saliva-derived gDNA sequencing results. Mapped reads and fragment lengths, on the other hand, were found to be lower in saliva. Duplicated reads were also higher in saliva samples; however, this was statistically significant only for WES (Figure 2).

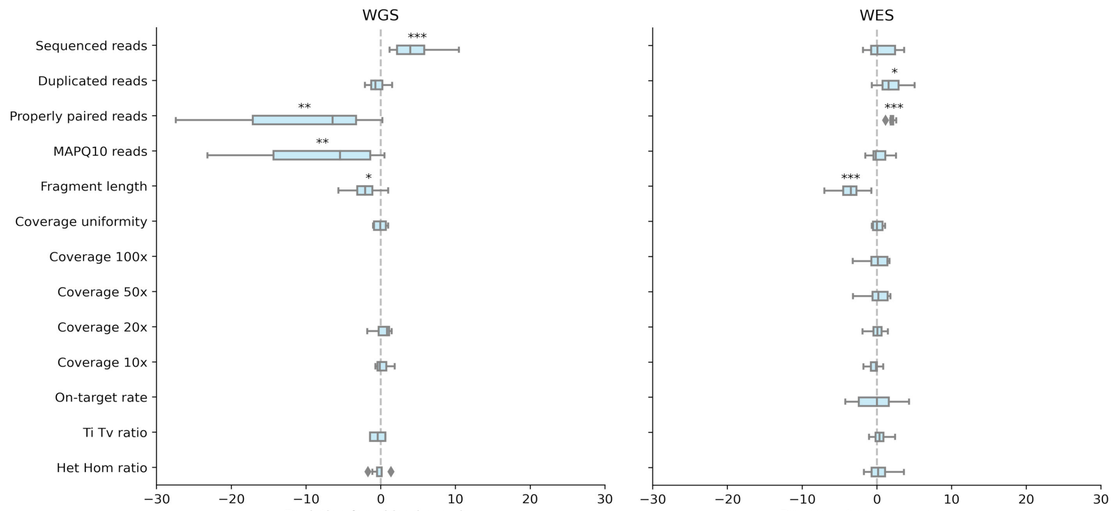

Figure 2. The quality control metrics of sequencing runs

Finally contamination detection was performed to assess the ratio of human and non-human sequencing reads in the saliva-derived gDNA samples.
Notably, both the iSeq and NovaSeq 6000 experiment revealed highly similar relative contamination rates, and this was typical for both the blood and the saliva samples (Supplementary Table).

Conclusion

Similarities in the sequencing accuracy and the distribution of different patterns of inaccuracies throughout the genome, together with results obtained when comparing other technical characteristics of the sequencing data, suggest that saliva-derived gDNA may be considered an equivalent material to blood-derived gDNA for WGS and WES analysis. Although microbiome-to-human misalignment in the saliva-derived samples cannot be equivocally ruled out, our results suggest that the effect does not deviate sequencing accuracy from values typically obtained using blood-derived gDNA.

Validated WGS and WES protocols proved saliva-derived gDNA as an equivalent to blood-derived gDNA for clinical and population genomic analyses

Katerina Kvapilova[1,2], Pavol Misenko[3], Jan Radvanszky[3,4,5,6], Ondrej Brzon[2], Jaroslav Budis[3,6,7], Juraj Gazdarica[3,6,7], Ondrej Pos[3,6], Marie Korabecna[8], *Martin Kasny[2,9], Tomas Szemes[3,5,6], Petr Kvapil[2], Jan Paces[1,10] and Zbynek Kozmik[11]

2 Institute of Applied Biotechnologies, Sluzeb 3056/4, 108 00 Prague, Czech Republic
* Correspondence: kasny@iabio.eu; Tel.: +420 739 394 364
(All affiliations in detail are presented in supplement)

Affiliations:

[1] Charles University, Faculty of Science, Albertov 6, 128 00 Prague, Czech Republic

[2] Institute of Applied Biotechnologies a.s., Sluzeb 4, 108 00 Prague, Czech Republic

[3] Geneton s.r.o., Ilkovicova 8, 841 04 Bratislava, Slovakia

[4] Institute of Clinical and Translational Research, Biomedical Research Centre, Slovak Academy of Sciences, Dubravska cesta 9, 814 39 Karlova Ves, Bratislava, Slovakia

