SMaHT Sample & File Nomenclature

Resource for the SMaHT Network Members
Version 1.2
Last modified: 2024-Oct-31

Description

The SMaHT sample and file names are the primary identifiers of biosamples and files generated by the Tissue Procurement Center (TPC) and Data Analysis Center (DAC) of the SMaHT Network (“Network”).

Overview

The SMaHT sample and file names contain identifiers that are unique and immovable, as well as semi-human readable codes that correspond to metadata. This document describes the naming schema and tables of codes for each metadata type that are included in sample and file names.

Naming schema:

[Project][Kit/Donor ID] - [Protocol ID] - [Tissue aliquot & Tissue core IDs] - [Sex & Age] - [Sequencing platform & Sequencing assay code] - [Center code] - [Unique Accession ID] - [Analysis info] . [File format]

Tissue

The donor, tissue, and sample-related metadata and identifiers are generated at TPC/NDRI and are used to name samples that are sent to GCCs and TTDs. Donor/tissue metadata fields are delimited by a hyphen ("-"). "#" indicates an integer number, and "A" indicates an alphabet letter.

For benchmarking tissues, the project is "**ST**".
For production tissues, the project is "**SMHT**".

Some tissues were provided to TTDs to test their protocols. For these non-benchmarking, non-production tissues, the project is "**SN**".

Schema:	**ProjectIDDonorKitID – TissueProtocolID – TissueAliquotIDCoreID**
Benchmarking tissues:	**ST### – #A – ###A#**
Production tissues:	**SMHT### – #A – ###A#**
Non-benchmarking, non-production tissues:	**SN### – #A – ###A#**

Note:

1. The majority of benchmarking tissues will be homogenized. At specific requests by certain GCCs, non-homogenized benchmarking tissues have been made available for sequencing.
2. All production tissues will *not* be homogenized.
3. For non-homogenized tissues (for both benchmarking or production), there will be tissue aliquots from each tissue, which then get sectioned into "tissue cores" or "tissue samples", which are larger than cores and smaller than tissue aliquots.
4. GCCs do not need to rename the samples received from TPC when submitting the metadata to DAC. They need to provide DAC with the metadata they received from TPC and any information related to sample processing performed at GCC, e.g. aliquot numbers for benchmark homogenate tissues; which aliquot(s) were used to create a library.

- **Project**: **ST** for a benchmarking tissue sample; **SMHT** for a production tissue sample.
- **Kit ID**: Provided by TPC/NDRI. It is a three-digit pre-generated kit identifier that references the donor. 001–150 for production tissues; 001–004 for benchmarking tissues.
- **Protocol ID**: Automatically generated by NDRI's Rhythm system. A single internal protocol group number is followed by a letter designation (A–ZZ) corresponding to a specific combination of tissue type and tissue preservation method (Table 1).
- **Tissue aliquot & core IDs**: Provided by TPC/NDRI. Aliquot ID refers to spatially numbered tissue aliquots from longitudinally bisected or non-bisected tissue samples. Serial tissue cores subsampled from within an aliquot are designated by a letter followed

by a number (Figures 1 and 2). No aliquots or cores will be available for buccal swab samples. No cores will be available for isolated fibroblasts and benchmarking tissue homogenate samples. Null values are represented as "**X**".

- For the homogenate benchmarking tissue (no core or tissue sample):
  - [Metadata from GCC] **Tissue aliquot** number will be provided by TPC, e.g., **ST### – #A – ###X**
  - [File names at DAC] **Tissue aliquot** and **Core ID** are both "**X**", e.g., **ST### – #A – XX**
    - To avoid having multiple homogenate aliquots associated with a file name.
- For the non-homogenate benchmarking tissue (in either core or tissue sample form):
  - [All] **ST### – #A – ###A#**

- **Core ID** for the core sample: ID is comprised of a letter between **A–F**, followed by a digit between **1–6** (ID is associated with the spatial information from within the individual tissue aliquot, - See Figure 1).
- **Core ID** for the tissue specimen: ID is comprised of a letter between **S–W** followed by **1–9** (no spatial information associated with the ID).
- **Core ID** for the homogenate benchmarking tissues: Null value for the ID, i.e., "**X**".

