**Research Projects between PoliMi and IEO-IIT**

**Stefano Ceri, Marco Masseroli, Fernando Palluzzi, Pietro Pinoli**

**April 2015**

The following document describes 6 joint PoliMi-IEO-IIT projects modeled and developed during the period April 2014 – April 2015. In every project, the GenoMetric Query Language has been used as a resource to store, pre-process, analyze, and retrieve genomic data produced at IEO-IIT.

# Project 1. 3D chromatin structure.

## 1. Biology

### 1.1 Biological description

ChIA-PET technology defines genomic regions where long-range interactions take place. Couples of genomic regions come to contact generating loops. In this project we want to assess if there is a correspondence between the distribution and the position of such loops across the genome and interesting/relevant regions, in order to assess if such chromatin structures influence gene expression in the regions proximal to the junction or in the loops. ChIA-PET data will be used to define CTCF-dependent loops, i.e. loops in which CTCF-binding regions are located at loop junctions. The left-flanking, loop, and right-flanking regions define physical chromatin domains, separated by CTCF. Loops can be analyzed by considering their spatial distribution and nested structures. The project will be performed by considering two different approaches.

* The first one is concerned with general questions that occur in the presence of 3D information on the genome; the purpose of this study is to define generic questions and observation methodologies (through the output viewer) that could in turn lead to the definition of a general-purpose service focused on 3D analysis, where each user can present her own files and then solve specific epigenetic questions, which will be defined as part of the project. Based on histone and RNAPII profiles, we can define the CTCF chromatin domains; these domains could be characterized by a different transcriptional status. For instance, we can put in contact distal enhancers and genes, by detecting putative enhancer regions, using H3K27ac and H3K4me1 histone profiles, and then by mapping RefSeq genes onto data to detect possible Enhancer-Gene interactions at CTCF-mediated loops. Then, through RNA-seq data, it is possible to examine the transcriptional status of such genes. Through RNA-seq data, it is possible to examine the transcriptional status of such domains.
* The second one can be considered as a specific question within the above setting. We consider two experimental datasets produced at IEO relative to in mouse embrio fibroblasts (MEFs), where the transcription factor NPM, which is supposed to be involved in the formation of CTCF-mediated loops, is either present or silenced; the analysis is focused on differentially examining the information at the loop ends in the two biological conditions, in order to confirm that NPM silencing should lead to less CTCF-mediated loops.

### 1.2. Biological questions (second part of the project):

* Do CTCF-mediated long-range interactions change in presence or absence of NPM?
* If they change, do this changes affect the expression levels, by acting locally on the CTCF insulator (i.e. the capability to isolate silenced regions) effect?

### References

L. Handoko et.al., CTCF-mediated functional chromatin interactome in pluripotent cells, Nature Genetics, 43:7, July 2011, 630-638 + Erratum and Supplemental Material (2)

L. A. Pennacchio, W. BickMore, A. Dean, M.A. Nobrega, G. Bejerano, Enhancers: five essential questions, Nature reviews/Genetics:14, April 2013, 288-294.

C. Y. McLean et. al., GREAT improves functional interpretation of *cis*-regulatory regions, Nature Biotechnology, 28:5, 2010, 495-501.

## 2. Data Availability

* Cell line: MEF (mm9)
* CTCF ChIA-PET, NPM/P35 knock-down
* Histone modifications NPM/P35 knock-down
* RNA-seq, NPM/P35 knock-down

## 3. Queries

#### 3.1.1 Data Preparation: dataset loading

ChIA-PET data is loaded inside the CHIAPET variable, with the following schema:

<chromosome, loopStart, loopEnd, read1, read2, score>

where read1 and read2 are the loop junction coordinates (i.e., the left loop is defined as [loopStart, read1], while the right loop is [read2, loopEnd]).

Variable REF contains RefSeq genes with their expression levels as TPM (Transcripts Per Million). On the base of RefSeq genes TSS, promoter coordinates are defined as [TSS – kL, TSS + kR]. In the following example, kL = 2000 and kR = 1000.

# Loading ChIA-PET data

CHIAPET = SELECT( dataType == 'ChIA-PET' AND treatment == 'p53ko' ) MM9\_CHIAPET;

CHIAPET\_NPM = SELECT( dataType == 'ChIA-PET' AND treatment == 'dkop53NPM' ) MM9\_CHIAPET;

# Loading expression data

REF = SELECT( dataType == 'RNA-seq' AND treatment == 'none' ) MM9\_RNASEQ;

PROMOTERS = PROJECT (true; start = start - 2000, stop = start + 1000) REF;

REF\_NPM = SELECT( dataType == 'RNA-seq' AND treatment == 'NPM' ) MM9\_RNASEQ;

PROMOTERS\_NPM = PROJECT (true; start = start - 2000, stop = start + 1000) REF\_NPM;

#### 3.1.2 Data Preparation: enhancer definition

After loading ChIA-PET data and defining promoter regions, we use ENCODE histone modifications H3K27ac, H3K4me1, and not H3K4me3, to define active enhancer regions. A region is said to be an active putative enhancer (PE) if these three modifications co-occur:

EHM = SELECT( cell == 'MEF' AND ( antibody == 'H3K27ac' OR antibody == 'H3K4me1') AND lab == 'LICR-m' ) MM9\_ENCODE\_BROAD;

PE = COVER(ALL, ANY) EHM;

#### 3.1.3 Data Processing: PE-Gene (PEG) interaction

Now that we defined promoters and PE regions, we want to assess if at least some of the loops in the CHIAPET dataset lead to CTCF-mediated enhancer-gene interactions. The following code enables us to annotate both junctions at the same time, considering PE regions at the left junction controlling promoters at the right junction. PROJECT arguments can be set for the opposite paradigm i.e., promoters on the left junction and PE regions on the right one.

