HTAN Sequencing Data Requirements

Author: *Jennifer Altreuter, DFCI*

Updated: Aug 27, 2024

Status: Approved

Terminology:

This document uses the following terms from IEFT RFC 2119

- **MUST / REQUIRED / SHALL**: ✅ (denotes absolute requirement)
- **MUST NOT / SHALL NOT**: ❌ (denotes absolute prohibition)
- **SHOULD / RECOMMENDED**: 🌟 (denotes recommendation)
- **SHOULD NOT / NOT RECOMMENDED**: 🤦‍♂️ (denotes not recommended)
- **MAY / OPTIONAL**: 🤷‍♀️ (denotes optional)

HTAN supports multiple types of sequencing data. This document is organized by assay type. Please see the table of contents below to navigate to a particular assay type.

General information about sequencing data levels is summarized in Table 1.

Table 1 HTAN Data Levels for Sequencing Data

Level	Definition	Example Data
1	Raw data	FASTQs, Unaligned BAMs
2	Aligned primary data	Aligned BAMs
3	Derived biomolecular data	Gene expression matrix files, VCFs, etc.
4	Sample-level summary data	t-SNE plot coordinates, etc.

Bulk DNA-seq

Level 1

✅ FASTQ files or Unaligned BAM files **MUST** be submitted for all sequencing data.

✅ Each FASTQ or Unaligned BAM file **MUST** have a single record (row) in the manifest.

Level 2

✅ Level 2 data **MUST** be submitted if alignment was performed.

Level 3

✅ Level 3 DNA-seq files **MUST** include a vcf file containing called variants.

🌟 Level 3 DNA-seq files **SHOULD** include a seg file if copy number variation was assessed.

🤷 Submission of maf files is **OPTIONAL.**

Bulk RNA-seq

Level 1

✅ FASTQ files or Unaligned BAM files **MUST** be submitted for all sequencing data.

✅ Each FASTQ or Unaligned BAM file **MUST** have a single record (row) in the manifest.

Level 2

✅ Level 2 data **MUST** be submitted if alignment was performed. (If a bam file is not produced because of the use of pseudo-aligners like Kallisto or Salmon, this level is not required.)

🌟 Alignment **SHOULD** use assembly version GRCh38 and the consensus GENCODE version. The HTAN Phase 2 consensus GENCODE version will be determined by HTAN Centers at the beginning of HTAN Phase 2.

Level 3

✅ Gene expression data **MUST** be in csv or market exchange format (mtx or mex)

✅ Genes **MUST** be identified using ENSEMBL gene identifiers e.g., ENSG00000242268.2.

🤷 Submission of multiple Level 3 data files for each sample, generated using different pipelines or analysis methods, or based on batch-correction is **OPTIONAL**.

Bulk Methylation-seq

Level 1

✅ FASTQ files or Unaligned BAM files **MUST** be submitted for all sequencing data.

✅ Each FASTQ or Unaligned BAM file **MUST** have a single record (row) in the manifest.

Level 2

✅ Level 2 data **MUST** be submitted if alignment was performed.

Level 3

✅ Level 3 data **MUST** represent derived data such as gene expression matrix files, VCFs, etc. .

Hi-C

Level 1

✅ FASTQ files or Unaligned BAM files **MUST** be submitted for all sequencing data.

✅ Each FASTQ or Unaligned BAM file **MUST** have a single record (row) in the manifest.

Level 2

✅ Level 2 data **MUST** be submitted if alignment was performed.

Level 3

🌟 Level 3 Hi-C **SHOULD** include files to represent Stripe peaks and topologically associated domains (bed), loop peaks (bedpe), and compartment values.

🌟 Level 3 Hi-C **SHOULD** include files to represent stripe strengths (bedGraph/bigwig).

Multiplexed CITE-seq

The multiplexed CITE-seq assay has multiple samples represented in one level 1 FASTQ file. To facilitate associating the FASTQ with all downstream files, a field called “Associated mRNA Library Data File ID” was created. In addition, this assay uses two sets of protein (antibody) tags: 1) tags for de-multiplexing 2) tags for identification of surface proteins. Information about the multiplexing methods and the protein libraries is also included in the level 1 manifest.

Level 1

✅ FASTQ files or Unaligned BAM files **MUST** be submitted for all sequencing data.

✅ Each FASTQ or Unaligned BAM file **MUST** have a single record (row) in the manifest.

