Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/At1_Genome/Analysis/2023-01-23.Methylation

PDF Version generated by

Katarina Stuart (z5188231@ad.unsw.edu.au)

on

Aug 23, 2023 @09:46 AM NZST

Table of Contents

2023-01-23.Methylation

Methylome and 5mc base calling

https://github.com/No1RoaldFan/sbai200_Thesis_2022

Paper's

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02510-z

https://www.nature.com/articles/s41467-021-23778-6

https://github.com/comprna/METEORE#pipeline

https://www.biorxiv.org/content/10.1101/2022.12.22.521577v1.full (methylation section)

Probably what I will use: https://nanopolish.readthedocs.io/en/latest/quickstart_call_methylation.html

do correlation but with the different runs to check?

amount along genome

variability along genome

/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/methylation/

#1.CALLS DATASET

#Example code for methylation calling

```
#!/bin/bash -e
#SBATCH --job-name=2023_01_30.methylation_basecalling.sl
#SBATCH --time=24:00:00
#SBATCH --partition=gpu
#SBATCH --mem=20G
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --account=uoa02613
#SBATCH --gpus-per-node=A100:1

module load ont-guppy-gpu/6.2.1

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta

guppy_basecaller --config /opt/nesi/CS400_centos7_bdw/ont-guppy-gpu/6.2.1/data/dna_r9.4.1_450bps_modbases_5mc_cg_sup.cfg --device auto --bam_out --recursive --compress_fastq --

guppy_basecaller --config /opt/nesi/CS400_centos7_bdw/ont-guppy-gpu/6.2.1/data/dna_r9.4.1_450bps_modbases_5mc_cg_sup.cfg --device auto --bam_out --recursive --compress_fastq --

guppy_basecaller --config /opt/nesi/CS400_centos7_bdw/ont-guppy-gpu/6.2.1/data/dna_r9.4.1_e8.1_modbases_5mc_cg_sup.cfg --device auto --bam_out --recursive --compress_fastq --alig

#guppy_basecaller --config /opt/nesi/CS400_centos7_bdw/ont-guppy-gpu/6.2.1/data/dna_r9.4.1_450bps_modbases_5mc_cg_sup.cfg  --device auto --bam_out --recursive --compress_fastq
```

#Used SAMools to sort and merge the resulting .bam files

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_01.methylation_merge_all.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=8
#SBATCH --profile task

module load SAMtools/1.15.1-GCC-11.3.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/merged

#generated list of bamfiles
find /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation -name "*bam" | grep "pass" >  methylation_ALL_bamfiles.txt

samtools merge -@ 8-o AcTris_vAus_ALL.sorted.bam -b methylation_ALL_bamfiles.txt
samtools index AcTris_vAus_ALL.sorted.bam

#generated list of bamfiles per flowcell type
find /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/LSK109_5mc_rd3 -name "*bam" | grep "pass" >  methylation_109_bamfiles.txt
find /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/LSK110_5mc -name "*bam" | grep "pass" >  methylation_110_bamfiles.txt
find /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/LSK112_5mc -name "*bam" | grep "pass" >  methylation_112_bamfiles.txt

for i in 109 110 112
do
echo processing $i
samtools merge -@ 8-o AcTris_vAus_${i}.sorted.bam -b methylation_${i}_bamfiles.txt
samtools index AcTris_vAus_${i}.sorted.bam
done
```

#2. Generating BedMethyl output using modbam2bed

#ONT tool to generate methylBed format outputs

From Mcclintok document:

```
# To activate this environment, use
#     $ conda activate /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba
# To deactivate an active environment, use
#     $ conda deactivate
```

```
#installed via conda (but not by me)
#epi2me github (https://github.com/epi2me-labs/modbam2bed) gives this as their conda install instructions.
mamba create -n modbam2bed -c bioconda -c conda-forge -c epi2melabs modbam2bed
```

```
To activate this environment, use
    $ mamba activate modbam2bed
To deactivate an active environment, use
    $ mamba deactivate
```

#Running with --aggregate flag produces a second output file in which F and R strand counts for the same CpG feature have been combined. I have used t

#including a -p flag will supply name to aggregate file (not done here).

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_03.methylation_bed_all.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/merged
module load Miniconda3
CONDA_BASE=$(conda info --base)
source ${CONDA_BASE}/etc/profile.d/conda.sh
conda activate /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba
#mamba init
conda activate modbam2bed



#interactive run
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/merged
mamba init
mamba activate modbam2bed

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta

modbam2bed -e -m 5mC -t 2 --aggregate --cpg $GENOME AcTris_vAus_ALL.sorted.bam  > AcTris_vAus_ALL_modbam2bed.out

for i in 109 110 112
do
modbam2bed -e -m 5mC -t 2 --aggregate --cpg $GENOME AcTris_vAus_${i}.sorted.bam  > AcTris_vAus_${i}_modbam2bed.out
done
```