The initial ChIA-PET data has the following schema:

<chromosome, loopStart, loopEnd> read1, read2, score

where the attributes in < . > denote the genomic region being processed, while the other attributes are its associated data values.

# Initial ChIA-PET data: <chromosome, loopStart, loopEnd> read1, read2, score

# Storing coordinates as data:

# <chromosome, loopStart, loopEnd> read1, read2, score, loopStart, loopEnd

JUNCTION\_L\_tmp = PROJECT(true; loopStart as left, loopEnd as right) CHIAPET;

# Moving to left junction:

# <chromosome, loopStart, read1> read1, read2, score, loopStart, loopEnd

JUNCTION\_L = PROJECT(true; right = read1) JUNCTION\_L\_tmp;

# Annotating PE-overlapping left-junctions:

# <chromosome, peStart, peEnd> read1, read2, score, loopStart, loopEnd

PEL\_tmp = JOIN(distance < 0, left) PE JUNCTION\_L;

# Storing coordinates as data:

# <chromosome, peStart, peEnd> read1, read2, score, loopStart, loopEnd, peStart, peEnd

PEL = PROJECT(true; peStart as left, peEnd as right) PEL\_tmp;

# Moving to right junction:

# <chromosome, read2, loopEnd> read1, read2, score, loopStart, loopEnd, peStart, peEnd

PRJ\_L = PROJECT(true; left = read2 AND right = loopEnd) PEL;

# Annotating promoter-overlapping right-junctions:

# <chromosome, promStart, promEnd> read1, read2, score, loopStart, loopEnd, peStart,

# peEnd

PRJ\_L\_tmp = JOIN(distance < 0, left) PROMOTERS PRJ\_L;

# Storing coordinates as data:

# <chromosome, promStart, promEnd> read1, read2, score, loopStart, loopEnd, peStart,

# peEnd, promStart, promEnd

PEG\_L\_tmp = PROJECT(true; promStart as left, promEnd as right) PRJ\_L\_tmp;

# Restoring original loop coordinates:

# <chromosome, loopStart, loopEnd> read1, read2, score, loopStart, loopEnd, peStart,

# peEnd, promStart, promEnd

PEG\_L = PROJECT(true; left = loopStart AND right = loopEnd) PEG\_L\_tmp;

In this example we used the variables CHIAPET and PROMOTERS. To annotate NPM samples, just substitute these variables with CHIAPET\_NPM and PROMOTERS\_NPM.

The final PEG\_L variable contains PEG interactions where PE regions are upstream the promoters they control.

#### 3.1.4 Data Processing: defining chromatin domains

Loops can be spatially represented as:

 L-junction R-junction

 ++++++++++++|----------X---------------------------------X----------|++++++++++++ L\_flanking (LFR) Loop\_region (LR) R\_flanking (RFR)

In this representation, the two junctions identify three different chromatin domains, namely left flanking, right flanking, and the loop region. These regions may be transcriptionally very different, due to chromatin conformation and the presence of distal regulatory elements (e.g. enhancers).

As initial ChIA-PET data we can use either PEG loops or whole-genome (WG) loops.

The GMQL code used to define these three regions from ChIA-PET data is defined below:

# Left Flanking Region (LFR) and Right Flanking Region (RFR)

# The upstream and downstream limits for LFRs and RFRs respectively are chosen

# arbitrarily. In this example they are both set at 5kbp, so that

# LFR = [loopStart – 5000, loopStart] and RFR = [loopEnd, loopEnd + 5000]

LFR = PROJECT (true; start = start - 5000, stop = start) CHIAPET;

LFR\_NPM = PROJECT (true; start = start - 5000, stop = start) CHIAPET\_NPM;

RFR = PROJECT (true; start = stop, stop = stop + 5000) CHIAPET;

RFR\_NPM = PROJECT (true; start = stop, stop = stop + 5000) CHIAPET\_NPM;

# Loop regions (LR)

# The loop region is defined as [read1, read2]

LR = PROJECT (true; left = read1 AND right = read2) CHIAPET;

LR\_NPM = PROJECT (true; left = read1 AND right = read2) CHIAPET\_NPM;

#### 3.1.5 Data Processing: checking transcriptional activity in different chromatin domains

Chromatin domains will be used as reference cenomic regions, while gene expression can be mapped onto them. Since gene expression is transcript-based, the resulting variable will contain transcript-based gene expression for each chromatin domain.