✅ Each FASTQ/unaligned-bam file **MUST** be associated with a mRNA Library Data File. The mRNA Library Data File lists the HTAN IDs for all mRNA Libraries associated with the FASTQ/unaligned-bam.

✅ Feature and hashing barcodes used to create the bam file (CITE-seq, hashing) **MUST** be indicated in “Single Cell Barcode Method Applied”. Multiple methods **MUST** be comma-separated.

Level 2

The multiplexed CITE-seq assay has multiple samples represented in one level 2 bam file.

✅ All HTAN Parent Biospecimen ID associated with a bam file **MUST** be provided. Multiple Parent Biospecimen IDs should be separated by commas.

✅ Each file **MUST** be associated with a mRNA Library Data File. The mRNA Library Data File lists the HTAN IDs for all mRNA Libraries associated with the alignment file.

Level 3

✅ All HTAN Parent Biospecimen ID associated with a file **MUST** be provided. Multiple Parent Biospecimen IDs should be separated by commas.

✅ Each file **MUST** be associated with a mRNA Library Data File. The mRNA Library Data File lists the HTAN IDs for all mRNA Libraries associated with the level 3 file.

Level 4

✅ All HTAN Parent Biospecimen ID associated with a file **MUST** be provided. Multiple Parent Biospecimen IDs should be separated by commas.

✅ Each file **MUST** be associated with a mRNA Library Data File. The mRNA Library Data File lists the HTAN IDs for all mRNA Libraries associated with the level 4 file.

scDNA-seq

Level 1

✅ FASTQ files or Unaligned BAM files **MUST** be submitted for all sequencing data.

✅ Each FASTQ or Unaligned BAM file **MUST** have a single record (row) in the manifest.

Level 2

✅ Level 2 data **MUST** be submitted if alignment was performed.

scATAC-seq

Level 1

✅ FASTQ files or Unaligned BAM files **MUST** be submitted for all sequencing data.

✅ Each FASTQ or Unaligned BAM file **MUST** have a single record (row) in the manifest.

Level 2

✅ Level 2 data **MUST** be submitted if alignment was performed.

Level 3

🌟 Level 3 scATAC-seq files **SHOULD** include a peak by cell matrix file (csv/mtx), a peak file (bed) and a fragments file (tsv/txt).

snmC-Seq

single-cell (nucleus) DNA methylome (methyl-C)

Level 1

✅ FASTQ files or Unaligned BAM files **MUST** be submitted for all sequencing data.

✅ Each FASTQ or Unaligned BAM file **MUST** have a single record (row) in the manifest.

Level 2

✅ Level 2 data **MUST** be submitted if it was performed.

Level 3

🌟 Level 3 snmC-seq files **SHOULD** include a bins by cell matrix file (csv/mtx), a peak summits file (bed) and a fragments file (tsv/txt).

scRNA-seq/snRNA-seq

Level 1

✅ FASTQ files or Unaligned BAM files **MUST** be submitted for all sequencing data.

✅ Each FASTQ or Unaligned BAM file **MUST** have a single record (row) in the manifest.

Level 2

✅ Level 2 data **MUST** be submitted if it was performed. (If a bam file is not produced because of the use of pseudo-aligners like Kallisto or Salmon, this level is not required.)

🌟 Alignment **SHOULD** use assembly version GRCh38 and the consensus GENCODE version. The HTAN Phase 2 consensus GENCODE version will be determined by HTAN Centers at the beginning of HTAN Phase 2.

Level 3 and Level 4

HDF5-backed AnnData (h5ad) is the preferred format for both Level 3 and 4 data with Level 3 being raw and normalized gene expression values and Level 4 being cell-type assignments (including tSNE / UMAP coordinates, cell types and any other features inferred about individual cells).

🌟 h5ad format **SHOULD** be used.

🌟 h5ad formatted files containing both level 3 and level 4 data **SHOULD** be submitted with a level3 Manifest.

🌟 h5ad formatted file **SHOULD** follow the cellxgene schema.

✅ Raw gene expression matrices **MUST** be submitted. For h5ad format, these are provided in raw.X

🌟 Normalized/"final" gene expression matrices **SHOULD** be submitted. Normalized/"final" gene counts are ***strongly recommended***. For h5ad format, these are provided in .X.

✅ Genes **MUST** be identified using ENSEMBL gene identifiers e.g., ENSG00000242268.2.

🙅‍♂️ Genes **SHOULD NOT** be filtered.

✅ One of the provided cell types **MUST** follow CL Ontology (https://www.ebi.ac.uk/ols4/ontologies/cl)