Comparing the individual runs: using binning

bin the genome

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/merged
ln -s /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta .
module load SAMtools/1.15.1-GCC-11.3.0
samtools faidx AcTris_vAus2.0.fasta
cut -f1,2 AcTris_vAus2.0.fasta.fai > sizes.genome
```

bin the mehylation bed files

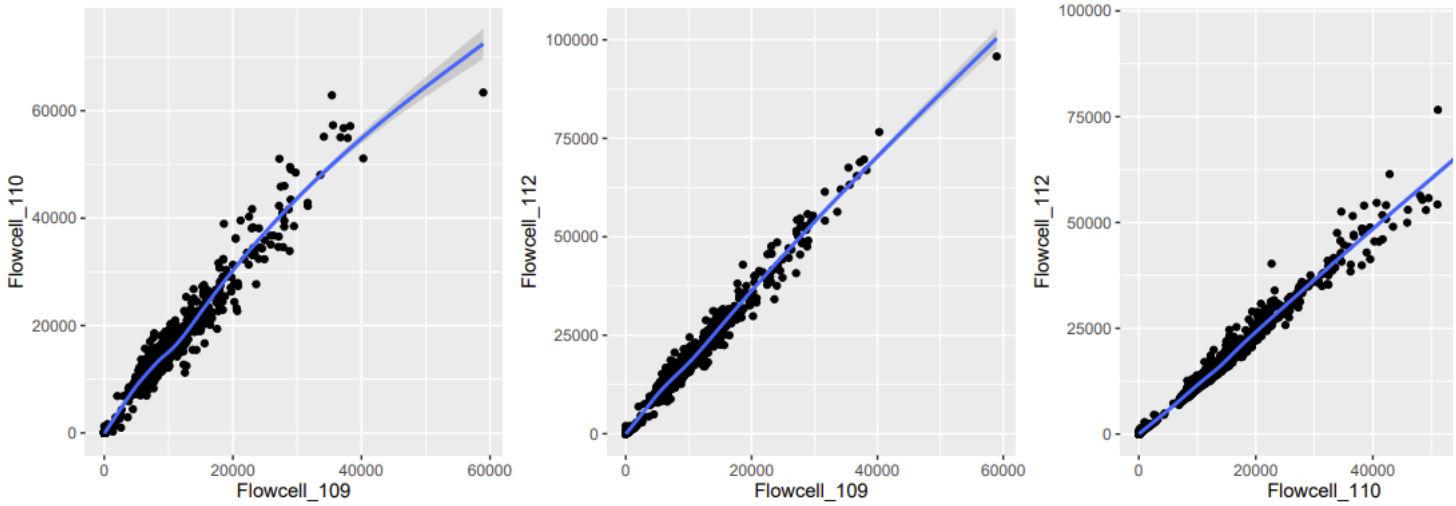
```
module load BEDTools/2.30.0-GCC-11.3.0
WIDTH=1000000
bedtools makewindows -g sizes.genome -w ${WIDTH} > AcTris_vAus2.0_${WIDTH}bps.bed
for i in ALL 109 110 112
do
bedtools coverage -a AcTris_vAus2.0_${WIDTH}bps.bed -b AcTris_vAus_${i}_modbam2bed.out -counts >  AcTris_vAus_${i}_${WIDTH}bps.txt
done
```

Combine binneed columns

```
paste -d' ' AcTris_vAus_ALL_1000000bps.txt AcTris_vAus_109_1000000bps.txt AcTris_vAus_110_1000000bps.txt AcTris_vAus_112_1000000bps.txt | sed 's/ /\t/g' | awk '{print $1"_"$2,$4,$8
```

plot

```
module load R/4.1.0-gimkl-2020a
R
library(ggplot2)
library(gridExtra)
library(grid)
library(lattice)


setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/merged")
file1<-read.table("AcTris_vAus2.0_1000000.5col.txt", sep=" ")

colnames(file1) <- c("POS","Meged_Runs", "Flowcell_109", "Flowcell_110", "Flowcell_112")

A1 <- ggplot(file1 , aes(x = Flowcell_109, Flowcell_110)) +
  geom_point() +
  stat_smooth()

B1 <- ggplot(file1 , aes(x = Flowcell_109, Flowcell_112)) +
  geom_point() +
  stat_smooth()