# Map expression data

LFR\_MAP = JOIN(distance < 0, left) LFR REF;

RFR\_MAP = JOIN(distance < 0, left) RFR REF;

LR\_MAP = JOIN(distance < 0, left) LR REF;

LFR\_MAP\_NPM = JOIN(distance < 0, left) LFR\_NPM REF\_NPM;

RFR\_MAP\_NPM = JOIN(distance < 0, left) RFR\_NPM REF\_NPM;

LR\_MAP\_NPM = JOIN(distance < 0, left) LR\_NPM REF\_NPM;

# Storing results

MATERIALIZE LFR\_MAP;

MATERIALIZE RFR\_MAP;

MATERIALIZE LR\_MAP;

MATERIALIZE LFR\_MAP\_NPM;

MATERIALIZE RFR\_MAP\_NPM;

MATERIALIZE LR\_MAP\_NPM;

## 4. Evaluation

#### 4.1 Gene expression at loops

Data can be collected according to different criteria:

* Focusing only on PEG loops or using WG loops
* Analyzing different domains separately
* Comparing the two condition on genes in common between the two condition for each domain
* Considering condition-specific genes for each domain
* Ranking data according their absolute gene expression in the two condition
* Ranking data according to the expression log-FC
* Ranking data according to the loop difference in the two conditions

As a first test, we explored data collected from PEG loop regions. From this first inspection, resulted that 542 genes show at least a 2-fold increase or decrease in the expression levels. Result data is reported in the attached file PEGresults.xlsx. The results file contains the following data:

* Loop coordinates and length for the two conditions
* Enhancer coordinates (preStart, peEnd) at junctions in the two conditions
* Gene names at junctions (junctionSymbol, npmJunctionSymbol) in the who conditions, with their expression levels
* Names and IDs of the genes within each loop, with their expression levels and log2-FC in between conditions (avgCount, npmAvgCount, and log2FC)

On the base of these data, we can proceed in three main directions: (i) looking at loops number and dimension, extracting the genes that are involved in such changes (reduction in number and length is observed from NPM-active versus non-NPM samples); (ii) extracting genes within loop regions having higher differences in gene expression between conditions; (iii) finding transcription factor binding patterns for the genes defined in step (ii). These three steps define lists of transcription factors with their target genes that may be further analyzed using network analysis.

# Project 2:

# Role of replication and transcription machineries in mutation.

## 1. Biology

**1.1 Biological description:**

Replication forks are the starting point of replication during DNA synthesis. They are characterized by bimodal signals in Repli-seq experiments. The replication process can be subdivided into time steps, corresponding to cell cycle phases. Then, the per-base amount of DNA produced at each step is measured in a Repli-seq assay. Here, the replication signal intensity is measured as s50. Each s50 value is calculated in a window of 50kb. Adjacent windows are considered all across the genome. Replication timing for one gene can be chosen either assigning the same s50 value as the TSS or calculating the average of all the s50 regions intersecting the gene. During time, we can observe a progressive increase of distance in the two peaks of the initial bimodal signal, until the moving peak merges with a neighboring peak from another replication fork. If we superimpose samples for each time step, we may see a sort of V-shape, depicted by the moving peaks. Based on the presence of a significant replication signal over time, we can distinguish in early and late replication regions. If s50 ≈ 0, the gene is considered to be at an earlier replication stag, while if s50 ≈ 1, the gene is at a later replication stage. These regions may be detected as a proportion of the overall replication signal over the N time steps. This proportion can be plotted as a sinusoid function of time, thus allowing us to clearly define regions of early and late replication.

It has been observed that transcription activity is down-expressed when replication is active. The hypothesis we want to test is that the observed down-regulation is a natural mechanism to avoid collisions between replication and translation machineries. Conversely, when an oncogene controlling a replication site is deregulated, the down-expressed in gene expression is not observed, suggesting a possible collision between the two molecular machineries. This phenomenon has been suggested as a necessary condition for DNA mutation in cancer.

**1.2. Biological questions**:

What we want to assess is whether early and late replication regions contain mutations in presence of deregulated oncogene activity and absence of down-expression. We consider all the mutations in the dataset. For each mutation, we consider the s50 value of the TSS of the intersecting gene and its expression level.

1. Are there mutations with common features?
2. Are mutations early replicating, late replicating, or there is no preference?
3. Are mutations associated with high expression levels, low expression levels, or there is no preference?
4. What is the average distance between each mutation and the nearest PML-RAR peak?
5. Does the nearest PML-RAR peak fall within the mutation-containing gene?

### References

G. I Dellino, D. Cittaro, R. Piccioni: Genome-wide mapping of human DNA-replication origins: Levels of transcription at OCR sites regulate origin selection and replication timing, Genome Research, Dec. 2012, 1-11.

Y. Jeon et. al., Temporal profiles of replication of human chromosomes, PNAS, May 2005, 6419-6424.

PA Greif et.al., Somatic mutations in acute promyelocytic leukemia (APL) identified by exome sequencing, Letters to the Editors, Leukemia, 2011, 1519-1522.

## 2. Data Availability

Cell line: PR9

**2.1. Available datasets.**

* AML mutation samples
* Protein binding data under PML\_RAR Zn induction 8h/24h
* Histone modification data under PML\_RAR Zn induction 8h/24h
* RNA-seq under PML\_RAR Zn induction 8h/24h
* Repli-seq under PML\_RAR Zn induction 8h/24h

## 3. Queries

#### 3.1.1 Data Preparation

GENES\_EXP = SELECT( PI == 'Dellino' ) HG19\_EXPRESSION;

S50 = SELECT( name == 's50' ) HG19\_REPLICATION;

MUTATIONS = SELECT( type == 'SNP' ) HG19\_TCGA\_MUTATIONS;

#### 3.1.2 Data Extraction

GENES\_MAP = MAP(AVG(score)) GENES\_EXP S50;

GENES\_MUT = MAP(COUNT) GENES\_MAP MUTATIONS;

MATERIALIZE GENES\_MUT;

**Project 3:**

**GTF2I Copy Number Variants (CNVs) in Chromosome 7, involved in Autism**

**Biological description:** CNVs localized in the long arm of chr7 may be involved in very different phenotypes affecting social behavior: “cocktail party” behavior (lack of behavioral restraints), and Williams syndrome, a disease characterized by autism-like social problems.

