Supplementary Materials

Evaluation of saliva as a source of accurate whole-genome and microbiome sequencing data

Anthony F. Herzig,[1,*] Lourdes Velo-Suárez,[1,2] Gaëlle Le Folgoc,[1] Anne Boland,[3,4] Hélène Blanché,[5,4] Robert Olaso,[3,4] Liana Le Roux,[6] Christelle Delmas,[7] Marcel Goldberg,[8] Marie Zins,[8] Franck Lethimonnier,[9] Jean-François Deleuze,[3,4,5,10], Emmanuelle Génin,[1,11,*]

1. Inserm, Univ Brest, EFS, UMR 1078, GGB, F-29200, Brest, France
2. Centre Brestois d'Analyse du Microbiote (CBAM), CHU Brest, F-29200, Brest, France
3. Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, F-91057, Evry, France
4. Laboratory of Excellence GENMED (Medical Genomics), F-75010, Paris, France
5. Fondation Jean Dausset - CEPH, F-75010, Paris, France
6. Centre d'Investigation Clinique 1412, Inserm, CHU Brest, F-29200, Brest, France
7. Pôle de Recherche Clinique, Inserm, F-75013 Paris, France
8. Inserm-Paris Saclay University UMS 011, Université de Paris, Villejuif, France
9. Alliance nationale pour les sciences de la vie et de la santé (Aviesan), Institut thématique multiorganisme, Technologies pour la santé, Inserm, F-75013, Paris, France
10. Centre de référence, d'innovation et d'expertise (CREFIX), US39, Commissariat à l'énergie atomique et aux énergies alternatives, F-91057, Evry, France
11. CHU Brest, F-29200, Brest, France

[*] Corresponding authors

Anthony Francis Herzig
anthony.herzig@inserm.fr
Inserm UMR 1078, 22 Avenue Camille Desmoulins, 29238 Brest, France

Emmanuelle Génin
emmanuelle.genin@inserm.fr
Inserm UMR 1078, 22 Avenue Camille Desmoulins, 29238 Brest, France

Quality Control

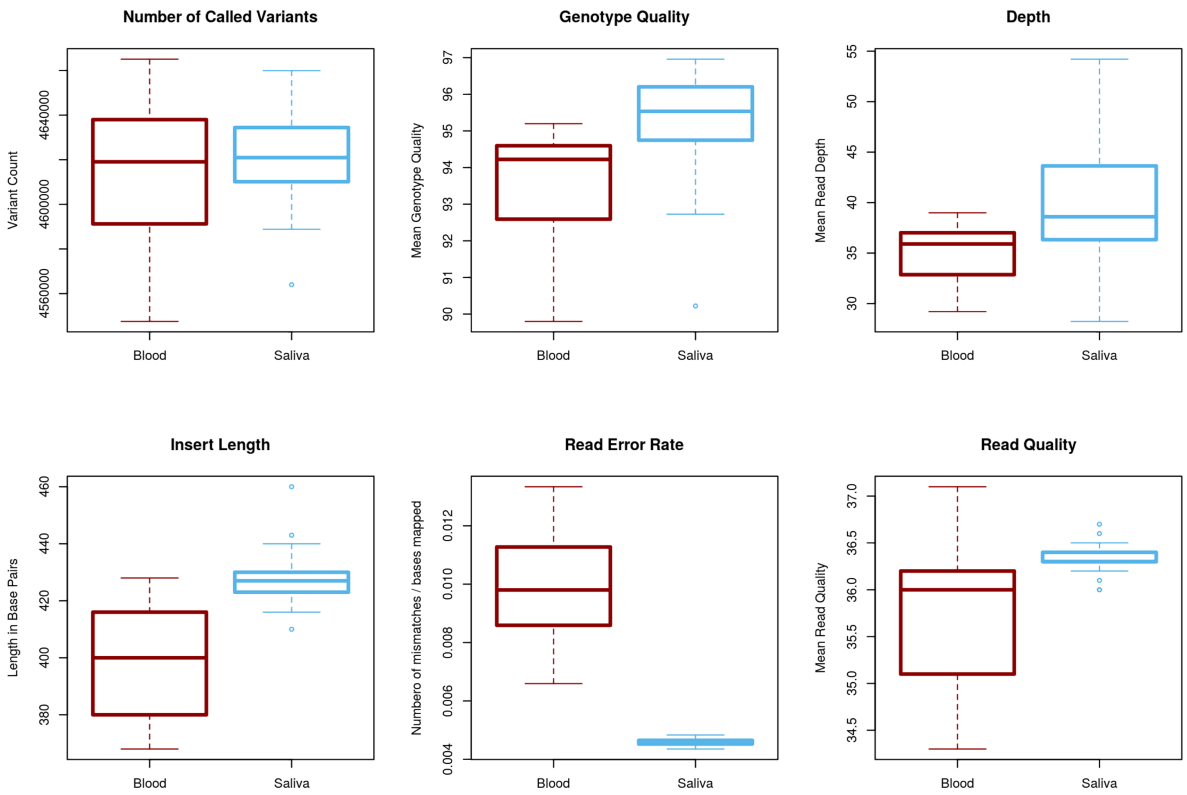
The quality control was performed by VCFProcessor (Ludwig, Marenne, & Génin, 2020) using the QC1078 setting. This involved the application of various criteria detailed below.

