A semi-automated pipeline for metadata cleanup

Fawaz Dabbaghie[1], Felipe Albrecht[1,2], Markus List[1]

E-mail: markus.list@mpi-inf.mpg.de

1. Max Planck Institute for Informatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany
2. Graduate School of Computer Science, Saarland Informatics Campus, 66123 Saarbrücken, Germany

Deutsches Epigenom Programm

Abstract

Metadata, which include details on samples, experimental protocols or data processing, are a cornerstone of reproducible research. However, keeping metadata in a structured and consistent form is challenging, in particular in large consortia that often span various labs with different conventions. This poses a serious problem, e.g. due to inconsistent use of terms (e.g. ChIPseq vs ChIP-seq) as well as formatting and encoding problems. Typically, these issues are solved in a laborious, manual fashion. As a result, analysis will often focus on a small set of terms, often neglecting potential confounding factors such as software versions, laboratory protocols or devices used.

In order to tackle this issue, we implemented a semi-automated pipeline with the following steps: (i) common formatting issues in existing metadata files are addressed; (ii) the user may interact with a unified version of the metadata through OpenRefine, a tool that facilitates interactive and user-friendly mass editing operations. This step yields a set of editing rules that are extracted for future operations; (iii) based on the cleaned up metadata, a controlled vocabulary is generated and complemented by user-defined regular expressions.

Completing this procedure once allows us to address previously encountered issues automatically when metadata for additional samples or experiments become available. Our pipeline, which consists of a series of freely available and well-documented python scripts (https://github.molgen.mpg.de/DEEP/tidy_meta_data/), allows data analysts to save substantial time. Moreover, the cleaned up metadata allows for potential confounding factors to be considered without additional effort. We successfully applied this strategy within the German Epigenome consortium DEEP and envision that our effort are useful to other IHEC members and beyond.

(i) Preparing and fixing metadata files

Original meta data files

file1.tsv

COMMENT	some comment
DONOT_ID	44_Mm01
EXPERIMENT_TYPE	ChIPseq

file2.tsv

STARTING_CELL_NR	50X10^6	
REPLICATE_ID	1	
COMMENT	some	comment

PROTOCOL_LINK	http://www.illumina.com/...
EXPERIMENT_TYPE	ChIP-seq
STARTING_CELL_NR	5E7

file3.tsv

Common issues with
- encoding
- column separation
- ...

file1.tsv

COMMENT	some comment
DONOR_ID	44_Mm01
EXPERIMENT_TYPE	ChIPseq

file2.tsv

STARTING_CELL_NR	50X10^6
REPLICATE_ID	1
COMMENT	some comment

PROTOCOL_LINK	http://www.illumina.com/...
EXPERIMENT_TYPE	ChIP-seq
STARTING_CELL_NR	5E7

file3.tsv

Files are merged into a joint table

(ii) Batch editing with OpenRefine

FILE_NAME	COMMENT	DONOR_ID	EXPERIMENT_TYP	STARTING_CELLS_NR	REPLICATE_ID	PROTOCOL_LINK
file1	some comment	44_Mm01	ChIPseq	[[[NA]]]	[[[NA]]]	[[[NA]]]
file2	some comment	[[[NA]]]	[[[NA]]]	50X10^6	1	[[[NA]]]
file3	[[[NA]]]	[[[NA]]]	ChIP-seq	5E7	[[[NA]]]	http://www.illumina.com/...

Source code and documentation available online

JSON operations

```
{
  "op": "core/mass-edit",
  "description": "Mass edit cells in column EXPERIMENT-TYPE",
  "engineConfig": {
    "mode": "row-based",
    "facets": []
  },
  "columnName": "EXPERIMENT-TYPE",
  "expression": "value",
  "edits": [
    {
      "fromBlank": false,
      "fromError": false,
      "from": [
        "ChIPseq"
      ],
      "to": "ChIP-seq"
    }
  ]
},
```

Batch Editing

User defines editing rules interactively

New rules are stored

FILE_NAME	COMMENT	DONOR_ID	EXPERIMENT_TYP	STARTING_CELLS_NR	REPLICATE_ID	PROTOCOL_LINK
file1	some comment	44_Mm01	ChIP-seq	[[[NA]]]	[[[NA]]]	[[[NA]]]
file2	some comment	[[[NA]]]	[[[NA]]]	50X10^6	1	[[[NA]]]
file3	[[[NA]]]	[[[NA]]]	ChIP-seq	50x10^6	[[[NA]]]	http://www.illumina.com/...

Existing rules are automatically applied to the table of merged metadata

file1.tsv

COMMENT	some comment
DONOR_ID	44_Mm01
EXPERIMENT_TYPE	ChIP-seq

file2.tsv

STARTING_CELL_NR	50X10^6
REPLICATE_ID	1
COMMENT	some comment

PROTOCOL_LINK	http://www.illumina.com/...
EXPERIMENT_TYPE	ChIP-seq
STARTING_CELL_NR	50x10^6

file3.tsv

cleaned up files are produced

(iii) Controlled vocabulary and reporting

Processed and user corrected meta data is used to build a controlled vocabulary. When additional files are processed unknown values will be highlighted to reveal potential problems.

Controlled Vocabulary

Report

Processed metadata files

User intervention

Key	Regular Expression	Accepted example
BIOLOGICAL_REPLICATE_ID	^[0-9]$	One number only
DONOR_ID	^\d{2}_\w+$	Accroding to DEEP naming Scheme

regex_dictionary.tsv

Users may specify regular expressions to control particular keys that specify e.g. a volume or a concentration.

COMMENT
PROTOCOL_LINK

black_keys.txt

Users add additional information about blacklisted keys that can be ignored.

DeepBlue EPIGENOMIC DATA SERVER

UNIVERSITÄT DES SAARLANDES

max planck institut informatik

BLUEPRINT epigenome

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung