Clinical trial curation

1.1. Immunotherapy datasets

1.1.1. Introduction

This documentation goes over the clinical trial data curation process in detail, using immunotherapy data.

1.1.2. Objective

The objective is to curate a clinical dataset into R's MultiAssayExperiment object. An example of a clinical data MultiAssayExperiment (MAE) object can be found in ORCESTRA.

Currently, a clinical data object contains the following data parts:

1. Clinical metadata: Contains patient/sample metadata.
2. Molecular profiles: Molecular assay data (Currently RNA-seq, SNV or CNA) which is formatted in either RangedSummarizedExperiment or regular SummarizedExperiment object.

1.1.3. Data Access

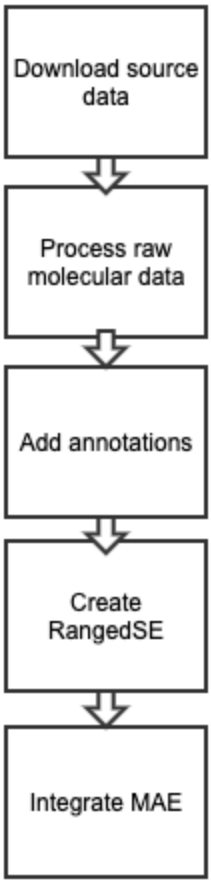
Public data

If the source is Pubmed, the raw omics files and clinical response metadata are available from Supplementary or external repository links in Data Availability section of the paper.

Private data

Private data such as PHI, clinical response might be available only upon request. Please contact the author(s) or whoever is responsible for requesting such data.

1.1.4. Data Processing Overview



An example of clinical data processing pipeline can be found here as a Snakemake pipeline.

Generally, an overall process of the curation follows the steps outlined below:

1. **Download source data**: Download data from publications or data repositories. The source data can be in various formats such as an Excel file, CSV or TXT.
2. **Process raw molecular data ,if available:** The RNA-seq processing from raw FASTQ is outlined in this article.
3. **Add annotations**: Ensure that genes, tissues and treatments are annotated with metadata available from external source and lab standardized columns.
4. **Create RangedSummarizedExperiment or SummarizedExperiment (SE) object:** For the molecular data, we prefer RangedSummarizedExperiment as it is compatible with GenomicRanges R package.
5. **Create MAE object:** Format downloaded data to the layout and structure that is favourable to creating a MAE object. Through this process, the source data is extracted from the source data format and formatted into a CSV or TSV file. Integrate molecular data to MAE.

1.1.5. Processing Clinical Metadata

The clinical data should be formatted into patient/sample ids as rows and attributes as column data. This will be added as **colData** of the SE or MAE object.

The following columns are mandatory and should be filled with NA if the data is not available to maintain consistency across ICB and non-ICB datasets:

Column name	Description
Patientid	This column contains unique patient identifiers
treatmentid	This column contains the treatment regimen of each patient. Individual drug names are separated by ":" and standardized based on the lab's nomenclature. For example, the drug combo "FAC" is represented as "5-fluorouracil:Doxorubicin:Cyclophosphamide"
response	This column contains the response status of the patients to the given treatment - Responders (R) and Non-responders (NR)
tissueid	Cancer type standardized based on the lab's nomenclature from Oncotree. Example: “Breast”
survival_time_pfs/ survival_time_os	The time starting from taking the treatment to the occurrence of the event of interest. The event name like "pfs", "os" must be appended to *survival_time* to differentiate the survival measure. Example for data in this column: “2.6”
survival_unit	The unit in which the survival time is measured. If the event is measured in other units such as “day”, or “year”, it must be converted to "month" for consistency

event_occurred_pfs /event_occurred_os	Binary measurement showing whether the event of interest occurred (1) or not (0). The event name like "pfs", "os" must be appended to *event_occurred* to differentiate the survival measure

Note: Common columns have to be the first set of columns appearing in the metadata followed by the rest of the columns. You could add other columns with the name in the source data, but the standard columns with the above mentioned names should be present.

If you are adding new columns based on restructured data from existing columns, please assign the lucid, self-explanatory column names.

The table below shows the other common columns across the 19 ICB datasets curated.

Column name	Description	type
sex	Sex of the patient - Male or Female	source
age	Age	source
cancer_type	Type of cancer tissue	source
histo	Histological info such as subtype	source
stage	Cancer stage	source
response.other.info	Same data as Responders (R) and Non-responders (NR)	source
recist	Annotated using RECIST. The most commonly used responses are CR,PR,SD, PD.	source
treatment	Drug target or drug name	source
dna	DNA sequencing type. eg: whole exome sequencing	source
rna	Type of rna processed data. eg: TPM	source
survival_type	PFS or OS or both (denoted by '/'). If both, added by in-lab curation	in-lab curation
TMB_raw	Tumor Mutation Burden raw values	in-lab curation
nsTMB_raw	-	in-lab curation
indel_TMB_raw	-	in-lab curation
indel_nsTMB_raw	-	in-lab curation
TMB_perMb	TMB per megabase (Mb) was performed as defined: *TMB = mutns/target.* With *mutns* = number of non-synonymous mutations; and *target* = target size of the sequencing See Supplementary Table S2 of https://pubmed.ncbi.nlm.nih.gov/36055464/	in-lab curation
nsTMB_perMb	-	in-lab curation
indel_TMB_perMb	-	in-lab curation
indel_nsTMB_perMb	-	in-lab curation
CIN	Calculated from CNA values	in-lab curation
CNA_tot	Sum of total CNA/coverage; calculated from CNA values	in-lab curation
AMP	Sum of total AMP/coverage; calculated from CNA values	in-lab curation
DEL	Sum of total DEL/coverage; calculated from CNA values	in-lab curation

