\*lDevelopment of a new annotator based on Ontologies like BRENDA and DISEASE ONTOLOGY.

After reading the paper (Funk et al., 2014) we decided to annotate GEO metadata with Concept Mapper.

The source code for an annotator based on ConceptMapper is available at <http://sourceforge.net/projects/bionlp-uima/files/nlp-pipelines/v0.5/>

The pipeline can be either personalized or used in a project adding the dependency in the pom file of the project.

I downloaded the nlp-pipelines-0.5 and imported the project into eclipse as a maven project. After the import into eclipse there is the need to set up maven setting a new maven configuration with

mvn compile test package

And setting the environment variable JAVA\_HOME (into eclipse) to the value returned by

export JAVA\_HOME=$(/usr/libexec/java\_home)

The pipeline to produce concept annotations is located in *edu.ucdenver.ccp.nlp.pipelines.conceptmapper.EntityFinder.java*.

Brenda Tissue Ontology has been downloaded and stored in

*/Users/egaleota/Documents/workspace/nlp-pipelines-0.5/nlp-pipelines-conceptmapper/src/test/resources/oboFiles*

To produce annotations from a directory of text files use the command in the nlp-pipelines directory:

*mvn -f nlp-pipelines-conceptmapper/pom.xml exec:java -Dexec.mainClass="edu.ucdenver.ccp.nlp.pipelines.conceptmapper.EntityFinder" -Dexec.args="<InputDir> <OutputDir> <ontology> <OBOfile> <DictDir>"*

The parameters required are as follows:

* <InputDir> - Directory of text files for input. The text files should be named by the identifier of the article and have a file extension of ".txt". The text files can be in ASCII or UTF-8 encoding but must be plain text. It is possible to use a different collection reader with xml input but it requires editing the source.
* <OutputDir> - Directory where annotations will be written. Output is one ".a1" file per input text file in the BioNLP/Brat format (further detail explained below).

More information about the output format can be seen at <http://2011.bionlp-st.org/home/file-formats>

Simply, the format is "entity# \t typeEntity \t spanBegin \t spanEnd \t coveredText".

The lines beginning with N capture the normalization

Examples from the Gene Ontology can be seen below:

T2 go\_term 232 244 pigmentation

N2 Reference T2 GO:0043473

T7 go\_term 1841 1862 embryonic development

N7 Reference T7 GO:0009790

* <ontology> - Specifies which ontology concepts are from. If one of the ontologies evaluated in Funk et al. 2014 the best parameter combinations are used. The ontologies with specific parameter combinations are: CHEBI - Chemical Entities of Biological Interest, SO - Sequence Ontology, GO - Gene Ontology, GO\_MF - GO Molecular Function, GO\_CC - GO Cellular Component, GO\_BP - GO Biological Process, PR - Protein Ontology, CL - Cell Ontology, NCBI\_TAXON - NCBI Taxonomy. If the specific ontology of interestes is not listed please use OBO. <OBOfile> - The obo ontology file. A large list of biomedical ontologies can be found at http://www.obofoundry.org/. <DictDir> - The directory where the Concept Mapper dictioary will be written.

RETRIEVAL OF ChIP-seq data

ChIP-seq data is available in SRAdb. For the case of this data we have chosen SRAdb and not GEOmetadb because it is possible to query experiments by library\_strategy==’ChIP-Seq’. Data is organized in different tables. In particular the experiment table contains the deepest level data (raw). The same data in GEO is available only in processed form, while we cannot retrieve raw sra or fasq files. Each experiment in SRA can belong to one or more samples (with SRS id) and can be retrieved in GEO as a single sample or a set of experiments in SRA can be associated to a single GEO sample.

For each SRA experiment, the corresponding sample information has been obtained and has been processed by ConceptMapper to identify BTO and DO terms. After the processing we will map SRA ids to GEO ids.

Most of the times the annotation in experiment\_attribute field is the link to the GEO sample the experiment refers to. In the june 2015 version of SRAdb we were able to find annotations for 21038 SRA experiment accessions of which 21037 had a correspondence in GEO. These experiments where mapped to 20913 unique GEO samples.

The tool to read metadata from SRAdb has been written in R (EpiGEO/CompletePipeline folder MetadataHandler.R). The program reads SRAdb samples and stores them into a directory in order to be processed by ConceptMapper. Furthermore for each SRA experiment the corresponding GEO sample id is provided.

