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

# get sra from SRR from SRA toolkit

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

wget --output-document sratoolkit.tar.gz http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-ubuntu64.tar.gz

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

# run the nexflow ChIP-seq pipeline by creating an environment under screen

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

screen -S chipseq

module load Anaconda2/5.3.0

wget -qO- https://get.nextflow.io | bash

module load Nextflow/19.10.0

conda create -n nextflow-env nextflow

conda activate nextflow-env

nextflow run nf-core/chipseq -c parameter\_SE.config -profile conda

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

# parameter setting of the config file

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

params.input = 'CG12\_1\_design.csv'

params.genome = 'GRCh38'

params.save\_reference = true

params.email = 'gloria.wu@med.lu.se'

prarms.email\_on\_fail = 'gloria.wu@med.lu.se'

params.project = 'lu2022-7-3'

params.single\_end = true

params.narrow\_peak = true

resume = true

workDir = '/home/gloriawu/snic2020-12-1/working/Cistrome'

params.max\_memory = '60.GB'

params.outdir = '/home/gloriawu/snic2020-12-1/ChIP/analysis2022/12\_1\_5573\_narrow'

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

# merge BAM files of replicates and call peaks

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

#!/bin/bash -eu

#SBATCH -A LU2022-7-3

#SBATCH -n 12

#SBATCH -t 08:00:00

#load modules

module load GCC/4.9.3-2.25 OpenMPI/1.10.2

module load icc/2015.3.187-GNU-4.9.3-2.25 impi/5.0.3.048

module load ifort/2015.3.187-GNU-4.9.3-2.25 impi/5.0.3.048

module load MACS2/2.1.0.20150731-Python-2.7.11

set -euxo pipefail

# set HOME

MYHOME=/home/gloriawu/snic2020-12-1/ChIP/analysis2022

# input files

BAM1=$MYHOME/34\_46025\_narrow/bwa/mergedLibrary/Cis\_GATA1\_34\_1\_R1.mLb.clN.sorted.bam

BAM2=$MYHOME/34\_46025\_narrow/bwa/mergedLibrary/Cis\_GATA1\_34\_1\_R2.mLb.clN.sorted.bam

BAM3=$MYHOME/erythroblast\_1\_narrow/bwa/mergedLibrary/GATA1\_eryth\_1\_R1.mLb.clN.sorted.bam

BAM4=$MYHOME/erythroblast\_1\_narrow/bwa/mergedLibrary/GATA1\_eryth\_1\_R2.mLb.clN.sorted.bam

control1=$MYHOME/34\_46025\_narrow/bwa/mergedLibrary/WT\_CONTROL\_R1.mLb.clN.sorted.bam

control2=$MYHOME/erythroblast\_1\_narrow/bwa/mergedLibrary/WT\_CONTROL\_R1.mLb.clN.sorted.bam

OUT=$MYHOME/GATA1/GATA1\_GE34\_merge\_

# run macs2 call peak

macs2 callpeak -t $BAM1 $BAM2 $BAM3 $BAM4 -c $control1 $control2 -g hs --outdir $MYHOME/GATA1 -n $OUT -B -q 0.01

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

# merge narrow peak files

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

#!/bin/bash -eu

#SBATCH -A LU2022-7-3

#SBATCH -n 6

#SBATCH -t 04:00:00

set -euxo pipefail

module load GCC/5.4.0-2.26

module load OpenMPI/1.10.3

module load impi/2017.1.132

module load BEDTools/2.26.0

cd /home/gloriawu/snic2020-12-1/ChIP/analysis2022/GATA1/

bedtools intersect -a Cis\_GATA1\_3\_peaks.narrowPeak -b Cis\_GATA1\_12\_1\_peaks.narrowPeak Cis\_GATA1\_12\_2\_peaks.narrowPeak GATA1\_GE34\_merge\_\_peaks.narrowPeak -c > Intersect\_with\_Cis\_3\_GATA1\_adult.bed

