HiFi Long-read WGS: Sequencing Performance of Buccal/Saliva Samples and Estimation of Non-human DNA Contamination from Unaligned HiFi Reads

Stuti Joshi[1]; Kenny Chen[1]; Tiantian Geier[1]; Erica Smith[1]; Sami Belhadj[1]; UCI–GREGoR[2]; Emmanuèle Délot[2]; Seth Berger[1]; Eric Vilain[2]; Rachid Karam[1]

[1]Ambry Genetics. 1 Enterprise, Aliso Viejo, CA; [2] Institute for Clinical and Translational Science, University of California, Irvine, CA

BACKGROUND

Oral collection of genomic material, e.g., buccal swabs or saliva, offer non-invasive, autonomous, and patient-accessible options for genetic testing. Compared to blood, these sample types are challenging for long-read whole genome sequencing (LR-WGS), yielding a lower mass of genomic DNA (gDNA), poorer integrity, and potential contamination from oral microbes. Here, we show the performance of LR-WGS in 129 buccal and 34 saliva samples that were sequenced at Ambry Genetics for the UCI-GREGoR site. We also show methods to estimate non-human DNA contamination from unaligned HiFi reads.

METHODS

Isolation: Blood samples were isolated using the PacBio Nanobind HT CBB kit. Buccal swabs and saliva samples were isolated using the Zymo Quick DNA HMW kit.
QC: gDNA integrity was evaluated on Femto Pulse. The Genomic Quality Number (GQN) ranges from 0-10, representing 0-100% of fragments above 10kb or 30kb (GQN10 or GQN30, respectively).
Library Preparation: gDNA underwent mechanical fragmentation, HiFi library preparation[1] and gel-based size selection on the LightBench instrument[2].
Sequencing and Analysis: Libraries were sequenced on PacBio Revio using SPRQ chemistry. Raw data was processed via PacBio's HiFi-human-WGS-WDL workflow[3]. Microbial k-mers were evaluated with fastv[4].

RESULTS I: Performance of Buccal/Saliva samples from gDNA isolation to sequencing

HMW gDNA Isolation from Buccal/Saliva

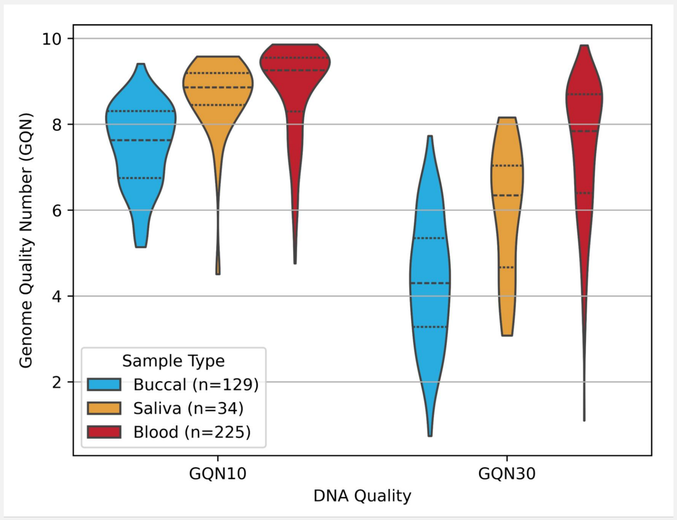

Figure 1. gDNA Quality Across Sample Types. Femto Pulse analysis revealed variable quality of HMW gDNA from oral specimen types. Acceptable quality for buccal and saliva was set at GQN10 ≥ 6.0 (Buccal: 91%, Saliva: 97%) and GQN30 ≥ 3.0 (Buccal: 78%, Saliva: 100%).

Table 1. Mean Genomic Quality Number By Sample Type

Sample Type	Mean GQN10	Mean GQN30
Buccal	7.5 (± 1.0)	4.3 (± 1.5)
Saliva	8.6 (± 0.9)	6.0 (± 1.5)
Blood	9.4 (± 0.6)	7.4 (± 1.5)