Steps:

1. Run the script 01\_MetadataHandler.R

The medatada has been stored into /Users/egaleota/epiMining/data/Rdataframes/ChIP\_raw\_metadata.rda

AIM OF THE ANNOTATION PROCESS

We would like to standardize the representation of GEO metadata to facilitate data retrieval in a programmatic way. To this aim we map ontology terms to GEO metadata. In particular, we are interested in tissues and cell lines, diseases and antibody targets. For tissues we can use Cell Line Ontology and Brenda Tissue Ontology. For diseases Disease Ontology is available. To further enhance the retrieval of cell lines we can also use a controlled vocabulary provided by ENCODE as a CV.ra file (Representational artifact format). Among other information, this controlled vocabulary contains ENCODE cell lines and Antibodies that can be integrated into the available ontologies to provide more details about cell lines and tissues.

CREATION OF DICTIONARY FOR CELL LINES AND ANTIBODIES

We have built a ConceptMapper dictionary for the ENCODE controlled vocabulary. Actually the supported semantic types are “Cell Line” and “Antibody”. There are different options to build this controlled vocabulary. In particular we have added the term, its label as preferred names. Other information, depending on the semantic type, like the vendorID or the and all the is controlled vocabulary that allowed us to identify also antibodies.

CONCEPT MAPPER TYPE SYSTEM

Conceptmapper is a UIMA component used to map text from different collections (Pubmed, medline, simple text files) to concepts from an ontology. The basic ConceptMapper has been extended and more Types have been defined by the ccp-nlp library to identify specific concepts from OBO ontologies.

Starting from the source code provided in nlp-pipelines 0.5 we tried to personalize it for our purposes.

After building the dictionary for Genes, Histone post translational modifications, diseases and tissues (cell lines) we use conceptmapper to retrieve Myc samples.

The first step of the pipeline includes the definition of the Types of annotations we want to have. This is done through a method called createConceptMapperTypeSystem().

The type system used for the ConceptMapper pipeline is made of different other type systems.

* The CCP\_TYPE\_SYSTEM includes
  + edu.ucdenver.ccp.nlp.core.uima.TypeSystem: this is defined in /Users/egaleota/Documents/workspace/ccp-nlp-3.3/ccp-nlp-uima/src/main/resources/edu/ucdenver/ccp/nlp/core/uima

And contains basic information for the definition of document properties, collection properties and annotations. For example SLOT mentions, class mentions, false positive properties, comments on the annotations and on the document

* The SENTENCE\_DETECTOR\_TYPE\_SYSTEM: this is for sentences. In this case we use the Sentence Type defined in org.cleartk.token.type.Sentence, which is a scored sentence
* The type system defined as CONCEPT\_MAPPER\_TYPESYSTEM edu.ucdenver.ccp.nlp.wrapper.conceptmapper.TypeSystem is available at Users/egaleota/Documents/workspace/ccp-nlp-3.3/ccp-nlp-wrapper-conceptmapper/src/main/resources/edu/ucdenver/ccp/nlp/wrapper/conceptmapper/TypeSystem.xml' defines an annotation type Ontology Term. OntologyTerm.xml is in the same location. It is composed by a primitive ccp-nlp dictTerm (defined in ccp-nlp core) with an additional ID (for the Ontology ID)
* analysis\_engine.primitive.DictTerm is a UIMA conceptmapper type defined for dictionary terms. Contains the canonical form of the concept, the enclosing span of text, the matched Text and the matched tokens
* To annotate tokens the Conceptmapper org.apache.uima.conceptMapper.support.tokenizer.TokenAnnotation is used. It includes the text describing the toke, a token type and a token class. Ontology terms are also tokenized into the pipeline

ENTITY FINDER

The entity finder takes as input

* the description of the type system
* the ontology
* the input directory from which read the files
* the output directory
* the obo file to create the dictionary
* the obo directoty where to store the dictionary
* the character encoding

CONCEPTMAPPER OPTIONS

Options can be configured in a class called ConceptMapperPermutationFactory. The allowed options are:

* Search strategy
  + ***CONTIGUOUS\_MATCH –*** *Longest match of contiguous tokens within enclosing span (as specified by SpanfeatureStructure)*
  + ***SKIP\_ANY\_MATCH -*** *longest match of not-necessarily contiguous tokens*
  + ***SKIP\_ANY\_MATCH\_ALLOW\_OVERLAP -*** *longest match of not-necessarily contiguous tokens*
* CaseMatch
  + CASE\_IGNORE: ignoreall - fold everything to lowercase for matching
  + CASE\_INSENSITIVE: insensitive - fold only tokens with initial caps to lowercase
  + CASE\_FOLD\_DIGITS: digitfold - fold all (and only) tokens with a digit
  + CASE\_SENTITIVE: sensitive - perform no case folding
* Stemmer
  + ***BIOLEMMATIZER -*** for the morphological analysis of biomedical literature. The tool focuses on the inflectional morphology of English and is based on the general English lemmatization tool MorphAdorner. The BioLemmatizer is further tailored to the biological domain through incorporation of several published lexical resources. It retrieves lemmas based on the use of a word lexicon, and defines a set of rules that transform a word to a lemma if it is not encountered in the lexicon. An innovative aspect of the BioLemmatizer is the use of a hierarchical strategy for searching the lexicon, which enables the discovery of the correct lemma even if the input Part-of-Speech information is inaccurate.
  + ***PORTER -*** The Porter stemming algorithm (or ‘Porter stemmer’) is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems.
  + ***NONE***
* StopWords
  + PUBMED
  + NONE
* OrderIndependentLookup
  + ON
  + OFF

If "True", token (as specified by [TokenAnnotation](https://uima.apache.org/d/uima-addons-current/ConceptMapper/ConceptMapperAnnotatorUserGuide.html#ConceptMapper.param.tokenannotation)) ordering within span (as specified by [SpanFeatureStructure](https://uima.apache.org/d/uima-addons-current/ConceptMapper/ConceptMapperAnnotatorUserGuide.html#ConceptMapper.param.spanfeaturestructure)) is ignored during lookup (i.e., "top box" would equal "box top"). Default is False.

* FindAllMatchesParamValues
  + YES
  + NO

If True, all dictionary matches are found within the span specified by [SpanFeatureStructure](https://uima.apache.org/d/uima-addons-current/ConceptMapper/ConceptMapperAnnotatorUserGuide.html#ConceptMapper.param.spanfeaturestructure), otherwise only the longest matches are found.

* SynonymType
  + ALL
  + EXACT\_ONLY

CONVERTING THE OUTPUT OF CONCEPTMAPPER FROM A1 FORMAT TO TAB DELIMITED FORMAT

The output files from conceptmapper can be converted using a script in perl ConvertA1forEvaluation.pl

IDENTIFICATION OF ENCODE ANTIBODIES

For each one of the SRA experiment’s metadata we aim to identify the target antibody.

From the cv.ra file provided by ENCODE we have extracted the antibodies and run the pipeline with the following command line

Mvn -f pom.xml exec:java

-Dexec.mainClass="iit.comp.genomics.SemanticMapping.EntityFinder" -Dexec.args="/Users/egaleota/epiMining/data/ChIPmetadata/ /Users/egaleota/epiMining/data/ChIPEncode\_Antibody/ ENCODE\_ANTIBODY /Users/egaleota/git/semanticmapping/SemanticMapping/src/main/resources/sourceDictionaries/cv.ra /Users/egaleota/git/semanticmapping/SemanticMapping/src/main/resources/conceptMapperDictionaries/"

The pipeline uses ENCODE\_ANTIBODY to specify that we want encode antibodies and uses the 30th configuration of the pipeline to run conceptmapper. We used the script ConvertA1forEvaluation to convert the results of ConceptMapper into a tabular format. Results are available at /Users/egaleota/epiMining/data/ChIPEncode\_Antibody/stdout

WE SHOULD MAP THE ENCODE ANTIBODIES TO OUR DICTIONARY WHERE POSSIBLE PERCHE’ PER ESEMPIO POL2 E’ INSERITA DUE VOLTE

IDENTIFICATION OF TARGET ANTIBODY USING R

The target antibody has been retrieved using a tailored R pipeline that uses the Bioconductor annotation packages org.Hs.eg.db and org.Mm.eg.db for Homo Sapiens and Mouse genes respectively. A list of histone modifications has been obtained from the Post translational modifications listed in the HIstome database available at <http://www.actrec.gov.in/histome/>

The script R takes the metadata associated to the ChIP-seq samples and first looks for the following keywords identifying the target molecule field:

*"hgn:", "chip epitope:", "antibody target:", "chip\_antibody:", "chip antibody:", "chip-antibody:", "antibodies:", "immunoprecipitate:", "chromatin ip against:", "antibody:", "chip antibody details:", "chip:", "chip ab:", "lymphoblast antibody:", "factor:", "experiment\_type:”*

In the case a keyword is found, we extract the subsequent words and match them against the dictionary of gene names and histone modifications

MERGING THE ANTIBODY RESULTS

We merged the annotations obtained using R with those obtained from ENCODE antibodies and finally obtained the antibody target for (of the total 21037 SRA experiments) 16953 (unidentifiable antibody for 4084 samples). Results are available at

/Users/egaleota/epiMining/data/Rdataframes/antibody\_annotation.rds

The number of unique antibody names is 790. We found 2493 samples used as control (Input), 360 (igg) and 4084 sample descriptions for which it wasn’t possible to determine the antibody (NA)

IDENTIFICATION OF DISEASE STATE AND HEALTHY OR NORMAL SAMPLES

With the first prototype using conceptmapper we are not able to identify healthy or normal states in disease. Furthermore we do not subset the sentence where possible by selecting the disease related keywords. We also don’t remove the generic terms like “disease”. We ran conceptmapper on the metadata and obtained annotations for 5323 /21037 samples. To find normal or healthy samples we used an R script which allowed us to identify 1896 Normal samples. Of the 5323, 492 were annotated as normal (so they can be considered as conceptmapper false positives).

Resulting data frame annotated with disease where this information was available can be retrieved at

/Users/egaleota/epiMining/data/Rdataframes/ChIP\_metadata\_disease.rds'

IDENTIFICATION OF TISSUES

We ran Conceptmapper with the same options as previously described. Then we obtained a list of annotations for 19152/21037 samples, with 613 different tissue and cell line types. Annotations of tissues and cell lines from BRENDA has been stored at

'/Users/egaleota/epiMining/data/Rdataframes/chip\_tissue\_annotation.rds'

IDENTIFICATION OF TISSUES AND CELL LINES WITH THE ENCODE DICTIONARY

We built a dictionary with Encode Cell lines using the same file cv.ra obtained from the UCSC website.

The dictionary has been built with the following command

-f pom.xml exec:java

-Dexec.mainClass="iit.comp.genomics.SemanticMapping.EncodeCVBuilder"

-Dexec.args="/Users/egaleota/git/semanticmapping/SemanticMapping/src/main/resources/sourceDictionaries/cv.ra /Users/egaleota/git/semanticmapping/SemanticMapping/src/main/resources/conceptMapperDictionaries/encode\_cells\_dict.xml 'Cell Line'"

We obtained a list of 447 ENCODE cell lines, some of which had the corresponding id in BTO or the indication of the tissue they come from.

Next we run the ConceptMapper Entity finder to extract elements from the encode dictionary. In this case we were able to retrieve 6527 annotations for 4917 unique samples with 58 cell line names and tissues.

**FILTERS THAT WE NEED TO APPLY TO ENCODE**

1. Need to remove the deprecated terms (like K562B which is k562)

**Extraction of Myc samples and H3K4me1 and H3K27ac samples**

Among the annotated elements we were able to identify 910 GEO samples for H3K4me1, 1095 GEO samples for H3K27ac and 100 Myc samples.

Computation of the semantic similarity between the samples annotated with tissue terms

We developed a procedure to compute the groupwise semantic similarity between samples annotated with Brenda Terms. The java application is SMComutationBTO\_groupwise.java and returns the LIN semantic similarity for each couple of annotated files (ignoring the ones that were not annotated by any BTO term). Results are stored into

/Users/egaleota/epiMining/data/LinGroupWiseTissueSimilarity.txt

**The case study of Myc**

We retrieved the ChIP-seq experiments where Myc TF has been used as an antibody.