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

# get fasta file on the merged bed files with hg38

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

#!/bin/bash -u

#SBATCH -A LU2022-7-3

#SBATCH -n 8

#SBATCH -t 8:00:00

set -euxo pipefail

module load GCC/5.4.0-2.26

module load OpenMPI/1.10.3

module load impi/2017.1.132

module load BEDTools/2.26.0

bedtools getfasta -name -fi /home/gloriawu/reference/hg38/hg38.fa -bed /home/gloriawu/snic2020-12-1/ChIP/analysis2022/GATA1/Intersect\_with\_Cis\_3\_GATA1\_adult.bed -fo /home/gloriawu/snic2020-12-1/ChIP/analysis2022/GATA1/intersect/Intersect\_with\_Cis\_3\_GATA1\_adult.fa

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

# Find GATA1 motif with FIMO from MEME suite

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

#!/bin/bash -e

#SBATCH -A LU2022-7-3

#SBATCH -n 8

#SBATCH -t 12:00:00

set -euxo pipefail

module load GCC/7.3.0-2.30

module load OpenMPI/3.1.1

module load MEME/5.0.4

fimo --thresh 1e-3 --max-stored-scores 1000000 --oc /home/gloriawu/snic2020-12-1/ChIP/analysis2022/GATA1/intersect/Intersect\_Cis\_3\_FIMO35\_3 MA0035.3.meme /home/gloriawu/snic2020-12-1/ChIP/analysis2022/GATA1/intersect/Intersect\_with\_Cis\_3\_GATA1\_adult.fa

## rename the file

mv fimo.tsv Intersect\_Cis\_3\_GATA1\_FIMO35\_3.tsv

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

# sort file structure using open Refine

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

#Load tsv file, create project

# split sequence\_name by ::

# delete empty column (sequence\_name 2)

# separate sequence\_name 4 by -

# rename sequence name column titles

# split sequence\_name 1 by /

# remove all other columns but sequence\_name 1 8, which is the ID for peaks

# rename sequence\_name 1 8 to sequence name

# remove the first two columns which are the motif\_id and motif\_alt\_id

# sort by sequence name

# move sequence\_name columne to the 4th column to allow typical bed files at the start

# make start and stop numbers

# create column to calculate the motif start and motif stop

# reorder columns

[

{

"op": "core/column-split",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "sequence\_name",

"guessCellType": true,

"removeOriginalColumn": true,

"mode": "separator",

"separator": ":",

"regex": false,

"maxColumns": 0,

"description": "Split column sequence\_name by separator"

},

{

"op": "core/column-removal",

"columnName": "sequence\_name 2",

"description": "Remove column sequence\_name 2"

},

{

"op": "core/column-split",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "sequence\_name 4",

"guessCellType": true,

"removeOriginalColumn": true,

"mode": "separator",

"separator": "-",

"regex": false,

"maxColumns": 0,

"description": "Split column sequence\_name 4 by separator"

},

{

"op": "core/column-rename",

"oldColumnName": "sequence\_name 3",

"newColumnName": "chr",

"description": "Rename column sequence\_name 3 to chr"

},

{

"op": "core/column-rename",

"oldColumnName": "sequence\_name 4 1",

"newColumnName": "peak\_start",

"description": "Rename column sequence\_name 4 1 to peak\_start"

},

{

"op": "core/column-rename",

"oldColumnName": "sequence\_name 4 2",

"newColumnName": "peak\_end",

"description": "Rename column sequence\_name 4 2 to peak\_end"

},

{

"op": "core/column-split",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "sequence\_name 1",

"guessCellType": true,

"removeOriginalColumn": true,

"mode": "separator",

"separator": "/",

"regex": false,

"maxColumns": 0,

"description": "Split column sequence\_name 1 by separator"

},

{

"op": "core/column-removal",

"columnName": "sequence\_name 1 1",

"description": "Remove column sequence\_name 1 1"

},

{

"op": "core/column-removal",

"columnName": "sequence\_name 1 2",

"description": "Remove column sequence\_name 1 2"

