Epigenomics, with focus on the methylome

Mehran Karimzadeh

June 21st, 2022

Outline

- Introduction to WGBS
- Introduction to submitting jobs in batch on Slurm
- Reviewing epigenomics
- Methylome workshop

Whole-genome bisulfite sequencing (WGBS)

- How WGBS is different than other types of DNA sequencing?
- What are the pros of WGBS?
- What are the cons of WGBS?
- What type of information can you get from WGBS?

Excercise

What do you think is the order of the following statements?

- Aligning with bwa-meth
- Copy fastq files from a shared folder
- Download the reference genome
- Extracting CpG methylation estimates
- Identifying differentially methylated regions
- Index the reference genome
- Investigating genes affected by differential methylation
- Trimming fastq files with trim galore
- Visualizing methylation at CpG islands

Introduction to high performance computing

Handling genomic datasets requires:

- Secure and efficient data storage
- Computational power for parallel processing and analysis
- Efficient use of hardware by people with access to the datasets
- How do you think a high performance computing server, such as SciNet teaching cluster, varies from your laptop?

How can we run a program on the cluster?

- Request an interactive session

```
salloc
samtools ...
```

- Write a bash script and submit using *sbatch*

```
echo -e '#!/bin/sh' > myBashScriptThatIWillSubmitToRunOnSlurm.sh
echo "samtools ..." >> myBashScriptThatIWillSubmitToRunOnSlurm.sh
sbatch -c 1 --mem=4G -t 1:00:00 myBashScriptThatIWillSubmitToRunOnSlu
```

- Assignment: write a bash script that for 100 times, will print your name and submit it to the Slurm cluster using *sbatch*

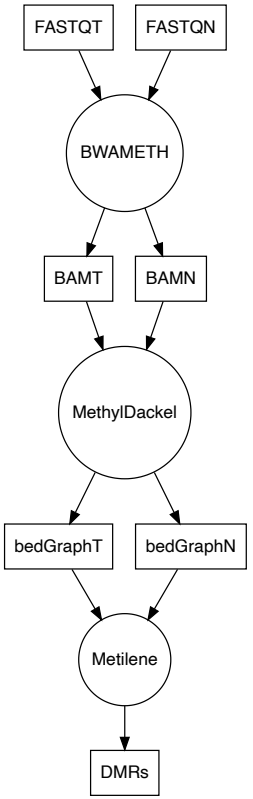
For loops in bash

```
for i in {0..99}
do
  echo "Variable i is set to $i"
done
```

What is epigenomics?

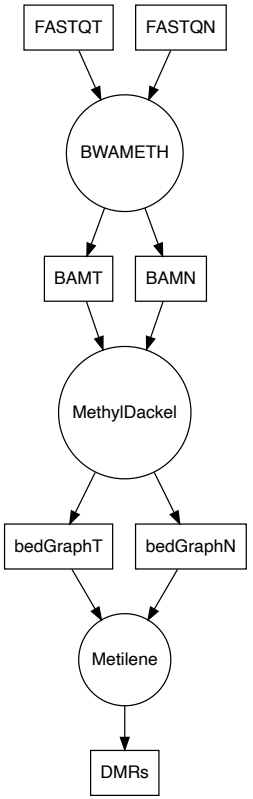
- All the cells in our body have the same genetic code, but they use it differently
- Epigenomics is the science of assessing cell type specific changes to DNA using high-throughut methods
    - Chromatin accessibility: ATAC-seq
    - Histone modifications: ChIP-seq
    - Transcription factor binding: ChIP-seq
    - Short and long range DNA interactions: Chromatin capture, HiC, Hi-ChIP, etc.
    - Chemical modifications of nucleotides: Methylation arrays, MeDIP-seq, and bisulfite sequencing

Bisulfite sequencing pipeline



Question

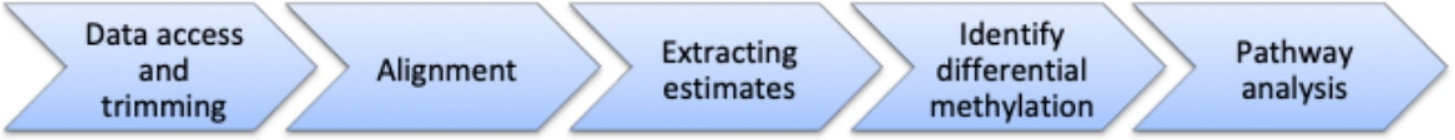
- Which important steps are missing here?



Quality control

- Each step of the process requires a quality control
- "FASTQC" to assess quality of FASTQ files before and after removing the adapters
- "samtools flagstat" to assess the percent of uniquely aligned reads
- Enrichment of differentially methylated regions in promoters, CpG islands, etc.

- Outline
  - Download the reference genome
  - Index the reference genome
  - Copy fastq files from a shared folder
  - Trimming fastq files with trim galore
  - Aligning with bwa-meth (copy the BAM files from a shared folder)
  - Extracting CpG methylation estimates
  - Visualizing methylation at CpG islands (supplementary slides)
  - Identifying differentially methylated regions
  - Investigating genes affected by differential methylation



Troubleshooting guide

Throughout this tutorial, you may see that you do not see the same output as the instructor.

Feel free to ask for help.

Some common reasons include:

- You are not logged into the teaching cluster.
    - You can type `echo $HOSTNAME` to see if it returns `teach01.scinet.local` or not
- Your session got disconnected and the environmental variables that you defined in the earlier steps are not initialized.
    - You can check if a variable is initialized by typing `echo $VARIABLENAME`. In the case of arrays, to see their elements, you can type `echo ${ARRAYNAME[@]}`.
    - If these are not initialized, nothing will be printed and that means you need to go back to the first occurence of the variable or array and re-run the command.

Indexing BWA-meth

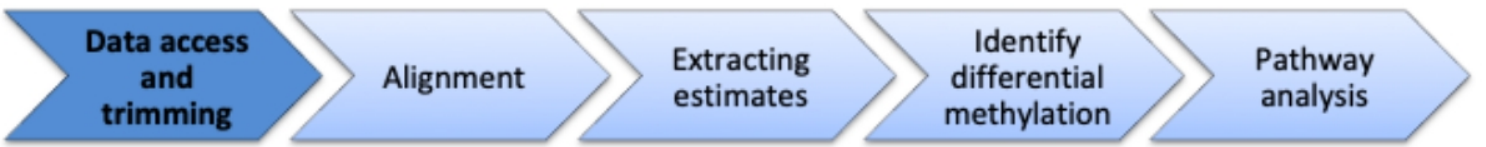
- Download FASTA file of chr22

```
mkdir -p $SCRATCH/Ref
cd $SCRATCH/Ref
wget ftp://ftp.ensembl.org/pub/release-96/fasta/homo_sapiens/dna/Homo
zcat Homo_sapiens.GRCh38.dna.chromosome.22.fa.gz | sed 's/>22/>chr22/
```

- Index FASTA file

```
salloc
module load anaconda3 gcc java fastqc cutadapt trimgalore bwa samtool
bwameth.py index Homo_sapiens.GRCh38.dna.chromosome.22.fa
```

- Expected run time is 3 minutes



Download CpG islands

- Visit http://genome.ucsc.edu
- Select the 3rd tool, *Table Browser*
- Select the correct genome assembly, and group *Regulation*
- Select the *CpG Islands* track
- Under position, type chr22
- Select the output format as *BED - browser extensible data*
- Save the file as: *hg38CpgIslandsForChr22.bed*
- Use scp to transfer the file to $SCRATCH/Datasets

```
# From the local computer
scp ~/Downloads/hg38CpgIslandsForChr22.bed username@teach.scinet.utor
```

Downloading necessary files

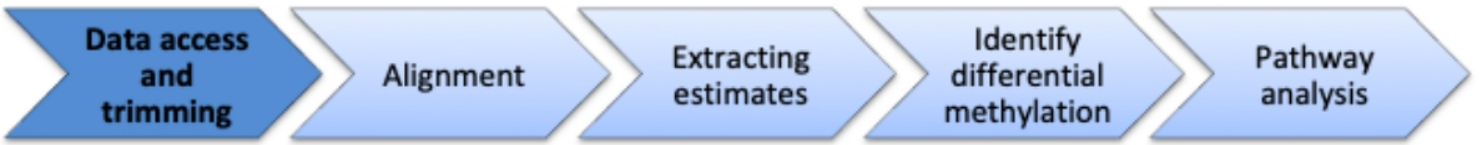
- The following directory contains all of the pipeline for chromosome 22:

```
/scratch/m/mhoffman/karimzad
```

- You can copy the fastq files from that directory to your scratch folder:

```
cp -rf /scratch/m/mhoffman/karimzad/newFastqFilesChr22 $SCRATCH
```

- This folder contains all of the files from the pipeline we process. You can copy them the same way if you have issues running the commands.