We were able to retrieve 100 unique GEO samples with Myc used as antibody. Of these, only for 83 samples the tissue or cell line has been identified. Missing matches include HEK293T and iMOP cells || cell type: immortalized multipotent otic progenitor cells. We have extracted a subset of Myc samples annotated as B-cell lymphoma. Subsequently we have also extracted samples of H3K27ac, H3K4me1, H3K4me3, and POL2. The results of alignement and peak calling using macs2 on hg19 genome have been stored at

/data/BA/public\_html/HTS-flow/DB/secondary/659/NARROW

and

/data/BA/public\_html/HTS-flow/DB/secondary/671/NARROW$

The following list contains the GEO samples taken into consideration

GSM1036404\_Myc\_B\_cell

GSM894058\_Myc\_B\_cell

GSM894060\_Myc\_B\_cell

GSM894059\_Myc\_B\_cell

GSM1133648\_H3K4me3\_B\_cell

GSM894068\_H3K4me3\_B\_cell

GSM1036412\_Pol2\_B\_cell

GSM803355\_Pol2\_B\_cell

GSM1254214\_H3K27ac\_B\_cell

GSM998994\_H3K4me1\_B\_cell

Of the Myc B-cell samples we only consider the GSM894058 (the other two refer to experiments where Myc has been repressed by Tetracicline at 24hs. The other two samples will be used to show that when one merges samples has to be careful to the experimental conditions.

We have also obtained samples for the same proteins or histone modifications profiled in a different tissue/cell line. In particular we considered the samples annotated as lung :

GSM894066\_H3K27ac\_Lung:

GSM621429\_H3K4me1\_Lung:

GSM621405\_H3K4me3\_Lung:

GSM1003607\_Myc\_Lung

For pol2 there was no sample annotated exactly with the term LUNG, so by relaxing the semantici similarity threshold we took the most similar POL2 sample (with a similarity value to MYC of 0.8532783) GSM1055822 annotated as fetal lung fibroblast.

EVALUATION DATASETS

We have defined two datasets to verify the quality of annotations both from Metamap and conceptMapper. The first dataset S1 contains 116 unique GEO samples (corresponding to 116 experiment\_accessions in SRA and 101 sample accessions in SRA). The dataset has been annota

|  |  |  |  |
| --- | --- | --- | --- |
| Number of GSMs | Number of  SRXs | Number of  SRSs | Unique sentences |
| 225 | 234 | 198 | 198 |
|  |  |  |  |

This table and the statistcs refer to the set of annotations as returned by the application of the annotators procedures on the 200 sample sentences without any post processing filtering of generic terms to reduce the false positive annotations.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP | FP | TN | FN | SEN | SP | P | NPV | F1 |
| MM full Tissues | 267 | 257 | 2 | 82 | 0.76 | 0.01 | 0.67 | 0.05 | 0.61 |
| MM keyw Tissues | 118 | 57 | 6 | 107 | 0.52 | 0.65 | 0.67 | 0.05 | 0.59 |
| MM full Diseases | 106 | 112 | 113 | 1 | 0.99 | 0.5 | 0.49 | 0.99 | 0.65 |
| MM keyword Diseases | 62 | 36 | 73 | 63 | 0.5 | 0.67 | 0.63 | 0.54 | 0.56 |
| CM Tissues | 312 | 33 | 15 | 34 | 0.9 | 0.31 | 0.9 | 0.31 | 0.9 |
| CM Diseases | 73 | 19 | 141 | 17 | 0.81 | 0.88 | 0.79 | 0.89 | 0.8 |

We applied also a filter to remove generic annotations. The filter in metamap was manual since none of the thresholds we tried to apply was able to correctly remove the generic terms like cells or tumor.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP | FP | TN | FN | SEN | SP | P | NPV | F1 |
| MM full Tissues | 261 | 50 | 2 | 12 |  |  |  |  |  |
| MM full Diseases | 92 | 57 | 113 | 1 |  |  |  |  |  |
| CM Tissues |  |  |  |  |  |  |  |  |  |
| CM Diseases |  |  |  |  |  |  |  |  |  |

After the application of the filters to remove the generic annotations we have performed a hierarchical clustering of the unique CUIs used to annotate the datasets. Clusters for metamap are available in the following folder