},

{

"op": "core/column-removal",

"columnName": "sequence\_name 1 3",

"description": "Remove column sequence\_name 1 3"

},

{

"op": "core/column-removal",

"columnName": "sequence\_name 1 4",

"description": "Remove column sequence\_name 1 4"

},

{

"op": "core/column-removal",

"columnName": "sequence\_name 1 5",

"description": "Remove column sequence\_name 1 5"

},

{

"op": "core/column-removal",

"columnName": "sequence\_name 1 6",

"description": "Remove column sequence\_name 1 6"

},

{

"op": "core/column-removal",

"columnName": "sequence\_name 1 7",

"description": "Remove column sequence\_name 1 7"

},

{

"op": "core/column-rename",

"oldColumnName": "sequence\_name 1 8",

"newColumnName": "sequence\_name",

"description": "Rename column sequence\_name 1 8 to sequence\_name"

},

{

"op": "core/column-removal",

"columnName": "motif\_id",

"description": "Remove column motif\_id"

},

{

"op": "core/column-removal",

"columnName": "motif\_alt\_id",

"description": "Remove column motif\_alt\_id"

},

{

"op": "core/column-move",

"columnName": "sequence\_name",

"index": 1,

"description": "Move column sequence\_name to position 1"

},

{

"op": "core/column-move",

"columnName": "sequence\_name",

"index": 2,

"description": "Move column sequence\_name to position 2"

},

{

"op": "core/column-move",

"columnName": "sequence\_name",

"index": 3,

"description": "Move column sequence\_name to position 3"

},

{

"op": "core/text-transform",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "start",

"expression": "value.toNumber()",

"onError": "keep-original",

"repeat": false,

"repeatCount": 10,

"description": "Text transform on cells in column start using expression value.toNumber()"

},

{

"op": "core/text-transform",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "stop",

"expression": "value.toNumber()",

"onError": "keep-original",

"repeat": false,

"repeatCount": 10,

"description": "Text transform on cells in column stop using expression value.toNumber()"

},

{

"op": "core/column-addition",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"baseColumnName": "peak\_start",

"expression": "grel:value+cells['start'].value",

"onError": "set-to-blank",

"newColumnName": "FIMO\_motif\_start",

"columnInsertIndex": 2,

"description": "Create column FIMO\_motif\_start at index 2 based on column peak\_start using expression grel:value+cells['start'].value"

},

{

"op": "core/column-addition",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"baseColumnName": "peak\_start",

"expression": "grel:value+cells['stop'].value",

"onError": "set-to-blank",

"newColumnName": "FIMO\_motif\_stop",

"columnInsertIndex": 2,

"description": "Create column FIMO\_motif\_stop at index 2 based on column peak\_start using expression grel:value+cells['stop'].value"

},

{

"op": "core/column-reorder",

"columnNames": [

"chr",

"FIMO\_motif\_start",

"FIMO\_motif\_stop",

"sequence\_name",

"peak\_start",

"peak\_end",

"start",

"stop",

"strand",

"score",

"p-value",

"q-value",

"matched\_sequence"

],

"description": "Reorder columns"

}

]

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

# change tsv to bed and make sure no weird characters

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

mv Intersect\_Cis\_3\_GATA1\_FIMO35\_3\_sort.tsv Intersect\_Cis\_3\_GATA1\_FIMO35\_3\_sort.bed

cat Intersect\_Cis\_3\_GATA1\_FIMO35\_3\_sort.bed | tr -d '\r' > Intersect\_Cis\_3\_GATA1\_FIMO35\_3\_sorted.bed

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

# intersect FIMO and JASPAR

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

# get rid of header

tail -n +2 Intersect\_Cis\_3\_GATA1\_FIMO35\_3\_sorted.bed > file.tmp && mv file.tmp Intersect\_Cis\_3\_GATA1\_FIMO35\_3\_sort\_noheader.bed