What are these samples

- H1-hESC is a human embryonic stem cell line which has been profiled extensively by the ENCODE consortium
- The left ventricle embryonic tissue is obtained from a human embryo
- By comparing these two tissues, we *may* identify which regions of chr22 must be (de)methylated for differentiating the embryonic stem cell towards a heart muscle progeny.

Trim the FASTQ files

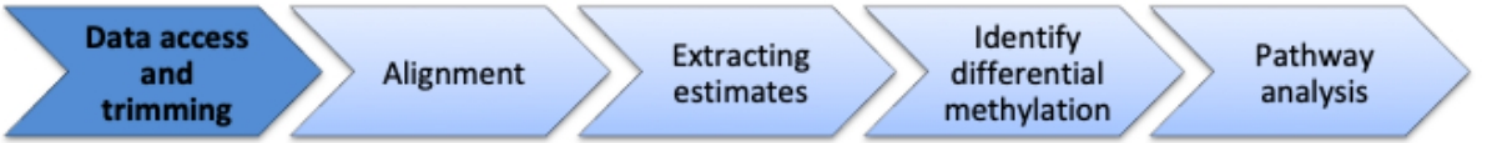
- Write a script to trim paired-end FASTQ files with trim galore in a new folder called trimmedFastqs

```
LOGDIR=$SCRATCH/Logs # Creates a new variable
SCRIPTDIR=$SCRATCH/Scripts # For scripts that run different programs
mkdir -p $SCRIPTDIR $LOGDIR # Creating multiple folders simultaneous
cd $SCRIPTDIR # Change diretory
FASTQDIR=$SCRATCH/newFastqFilesChr22 # PAth to our untrimmed fastq f
FQFOLDERS=($(ls $FASTQDIR)) # Arrays
OUTMAIN=$SCRATCH/trimmedFastqsChr22
for FQFOLDER in ${FQFOLDERS[@]}
do
  FQ1=$FASTQDIR/$FQFOLDER/$FQFOLDER\__1.fastq.gz
  FQ2=$FASTQDIR/$FQFOLDER/$FQFOLDER\__2.fastq.gz
  OUTDIR=$OUTMAIN/$FQFOLDER
  mkdir -p $OUTDIR
  echo -e '#!/bin/sh' > $SCRATCH/Scripts/$FQFOLDER\_TrimGalore.sh
  echo "module load anaconda3 gcc java fastqc cutadapt trimgalore bwa
  echo "trim_galore --fastqc --paired --gzip -o $OUTDIR $FQ1 $FQ2" >>
  sbatch -c 1 -t 1:00:00 -e $LOGDIR/$FQFOLDER\_TrimGalore.%A.ERR -o
done
```

FASTQC reports

- Use *scp* to download fastqc files to your local computer

```
STUDENTID=05
scp -r mmg3003student$STUDENTID@teach.scinet.utoronto.ca:/scratch/t/t
```