C1 <- ggplot(file1 , aes(x = Flowcell_110, Flowcell_112)) +
  geom_point() +
  stat_smooth()

pdf("At1_methylation_regression.pdf", width=12, height=4)
grid.arrange(A1, B1, C1, ncol=3)
dev.off()
```



Clutering methylation sites

https://github.com/jsh58/DMRfinder

download

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
git clone https://github.com/jsh58/DMRfinder.git
```

find methylation regions

```
#!/bin/bash -e
```

```
#SBATCH --job-name=2023_03_22.methylation_regions.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=20GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/merged

DIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/DMRfinder

wc -l AcTris_vAus_ALL_modbam2bed.out #18501863 is the total number of cpg sites

#filter for minimum coverage of 5
awk '$10>4' AcTris_vAus_ALL_modbam2bed.out > AcTris_vAus_ALL_modbam2bed.out_coverage5

#chrom | CpG pos | 5mc | 5mc | modified | unmodified count
#input for below
#hashed out because -r does the work below
#awk -v OFS="\t" '{print $1,$2,$4,$4,$13,$12}' AcTris_vAus_ALL_modbam2bed.out > AcTris_vAus_ALL_modbam2bed.out.6col
#awk -v OFS="\t" '{print $1,$2,$4,$4,$13,$12}' AcTris_vAus_ALL_modbam2bed.out_coverage5 > AcTris_vAus_ALL_modbam2bed.out_coverage5.6col

python ${DIR}/combine_CpG_sites.py -r 5 -c 15 -v -o combined.bed AcTris_vAus_ALL_modbam2bed.out.6col

#create column with density of CPG for each region (to account for the fact that larger regions will have more room for CPGs
tail -n +2 combined.bed | awk -v OFS="\t" '{print $0, ($6/$5)}' > methylation_regions_density.bed #wc -l 175595
```

Columns are (from manual):

chrom chromosome name

chromStart 1-based position of the cytosine in the CpG

chromEnd end of region

CpG : total count of CpG sites in window

unmethylated: count of total reads (across the CpG sites)

methylated: count of methylated reads (across the CpG sites)

Some quick calulations for paper: (discarded)

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/merged

module load R/4.1.0-gimkl-2020a
R
library("dplyr")
setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/merged")
txt <- read.table("methylation_regions_density.bed", header = F)
macro<-c("Superscaffold_chr1","Superscaffold_chr1A", "Superscaffold_chr2", "Superscaffold_chr3", "Superscaffold_chr4", "Superscaffold_chr5a", "Superscaffold_chr5b" , "Superscaffold_ch
sex <- c("Superscaffold_chrZ")
micro<-
c("Superscaffold_chr4A","Superscaffold_chr6","Superscaffold_chr7","Superscaffold_chr8","Superscaffold_chr9","Superscaffold_chr10","Superscaffold_chr11","Superscaffold_chr12","Supers

txt_macro<- filter(txt, V1 %in% macro)

txt_micro <- filter(txt, V1 %in% micro)
txt_sex <- filter(txt, V1 %in% sex)

#cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra/braker_gemoma_combined_katmanual_brakerforce_renamed.gtf
#grep -v "^#" braker_gemoma_combined_katmanual_brakerforce_renamed.gtf | cut -f1-5 | grep "gene" > braker_gemoma_combined_katmanual_brakerforce_renamed.geneonly.txt

gene <- read.table("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra/braker_gemoma_combined_katmanual_brakerforce_renamed.geneonly.txt", header = F
gene_macro<- filter(gene, V1 %in% macro)
gene_micro <- filter(gene, V1 %in% micro)
```

Additional quick calculations for the new version

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/methylation/merged/genes

module load BEDTools/2.30.0-GCC-11.3.0
GFF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/annotation/tsebra/braker_gemoma_combined_katmanual_brakerforce_renamed.gtf

#grab genes. note the strand
awk '$3=="gene"' $GFF | awk -v FS='\t' -v OFS='\t' '{print $1, $4, $5, $7}' > genes.bed
#create 5th column that is upstream of + or - located genes
awk '{ if ($4 == "+") $5 = $2 - 5000 ":"$2""; else if ($4 == "-") $5 = ""$3":" $3 + 5000; print $0 }' genes.bed  > genes_5kb_upstream.txt
#creating bed file of 5kb regions upstream of gene
grep -v "-" genes_5kb_upstream.txt | cut -d' ' -f 1,5 | sed -e 's/:\| /\t/g' > genes_5kb_upstream.bed

#calculate overlap: GENE
bedtools intersect -wb -b genes_5kb_upstream.bed -a ../methylation_regions_density.bed  > gene_overlap.txt
datamash mean 7 < gene_overlap.txt #0.35831809454435

#calculate overlap: genome wide
datamash mean 7< ../methylation_regions_density.bed #0.48109839837325

#calculate overlap: TEs
TE=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/repeats/earlgrey/AcTris_vAus2.0_earlgrey/acridotheresTristis_EarlGrey/acridotheresTristis_summaryFiles/acridot
cut -f -3 $TE | awk '{ if ($3 < $2) { t = $2; $2 = $3; $3 = t } print }' | sed 's/ /\t/g' > tes.bed
bedtools intersect -wb -b tes.bed -a ../methylation_regions_density.bed  > TE_overlap.txt
datamash mean 7 < TE_overlap.txt #0.62975655540597
```