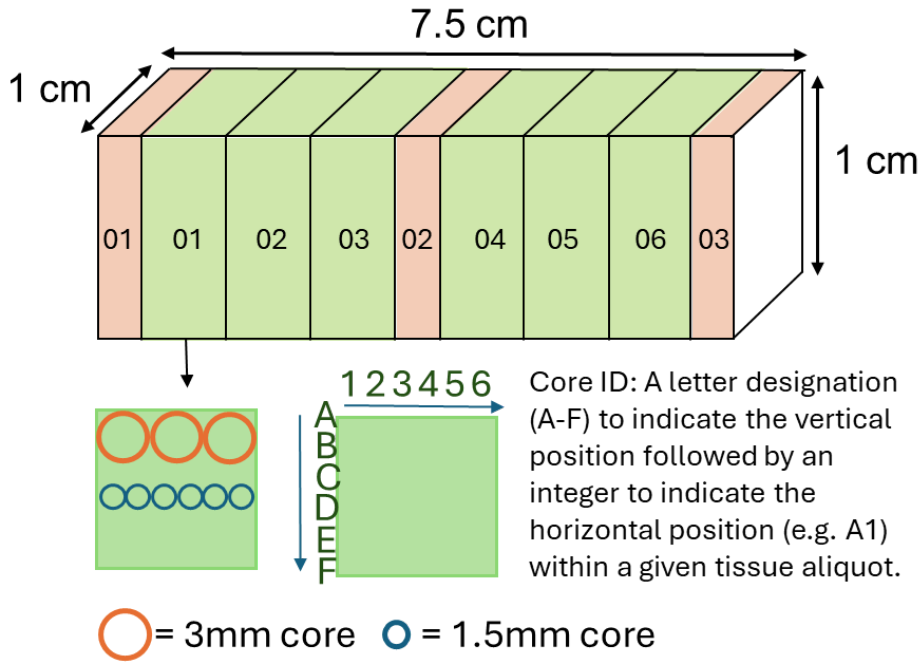

Figure 1. Example of spatially numbered tissue aliquots from non-bisected tissue samples. Fixed (pink) or frozen aliquots (green) are numbered separately. "**X**" represents null values.

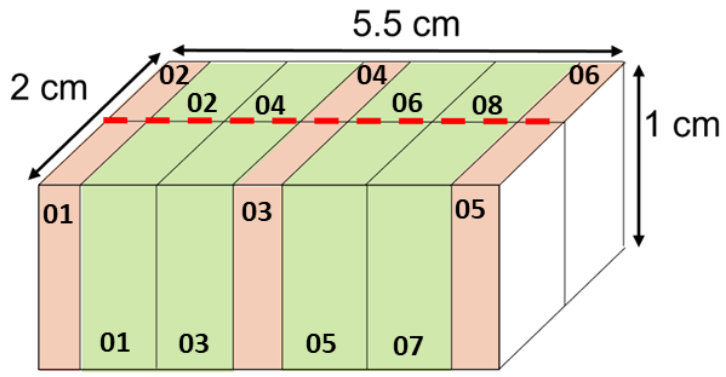

Figure 2. Example of spatially numbered tissue aliquots from longitudinally bisected tissue samples. Fixed (pink) or frozen (green) aliquots are numbered separately, in sequential order starting from the superior end of the sample.

Table 1A. Protocol IDs for SMaHT **benchmarking tissues.**

Sample Protocol ID	Tissue Name	Preservation	Notes
1A	Liver	Snap Frozen	Homogenates and non-homogenate samples
1B	*unassigned*	*N/A*	
1C	Liver	Fixed	
1D	Lung	Snap Frozen	Homogenates and non-homogenate samples
1E	*unassigned*	*N/A*	
1F	Lung	Fixed	
1G	Colon	Snap Frozen	Homogenates and non-homogenate samples
1H	*unassigned*	*N/A*	
1I	Colon	Fixed	
1J	Skin	Snap Frozen	Collected but not utilized for study; homogenization attempted but not successful.
1K	Skin	Snap Frozen	Cores made (intact tissue)

1L	Skin	Fixed	
1M/N/O/P	*unassigned*	*N/A*	
1Q	Brain, Frontal Lobe	Snap Frozen	Homogenates and non-homogenate samples

Table 1B. Protocol IDs for SMaHT **production tissues**.