[5] Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Ilkovicova 3278/6, 841 04 Karlova Ves, Bratislava, Slovakia

[6] Comenius University Science Park, Comenius University, Ilkovicova 8, 841 04 Karlova Ves, Bratislava, Slovakia

[7] Slovak Centre for Scientific and Technical Information, Lamacska cesta 8A, 811 04 Stare Mesto, Bratislava, Slovakia

[8] Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University and General University Hospital in Prague, Albertov 4, 128 00 Prague, Czech Republic

[9] Department of Botany and Zoology, Faculty of Science, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic

[10] Institute of Molecular Genetics of the Czech Academy of Sciences, Laboratory of Genomics and Bioiformatics, Videnska 1083, 142 20 Prague, Czech Republic

[11] Institute of Molecular Genetics of the Czech Academy of Sciences, Laboratory of Transcriptional Regulation, Videnska 1083, 142 20 Prague, Czech Republic

Supplementary Table. Sample/isolate QC parameters, WES/WGS libraries QC parameters, iSeq pre-sequencing and NovaSeq sequencing.

Sample/Isolate QC Parameters			WES Library QC Parameters for 1WB and 1S plex of 10 Libraries			WGS Library QC Parameters			iSeq		NovaSeq		
ID	c [ng/µl]	A260/280	Average Fragment Length at Fragment Range 200-800 bp [bp]	c [ng/µl]	c [nM]	Average Fragment Length at Fragment Range 200-9000 bp [bp]	c [ng/µl]	c [nM]	Total PE Reads	% of Human Reads Mapped [%]	Total PE Reads	Total Human PE reads	% of Human Reads Mapped [%]
WGS_01_WB	29.9	1.76	410	44.8	165.6	849	4.65	10.54	348 878	96	748 231 614	718 302 349	96
WGS_02_WB	35.5	1.88				814	6.59	15.43	338 090	96	858 214 586	823 886 003	96
WGS_03_WB	27.3	1.82				874	5.52	12.37	378 636	96	840 710 016	807 081 615	96
WGS_04_WB	28.5	1.87				836	5.47	12.57	348 014	96	885 780 514	850 349 293	96
WGS_05_WB	41.7	2.20				838	5.86	13.69	333 278	96	747 757 826	717 847 513	96
WGS_06_WB	22.6	1.86				823	5.34	12.44	291 576	96	969 512 698	930 732 190	96
WGS_07_WB	39.3	1.81				794	4.57	11.24	377 970	96	809 290 530	776 918 909	96
WGS_08_WB	26.8	1.85				825	7.20	17.42	353 760	96	877 294 582	842 202 799	96
WGS_09_WB	27.7	1.95				793	4.77	11.80	401 750	96	913 350 606	876 816 582	96
WGS_10_WB	44.6	1.83				763	7.72	19.72	341 794	96	853 134 596	819 009 212	96
WGS_01_S	28.4	1.88	388	37.5	146.4	1 324	3.14	6.23	509 314	94	971 005 578	912 745 243	94
WGS_02_S	34.5	1.92				1 265	3.95	8.53	505 884	70	1 255 819 802	879 073 861	65
WGS_03_S	75.2	2.25				1 057	3.54	7.64	351 974	89	1 127 165 422	1 003 177 226	88
WGS_04_S	38.5	1.93				1 332	1.90	4.18	466 744	81	1 240 438 380	1 004 755 088	78
WGS_05_S	29.9	1.87				1 511	1.89	4.00	542 622	61	1 461 669 022	877 001 413	54
WGS_06_S	64.1	1.83				948	4.63	10.54	333 330	95	1 001 619 454	951 538 481	94
WGS_07_S	46.3	1.72				1 411	2.75	5.39	511 142	72	1 119 801 438	806 257 035	70
WGS_08_S	52.2	1.92				1 197	4.63	9.65	446 348	64	1 573 635 498	991 390 364	58
WGS_09_S	31.8	1.91				1 443	2.32	4.70	771 842	92	931 465 120	856 947 910	91
WGS_10_S	26.5	1.83				1 175	4.30	9.42	536 142	85	999 442 302	849 525 957	84