LR-WGS Performance of HiFi Libraries

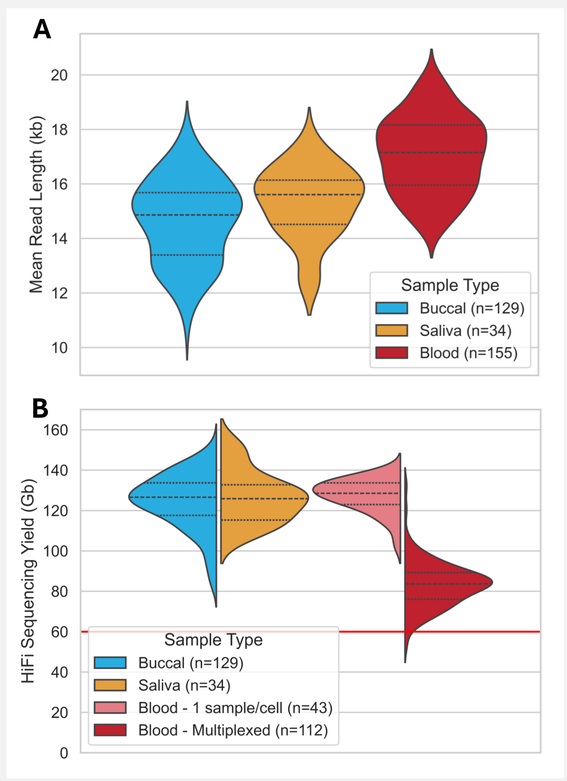

Figure 2. HiFi Sequencing Read Length and Yield. A) Mean read lengths across sample types. **B)** HiFi sequencing yield with SPRQ chemistry. Buccal and saliva samples were sequenced as 1 sample/cell, while blood samples were sequenced either as 1 sample/cell or 6 samples/4 cells.

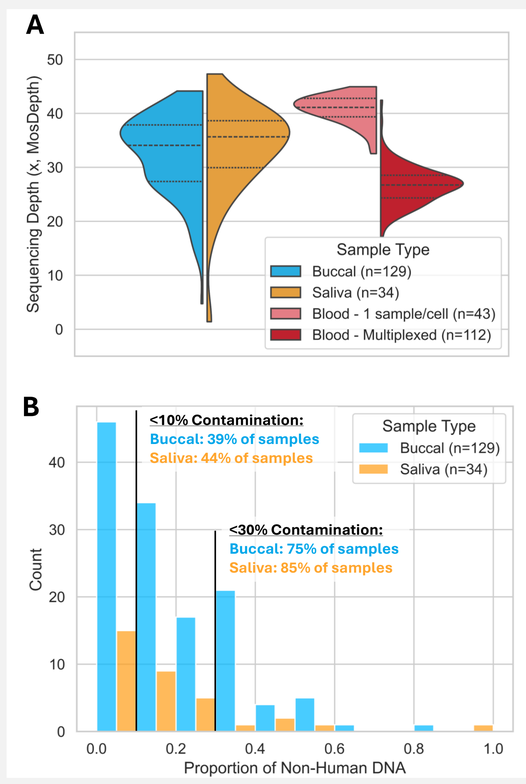

Figure 3. Post-alignment Sequencing Depth and Non-Human Content. A) Observed mean sequencing depth of the human genome. **B)** Frequency distribution of non-human content observed in samples from oral collection methods.

Key Observations:

- **Read Length:** Sequencing read lengths of buccal (14.7 ± 1.5 kb), and saliva (15.3 ± 1.3 kb) were shorter than blood (17.1 ± 1.4 kb).
- **Yield:** Revio sequencing with the improved SPRQ chemistry achieved 125 ± 13 Gb HiFi yield per SMRT Cell. 6-plex multiplexing, as tested with blood samples, achieved 83 ± 11 Gb, sufficient to meet 60 Gb threshold (~20x depth).
- **Sequencing Depth:** Compared to blood, oral sample types showed increased variability in mean sequencing depth when aligned to the hg38 reference genome. This could be attributed to non-human contamination from the oral microbiome.

RESULTS II: Estimating Non-human DNA contamination in Buccal/Saliva samples from unaligned HiFi reads