# sort

#!/bin/bash -eu

#SBATCH -A LU2022-7-3

#SBATCH -n 6

#SBATCH -t 04:00:00

set -euxo pipefail

module load GCC/5.4.0-2.26

module load OpenMPI/1.10.3

module load impi/2017.1.132

module load BEDTools/2.26.0

cd /home/gloriawu/snic2020-12-1/ChIP/analysis2022/GATA1/intersect

bedtools sort -i Intersect\_Cis\_3\_GATA1\_FIMO35\_3\_sort\_noheader.bed > Intersect\_Cis\_3\_GATA1\_FIMO35\_3\_noheader\_sort.bed

bedtools sort -i MA0035.3\_GATA1\_chr\_hg38.bed > MA0035.3\_GATA1\_chr\_hg38\_sort.bed

# merge

#!/bin/bash -eu

#SBATCH -A LU2022-7-3

#SBATCH -n 6

#SBATCH -t 04:00:00

set -euxo pipefail

module load GCC/5.4.0-2.26

module load OpenMPI/1.10.3

module load impi/2017.1.132

module load BEDTools/2.26.0

cd /home/gloriawu/snic2020-12-1/ChIP/analysis2022/GATA1/intersect

bedtools intersect -wa -wb -a Intersect\_Cis\_3\_GATA1\_FIMO35\_3\_noheader\_sort.bed -b MA0035.3\_GATA1\_chr\_hg38\_sort.bed > Intersect\_with\_Cis\_3\_GATA1\_FIMO\_JASPAR.bed

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

# clean up with open refine for the FIMO-JASPAR file

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

# Load file, 0 line as column headers, create project

# make numbers for the variables need to be calculated

# create column to calculate the difference in base pairs between FIMO and JASPAR motif start sites

# rename the columns

# delete unnecessary columns

# rearrange the columns

[

{

"op": "core/text-transform",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "Column 2",

"expression": "value.toNumber()",

"onError": "keep-original",

"repeat": false,

"repeatCount": 10,

"description": "Text transform on cells in column Column 2 using expression value.toNumber()"

},

{

"op": "core/text-transform",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "Column 15",

"expression": "value.toNumber()",

"onError": "keep-original",

"repeat": false,

"repeatCount": 10,

"description": "Text transform on cells in column Column 15 using expression value.toNumber()"

},

{

"op": "core/column-addition",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"baseColumnName": "Column 2",

"expression": "grel:value-cells['Column 15'].value",

"onError": "set-to-blank",

"newColumnName": "diff\_FIMO\_Jaspar",

"columnInsertIndex": 2,

"description": "Create column diff\_FIMO\_Jaspar at index 2 based on column Column 2 using expression grel:value-cells['Column 15'].value"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 1",

"newColumnName": "chr",

"description": "Rename column Column 1 to chr"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 2",

"newColumnName": "FIMO\_start",

"description": "Rename column Column 2 to FIMO\_start"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 3",

"newColumnName": "FIMO\_stop",

"description": "Rename column Column 3 to FIMO\_stop"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 4",

"newColumnName": "sequence\_name",

"description": "Rename column Column 4 to sequence\_name"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 5",

"newColumnName": "peak\_start",

"description": "Rename column Column 5 to peak\_start"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 6",

"newColumnName": "peak\_end",

"description": "Rename column Column 6 to peak\_end"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 7",

"newColumnName": "start",

"description": "Rename column Column 7 to start"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 8",

"newColumnName": "stop",

"description": "Rename column Column 8 to stop"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 9",

"newColumnName": "strand",

"description": "Rename column Column 9 to strand"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 10",

"newColumnName": "FIMO\_score",

"description": "Rename column Column 10 to FIMO\_score"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 11",

"newColumnName": "FIMO\_pValue",

"description": "Rename column Column 11 to FIMO\_pValue"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 12",

"newColumnName": "FIMO\_qValue",

"description": "Rename column Column 12 to FIMO\_qValue"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 13",