Genotypes were set to missing when:

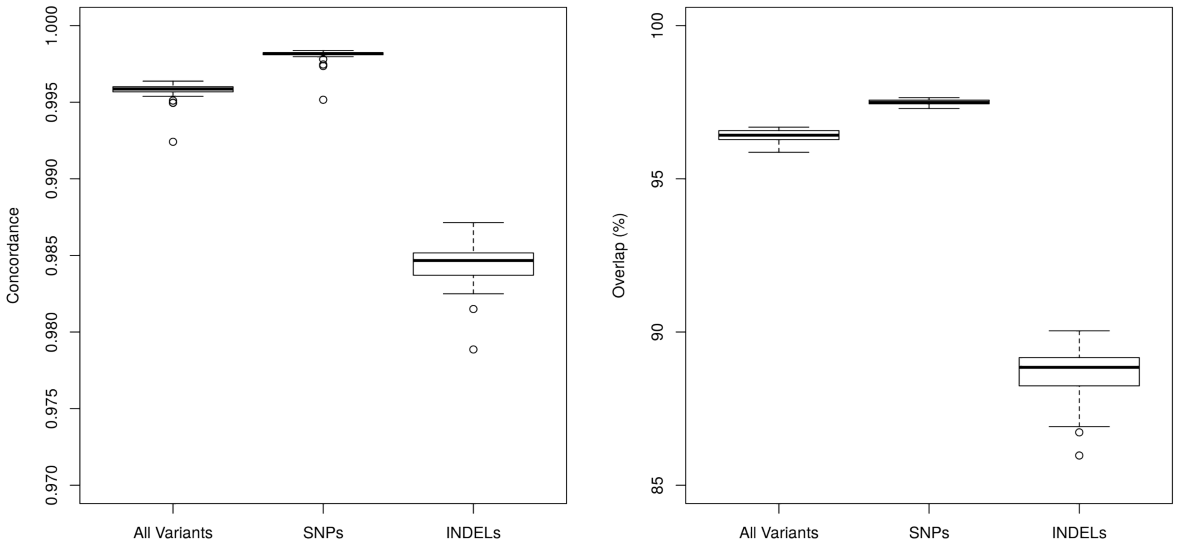
- Depth (DP) < 10
- Genotype Quality (GQ) < 20

Subsequently, variants were excluded (for all individuals) using the following criteria for various summary statistics (measured across all samples) generated by GATK v.3.8 (DePristo et al., 2011).