Protocol ID	Tissue Name for Container	Preservation
3A	Blood, Whole	Snap Frozen
3B	Buccal Swab	Fresh
3C	Esophagus	Frozen
3D	Esophagus	Fixed
3E	Colon, Ascending	Snap Frozen
3F	Colon, Ascending	Fixed
3G	Colon, Descending	Snap Frozen
3H	Colon, Descending	Fixed
3I	Liver Sample	Snap Frozen
3J	Liver Sample	Fixed
3K	Adrenal Gland, Left	Snap Frozen
3L	Adrenal Gland, Left	Fixed
3M	Adrenal Gland, Right	Snap Frozen
3N	Adrenal Gland, Right	Fixed
3O	Aorta, Abdominal	Snap Frozen
3P	Aorta, Abdominal	Fixed
3Q	Lung	Snap Frozen
3R	Lung	Fixed
3S	Heart, LV	Snap Frozen
3T	Heart, LV	Fixed
3U	Testis, Left	Snap Frozen
3V	Testis, Left	Fixed
3W	Testis, Right	Snap Frozen

3X	Testis, Right	Fixed
3Y	Ovary, Left	Snap Frozen
3Z	Ovary, Left	Fixed
3AA	Ovary, Right	Snap Frozen
3AB	Ovary, Right	Fixed
3AC*	Dermal Fibroblast	Cultured Cells
3AD	Skin, Calf	Snap Frozen
3AE	Skin, Calf	Fixed
3AF	Skin, Abdomen	Snap Frozen
3AG	Skin, Abdomen	Fixed
3AH	Muscle	Snap Frozen
3AI	Muscle	Fixed
3AJ	Brain	Fresh
3AK	Frontal Lobe, Brain, Left hemisphere	Snap Frozen
3AL	Temporal Lobe, Brain, Left hemisphere	Snap Frozen
3AM	Cerebellum, Brain, Left hemisphere	Snap Frozen
3AN	Hippocampus, Brain, Left hemisphere	Snap Frozen
3AO	Hippocampus, Brain, Left hemisphere	Snap Frozen

*Fibroblasts isolated from fresh calf skin.

Cell line

For benchmarking samples that are not distributed by TPC (e.g. UW, Yale), Kit/Sample ID is designated by DAC as shown in Table 2. As the TPC-based protocol and tissue aliquots do not apply to cell line cultures, the schema for the benchmarking cell lines is the following: ("**A**" indicates an alphabet letter, and "**X**" indicates a null value.

For benchmarking cell line (Table 2): **SMHTAAAA – X – X**

Table 2. Benchmarking cell line codes.

Kit/Sample ID	**Cell line description**
COLO829T	COLO829 tumor cell line
COLO829BL	COLO829BL normal lymphoblast cell line

COLO829BLT50	COLO829 1:50 admixture
HAPMAP6	Cell admixture of six HapMap cell lines
LBLA2	LB-LA2 fibroblast cell line
LBIPSC1	iPSC line from clone #1 derived from the LB-LA2 fibroblast cell line
LBIPSC2	iPSC line from clone #2 derived from the LB-LA2 fibroblast cell line
LBIPSC4	iPSC line from clone #4 derived from the LB-LA2 fibroblast cell line
LBIPSC52	iPSC line from clone #52 derived from the LB-LA2 fibroblast cell line
LBIPSC60	iPSC line from clone #60 derived from the LB-LA2 fibroblast cell line

Additional sample/assay metadata in the name

Additional sample- and assay-related metadata are delimited by a hyphen ("-") as shown below.