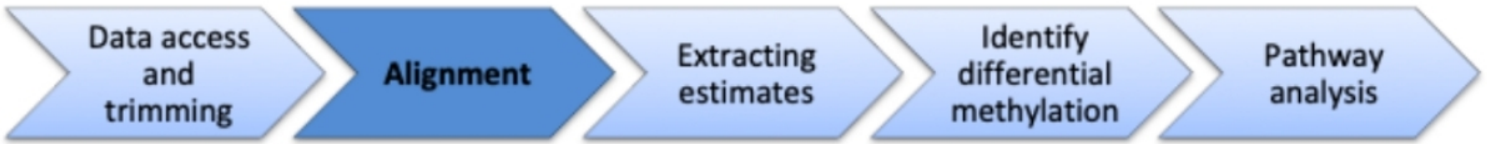
Align with BWA-Meth

- Write a script to generate _Align.sh scripts for aligning fastq files and submit them to cluster with sbatch

```
cd $SCRIPTDIR
REF=$SCRATCH/Ref/Homo_sapiens.GRCh38.dna.chromosome.22.fa
FASTQDIR=$SCRATCH/trimmedFastqsChr22
BAMDIR=$SCRATCH/trimmedAlignedBamsChr22
mkdir -p $BAMDIR
mkdir -p $SCRIPTDIR
SAMPLES=($(ls $FASTQDIR))
for SAMPLE in ${SAMPLES[@]}
do
  FQ1=$(ls $FASTQDIR/$SAMPLE | grep val_1.fq.gz)
  FQ2=$(ls $FASTQDIR/$SAMPLE | grep val_2.fq.gz)
  echo -e '#!/bin/sh' > $SCRIPTDIR/$SAMPLE\_Align.sh
  echo "module load anaconda3 gcc java fastqc cutadapt trimgalore bwa
  echo "bwameth.py --reference $REF $FASTQDIR/$SAMPLE/$FQ1 $FASTQDIR/
  # sbatch -c 1 -t 4:00:00 -e $LOGDIR/$SAMPLE\_Align.%A.ERR -o $LOGDI
done
```

Copy the aligned bam files

- It takes 4 hours of CPU time to align the FASTQ files to chr22.
- Assignment: Similar to FASTQ files, copy the folder containing bam files of chr22 to your *$SCRATCH* NOW!



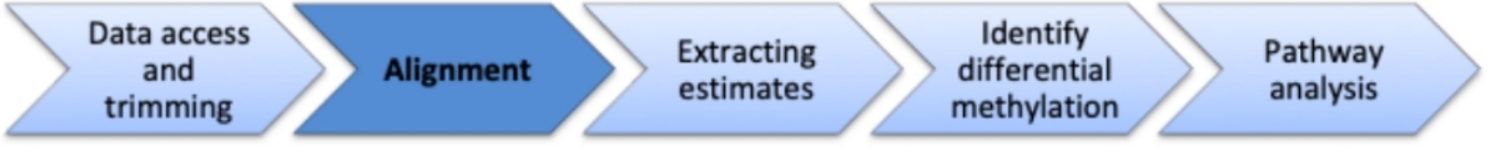
In-class assignment

- Match the following phrases to either of Methylation array, whole genome bisulfite sequencing, or ChIP-seq
  - Alignment
  - Bisulfite treatment
  - Mutations modifying the epigenome
  - EWAS
  - Fluorescence imaging
  - Anti-body cross-reactivity
  - Sex-specific batch effects
  - Sensitive to O_3 levels

Sort and index bam files

- MethylDackel requires sorted and indexed bam files
- Write a script to sort and index each bam file

```
cd $SCRIPTDIR
BAMDIR=$SCRATCH/trimmedAlignedBamsChr22
BAMFILES=($(ls $BAMDIR | grep .bam | grep -v bam.bai | grep -v sorted
for BAMFILE in ${BAMFILES[@]}
do
  SAMPLENAME=$(echo $BAMFILE | sed 's/.bam//')
  echo -e '#!/bin/sh' > $SCRATCH/Scripts/$SAMPLENAME\_sortAndIndex.sh
  echo "module load anaconda3 gcc java fastqc cutadapt trimgalore bwa
  echo "samtools sort $BAMDIR/$BAMFILE -o $BAMDIR/$SAMPLENAME\_sorted
  echo "samtools index $BAMDIR/$SAMPLENAME\_sorted.bam" >> $SCRATCH/$
  sbatch -c 1 -t 1:00:00 -e $LOGDIR/sortIndex.%A.ERR -o $LOGDIR/sortI
done
```