- Allele Balance for Heterozygous calls (ABHet) oustide of the range [0.25,0.75]
- Quality-By-Depth < 2
- MQRankSum < -12.5 (Z-score from Wilcoxon rank sum test of Alt vs. Ref read mapping qualities)
- Mapping Quality (MQ) < 40 for SNPs or < 10 for INDELS.
- Strand Bias odds ratio > 3 for SNPs or > 10 for INDELS
- Fisher's exact test for strand bias (phred scaled p-value) > 60 for SNPs or > 200 for INDELS
- HQRatio < 0.8
- Inbreeding Coefficient (estimated) < -0.8
- Callrate < 0.9

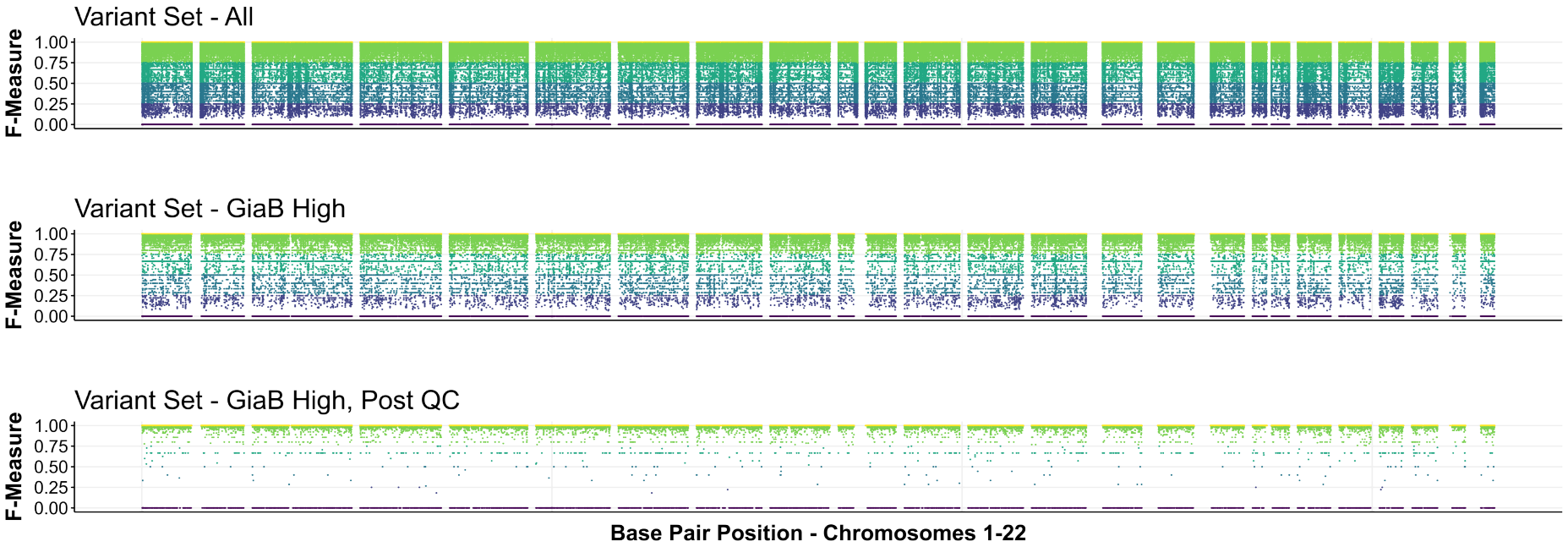


Supplementary Figure 1
Summary statistics relating to the quality of genomic data from blood (red) and saliva (blue). From top left and going clockwise: Number of variants called in single sample VCFs, mean GQ (Genotype Quality), mean read depth (DP), mean read quality measured from bam files, estimated read error rates estimated from bam files, and mean length of read inserts (distance between the two paired reads). Statistics were calculated using the GATK Haplotype Caller v.3.8 (DePristo et al., 2011) and samtools (Li et al., 2009).



Supplementary Figure 2

Percentage of variants identified in both saliva and blood. Overlap of variant calls and concordance of genotypes for overlapping variants are presented respectively on the left and right, respectively. One statistic is calculated per individual and then boxplots are constructed across the 39 individuals in the study. Results presented are for all variants, Single Nucleotide Polymorphisms (SNPs), and Insertions/Deletions (INDELs).