Correlation Between CpG Methylation Probabilities and Non-Human Content

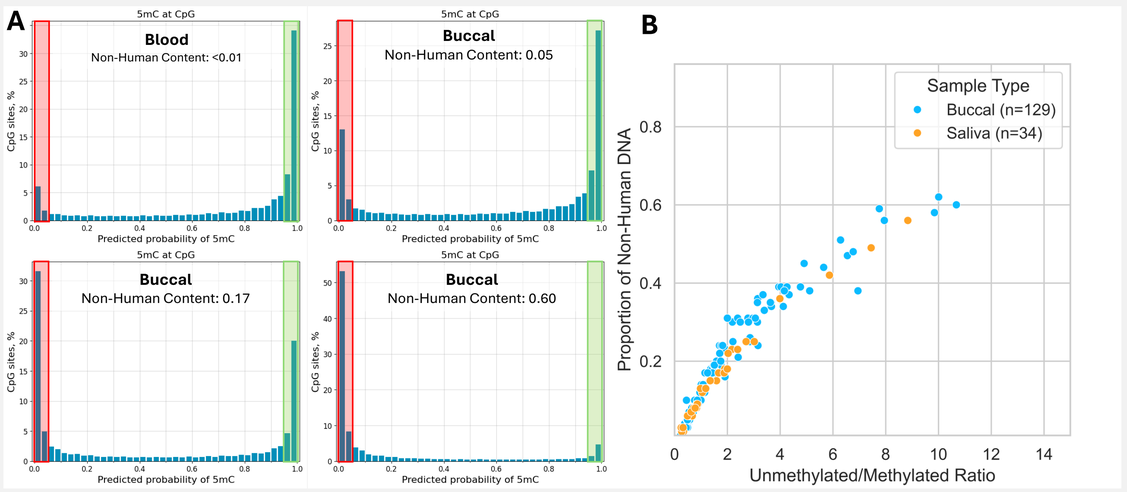

Figure 4. Comparison of 5mC Methylation Probabilities with Observed Non-Human Content. A) Example probability distributions of 5mC modifications at CpG sites. A ratio was calculated between the number of likely unmethylated sites (red) and likely methylated sites (green). **B)** Unmethylated vs Methylated ratios demonstrated a logarithmic correlation with the proportion of non-human DNA observed in buccal and saliva samples.

Key Observations

- 5mC levels were a good predictor of non-human DNA content, particularly below 40% contamination.
- K-mer analysis using fastv could identify and count microbe reads, allowing a more accurate estimation of sequencing depth.
- These could potentially be leveraged to estimate non-human DNA content after isolation to supplement QC and inform multiplexing.

K-mer-Corrected Analysis of Non-Human Content

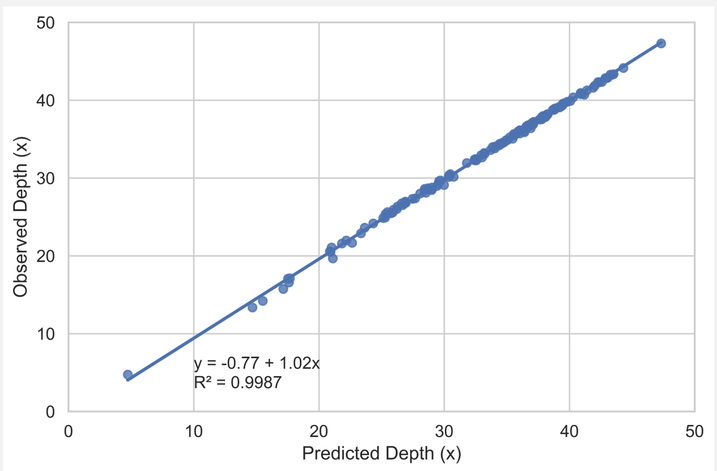

Figure 5. Correlation of Microbial K-mer Correction with Observed Post-Alignment Depth. Predicted coverage depth of 124 buccal samples and 11 saliva samples was calculated by removing reads corresponding to k-mers of oral microbial contaminants. Predicted depth showed a strong linear correlation with the observed post-alignment.

TAKE HOME POINTS

1. gDNA isolation of buccal and saliva samples enabled sufficient HMW DNA for LR-WGS sequencing. Post-alignment analysis revealed variable contamination levels from the oral microbiome, which reduced mean sequencing depth.
2. Analysis of 5mC levels and non-human k-mers offer means to estimate contamination, prior to alignment to the human genome.
3. Further work is needed to quantify contamination before sequencing, which could identify suboptimal samples and enable plexing.

REFERENCES

1. HiFi Prep 96 Procedure: https://www.pacb.com/wp-content/uploads/Procedure-checklist-Preparing-whole-genome-libraries-using-the-HiFi-prep-kit-96.pdf
2. Size selection using the LightBench: https://www.pacb.com/wp-content/uploads/Technical-note-Gel-cassette-size-selection-methods-for-WGS-HiFi-libraries.pdf
3. https://github.com/PacificBiosciences/HiFi-human-WGS-WDL
4. Shen C. et al. Briefings in Bioinformatics 2021