**Data**: ChIP-seq samples have been prepared at IEO (in-house data) for TFs GTF2i and LSD1, thought to have a major role in CNV-derived behavioral aberrations. ChIP-seq data is collected both for cases (patients with 1 and 3 copies of the chr7 long arm region) and normal people (2 copies).

In-house RNA-seq data is also collected for cases and controls in order to monitoring gene expression.

**Goal**: correlating differential TF binding regions for GTF2i and LSD1 (ChIP-seq data) with gene expression. We could also integrate ENCODE data.

**GMQL code**:

# Select experiments

#Note: this data should come from Testa group, and the metadata will be set accordingly

GTF2i\_1 = SELECT(antibody==’Gtf2i’, cell==‘1\_1’) GTF2I;

GTF2i\_3 = SELECT(antibody==’Gtf2i’, cell==‘3\_3’) GTF2I;

GTF2i = SELECT(antibody==’Gtf2i’, cell==’2\_2’) GTF2I;

LSD1\_1 = SELECT(antibody==’Lsd1’, cell==‘1\_1’) GTF2I;

LSD1\_3 = SELECT(antibody==’Lsd1’, cell==‘3\_3’) GTF2I;

LSD1 = SELECT(antibody==’Lsd1’, cell==’2\_2’) GTF2I;

# Find peaks that are specific for a given case

GTF2i\_1\_ONLY = JOIN(minDistance and distance >= 0, PROJECT\_LEFT\_DISTINCT) GTF2i\_1 GTF2i;

GTF2i\_3\_ONLY = JOIN(minDistance and distance >= 0, PROJECT\_LEFT\_DISTINCT) GTF2i\_3 GTF2i;

LSD1\_1\_ONLY = JOIN(minDistance and distance >= 0, PROJECT\_LEFT\_DISTINCT) LSD1\_1 LSD1;

LSD1\_3\_ONLY = JOIN(minDistance and distance >= 0, PROJECT\_LEFT\_DISTINCT) LSD1\_3 LSD1;

# Search for differentially expressed genes overlapping differentially bound regions

# Now we can do it only having it already as an annotation

DEG1 = SELECT(type ==’gene’ AND provider==’Pierre’ AND cell==’1copy’) ANNOTATION;

DEG3 = SELECT(type ==’gene’ AND provider==’Pierre’ AND cell==’3copy’) ANNOTATION;

DEG\_GTF2i\_1 = JOIN(distance < 0, left) DEG1 GTF2i\_1\_ONLY;

DEG\_GTF2i\_3 = JOIN(distance < 0, left) DEG3 GTF2i\_3\_ONLY;

DEG\_LSD1\_1 = JOIN(distance < 0, left) DEG1 LSD1\_1\_ONLY;

DEG\_LSD1\_3 = JOIN(distance < 0, left) DEG3 LSD1\_3\_ONLY;

# Project 4:

# Analysis of unknown P53 binding loci.

## 1. Biology

### 1.1 Biological description

P53 is an oncogene, in particular a tumor suppressor, located inside chromosome 11 in mouse, involved in apoptosis, genomic stability, and inhibition of angiogenesis. We have a core dataset consisting of six P53 ChIP-seq samples from mouse cell line CH12, in which a significant proportion of the binding signal is located in unknown genomic areas; i.e. unannotated regions. Our goal is to characterize at least part of these unknown binding regions. Firstly, we tried to map these unknown areas to RefSeq. Secondly, we tried to enlarge each P53 binding region by a window size of 10kb (i.e.; [region – 5kb, region + 5kb]). In both cases, we want to highlight the presence of annotations (genes and/or enhancers) and TF binding co-occurrence (i.e.; experimental ChIP-seq data for cell line CH12 from the ENCODE project). In addition, we pointed out which of the unknown binding loci are present in open chromatin regions.

### 1.2 Biological question

**Current question:** Characterize unknown P53 binding loci using annotations (RefSeq Genes, UCSC Known Genes, and Vista Enhancers), ENCODE ChIP-seq data and DNase-seq data from mouse cell line CH12.

Specifically, we target the questions below:

1. Map P53 unknown binding regions to RefSeq
2. Map P53 unknown binding regions to UCSC Known Genes
3. Map P53 unknown binding regions to Vista Enhancers
4. Map P53 unknown binding regions to putative enhancer regions
5. Find ChIP-seq samples from ENCODE mouse cell line CH12 overlapping unknown P53 binding regions
6. Find ChIP-seq samples from ENCODE mouse cell line CH12 overlapping unknown P53 binding regions enlarging the binding regions by a window of 10kb
7. Map P53 unknown binding regions to mouse CH12 DNase-seq experiments
8. Map expression data (RNA-seq) inside genes overlapping P53 binding regions

## 2. Data Availability

### Available datasets

* Variable UNKNOWN (dataset fernando.mm9p53):

Core dataset of 6 P53 ChIP-seq samples

* + The core dataset is available at:

genomic\_computing\_shared\progetti\progetto3\data\_bed\_files

* ENCODE TF and HM ChIP-seq data, and DNase-seq data for mouse cell line CH12.