/Users/egaleota/epiMining/data/images

**TODO**

* Identification of target antibody in ChIP-seq metadata.
  + First we use ENCODE antibody dictionary with ConceptMapper
  + Second we use the procedure in R that uses Homo sapiens and Mus musculus bioconductor annotations
  + Need to identify specifically
    - Myc samples
    - H3K4me1
    - H3K27ac
  + Need to build a procedure that uses HUGO nomenclature and HIStome for building a dictionary for conceptmapper
* Identification of disease state for ChIP-seq, 450K data and Gene expression data
  + To identify diseases with disease ontology we can use the ConceptMapper. The issue here is to provide a module for identification of negations and healthy samples. It is possible to use GEOquery to retrieve GEO Datasets wich are already annotated with that information to assess the quality of the annotations
* Identification of tissues and cell lines ChIP-seq, 450k data and Gene expression data
  + To identify tissues and cell lines it is possible to use the ENCODE cell semantic type. It also provides links to the closest BTO or CL term that can be used to map the results to BTO
  + Than we use BTO itself with conceptmapper
* Validation of the annotation process
  + Assessment of the best performing ConceptMapper Settings (Can we use the evaluation pipeline?) for each specific case (tissue mentions and disease)
    - we can use the GDS dataset
    - We can use the manual annotated Myc dataset
* Retrieve gene expression data corresponding to the tissues and diseases considered into the 450K
* Add to conceptmapper the ability to retrieve specific pieces of text (labeled information like disease: )
* Write the paper

CONCEPT MAPPING STEPS

1 – Sentence detection, tokenization, lexical variant generation,

2 – Annotation of concepts from ontologies

3 – Filtering out of generic terms

4 – Filter to take only the most specific concept if two or more in the same hierarchy are used to annotate the same samples (This can be done creating a graph with SML)

5 – We also want to have partial matches for cells for example [Term]

id: BTO:0002181

name: HEK-293T cell

namespace: BrendaTissueOBO

def: "A highly transformed human renal epithelial line expressing two viral oncogenes, adenovirus E1a and SV40 large T antigen." [PMID:7553648]

synonym: "293 T cell" RELATED []

synonym: "293-T cell" RELATED []

synonym: "293T cell" RELATED []

synonym: "HEK293T cell" RELATED []

is\_a: BTO:0000067 ! kidney cell line

this term is not retrieved by Metamap when the sample description looks like

CELL: HEK293T

Instead imOP cells are not in the set of BTO terms

CCP NLP Library

Some of the GEO samples have been re-analyzed but refer to the same biological sample (I discovered this fact from the description of one geo sample which contained the information #Relationships: reanalyzed by # (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM822290>)

We are using the Conceptmapper wrapper and the type system provided by these people

MYC STUDY CASE

For what concerning Myc we are downloading the data of 15 human samples annotated with either B-cell lymphoma or B-lymphocyte.

Some of the files are in bed format, while others are in wig format. To convert wig to bed files I downloaded BEDOPS and ran the command

wig2bed < GSM1036404\_02242012\_C06TJACXX\_2.TTAGGC.wig > GSM1036404\_02242012\_C06TJACXX\_2.TTAGGC.bed

/Users/egaleota/Downloads/sratoolkit.2.5.2-mac64/bin/fastq-dump \*.sra

ENCODE Metadata download to check quality of my annotations

The case study of Myc revisited

We hypothesize that samples annotated with the same or very semantically close tissue or cell line and disease should have very similar profiles of regulatory elements (e.g similar chromatin states and similar transcription factors binding patterns) with respect to semantically distant elements. To test this hypothesis we retrieved the samples annotated with the term BTO:0001518 (B-lymphoma cell line) we get the following table) and the disease DOID:707 B-cell lymphoma.

Files are available at

/Users/egaleota/epimining/data/b\_cell\_lymphoma

'/data/BA/egaleota/MYC/b\_cell\_lymphoma.txt'