[Project][Kit/Donor ID] - [Protocol ID] - [Tissue aliquot & Tissue core IDs] - [Sex & Age] - [Sequencing platform & Sequencing assay code] - [Center code] - [Unique Accession ID] - [Analysis info] . [File format]

- **Sex & Age**: Sex and age of the donor - "M" (for male) or "F" (for female), followed by the donor's age in years. "N" and "N" will be used when sex or age is unknown or not applicable (e.g. cell line mixture).
- **Sequencing platform & assay code**: A code consisting of a single alphabet corresponding to a sequencing platform, followed by a 3-digit number designated for a specific experimental assay type (Tables 3A and 3B).
- **Center code**: Letters corresponding to the centers, i.e. GCCs and TTDs that generate original/raw sequencing data and submit to DAC (Table 4).

Table 3A. Sequencing platform codes.

SMaHT code	Sequencing platform
A	Illumina NovaSeq X
B	PacBio Revio HiFi
C	Illumina NovaSeq 6000
D	ONT PromethION 24
E	ONT PromethION 2 Solo
F	ONT MinION Mk1B

G	Illumina HiSeq X
H	Illumina NovaSeq X Plus
I	BGI DNBSEQ-G400
J	Element AVITI
K	Illumina NextSeq 2000

Table 3B. Experimental assay codes.

Code	**Assay Name**	**Description**
000		(Null or not-applicable)
[001-100: DNA-based assays]		
001	WGS	DNA, PCR-free, Bulk, Whole genome sequencing (WGS)
002	PCR WGS	DNA, PCR, Bulk, WGS
003	Ultra-Long WGS	DNA, PCR-free, Bulk, Ultra-Long WGS
004	Fiber-seq	DNA, PCR-free, Bulk, Fiber-seq
005	Hi-C	DNA, Bulk, Hi-C
006	Bulk NTSeq	DNA, Bulk, NTSeq
007	CODEC	DNA, Bulk, Duplex-seq, CODEC
008	Bot-seq	DNA, Bulk, Duplex-seq, Bot-seq
009	NanoSeq	DNA, Bulk, Duplex-seq, NanoSeq
010	scNanoSeq	DNA, Single-cell, Duplex-seq, scNanoSeq
011	DLP+	DNA, Single-cell, DLP+
012	Microbulk MALBAC WGS	DNA, Microbulk, MALBAC-amplified WGS
013	Single-cell MALBAC WGS	DNA, Single-cell, MALBAC-amplified WGS
014	Microbulk PTA WGS	DNA, Microbulk, PTA-amplified WGS
015	Single-cell PTA WGS	DNA, Single-cell, PTA-amplified WGS

016	scDip-C	DNA, Single-cell, scDip-C
017	CompDuplex-seq	DNA, Bulk, Duplex-seq, CompDuplex-seq
018	scCompDuplex-seq	DNA, Single-cell, Duplex-seq, scCompDuplex-seq
019	Strand-seq	DNA, Bulk, Strand-seq
020	scStrand-seq	DNA, Single-cell, scStrand-seq
021	HiDEF-seq	DNA, Bulk, Duplex-seq, HiDEF-seq
022	HAT-seq	DNA, Bulk, HAT-seq
023	Microbulk HAT-seq	DNA, Microbulk, PTA-amplified HAT-seq
024	scHAT-seq	DNA, Single-cell, PTA-amplified, HAT-seq
025	VISTA-seq	DNA, Bulk, Duplex-seq, VISTA-seq
026	Microbulk VISTA-seq	DNA, Microbulk, Duplex-seq, VISTA-seq
027	scVISTA-seq	DNA, Single-cell, Duplex-seq, VISTA-seq
028	TEnCATS	DNA, Bulk, TEnCATS
029	L1-ONT	DNA, Bulk, L1-ONT
[101-200: RNA-based assays]		
101	RNA-seq	RNA, Bulk, RNA-seq
102	Kinnex	RNA, Bulk, Kinnex
103	snRNA-seq	RNA, Single-cell, snRNA-seq
104	STORM-seq	RNA, Single-cell, STORM-seq
105	Tranquil-seq	RNA, Single-cell, Tranquil-seq
[201-300: Chromatin-based assays]		
201	ATAC-seq	Chromatin, Bulk, ATAC-seq
202	CUT&Tag	Chromatin, Bulk, CUT&Tag
203	varCUT&Tag	Chromatin, Bulk, varCUT&Tag
204	sc-varCUT&Tag	Chromatin, Single-cell, sc-varCUT&Tag

Table 4. SMaHT data generation center codes.