## 3. Queries

### Data pre-processing and extraction is divided in eight different steps. The goal is to discover hidden region patterns in unknown P53 data. Queries 3.1 and 3.2 are used to characterize at least part of the unknown binding regions, using UCSC known genes and putative enhancer regions, defined by the overlap of typical enhancer-specific histone marks. An annotation-based enhancer definition is provided by step 3.3, using the Vista enhancer collection, although this set is quite limited. Gene-enhancer couples are further investigated using different association strategies: assigning the closest enhancer region to a gene (query 3.6), checking for CTCF or P300 binding (query 3.7), and finally identifying at least partial overlap with open chromatin region.

Query 3.3 is used to detect binding co-occurrence inside each P53 binding region. Step 3.4 is an extended version of 3.3, in which a window of 5kb from each P53 binding region terminus is considered.

### 3.1 Step 1: map P53 binding regions onto UCSC Known Genes

UNKNOWN = SELECT( cell == 'CH12' AND dataType == 'ChipSeq' AND antibody == 'p53' ) MM9\_P53unknown;

# Retrieve UCSC known genes

KNOWN = SELECT( name == 'knownGenes' ) fernando.knownGenes;

# Map the unknown P53 binding regions onto UCSC known genes

KNOWN\_MAP\_tmp = MAP(COUNT) KNOWN UNKNOWN;

# Retain only those UCSC known genes having at least one P53 peak

KNOWN\_MAP = PROJECT(count>0) KNOWN\_MAP\_tmp;

MATERIALIZE KNOWN\_MAP;

### 3.2 Step 2: map P53 binding regions onto putative enhancer (PE) regions

# In this case, enhancers are predicted, on the base of enhancer-specific

# histone modification (H3K27ac, H3K4me1, H3K4me3).

# Histone modification H3K27ac is compulsory.

# The UNKNOWN variable is the same used in query 3.1

UNKNOWN = SELECT(cell == 'CH12' AND (name == '1uMDOXO' OR name == '5uMNutl')) fernando.mm9p53;

HM1 = SELECT( cell == 'CH12' AND ( antibody == 'H3K4me1' OR antibody == 'H3K4me3' )

 AND lab == 'LICR-m' ) MM9\_ENCODE\_BROAD;

HM2 = SELECT( cell == 'CH12' AND antibody == 'H3K27ac' AND lab == 'LICR-m')

 MM9\_ENCODE\_BROAD;

# A putative enhancer is a region covered at least by a H3K27ac binding signal

# Other signals (H3K4me1, H3K4me3) only strengthen up the enhancer evidence

PE = COVER(1, 2) HM1, HM2;

# Map the unknown P53 binding regions onto PE regions

PE\_MAP\_tmp = MAP(COUNT) PE UNKNOWN;

# Retain only those PE regions having at least one P53 peak

PE\_MAP = PROJECT(count>0) PE\_MAP\_tmp;

MATERIALIZE PE\_MAP;

### 3.3 Step 3: Map P53 binding regions onto Vista Enhancers

# The UNKNOWN variable is the same used in query 3.1

# Retrieve enhancer from Vista database

VISTA = SELECT( name == 'VistaEnhancers' ) HG19\_BED\_ANNOTATION;

# Map the unknown P53 binding regions onto Vista enhancers

VISTA\_MAP\_tmp = MAP(COUNT) VISTA UNKNOWN;

# Retain only those Vista enhancers having at least one P53 peak

VISTA\_MAP = PROJECT(count>0) VISTA\_MAP\_tmp;

MATERIALIZE VISTA\_MAP;

**3.4 Step 4:**

**ENCODE mouse cell line CH12 transcription factor-overlapping unknown P53 binding regions**

# The UNKNOWN variable is the same used in query 3.1

# Retrieve transcription factor and histone modification data from cell line CH12

MM9\_BROAD = SELECT( cell == 'CH12' AND dataType == 'ChipSeq' ) MM9\_ENCODE\_BROAD;

MM9\_NARROW = SELECT( cell == 'CH12' AND dataType == 'ChipSeq' ) MM9\_ENCODE\_NARROW;

MM9\_CHIPSEQ = UNION MM9\_BROAD MM9\_NARROW

# Map ChIP-seq data onto unknown P53 binding regions

CHIPSEQ\_MAP\_tmp = MAP(COUNT) UNKNOWN MM9\_CHIPSEQ

# Retain only those Vista enhancers having at least one P53 peak

CHIPSEQ\_MAP = PROJECT(count > 0) CHIPSEQ\_MAP\_tmp

MATERIALIZE CHIPSEQ\_MAP

**3.5. Step 5:**

**ENCODE mouse cell line CH12 transcription factor-overlapping unknown P53 10kb-extended binding regions**

# In this case, we want to map ENCODE mm9 ChIP-seq data.

# The extension include 5kb upstream and 5kb downstream each region.

# ALL\_EXP data is the same used for query 3.4

DATI1\_10k = PROJECT(left = left – 5000, right = right + 5000) M1

DATI2\_10k = PROJECT(left = left – 5000, right = right + 5000) M2

DATI3\_10k = PROJECT(left = left – 5000, right = right + 5000) M3

DATI4\_10k = PROJECT(left = left – 5000, right = right + 5000) M4

DATI5\_10k = PROJECT(left = left – 5000, right = right + 5000) M5

DATI6\_10k = PROJECT(left = left – 5000, right = right + 5000) M6

M1\_10k = MAP(COUNT) DATI1\_10k ALL\_EXP;