"newColumnName": "FIMO\_matched\_sequence",

"description": "Rename column Column 13 to FIMO\_matched\_sequence"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 15",

"newColumnName": "JASPAR\_start",

"description": "Rename column Column 15 to JASPAR\_start"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 16",

"newColumnName": "JASPAR\_stop",

"description": "Rename column Column 16 to JASPAR\_stop"

},

{

"op": "core/column-removal",

"columnName": "Column 17",

"description": "Remove column Column 17"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 18",

"newColumnName": "JASPAR\_score",

"description": "Rename column Column 18 to JASPAR\_score"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 14",

"newColumnName": "chromosome",

"description": "Rename column Column 14 to chromosome"

},

{

"op": "core/column-reorder",

"columnNames": [

"chr",

"FIMO\_start",

"FIMO\_stop",

"FIMO\_score",

"FIMO\_pValue",

"FIMO\_qValue",

"FIMO\_matched\_sequence",

"chromosome",

"JASPAR\_start",

"JASPAR\_stop",

"JASPAR\_score",

"diff\_FIMO\_Jaspar",

"sequence\_name",

"peak\_start",

"peak\_end",

"start",

"stop",

"strand"

],

"description": "Reorder columns"

}

]

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

# clean up with open refine for the intersect file with overlaps

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

# Load file, 0 line as column headers, create project

# split column 4

# remove all other column 4s but keep column 4 8, which is the peak ID

# remove column 6 which is single .

# rename all fields

[

{

"op": "core/column-split",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "Column 4",

"guessCellType": true,

"removeOriginalColumn": true,

"mode": "separator",

"separator": "/",

"regex": false,

"maxColumns": 0,

"description": "Split column Column 4 by separator"

},

{

"op": "core/column-removal",

"columnName": "Column 4 1",

"description": "Remove column Column 4 1"

},

{

"op": "core/column-removal",

"columnName": "Column 4 2",

"description": "Remove column Column 4 2"

},

{

"op": "core/column-removal",

"columnName": "Column 4 3",

"description": "Remove column Column 4 3"

},

{

"op": "core/column-removal",

"columnName": "Column 4 4",

"description": "Remove column Column 4 4"

},

{

"op": "core/column-removal",

"columnName": "Column 4 5",

"description": "Remove column Column 4 5"

},

{

"op": "core/column-removal",

"columnName": "Column 4 6",

"description": "Remove column Column 4 6"

},

{

"op": "core/column-removal",

"columnName": "Column 4 7",

"description": "Remove column Column 4 7"

},

{

"op": "core/column-removal",

"columnName": "Column 6",

"description": "Remove column Column 6"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 1",

"newColumnName": "chr",

"description": "Rename column Column 1 to chr"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 2",

"newColumnName": "peak\_start",

"description": "Rename column Column 2 to peak\_start"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 3",

"newColumnName": "peak\_end",

"description": "Rename column Column 3 to peak\_end"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 4 8",

"newColumnName": "sequence\_name",

"description": "Rename column Column 4 8 to sequence\_name"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 5",

"newColumnName": "peak\_score",

"description": "Rename column Column 5 to peak\_score"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 7",

"newColumnName": "signal\_value",

"description": "Rename column Column 7 to signal\_value"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 8",

"newColumnName": "pValue",

"description": "Rename column Column 8 to pValue"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 9",

"newColumnName": "qValue",

"description": "Rename column Column 9 to qValue"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 10",

"newColumnName": "peak\_summit",

"description": "Rename column Column 10 to peak\_summit"

},

{

"op": "core/column-rename",

"oldColumnName": "Column 11",

"newColumnName": "overlap",

"description": "Rename column Column 11 to overlap"

}

]

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

# clean up the data before working with intersect cis adult file to annotate in R

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

# first cut the header

tail -n +2 Intersect\_with\_Cis\_3\_GATA1\_adult\_sort.bed > file.tmp && mv file.tmp Intersect\_with\_Cis\_3\_GATA1\_adult\_sort.bed