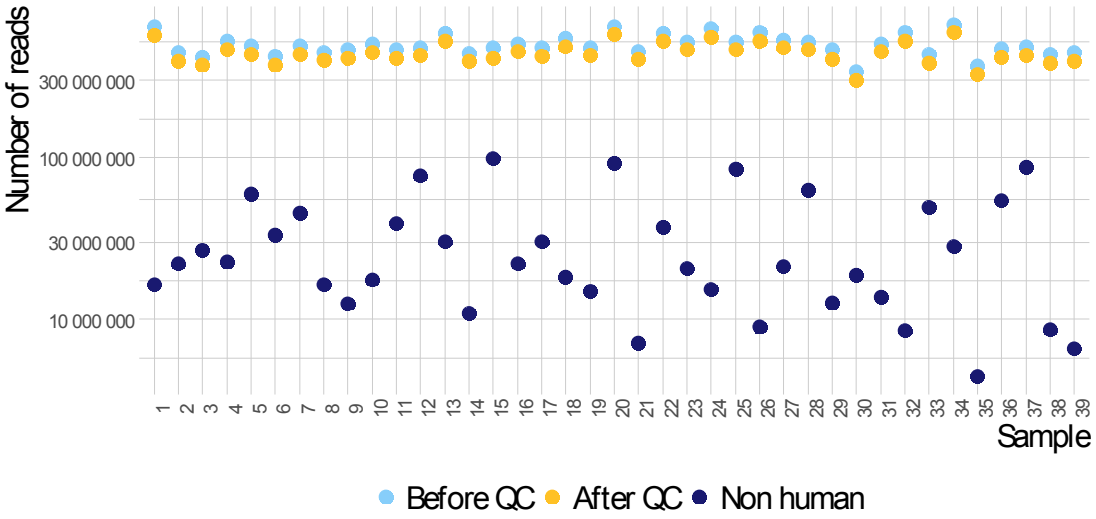


Supplementary Figure 3

Equivalent to Figure 1(b) in the main text but for all 22 autosomal chromosomes.

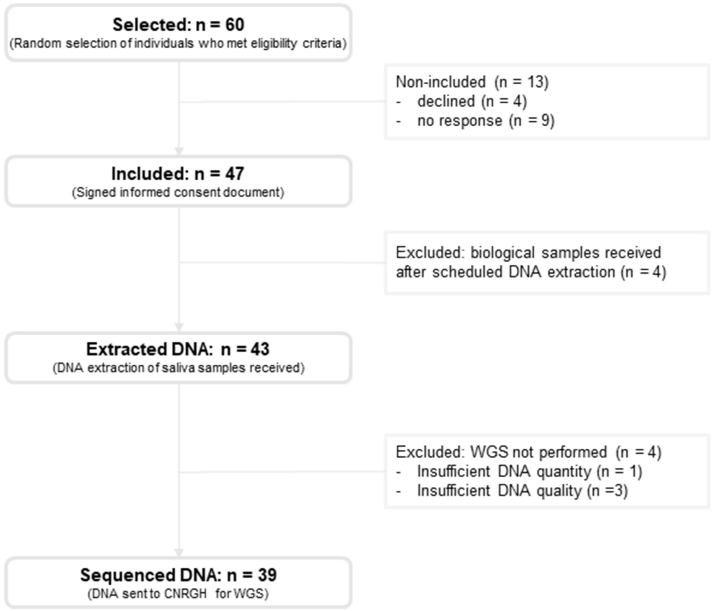
GiaB – Genome in a Bottle, a project which established lists of genomic regions that are commonly sequenced with either High or Low confidence.

QC – Quality control, the last panel is based on data after the application of quality control thresholds.



Supplementary Figure 4
For each individual in the study, the number of total reads before and after quality control as well as number of non-human reads (used for the analysis of salivary microbiomes) are presented.

GAZEL-ADN Recruitment



Supplementary Figure 5
Recruitment of individuals in to the GAZEL-ADN pilot project.

References

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. doi: 10.1038/ng.806

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Ludwig, T. E., Marenne, G., & Génin, E. (2020). VCFProcessor. Http://lysine.univ-brest.fr/vcfprocessor/index.html. Accessed 08/10/2020.