Run MethylDackel

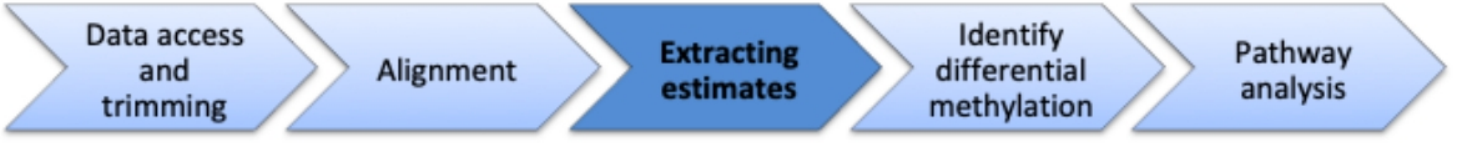
- MethylDackel uses BAM files to extract cytosine methylation counts
- Run a script to run MethylDackel files for each BAM file

```
cd $SCRIPTDIR
BAMDIR=$SCRATCH/trimmedAlignedBamsChr22
OUTMAIN=$SCRATCH/methylDackelOutputChr22
BAMFILES=($(ls $BAMDIR | grep sorted | grep -v bai | grep bam))
REF=$SCRATCH/Ref/Homo_sapiens.GRCh38.dna.chromosome.22.fa
for BAMFILE in ${BAMFILES[@]}
do
  SAMPLENAME=$(echo $BAMFILE | sed 's/_sorted.bam//')
  OUTDIR=$OUTMAIN/$SAMPLENAME
  mkdir -p $OUTDIR
  echo -e '#!/bin/sh' > $SCRIPTDIR/MethylDackel_$SAMPLENAME.sh
  echo "module load anaconda3 gcc java fastqc cutadapt trimgalore bwa
  echo "MethylDackel extract --fraction --mergeContext $REF $BAMDIR/$
  sbatch -c 1 -t 1:00:00 -e $LOGDIR/Meth.%A.ERR -o $LOGDIR/Meth.%A.LO
done
```

Explore the output of MethylDackel

- What does each column of MethylDackel output represent?

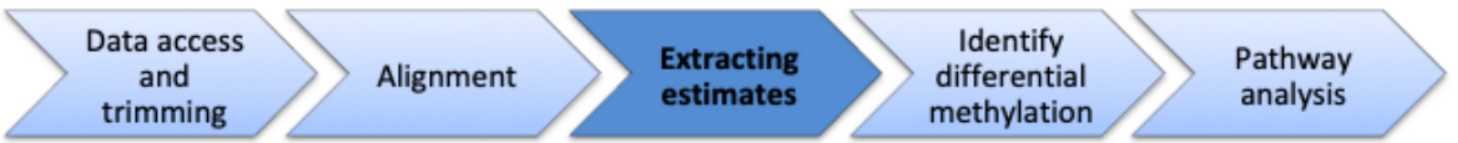
```
track type="bedGraph" description="/scratch/t/teachmmg3003/mmg3003ta
22      10513864        10513865        0
22      10513906        10513907        1
22      10515169        10515170        0
```



bedGraph is not efficient

- bedGraph is a user-readable file format
- Storing genomic signal in bedGraph format takes too much space and is computationally inefficient for random data retrieval
- bigWig format, however, can store and retrieve genomic signals efficiently
- Here we will download a program called bedGraphToBigWig and use it to convert bedGraph files

```
mkdir -p ~/software/bin
cd ~/software/bin
wget http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/bedGraphTo
# Give yourself permission to run this program
chmod u+x bedGraphToBigWig
```



Finding size of chromosomes

- bedGraphToBigWig requires a file with information of how long each chromosome is

```
cd ~/software/bin
wget http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/fetchChror
chmod u+x fetchChromSizes
./fetchChromSizes hg38 > $SCRATCH/Ref/hg38.chromsizes
```

- How else can we extract chromosome sizes from a fasta file?

Convert bedGraph to bigWig

- Write a script to convert output of MethylDackel from bedGraph to bigWig

```
salloc
MAINDIR=$SCRATCH/methylDackelOutputChr22
SAMPLES=($(ls $MAINDIR))
for SAMPLE in ${SAMPLES[@]}
do
  BDG=$(ls $MAINDIR/$SAMPLE | grep bedGraph)
  BW=$(echo $BDG | sed 's/bedGraph/bigWig/')
  ~/software/bin/bedGraphToBigWig $MAINDIR/$SAMPLE/$BDG $SCRATCH/Ref/
done
```

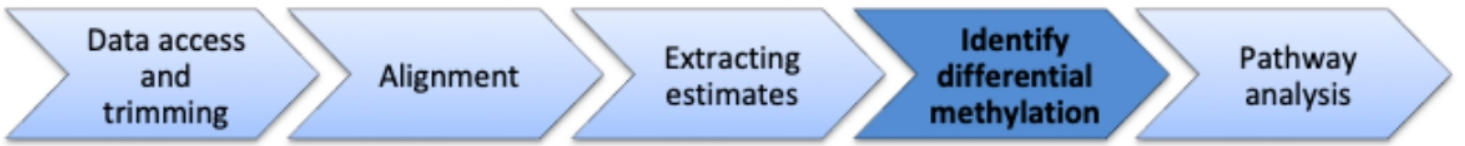
Identify differentially methylated regions