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GSE accession | LAB | experiment\_accession | Run accession | GEO\_accession | official\_antibody |
| GSE30726 | Charité Universitätsmedizin | SRX095393 | SRR341181 | GSM762707 | MYC |
| GSE30726 | Charité Universitätsmedizin | SRX095394 | SRR341182 | GSM762708 | MYC |
| GSE30726 | Charité Universitätsmedizin | SRX095395 | SRR341183 | GSM762709 | MYC |
| GSE30726 | Charité Universitätsmedizin | SRX095396 | SRR341184 | GSM762710 | MYC |
| GSE30726 | Charité Universitätsmedizin | SRX095397 | SRR341185 | GSM762711 | MYC |
| GSE51004  GSE51011 | Istituto Italiano di Tecnologia | SRX353807 | SRR993691 | GSM1234499 | MYC |
| GSE51004  GSE51011 | Istituto Italiano di Tecnologia | SRX353808 | SRR993692 | GSM1234500 | MYC |
| GSE51004  GSE51011 | Istituto Italiano di Tecnologia | SRX353809 | SRR993693 | GSM1234501 | MYC |
| GSE51004  GSE51011 | Istituto Italiano di Tecnologia | SRX353810 | SRR993694 | GSM1234502 | POLR2A |
| GSE51004  GSE51011 | Istituto Italiano di Tecnologia | SRX542553 | SRR1286936 | GSM1386342 | MYC |
| GSE51004  GSE51011 | Istituto Italiano di Tecnologia | SRX542554 | SRR1286937 | GSM1386343 | MYC |
| GSE51004  GSE51011 | Istituto Italiano di Tecnologia | SRX542555 | SRR1286938 | GSM1386344 | POLR2A |
| GSE51004  GSE51011 | Istituto Italiano di Tecnologia | SRX542556 | SRR1286939 | GSM1386345 | H3K4ME3 |
| GSE51004  GSE51011 | Istituto Italiano di Tecnologia | SRX542557 | SRR1286940 | GSM1386346 | H3K27AC |

In the figure

data/BA/egaleota/MYC/B\_cell\_lymphomas.pdf the overlap of peaks is reported for the samples analysed. Figure /data/BA/egaleota/MYC/B\_cell\_lymphomas\_promoters.pdf contains the overlap at promoter regions considering 2000 bp upstream and downstream the TSSs.

We then decided to semantically expand our initial dataset with semantically similar samples. In particular we selected the subset of ChIP-seq samples where the target was one of the four targets of the initial set of samples and had an intrinsic LIN sample’s similarity in ]0.9,1[.

Results are available in file /Users/egaleota/epiMining/data/Rdataframes/b\_cell\_similars.txt

And in file

/data/BA/egaleota/MYC/data/BA/egaleota/MYC/b\_cell\_similars.txt

In this way we were able to retrieve other 50 samples

* H3K27AC 17
* H3K4ME3 12
* MYC 5
* POLR2A 16

We

To test this hypothesis we considered several Myc samples from different experiments, in different laboratories, annotated with the Brenda term ‘B-lymphocytes’ and Disease ontology term ‘B-cell lymphoma’ and Myc samples in ‘Lovo cell’ with ‘Colon carcinoma’. Myc peaks co-occur with regulatory marks characterizing active states of the chromatin (promoters and enhancers) such as H3K4me1, H3K4me3, H3K27ac and with Pol225. Thus we also retrieved samples for these targets. When, for a given mark, we couldn’t find a sample exactly annotated with the tissue or disease of interest, the semantically closest sample has been retrieved. The following table summarizes the analyzed samples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SRA | GSM | TARGET | TISSUE TERM | DISEASE |
| SRX204411 | GSM1036404 | MYC | B-lymphocyte | B-cell lymphoma |
| SRX129067 | GSM894059 | MYC | B-lymphocyte | B-cell lymphoma |
| SRX129066 | GSM894058 | MYC | B-lymphocyte | B-cell lymphoma |
| SRX129068 | GSM804060 | MYC | B-lymphocyte | B-cell lymphoma |
| SRX275401 | GSM1133648 | H3K4me3 | B-lymphocyte | Diffuse large B-cell lymphoma |
| SRX129076 | GSM894068 | H3K4me3 | B-lymphocyte | B-cell lymphoma |
| SRX370348 | GSM1254214 | H3K27ac | B-lymphocyte | Diffuse large B-cell lymphoma |
| SRX183896 | GSM998994 | H3K4me1 | B-lymphocyte | Unknown [Presumed Healthy] |
| SRX204419 | GSM1036412 | Pol2 | B-lymphocyte | B-cell lymphoma |
| SRX100400 | GSM803355 | Pol2 | B-lymphocyte | Unknown |
| SRX361891 | GSM1242275 | MYC | LoVo cell, colon | Colon adenocarcinoma |
| SRX286204 | GSM1146450 | H3k4me3 | Colon | Colon carcinoma |
| SRX398300 | GSM1296643 | H3K27ac | HT-29 cell, Colon | Colon carcinoma |
| SRX128155 | GSM889410 | Pol2 | Colonic adenocarcinoma cell | Colon cancer |

Raw reads of the samples were aligned to the hg19 genome using BWA and peaks were called using MACS.