1.1.6. Processing Molecular Data

The raw omics data files are obtained and processed in the lab. If the raw files are not available, processed data is used. Exceptions are Mutation data where only processed data is used to avoid ambiguity around matched normals.

In general, all molecular data should be formatted into genes (eg: transcript IDs for RNA profiling) as rows and patient/sample IDs as columns.

1.1.6.1. RNA-seq data

First and foremost, **the RNA-seq data should be at gene-level and in TPM**. The TPM value should be log transformed with log2(TPM) + 0.001.

If the TPM values are not available, but counts values are available, you could use the following formula to convert counts value to TPM:

```
GetTPM <- function(counts, gene_size) {
    x <- counts/gene_size
    return(t(t(x)*1e6/colSums(x)))
}
```

If available, counts and transcript-level data (isoforms) should also be included.

1.1.6.2. SummarizedExperiment Object

Each molecular data needs to be formatted into a SummarizedExperiment (or RangedSummarizedExperiment) object.

At minimum, SummarizedExperiment requires:

1. **colData** (the patient metadata) formatted in patient/sample IDs as rows and attribute data as columns.
2. **assay** (expression values) formatted in gene/transcript IDs as rows and patient/sample IDs as columns.
3. **rowData** (gene metadata) is gene metadata for the genes that exist in the assay, formatted as gene/transcript IDs as rows and attributes as columns. More details on the gene metadata below.

1.1.7. Annotation

Lab standardized annotation data are stored in BHKLab-Pachyderm's Annotation repository.

1.1.7.1. Gene Annotations

Gene metadata is obtained from Gencode annotations. We have a few versions of Gencode annotation data available in .RData files. An .RData file includes data frames that contains gene and transcript information such as features_gene, features_transcript and tx2gene. Some of the available gene annotations include:

Gencode v19

Gencode v40

Note: Please use the most recent version for your gene annotations from this repository. The version of Gencode must be decided after checking the reference genome. Follow Gene curation SOP for detailed steps

1.1.7.2. Drug Annotations

For clinical data, drug annotations are performed in case-by-case basis. For immunotherapy treatments, both instances such as anti-"target" (eg: anti-CTLA4) and monoclonal antibody brand names can be present. Please follow the Drug curation SOP to correctly annotate such cases using the standard lab files in the Annotation repository.

1.1.7.3. Tissue Annotations

For tissue annotations that cannot be mapped using Tissue curation SOP to the standard lab files in the Annotation repository., manual review needs to be performed in case-by-case basis.

1.2. Non-Immunotherapy datasets

1.2.1. Objective

While most steps overlap between immunotherapy and non-immunotherapy dataset curation, it is important to understand the differences. The following details focuses on the current data elements and finally the differences.

Currently, a non-ICB clinical dataset is curated into R's SummarizedExperiment (SE) object and not MAE because of the absence of multiple omics data. An example of a clinical data SE object can be found in ORCESTRA. Sample code can be found on Github - https://github.com/bhklab/Clinical-Trial-SE/blob/master/ClinicalTrial_SE_curation.Rmd

1.2.2. Curation

A non-ICB clinical data object contains the following data parts:

1. Expression values or Assay data

2. Patient metadata or ColData

1.2.3. Expression values or Assay data

Assay data contains genomic profiles of the patients. The data is usually processed in-house from raw files or in some instances, published processed data is used directly. For instance, gene expression profiles of the patients are typically generated by either microarray or RNA-seq platforms. In the BHK lab, we use Robust Multiarray Averaging (RMA) and CDF files from Brainarray for processing microarray data, and the Kallisto method for processing RNA-seq data, as mentioned in immunotherapy curation.

1.2.4. Patient metadata or ColData

Any data pertaining to the samples or clinical response can be included in the Phenodata object. This is either fetched from public platforms like GEO if the data is public or upon request in case of confidentiality. Metadata sections in the SE objects include a few mandatory columns which are populated either by information from the other columns or the original published paper. NA is used to fill out columns for which no information is found. Each SE object includes additional metadata that may or may not be available in other SE objects.

Mandatory columns are the same as immunotherapy **colData**

1.2.5. Gene metadata or rowData

Similar to immunotherapy datasets, gene metadata for non-immunotherapy datasets is also obtained from Gencode annotations. "Ensembl.v99.annotation.RData" from "Gencode.v33.annotation.RData" is used for curating rowData in non-immunotherapy datasets. Annotation data are available in BHKLab-Pachyderm's Annotation repository.