- There are various software for identifying differentially methylated regions
- Here we will use https://dx.doi.org/10.1101%2Fgr.196394.115
- Metilene requires a union file of bedGraphs we generated earlier with MethylDackel with the following columns:

```
Chrom  Start  End  G1_1  G1_2  G2_1  G2_2
```

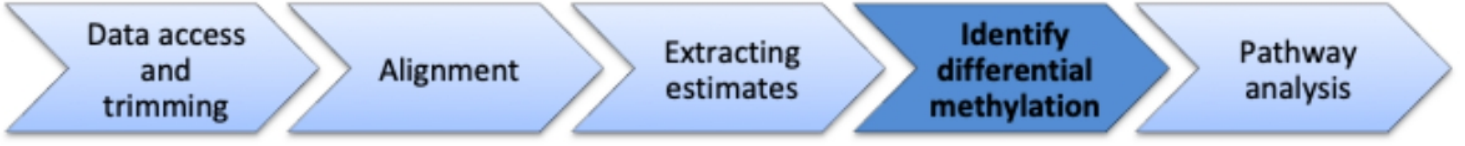
- We can generate the input file this way:

```
MAINDIR=$SCRATCH/methylDackelOutputChr22
SAMPLES=($(ls $MAINDIR))
BGS=()
HEADER=(chr start end)
for SAMPLE in ${SAMPLES[@]}
do
  HEADER+=($SAMPLE)
  BG=$(ls $MAINDIR/$SAMPLE | grep bedGraph)
  BGS+=($MAINDIR/$SAMPLE/$BG)
done
module load gcc/7.3.0 bedtools
echo -e ${HEADER[@]} | tr " " "\t" > $SCRATCH/methylDackelOutputChr22
bedtools unionbedg -i ${BGS[@]} >> $SCRATCH/methylDackelOutputChr22/r
```

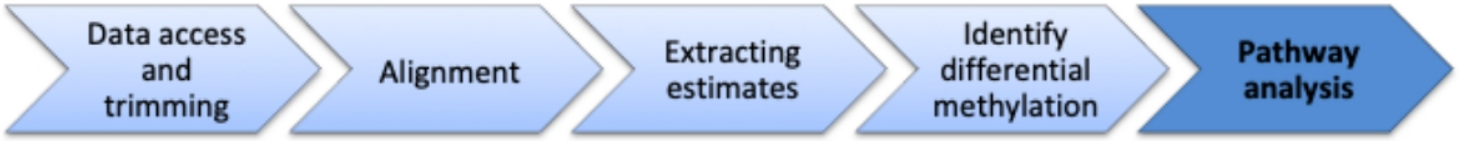


Metilene

```
module load metilene
OUTDIR=$SCRATCH/metileneOutputChr22
mkdir -p $OUTDIR
echo -e "Chrom\tStart\tEnd\tqVal\tmeanDiff\tnumCpgs\tpMWU\tp2DKS\tmea
metilene -a "H1-hESC" -b "leftVentricle" $SCRATCH/methylDackelOutputC
```



Exploring DMRs in the genome browser



In-class assignment - open question

- A recent study used DNA methylation arrays to compare white blood cells in 2,312 healthy individuals and 1,322 individuals with alzheimer's disease. They identified 5 methylation probes at the vicinity of NFE2 transcription factor with increased methylation in most Alzheimer's disease patients.
- What do you think are the key points to investigate before conclusing these methylation probes as biomarkers of alzheimer's disease?

Assignment

- In 200 words, describe how Metilene works
- For each of the following, choose the best format (BED, bedGraph, bigWig)
    - Reporting list of differentially methylated regions to a collaborator
    - Hosting a genome-wide signal track for chromatin accessibility
    - Reporting state of methylation for all CpGs in a single differentially methylated region to a collaborator

Continues on next slide:

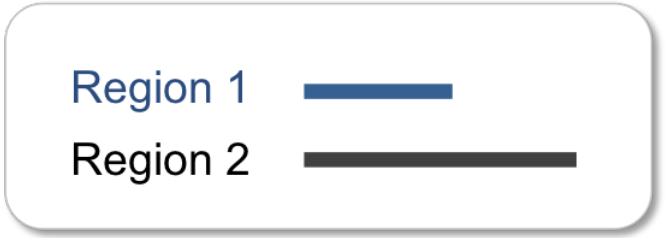
Assignment

- Repeat the same analysis for chr21.
  - Provide an annotated script which explains how you accomplished each of the steps.
  - List statistically significant differentially methylated regions
  - Which genes are likely to be affected by these changes in DNA methylation?
- Choose one of the following methods (ATAC-seq, ChIP-seq, or Hi-C) and describe how the method is performed experimentally (200 words limit).