In the following table we report for each sample the number of peaks obtained

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SRA | TARGET | Number of peaks in promoters | Number of peaks | Tissue class |
| SRX204411 | MYC | 8523 | 11938 | B-lymphocyte |
| SRX129067 | MYC | 2935 | 4382 |
| SRX129066 | MYC | 13830 | 21296 |
| SRX129068 | MYC | 10399 | 14360 |
| SRX275401 | H3K4me3 | 21803 | 41360 |
| SRX129076 | H3K4me3 | 19010 | 32926 |
| SRX370348 | H3K27ac | 18848 | 40733 |
| SRX204419 | Pol2 | 13598 | 17341 |
| SRX100400 | Pol2 | 22090 | 35107 |
| SRX361891 | MYC | 316 | 1453 | Colon |
| SRX286204 | H3k4me3 | 22727 | 28354 |
| SRX398300 | H3K27ac | 28083 | 53165 |
| SRX128155 | Pol2 | 4943 | 7106 |

The following table reports the experiment details of each sample

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| GSM | TARGET | GSE experiment | Title | LAB |
| GSM1036404 | MYC | GSE42262 | Chromatin occupancy of key transcriptional regulators in P493-6 cells with various levels of c-Myc expression | Young Lab |
| GSM894059 (T0) | MYC | GSE36354 | Transcriptional Amplification in Tumor Cells with Elevated c-Myc | Young Lab |
| GSM894058 (T24) | MYC |
| GSM804060 (T1) | MYC |
| GSM1133648 | H3K4me3 | GSE46663 | Discovery and Characterization of Super-Enhancer Associated Dependencies in Diffuse Large B-Cell Lymphoma | Bradner Lab |
| GSM894068 | H3K4me3 | GSE36354 | Transcriptional Amplification in Tumor Cells with Elevated c-Myc | Young Lab |
| GSM1254214 | H3K27ac | GSE46663 | Discovery and Characterization of Super-Enhancer Associated Dependencies in Diffuse Large B-Cell Lymphoma | Bradner Lab |
| GSM1036412 | Pol2 | GSE42262 | Chromatin occupancy of key transcriptional regulators in P493-6 cells with various levels of c-Myc expression | Young Lab |
| GSM803355 | Pol2 | GSE32465 | Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB | ENCODE DCC |
| GSM1242275 | MYC | GSE51290 | Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites [S-phase/M-phase synchronized cells | Taipale lab |
| GSM1146450 | H3k4me3 | GSE47190 | ChIP-seq of MYST acetyltransferases | CHUQ |
| GSM1296643 | H3K27ac | GSE53602 | ChIP-Seq of Transcriptional Components in Colon carcinoma | Young Lab |
| GSM889410 | Pol2 | GSE36349 | Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes | Wilhelmina Children's Hospital, University Medical Center Utrecht |

We divided the hg19 genome in windows of size 50k.

When running ChIP-seq analysis in HTS-flow we have aligned reads to the hg19 genome with bwa (default parameters). In the first set of aligned files

SRX702147 - SRR1576444

SRX702148 - SRR1576445

SRX702149 - SRR1576446

SRX702150 - SRR1576447

SRX702151 - SRR1576448

Could not be aligned because fastq-dump wasn’t able to retrieve some sequences.

SRX111321 – SRR389094 failed during alignement with bwa

SRX111320 - SRR389093

For the remaining samples we called peaks with MACS2 (default options). The peak calling failed in building the model (Too few paired peaks)

SRR027918 - SRX011574

SRR341182 -

SRR357471 -

SRR547939 -

SRR547940 -

To check the quality of the alignment with HTSflow I copied the table providing statistics in /Users/egaleota/epiMining/data/RDataframes/HTSflow\_alignment\_quality.txt

And

/data/BA/egaleota/MYC/HTSflow\_alignment\_quality\_human.txt

I also added peak statistics in the file

**/data/BA/egaleota/MYC/HomoSapiensAlPeaks.txt**