M2\_10k = MAP(COUNT) DATI2\_10k ALL\_EXP;

M3\_10k = MAP(COUNT) DATI3\_10k ALL\_EXP;

M4\_10k = MAP(COUNT) DATI4\_10k ALL\_EXP;

M5\_10k = MAP(COUNT) DATI5\_10k ALL\_EXP;

M6\_10k = MAP(COUNT) DATI6\_10k ALL\_EXP;

MATERIALIZE M1\_10k;

MATERIALIZE M2\_10k;

MATERIALIZE M3\_10k;

MATERIALIZE M4\_10k;

MATERIALIZE M5\_10k;

MATERIALIZE M6\_10k;

### 3.6 Step 6: extract the closest PE region for each P53-bound gene

# Variables KNOWN (known genes) and PE (putative enhancers) come from queries

# 3.1 and 3.2, respectively

CLOSEST = JOIN(FIRST(1) AFTER DISTANCE 1000, right\_distinct) GENE PE

MATERIALIZE CLOSEST;

### 3.7 Step 7: Map CTCF and P300 binding onto PE regions and genes

# Variables KNOWN (known genes) and PE (putative enhancers) come from queries

# 3.1 and 3.2, respectively

# CHIPSEQ\_MAP sample comes from query 3.4

CTCF = SELECT( antibody == 'CTCF' ) MM9\_CHIPSEQ;

P300 = SELECT( antibody == 'P300' ) MM9\_CHIPSEQ;

# CTCF/P53 or P300/P53 binding genes or PE regions

GENE\_CTCF = JOIN(DISTANCE < 0, left\_distinct) KNOWN CTCF;

GENE\_P300 = JOIN(DISTANCE < 0, left\_distinct) KNOWN P300;

PE\_CTCF = JOIN(DISTANCE < 0, left\_distinct) PE CTCF;

PE\_P300 = JOIN(DISTANCE < 0, left\_distinct) PE P300;

# Enhancer-interacting regions

PE\_INT\_CTCF = JOIN(FIRST(1) AFTER DISTANCE 1000, left\_distinct) PE\_CTCF KNOWN\_CTCF;

PE\_INT\_P300 = JOIN(FIRST(1) AFTER DISTANCE 1000, left\_distinct) PE\_P300 KNOWN\_P300;

**3.8. Step 8: map P53 binding regions onto ENCODE CH12 DHS sites (DNase-seq regions)**

OPEN = SELECT( cell == 'CH12' AND dataType == 'DnaseSeq' ) MM9\_ENCODE\_BROAD;

OPEN\_MAP\_tmp = MAP(COUNT) OPEN UNKNOWN;

OPEN\_MAP = PROJECT(count>0) OPEN\_MAP\_tmp;

MATERIALIZE OPEN\_MAP;

## 4. Evaluation

Here we report seven data processing steps. The general problem we want to address is the characterization of unknown P53 binding regions. The first three steps consist in mapping unknown data against known annotations to assess if there are unknown peaks that have been recently annotated. We used the latest version of the UCSC knownGene annotation set, based on RefSeq, GenBank, and UniProt, to cover a larger amount of gene annotations.

The second step consists in mapping the unknown data onto putative enhancer (PE) regions. Putative enhancers are defined as regions where the enhancer-specific histone modifications H3K27ac, H3K4me1, and H3K4me2 co-occur. The third query maps the unknown regions against annotations in the Vista enhancer database. This procedure is used to validate at least part of the PEs. Datasets from queries 3.1 and 3.2 revealed the presence of possible interaction patterns patterns betweem genes and PEs, as shown in the following heatmaps.

 

The two pictures above report regions by rows and samples by columns. On the left, we reported 39 P53 unknown binding regions as belonging to known genes. Color shading accounts for the number of regions in the unknown dataset supporting the corresponding gene region in the annotation dataset: from 0 (lightest green) to 3 (darkest green). On the right, 55 PE regions are reported: light blue means no enhancer region in the sample, dark blue means one enhancer region.

Protein-mediated loops can be used to recover enhancer-gene relationships. Transcription factor binding site co-occurrence may provide useful information about the possible function of unknown P53 minding sites, revealing the presence of protein complexes or protein-mediated chromatin loops (i.e. looping interactions). TFBS co-occurrence is defined in query 3.4. Query 3.5 is a variation of 3.4, in which each unknown P53 binding site is extended by 5kb from each terminus. This information is used to recover possible gene-PE interactions. This is possible using three different criteria: (i) distance-based rules, (ii) TF co-binding, and (iii) co-occurrence in open chromatin regions (DNase-seq). Typical enhancer-gene interactions involve DNA-binding proteins, such as CTCF or P300, at specific average distances (e.g. typical long-range interactions take place at an average distance of 120kb). We may associate putative enhancers with their target genes by mapping CTCF and P300 binding profiles onto PE and gene regions extracted from the unknown dataset, as reported in query 3.6. A further validation of PE-gene interacting regions can be obtained using DNase-seq samples. Each DNase-seq sample contains regions called DHS (DNase Hypersensitive Sites) defining regions of open chromatin. Open chromatin conformation is a hallmark of actively transcribed DNA sequences, supporting the evidence of a true enhancer-gene interaction.