Code	Category	Institute	Contact PI
bcm	GCC	Baylor College of Medicine	Richard Gibbs
broad	GCC	Broad Institute	Kristin Ardlie
nygc	GCC	New York Genome Center	Soren Germer
uwsc	GCC	University of Washington & Seattle Children's Hospital	Jimmy Bennett
washu	GCC	Washington University in St. Louis	Ting Wang

Code	Category	Institute	Contact PI
bcm1	TTD	Baylor College of Medicine	Chuck Zong
bcm2	TTD	Baylor College of Medicine	Fritz Sedlazeck
bch1	TTD	Boston Children's Hospital	Christopher Walsh
bch2	TTD	Boston Children's Hospital	Sangita Choudhury
broad1	TTD	Broad Institute	Fei Chen
cwru	TTD	Case Western Reserve University	Fulai Jin
dfci	TTD	Dana-Farber Cancer Institute	Kathleen Burns
mayo	TTD	Mayo Clinic	Alexej Arbyzov
nyu	TTD	New York University	Gilad Evrony
stfd	TTD	Stanford University	Alexander Urban
umass	TTD	University of Massachusetts	Thomas Fazzio
umich	TTD	University of Michigan	Ryan Mills
uutah	TTD	University of Utah	Gabor Marth
wcnygc	TTD	Weill Cornell Medicine & New York Genome Center	Dan Landau

Code	Category	Institute	Contact PI
dac	DAC	Harvard Medical School	Peter Park

Code	Category	Institute	Contact PI
tpc	TPC	National Disease Research Interchange (NDRI)	Thomas Bell

Additional file information in the name

The file and analysis-related metadata and identifiers are generated at DAC.

[Project][Kit/Donor ID] - [Protocol ID] - [Tissue aliquot & Tissue core IDs] - [Sex & Age] - [Sequencing platform & Sequencing assay code] - [Center code] - [Unique Accession ID] - [Analysis info] . [File format]

- **Unique Accession ID**: Automatically generated unique accession ID with the "SM" prefix (for SMaHT) for a file stored at DAC. This ID can also be used as an identifier to search the file via the SMaHT data portal.
- **Analysis info and File format**: Information about the analysis used to generate the file can be added here using the following format, delimited by an underscore ("_").
    - For unaligned raw sequence files, a null value, "**X**", is set for the analysis info field.
        - **[Tool name and version; separated by "_"] _ [Genome version] _ [Variant type / transcript isoform expression] _ [Other info]**

      e.g. Unaligned FASTQ
      …–SM192K2-X.fastq.gz
      e.g. Unaligned BAM
      …–SM192K2-X.bam
      e.g. BAM aligned against GRCh38 using Sentieon BWA-MEM
      …–SMRO2AR-sentieon_bwamem_202308.01_GRCh38.aligned.sorted.bam
      E.g. VCF of somatic calls using Strelka2 called against the GRCh38 reference genome
      …–SM092E4-strelka2_4.14_GRCh38_snv.vcf
        - **Tool name and version (mandatory):** Software/Computational tool name and version used to generate the file, e.g. bwamem_0.7.17.
        - **Genome version (optional):** Reference genome version used for aligned sequence files, or variant type for VCF files (Table 4A).
        - **Variant type or gene expression (optional;** Table 4B**)**.
        - **Additional key info (optional)** can be added, delimited by an underscore ("_"). Other information should include the GENCODE version for RNA-seq data.

Table 5. Genome version (A) and variant type (B) tables.

