Information/metadata needed to store data at ODCF

To register your data at ODCF several metadata has to be provided. If your data was sequenced at the GPCF we will receive most of the required data automatically. If you want to import older data from the GPCF, that has already been removed from "mid-term" storage, or data sequenced elsewhere please fill in the metadata sheet you can find in the download area of OTP (https://otp.dkfz.de/otp/info/templates).

Important: If you want to import external or old data please inform us before you transfer the files to the file system. Data can only be imported from specific locations. We will set up a folder to where you can copy your data. The import will be performed from that location and this location needs to be named in the metadata file.

The following metadata is mandatory:

Sample:

PROJECT: name of the ODCF project
SAMPLE_ID: name of the sample (needs to be unique)
SEQUENCING_TYPE: RNA, WGS, EXOME, ChIP-Seq, WGBS, ATAC, …
BASE_MATERIAL: indicate if the data is singe-cell data
PATIENT_ID: (Patient Pseudonym) name of individual or cell line
BIOMATERIAL_ID: Sample Type (CONTROL, TUMOR, TREATMENT)
SPECIES: mouse, human, … (if available).
GENDER: sex of patient, mouse, ... (if available)

File specific:

FASTQ_FILE: absolute path of FASTQ file on the file system
MD5: md5sum of the FASTQ file
MATE: Read 1 or 2 or index read
LIBRARY_LAYOUT: single-end or paired-end

Technical metadata:

RUN_ID: if not available please create one from date and sequencing center
RUN_DATE: date of the run (please make sure to separate the numbers by "-")
LANE_NO: number of the lane
BARCODE: barcode of the sample
INSTRUMENT_MODEL: instrument model (name and series) of the sequencer (e.g. HiSeq4000)
INSTRUMENT_PLATFORM: used platform of the sequencer (usually Illumina)
SEQUENCING_KIT: kit used for sequencing / Illumina chemistry
PIPELINE_VERSION: version of bcl2fastq

Sequencing location:

CENTER_NAME: name of sequencing center
ILSE_NO: required when data is sequenced at the DKFZ-GPCF

Sequencing type specific:

LIB_PREP_KIT: Lab protocol to prepare the sample. This mainly determines for us, which adapter sequence should be trimmed or which target file should be used, if applicable (for 'EXON' or 'WHOLE_GENOME_BISULFITE' or 'WHOLE_GENOME_BISULFITE_TAGMENTATION' or 'RNA').
ANTIBODY_TARGET: only for ChIP-Seq or CUT-nRUN data --> leave empty otherwise.
TAGMENTATION_BASED_LIBRARY: "true" if data is tagmentation data. This information is currently only used for bisulfite sequencing and will cause a different implementation of duplicate marking during alignment.
CUSTOMER_LIBRARY: for 'TAGMENTATION' data only

Optional metadata:

ANTIBODY, PHENOTYPE, BASE_COUNT, READ_COUNT, CYCLE_COUNT

Patient ID

The Patient ID is a project-wide unique ID for the individual (patient, animal, …) from which the sample is derived. It allows us to handle the data without knowledge or disclosure of patient identifying information und should obviously never contain any resemblance to the patients' real name. However, we do not only handle patient data. In more abstract terms the patient pseudonym field identifies a biological source for which a direct comparison of genetic information is reasonable. Examples of non-patient data are cell-lines, mouse individuals or inbred mouse strains (pooled mice).

If you have cell-line data the PID in most cases represents the cell-line (e.g. HeLa, Jurkat, HepG2). **Please do not encode sequencing type information in patient identifiers.** The PID should only contain alphanumeric characters ([a-zA-Z0-9]) or '-' characters.

Important: If we offer variant calling for your data, variants can only be called against samples of the same PID.

Biomaterial ID / Sample Type

Biomaterial ID is also called ‘Sample Type’. It represents biological material from which the data has been derived. In the simplest case, the sample type consists of a descriptive name and an ordinal number, for instance “tumor01”, “liver01”, “control02”, or “blood12”. This case is usually applicable for human or mouse data. We can run comparative analyses between multiple sample types only for human data, the same PID and only if one of the samples is categorized as “disease” while the other is categorized as “control”. These disease/control categories are associated with each sample type.

Often, additional information should be encoded in the sample type, in particular if you have additional experimental factors. For instance, you may have for the same cell line multiple phenotypes, timepoints, conditions, or treatments.

The sample type will appear on the file system and is therefore restricted to contain only alphanumeric characters ([a-zA-Z0-9]) or '-' characters. In particular, you should not use underscores '_'!

Library Preparation Kit

The library preparation kit is mandatory for several sequencing types, because processing relevant information is derived from them:

- EXOME: coverage / target region BED file is
- RNA & WHOLE_GENOME_BISULFITE: adapter sequence to be trimmed

Further information:

1. Please feel free to have a look at our FAQs: https://wiki.odcf.dkfz.de/pub/faq
2. All project related forms can be downloaded here: https://otp.dkfz.de/otp/info/templates
3. Further information regarding OTP was published in 2017: Reisinger, E. et al., Journal of Biotechnology (2017): http://dx.doi.org/10.1016/j.jbiotec.2017.08.006)

OFFICE ADDRESS

German Cancer Research Center (DKFZ)
Omics IT and Data Management (ODCF)
Mathematikon - Berliner Straße 41
D-69120 Heidelberg, Germany

All information in this flyer is subject to change. Please contact odcf-service@dkfz.de for up-to-date information about our services.

Version: 2020-01

Data Management at ODCF

Metadata

Omics IT and Data Management Core Facility