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

# R program script for annotating, joining, and filtering

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

# load files

Intersect\_with\_Cis\_3\_GATA1\_adult\_sort.bed <- read.delim("~/Documents/LU/ENCODE/ENCODE/analysis2022/GATA1/intersect/OpenRefine/Intersect\_with\_Cis\_3\_GATA1\_adult\_sort.bed.tsv")

# load packages

library(rtracklayer)

library(Rsamtools)

library(BiocManager)

BiocManager::install("TxDb.Hsapiens.UCSC.hg38.knownGene")

txdb <- TxDb.Hsapiens.UCSC.hg38.knownGene::TxDb.Hsapiens.UCSC.hg38.knownGene

library(ChIPpeakAnno)

library(ChIPseeker)

library(tidyverse)

library(dplyr)

# try to annotate the file after overlap

Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak <- readPeakFile("Intersect\_with\_Cis\_3\_GATA1\_adult\_sort.bed")

Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate <- annotatePeak(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak, tssRegion = c(-5000, 5000), TxDb = txdb, annoDb = "org.Hs.eg.db")

Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df <- data.frame(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate)

View(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df)

# name the columns

names(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df)[1] <-'Chr'

names(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df)[6] <-'sequence\_name'

names(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df)[7] <-'peak\_score'

names(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df)[8] <-'signal\_value'

names(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df)[9] <-'pValue'

names(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df)[10] <-'qValue'

names(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df)[11] <-'peak\_summit'

names(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df)[12] <-'overlap'

write.csv(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df, file="Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate.csv")

# load the file from FIMO\_JASPAR

# innerjoin by sequenc\_name

Intersect\_Cis\_3\_GATA1\_FIMMO\_JASPAR\_annotate <- inner\_join(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df,Intersect\_with\_Cis\_3\_GATA1\_FIMO\_JASPAR\_sort, by ='sequence\_name')

View(Intersect\_Cis\_3\_GATA1\_adult\_sort\_peak\_annotate\_df)

View(Intersect\_with\_Cis\_3\_GATA1\_FIMO\_JASPAR\_sort)

View(Intersect\_Cis\_3\_GATA1\_FIMMO\_JASPAR\_annotate)

# load gene symbol file

bloodgroupgenesymbol <- read.csv("~/Documents/LU/ENCODE/ENCODE/bloodgroupgenesymbol.csv", sep="")

View(bloodgroupgenesymbol)

# semijoin with blood group genes

Intersect\_Cis\_3\_GATA1\_FIMO\_JASPAR\_annotate\_blood <- semi\_join(Intersect\_Cis\_3\_GATA1\_FIMMO\_JASPAR\_annotate, bloodgroupgenesymbol, by = 'SYMBOL')

View(Intersect\_Cis\_3\_GATA1\_FIMO\_JASPAR\_annotate\_blood)

write.table(Intersect\_Cis\_3\_GATA1\_FIMO\_JASPAR\_annotate\_blood, file ="Intersect\_Cis\_3\_GATA1\_FIMO\_JASPAR\_annotate\_blood.tsv", sep = '\t', row.names = FALSE, col.names = TRUE)

count(Intersect\_Cis\_3\_GATA1\_FIMO\_JASPAR\_annotate\_blood, overlap >='3')

count(Intersect\_Cis\_3\_GATA1\_FIMO\_JASPAR\_annotate\_blood, overlap ='2')

count(Intersect\_Cis\_3\_GATA1\_FIMO\_JASPAR\_annotate\_blood, overlap)

count(Intersect\_Cis\_3\_GATA1\_FIMO\_JASPAR\_annotate\_blood, diff\_FIMO\_Jaspar)

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

# clean up the annotated data by OpenRefine

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

# Load file and create project

# compare the fields for the peak start and end to make sure it’s the same

# all 359 end at the same position, and 1 bp difference from the start site

# remove unnecessary fields

# rename fields

[

{

"op": "core/column-removal",

"columnName": "strand.x",

"description": "Remove column strand.x"

