Human Pangenome Reference Consortium
Level 1 Quality Metrics (YR1 Data Production) – Draft v1.1 November 3, 2020

The goal of the HPRC Level 1 data release policy is to encourage the timely public release of high-quality data types used in our assembly and evaluation production. Here we provide the guidelines and quality metrics of our lymphoblastoid cell lines (LCL lines), data production and storage that must be met before our formalized data release. This policy will be drafted and updated by the HPRC Administration and Coordination Working Group and approved by the HPRC steering committee.

LCL expansion and quality metrics: To minimize introduced sequence variation due to extended cell culture we limit our selection to low passage lines. LCLss are initially frozen at the lowest possible passage, in order to maximize growth potential and reduce cell senescence that can arise with higher passages. Following cryopreservation of the distribution bank, one frozen ampoule is recovered (at "passage-0") for quality control (QC) and viability testing by the Cell Culture Laboratory. A series of rigorous QC steps have been put in place to ensure each cell line is correctly identified and authenticated, free of contamination, and of the highest quality.. Typically, a portion of each original blood sample is retained unprocessed in the submitted tube for QC purposes; this aliquot is the primary identifier (specimen QC sample) and serves as the standard or baseline against which all subsequent expansions and samples are compared.

The following QC steps are used to evaluate cell culture contamination and viability:

- QC testing for bacterial and fungal contamination: cell cultures are grown and frozen in antibiotic-free media to aid in the detection and prevention of contamination.
- QC testing for mycoplasma contamination: cell lines are tested for mycoplasma contamination at freeze recovery via a sensitive and quantitative real-time PCR assay available from Applied Biosciences, Inc.
- QC testing for viability: recovered cultures are examined for viability and growth potential. LCLs must double within four days.

Cell pellets are expanded to support all standard production data types across the four HPRC production centers and banked in duplicate at one time, issuing an official lot freeze. At this time samples of the cells are submitted for two rounds of QC: cytogenetic characterization and microarray to evaluate introduced structural variation.

Cytogenetic Characterization of Cell Lines: Evaluation of G-banded metaphase chromosomes is used to verify that newly established and/or expanded cell cultures do not exhibit chromosomal abnormalities due to the cell culturing process. Chromosomes are counted per metaphase cell and individual and sex chromosomes are verified. Cytogenomic data is reviewed by Coriell's trained cytogenetic technologists andan external board certified Cytogenomics Consultant. A signed report is generated for each cell line and includes cytogenetic results that follow the International System for Human Cytogenetic Nomenclature (ISCN) format. The report also includes at least 2 images of

G-banded karyograms. Slides with dropped metaphases for analysis will be kept for one month, and copies of digital records will be kept for ten years.

Cytogenetic characterization of cell lines includes the following steps:
Twenty (20) cells are counted, and of these, at least five(5) are analyzed in their entirety to ensure all dark and light bands are present.

- If all cells are normal, the line is classified as normal
- If one (1) cell is abnormal, the line is classified as normal
- If two or more cells are abnormal, and the abnormalities for each cell are different, the line is classified as normal
- If two (2) cells of the same abnormality are present, the finding is reportable and the line is classified as abnormal

-When cells are karyotyped, chromosomes are cut out and paired next to each other to facilitate in-depth band-by-band analysis.

Chromosomal microarray: DNA extracted from cell cultures is used for downstream chromosomal microarray analysis. This analysis affords the opportunity to query the entire genome at high resolution for copy number variations. Samples are genotyped with the Illumina OMNI2.5 Human SNP Array (InfiniumOmni2-5-8v1-3_A1) offers ~2.4 million fixed markers, which allows the detection of loss of heterozygosity, and large CNVs (i.e., deletions or duplications) as small as 100 kbp neither of which is detectable by G-banding. This allows us to study each sample relative to a comprehensive set CNVs identified in large cohorts of samples (n=29,085 cases) with known neurodevelopmental disorders (Coe et al, Nature Genetics 2014) to flag likely pathogenic variation. Microarray data are generated at CHOP and analyzed using PennCNV and cnvPartition for copy number changes (data analysis performed in the Eichler Lab, UW Genome Sciences). The results will be reviewed alongside the G-banded karyotyping results and all cases are reviewed by the HPRC.

Level-1 Data Quality Metrics
Cell pellets that pass all QC are shipped directly to our four HPRC data production centers. Data generation for each line have quality metrics, or minimum thresholds specific to each platform, as outlined below:

- **PacBio HiFi Data**: Target library size of ~20 kb for ~17 kb reads, 4 SMRT Cells 8M to obtain a minimum of 30X coverage (96 Gbp) in >Q20 HiFi reads.

- **Oxford Nanopore PromethION (Ultralong Data)**: Median alignment identity per read ≥94% (Guppy 4.0.11). Average read N50 per flow cell of >60 kb. Average total coverage with 3 flow cells, >15x. Average coverage in >100 kb read length (3 flow cells) is >6x and average coverage in >200 kb read length (3 flow cells) is >1.4x.

- **UCSC Hi-C Omni-C Data:** At least two PCR replicates will be generated from Omni-C ligation products for each sample. Total library complexity will be at least 600 million when calculated using Picard Tools EstimateLibraryComplexity . Less than 35% of read pairs will span less than 500 bp, and total median genome coverage will be greater than or equal to 15X per library at 200 million 2x150 read pairs.

- **BioNano Optical Maps:** Label density must be observed 10-20 labels within each 100 kb window, yield of fragments >= 150 kbp will be at least 320 Gb, with an N50 of at least 250kb. Genome map alignments must be at least 50 Mb.

Level-1 Data Release

HPRC Level-1 datasets that pass minimum quality thresholds are considered to be high quality and move into the next phase of data public release. Sequence data, associated quality base information, annotation files, and cytogenetic images/microarray results are uploaded to a public resources HPRC cloud-storage (AnVIL 'AnVIL_HPRC' workspace and/or Amazon AWS s3 public resource storage. If quality issues are noted during the initial assembly production of Level-2 'reference-grade' data, then that information will be reported and hosted in the same location.