(A)

Reference Genome	**Code**
GRCh38 without ALT contigs	GRCh38
T2T CHM13	CHM13
Donor-specific genome assembly	DSA

(B)

Variant Type or other Data Type	**Code**
Substitutions, short insertions and deletions	snv
Copy number variants	cnv
Structural variants including large insertions and deletions, duplications, inversions, translocations	sv
Mobile element insertions	mei
Donor-specific genome assembly	dsa
Gene expression level	gene
Transcript isoform expression level	isoform

Benchmarking Cell Line Names

1. COLO829
   `SMHTCOLO829T - X - X - M45`
2. COLO829BL
   `SMHTCOLO829BL - X - X - M45`
3. COLO829 1:50 mixture
   `SMHTCOLO829BLT50 - X - X - M45`
4. HapMap mixture
   `SMHTHAPMAP6 - X - X - NN`
5. iPSC / fibroblast, e.g. LB-IPSC 1
   `SMHTLBIPSC1 - X - X - M29`

File Name Examples

[Note that space was added to make the delimiters (i.e. "-" and "_") more visible]

[Project][Kit/Donor ID] - [Protocol ID] - [Tissue aliquot & Tissue core IDs] - [Sex & Age] - [Sequencing platform & Sequencing assay code] - [Center code] - [Unique Accession ID] - [Analysis info] . [File format]

Bulk WGS on Illumina NovaSeqX - FASTQ file of the COLO829T-COLO829BL 1:50 admixture from BCM-GCC:

SMHTCOLO829BLT50 - X - X - M45 - A001 - bcm - SM123JK0 - X . fastq.gz

FiberSeq on PacBio HiFi - unaligned BAM file of the COLO829 tumor cell line from UW-GCC:

SMHTCOLO829T - X - X - M45 - B003 - uwsc - SM89QQ3LA - X . bam

FiberSeq on PacBio HiFi - BAM file of the COLO829 tumor cell line from UW-GCC, aligned to GRCh38, sorted by read name, and phased:

SMHTCOLO829T - X - X - M45 - B003 - uwsc - SM57S7SLA - pbmm2.6_GRCh38 . aligned.sorted.phased.bam

Bulk WGS on Ultima Genomics - SNVs called using Strelka2 from COLO829 1:50 admixture against COLO829BL:

SMHTCOLO829BLT50 - X - X - M45 - F001 - nygc - SM29A1QR7 - strelka2.9_GRCh38_snv_cmpToCOLO829BL . vcf

Bulk WGS on Illumina NovaSeqX - FASTQ file of the HapMap mixture from Broad:

SMHTHAPMAP6 - X - X - NN - A001 - broad - SM89XAY38 - X. fastq.gz

- **BAM file from the FASTQ file above:**
  SMHTHAPMAP6 - X - X - NN - A001 - broad - SM33A9K6R - sentieon_bwamem_202308.01_GRCh38 . aligned.sorted.bam
- **VCF that contains HapMap mixture-specific variants identified by running Mutect 2 on the BAM file above:**
  SMHTHAPMAP6 - X - X - NN - A001 - broad - SM3R123U9 - mutect4.1.4.1_GRCh38_snv . vcf

- In the current plan, for VCF files that are generated by GCCs and TTDs, and submitted to DAC, DAC will not rename the file names according to the SMaHT nomenclature. DAC will keep the original file names provided by the submitters, and the associated metadata will indicate who the submitting group is.

DSA - COLO829 1:50 admixture against COLO829BL:
SMHTCOLO829BLT50 - X - X - M45 - F001 - nygc - SM29A1QR7 - hifiasm2.4 . fasta

Accompanying chain file for DSA COLO829 1:50 admixture against COLO829BL:
SMHTCOLO829BLT50 - X - X - M45 - F001 - nygc - SM29A1QR7 - softwareVersion_GRCh37ToDSA (SourceToTargetGenome) . chain.gz

Bulk WGS on Illumina - BAM file of the benchmarking tissue from UW-GCC:
ST002 - 1G - XX - M74 - A001 - uwsc - SMAFIQUWOL9M - sentieon_bwamem_202308.01_GRCh38.aligned.sorted . bam