HPP/HPRC Metadata Proposal

15 Oct. 2020

Introduction: *The Human Pangenome Project utilizes and generates a variety of data, and that data should be made available (along with rich metadata) in as many public repositories as possible. This document provides a proposal for the metadata that will be provided by all data generation facilities in order for it to be uploaded to public repositories such as SRA and Gen3 by UCSC.*

The inclusion of metadata allows for the creation of artificial cohorts with data stored in Gen3/AnVIL as well as in SRA -- which uploads all metadata to AWS and GCP for querying in Athena and BigQuery, respectively.

The SRA and Gen3 require a minimal set of metadata, but do not provide guidance on what additional attributes to include, or how to define the attributes consistently across sequencing centers and consortia. This document is also written to help start the process of defining a flexible, but consistent set of metadata for the HPP's raw data types.

SRA Required Fields:

The following fields are required for all sequencing file types deposited to SRA (according to the SRA's library and metadata terms sheet).

Attribute	Description	Allowed Values	Example Value
sample_name	BioSample name	Must match name entered for BioSample	SAMN14611343
library_ID	Short identifier for library ID	Free text with no space characters	H_IJ-lib4.1
title	Short description to identify the dataset	{sequencing type} Sequencing of {sample alias}	PacBio Sequel II Sequencing of NA12878
library_strategy	General approach to library preparation	{WGS | OTHER}	WGS
library_source	Source of sequencing/library material	{GENOMIC}	GENOMIC
library_selection	Method of selection of library source material	{RANDOM | MNase | size fractionation}	size fractionation
library_layout	Format of sequence reads	{single | paired}	single

platform	Sequencing instrument manufacturer	{ILLUMINA | PACBIO_SMRT | OXFORD_NANOPORE}	PACBIO_SMRT
instrument_model	Model of instrument	{PacBio Sequel II | PromethION | NextSeq 500 | NextSeq 550 | Illumina NovaSeq 6000}	PacBio Sequel II
design_description	Brief description of materials and methods	Free Text	HiFi sequencing of 20kb fractionated gDNA
filetype	Type of file uploading	{fastq, fast5, tar, bam, cram}	bam
filename	File the metadata refers to	Free Text (without any spaces)	H_IJ-lib4.1.bam

Gen3 (Additional) Required Fields:

The following fields are required to be added to the HPP data model for indexing in Gen3 according to the Gen3/AnVIL data dictionary. These attributes are required in addition to the SRA attributes.

Attribute	Description	Allowed Values	Example Value
md5sum	Md5 of file (note: must be calculated in unchunked state)	32 character hash	a2631443d5363d4450543f145cd15734
file_size	Size of uploaded file	Integer with number of bytes (not in MB, GB, or TB)	199829389
data_type	Content of data file	{aligned reads| unaligned reads| analysis supplement}	unaligned reads

PacBio HiFi Supplemental Metadata:

Proposed supplemental metadata to include with all PacBio submissions.

Attribute (Column)	Description	Allowed Values	Example Value
n50_read_length	N50 read length	Integer (no commas or special characters)	50000

read_len_25_percentile	25th percentile of read length distribution	Integer (no commas or special characters)	10000
read_len_50_percentile	50th percentile of read length distribution	Integer (no commas or special characters)	15000
read_len_75_percentile	75th percentile of read length distribution	Integer (no commas or special characters)	20000
dna_extraction_method	DNA extraction kit	{ Circulomics NanoBind CBB Big DNA | Qiagen Gentra Puregene | Qiagen MagAttract HMW | }	Qiagen MagAttract HMW
shear_method	Approach to initial DNA fragmentation	{Megaruptor 1 | Megaruptor 3 | g-TUBE | needle shear | no shear} {opt. free text settings}	Megaruptor 3
shear_method_notes	Free text description of shear method settings	Free Text	Setting 30 then setting 31 to peak size of 18 kb
size_selection	Approach to final library size-selection	{SageELF | BluePippin | SageHLS | AMPure | Circulomics SRE | no SS} {opt. free text settings}	SageELF
size_selection_notes	Free text description of size selection settings	Free Text	1-18kb protocol, fractions 2 and 3
yield	Total number of bases (Divide by genome size to get average coverage)	Integer (no commas or special characters)	3000000000
yield_q20	Total number of bases > Q20 (Divide by genome size to get average coverage)	Integer (no commas or special characters)	2800000000
total_reads	Total number of reads	Integer (no commas or special characters)	280000
reads_q20	Total number of reads with Q>20	Integer (no commas or special characters)	270000
ccs_algorithm	Version of consensus sequence generation algorithm	Free Text	8.0.0.80529

polymerase_version	Sequel2 polymerase version	{ P2.0 | P2.1 }	P2.0
seq_plate_chemistry_version	Sequel2 sequencing plate chemistry	{ C2.0 }	C2.0
generator_facility	Facility that created the sequencing reads	Free Text	Washington University
generator_contact	Contact person for problems/inquiries about the data	Free Text	beth@suggestion.edu
notes	Free text used to flag major issues	Free Text	Sample has contaminating reads from E.coli

Hi-C Supplemental Metadata:

Proposed supplemental metadata to include with all Hi-C/Omni-C submissions.

Attribute (Column)	Definition	Allowed Values	Example Value
repl_per_ligation	Number of pcr replicates per ligation product	Integer (no commas or special characters)	2
yield	Total number of bases (Divide by genome size to get average coverage)	Integer (no commas or special characters)	300000000
total_reads	Total number of reads (including forward and reverse)	Integer (no commas or special characters)	280000
hic_type	Type of Hi-C chemistry	{Omni-C}	Omni-C
generator_facility	Facility that created the sequencing reads	Free Text	Washington University
generator_contact	Contact person for problems/inquiries about the data	Free Text	beth@suggestion.edu
notes	Free text used to flag major issues	Free Text	Sample has contaminating reads from E.coli

Oxford Nanopore Supplemental Metadata:

Proposed supplemental metadata to include with all Oxford Nanopore submissions.

Attribute	Definition	Allowed Values	Example Value

(Column)			
n50_read_length	N50 read length	Integer (no commas or special characters)	50000
read_len_25_percentile	25th percentile of read length distribution	Integer (no commas or special characters)	10000
read_len_50_percentile	50th percentile of read length distribution	Integer (no commas or special characters)	15000
read_len_75_percentile	75th percentile of read length distribution	Integer (no commas or special characters)	20000
shear_method	Approach to initial DNA fragmentation	{Megaruptor 1 | Megaruptor 3 | g-TUBE | needle shear | no shear} (opt. settings free text)	Megaruptor 3 setting 30
size_selection	Approach to final library size-selection	{SageELF | BluePippin | SageHLS | AMPure | Circulomics SRE | no SS} {opt. free text settings}	SageELF 1-18kb
seq_kit	Sequencing kit used for data generation	Free Text	SQK-LSK109
basecaller_version	Basecaller version	Free Text	V3.2.4
yield	Total number of bases (Divide by genome size to get average coverage)	Integer (no commas or special characters)	3000000000
total_reads	Total number of reads	Integer (no commas or special characters)	280000
generator_facility	Facility that created the sequencing reads	Free Text	Washington University
generator_contact	Contact person for problems/inquiries about the data	Free Text	beth@suggestion.edu
notes	Free text used to flag major issues	Free Text	Sample has contaminating reads from E.coli