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]; John Harting[1], Yun-Hua Hsiao[1], Bojan Losic[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

Introduction: Oral collection of genomic material, e.g. buccal swab 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 less genomic DNA (gDNA), poorer integrity, and potential contamination from oral microbes. Here, we show the performance of LR-WGS in 129 buccal samples.

Methods: Buccal samples were isolated using the Zymo Quick DNA HMW kit, adjusting reagent volumes to support a higher input volume. DNA integrity was evaluated on Femto Pulse based on Genomic Quality Number (GQN). HiFi library preparation was conducted following the manufacturer's protocol, including gel-based size selection. Libraries were sequenced on PacBio Revio instruments using SPRQ chemistry. Raw data was processed via PacBio's HiFi-human-WGS-WDL workflow.

Results: Analysis of isolated gDNA demonstrated varying degrees of quality and degradation. 91% of samples demonstrated GQN10 > 6.0 (mean=7.5±1.0), while 78% of samples demonstrated GQN30 >3.0 (mean=4.3±1.5). HiFi sequencing achieved mean read lengths of 14.9±1.7 kb, compared to 17.1±1.4 kb in blood-derived libraries. Alignment to the human genome revealed reduced overall sequencing depth from microbial contamination (mean=19±16% contamination). Post-alignment, libraries averaged 32.4±7.4x coverage, with 94% of samples achieving ≥20x coverage. We compared two methods to evaluate the relationship of microbial contamination to sequencing outcomes. Levels of 5mC modifications were inversely correlated to the observed proportion of non-human DNA. Separately, contamination estimates calculated from microbial k-mer counts correlated with post alignment depth (R^2=0.9992).

Conclusions: Despite sample-to-sample variability, our gDNA isolation method enabled sufficient HMW DNA for LR-WGS sequencing. Analysis of 5mC levels or non-human k-mers offer potential predictive approaches to estimate microbial contamination before aligning to the human genome. Further work is required to assess if contamination levels can be experimentally calculated prior to loading SMRT cells in order to optimize loading and avoid sequencing quality failures.