**Project 5:**

**Transcription factors co-occurrence with TEAD binding sites.**

**Biological Description**: The group is really interested in YAP bindings, but YAP cannot be traced; however it binds to TEAD which in turns binds to the genome. Find the co-occurence of TEAD transcription factor together with other transcription factors. We should look in an area which is around 2K bases from the TSS of genes (with genes having multiple TSS I recall a discussion which I think was concluded with taking the first TSS, SC). Looking with TFs, look at histone marks h3k21k? h3k43h? (I didn’t take good notes, SC). Scoring should go to TFs which are most present in these areas (so the answer is a scoring of 25 TS? Please confirm, SC)

### References

J.Yan et. al., Transcription Factors Binding in Huuman Cells occurs in Dense Clusters formed around Cohesin Anchor Sites, Cell 154, Aug. 2013, 801-813 and Supplemental material.

**Further details**:

In each of those cell lines, for each TEAD binding site, we want to see if other transcription factors bind within an interval (-2kb,+2kb). This is done in order to identify patterns of transcription factors that co-bind with TEAD. Furthermore, we want to split these patterns in three categories:

• Within a gene promoter (TSS-2kb, TSS+2kb)

• Within a gene body

• Intergenic regions

Then, given a pattern < TEAD, TF1, TF2, …, TFn> we want to tell if the sub-patterns <TF1, TF2, …, TFn> is common in the genome or if it shows a significant enrichment near TEAD peaks.

**Data:** 8 cell lines in Encode ChIP-seq datasets, 411 TF or HM after replica analysis (replica should either be merged or some should be eliminated; by now they are merged).

#######################

##

## This GMQL script should be the complete ideal one for the TEAD project.

## - At present, it does not work in this form due to the still partial working implementation of GMQL

## \*\* marks what is not functioning at present

## - Indented instructions could be avoided, but are included to generate also partial datasets that can facilitate

## the following data analysis

## - Note that calculation of number of regions obtained by an operation is very important to evaluate the operation result

##

##

## TEAD project

##

## 1) Select all samples including peaks (view = 'Peaks') of all (i.e. both TF and HM) ChIP-seq experiments

## (dataType = 'ChipSeq') of all the cell lines with TEAD ChIP-seq sample (antibody\_target == 'TEAD4')

## in ENCODE, excluding "Uniform" experiments (setType = exp), experiments with relevant drug treatment,

## and obsolete experiments (objStatus == '\*').

## 2) Annotate each peak of these sample as overlapping with a promoter, a gene but not with a promoter,

## or only with the remaining DNA regions (i.e. not overlapping either with a promoter or a gene).

## 3a) Identify which peaks of each of these samples overlap with TEAD peaks in the same cell line.

## 3b) Quantify the number of peaks of each selected ChIP-seq sample that overlap with a TEAD peak in the same sell line.

##

##

## Data selection and preparation

##

## All ChIP-seq samples of cell lines that include a TEAD ChIP-seq sample are selected and replicates are merged.

## Samples obtained from specimen treated with not ininfluent treatment (such as EtOH, DMSO, or Vehicle) are not selected

##

## Campaner provided promoter and gene lists are also selected

##

TEAD\_rep = SELECT(dataType == 'ChipSeq' AND view == 'Peaks' AND setType == 'exp' AND antibody\_target == 'TEAD4') \*\*HG19\_ENCODE\_PEAK;

## This public dataset (HG19\_ENCODE\_PEAK) should be obtained with a better repository creation

HM\_TF\_rep\_0 = \*\*SELECT(dataType == 'ChipSeq' AND view == 'Peaks' AND setType == 'exp' AND left->cell == right->cell

 AND NOT (treatment == 'BPA\_100nM' OR

 treatment == 'DEX\_100nM' OR treatment == 'DEX\_500pM' OR treatment == 'DEX\_50nM' OR treatment == 'DEX\_5nM' OR

 treatment == 'Estradiol\_100nM\_1hr' OR treatment == 'Estradiol\_10nM' OR treatment == 'Estradiol\_10nM\_30m' OR treatment =='Estradiol\_ctrl\_0hr' OR

 treatment == 'forskolin' OR

 treatment == 'Genistein\_100nM' OR

 treatment == 'insulin' OR

 treatment == 'pravastatin' OR

 treatment == 'serum\_free\_media' OR treatment == 'serum\_starved\_media' OR treatment == 'serum\_stimulated\_media' OR

 treatment == 'IFNa30' OR treatment == 'IFNa4h' OR treatment == 'IFNa6h' OR treatment == 'IFNg30' OR treatment == 'FNg4h' OR treatment == 'IFNg6h' OR

 treatment == 'estrogen' OR

 treatment == 'MNaseD')

 AND NOT (objStatus == '\*')

 ## This last (AND NOT (objStatus == '\*')) should be avoided with a cleaner repository creation

) HG19\_ENCODE\_PEAK TEAD\_rep;

HM\_TF\_rep = \*\*AGGREGATE(\_Region\_number AS COUNT) HM\_TF\_rep\_0;