},

{

"op": "core/text-transform",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "end",

"expression": "value.toNumber()",

"onError": "keep-original",

"repeat": false,

"repeatCount": 10,

"description": "Text transform on cells in column end using expression value.toNumber()"

},

{

"op": "core/text-transform",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "peak\_end",

"expression": "value.toNumber()",

"onError": "keep-original",

"repeat": false,

"repeatCount": 10,

"description": "Text transform on cells in column peak\_end using expression value.toNumber()"

},

{

"op": "core/column-addition",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"baseColumnName": "peak\_end",

"expression": "grel:value-cells['end'].value",

"onError": "set-to-blank",

"newColumnName": "diff\_peak\_ends",

"columnInsertIndex": 37,

"description": "Create column diff\_peak\_ends at index 37 based on column peak\_end using expression grel:value-cells['end'].value"

},

{

"op": "core/text-transform",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "peak\_start",

"expression": "value.toNumber()",

"onError": "keep-original",

"repeat": false,

"repeatCount": 10,

"description": "Text transform on cells in column peak\_start using expression value.toNumber()"

},

{

"op": "core/text-transform",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"columnName": "start.x",

"expression": "value.toNumber()",

"onError": "keep-original",

"repeat": false,

"repeatCount": 10,

"description": "Text transform on cells in column start.x using expression value.toNumber()"

},

{

"op": "core/column-addition",

"engineConfig": {

"facets": [],

"mode": "row-based"

},

"baseColumnName": "start.x",

"expression": "grel:value-cells['peak\_start'].value",

"onError": "set-to-blank",

"newColumnName": "diff\_peak\_starts",

"columnInsertIndex": 2,

"description": "Create column diff\_peak\_starts at index 2 based on column start.x using expression grel:value-cells['peak\_start'].value"

},

{

"op": "core/column-removal",

"columnName": "diff\_peak\_starts",

"description": "Remove column diff\_peak\_starts"

},

{

"op": "core/column-removal",

"columnName": "diff\_peak\_ends",

"description": "Remove column diff\_peak\_ends"

},

{

"op": "core/column-removal",

"columnName": "peak\_start",

"description": "Remove column peak\_start"

},

{

"op": "core/column-removal",

"columnName": "peak\_end",

"description": "Remove column peak\_end"

},

{

"op": "core/column-removal",

"columnName": "start.y",

"description": "Remove column start.y"

},

{

"op": "core/column-removal",

"columnName": "stop",

"description": "Remove column stop"

},

{

"op": "core/column-rename",

"oldColumnName": "start.x",

"newColumnName": "peak\_start",

"description": "Rename column start.x to peak\_start"

},

{

"op": "core/column-rename",

"oldColumnName": "end",

"newColumnName": "peak\_end",

"description": "Rename column end to peak\_end"

},

{

"op": "core/column-rename",

"oldColumnName": "width",

"newColumnName": "peak\_width",

"description": "Rename column width to peak\_width"

},

{

"op": "core/column-rename",

"oldColumnName": "signal\_value",

"newColumnName": "peak\_signal\_value",

"description": "Rename column signal\_value to peak\_signal\_value"

},

{

"op": "core/column-rename",

"oldColumnName": "pValue",

"newColumnName": "peak\_pValue",

"description": "Rename column pValue to peak\_pValue"

},

{

"op": "core/column-rename",

"oldColumnName": "qValue",

"newColumnName": "peak\_qValue",

"description": "Rename column qValue to peak\_qValue"

},

{

"op": "core/column-rename",

"oldColumnName": "strand.y",

"newColumnName": "strand",

"description": "Rename column strand.y to strand"

},

{

"op": "core/column-rename",

"oldColumnName": "chr",

"newColumnName": "FIMO\_chr",

"description": "Rename column chr to FIMO\_chr"

},

{

"op": "core/column-rename",

"oldColumnName": "chromosome",

"newColumnName": "JASPAR\_chr",

"description": "Rename column chromosome to JASPAR\_chr"

}

]