HM\_TF\_0 = COVER(1, ANY; \*\*AVG(signal); \*\*STD(signal); GROUP BY cell, antibody\_target) HM\_TF\_rep;

HM\_TF = \*\*AGGREGATE(\_Region\_number\_cover AS COUNT) HM\_TF\_0;

MATERIALIZE HM\_TF;

## This following part can be avoided if the annotation regions are counted at repository creation and this value included in their metadata

PROM\_0 = SELECT(original\_provider == 'Campaner' AND annotation\_type == 'promoter') HG19\_BED\_ANNOTATION;

PROM = \*\*AGGREGATE(\_Region\_number AS COUNT) PROM\_0;

GENE\_0 = SELECT(original\_provider == 'Campaner' AND annotation\_type == 'gene') HG19\_BED\_ANNOTATION;

GENE = \*\*AGGREGATE(\_Region\_number AS COUNT) GENE\_0;

 MATERIALIZE PROM;

 MATERIALIZE GENE;

##

## Data annotation

##

## Each selected ChIP-seq peak is univocally assigned to one of the following classes:

## - PROM: overlapping with a promoter

## - GENE: overlapping with a gene but not with a promoter

## - OTHER: not overlapping either with a promoter or a gene

##

HM\_TF\_PROMonly\_0 = JOIN(DISTANCE < 0, project\_left\_distinct) HM\_TF PROM;

HM\_TF\_PROMonly\_1 = \*\*AGGREGATE(\_Region\_number AS COUNT) HM\_TF\_PROMonly\_0;

HM\_TF\_PROMonly = \*\*PROJECT(jaccard < 2.0; class AS 'PROM') HM\_TF\_PROMonly\_1;

 MATERIALIZE HM\_TF\_PROMonly;

HM\_TF\_PROMnot\_0 = DIFFERENCE(left->cell == right->cell AND left->antibody\_target == right->antibody\_target) HM\_TF HM\_TF\_PROMonly;

HM\_TF\_PROMnot = \*\*AGGREGATE(\_Region\_number\_diff AS COUNT) HM\_TF\_PROMnot\_0;

HM\_TF\_PROMnot\_GENEonly\_0 = JOIN(DISTANCE < 0, left\_distinct) HM\_TF\_PROMnot GENE;

HM\_TF\_PROMnot\_GENEonly\_1 = \*\*AGGREGATE(\_Region\_number\_diff AS COUNT) HM\_TF\_PROMnot\_GENEonly\_0;

HM\_TF\_PROMnot\_GENEonly = \*\*PROJECT(jaccard < 2.0; class AS 'GENE') HM\_TF\_PROMnot\_GENEonly\_1;

 MATERIALIZE HM\_TF\_PROMnot\_GENEonly;

HM\_TF\_PROMnot\_GENEnot\_0 = DIFFERENCE(left->cell == right->cell AND left->antibody\_target == right->antibody\_target) HM\_TF\_PROMnot HM\_TF\_PROMnot\_GENEonly;

HM\_TF\_PROMnot\_GENEnot\_1 = \*\*AGGREGATE(\_Region\_number\_difference AS COUNT) HM\_TF\_PROMnot\_GENEnot\_0;

HM\_TF\_PROMnot\_GENEnot = \*\*PROJECT(jaccard < 2.0; class AS 'OTHER') HM\_TF\_PROMnot\_GENEnot\_1;

 MATERIALIZE HM\_TF\_PROMnot\_GENEnot;

HM\_TF\_all\_0 = \*\*UNION(left->cell == right->cell AND left->antibody\_target == right->antibody\_target) HM\_TF\_PROMonly HM\_TF\_PROMnot\_GENEonly HM\_TF\_PROMnot\_GENEnot;

HM\_TF\_all = \*\*AGGREGATE(\_Region\_number\_union AS COUNT) HM\_TF\_all\_0;

MATERIALIZE HM\_TF\_all;

##

## Data extraction

##

## Annotated TEAD samples are extracted and for each TEAD peak,

## the number of peaks of each selected ChIP-seq sample that overlap with such TEAD peak is counted

##

TEAD\_all = SELECT(antibody\_target == 'TEAD4') HM\_TF\_all;

 MATERIALIZE TEAD\_all;

TEAD\_HM\_TF = MAP(left->cell == right->cell, COUNT) TEAD\_all HM\_TF;

MATERIALIZE TEAD\_HM\_TF;

 TEAD\_HM\_TF\_PROMonly = PROJECT(class == 'PROM') TEAD\_HM\_TF;

 MATERIALIZE TEAD\_HM\_TF\_PROMonly;

 TEAD\_HM\_TF\_PROMnot\_GENEonly = PROJECT(class == 'GENE') TEAD\_HM\_TF;

 MATERIALIZE TEAD\_HM\_TF\_PROMnot\_GENEonly;

 TEAD\_HM\_TF\_PROMnot\_GENEnot = PROJECT(class == 'OTHER') TEAD\_HM\_TF;

 MATERIALIZE TEAD\_HM\_TF\_PROMnot\_GENEnot;

###########################

- 743 samples considerati (incluse repliche)
- 411 samples analizzati (dopo merging delle repliche)    [on average about 2 MB each sample]
- 8 cell lines
- 161 HM o TF distinti analizzati (nel complesso delle 8 cell lines)