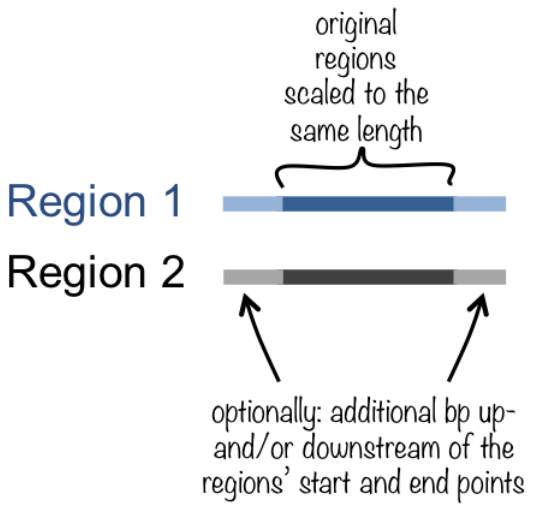
Supplementary slides

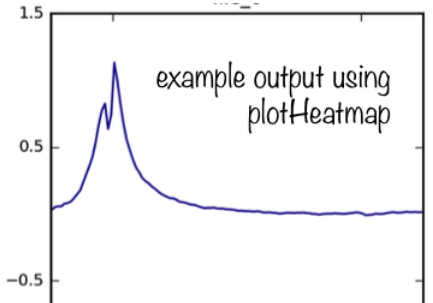
How can we explore hundreds of genomic regions for specific features, enrichments, etc.?

- DeepTools has a program called *computeMatrix*
- computeMatrix accepts signal files (e.g. in bigWig) and genomic region annotations (e.g. in BED or GTF) to calculate summary statistics
- computeMatrix has two modules:
    - reference-point: Obtains measures for entries of BED file (as reference) as well as their upstream and downstream
    - scale-regions: Calculates summary measures for BED files by shrinking each entry to a user-defined length

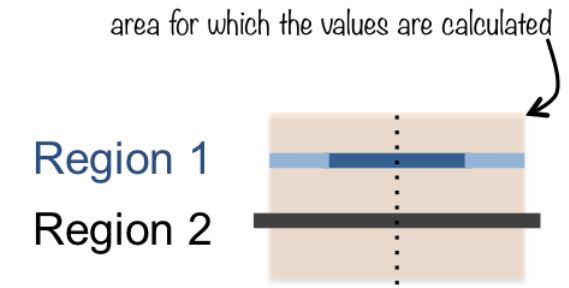


scale-regions

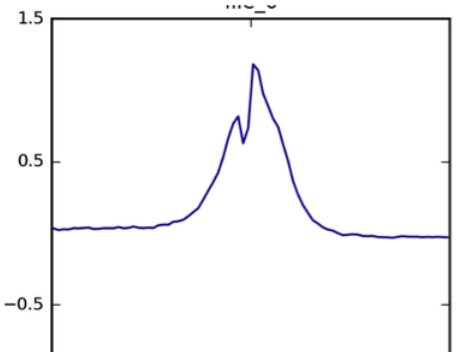




reference-point



1. regions are aligned at the selected reference point (here: center)
2. the specified numbers of bp are added up- and downstream of the reference point



How is methylation signal around CpG islands?

- Write a script to execute computeMatrix reference-point on CpG island BED file and the four bigWig files

```
MAINDIR=$SCRATCH/methylDackelOutputChr22
SAMPLES=($(ls $MAINDIR))
BWS=()
for SAMPLE in ${SAMPLES[@]}
do
  BW=$(ls $MAINDIR/$SAMPLE | grep bigWig)
  BWS+=($MAINDIR/$SAMPLE/$BW)
done
module load anaconda2/5.1.0 deeptools/3.2.1-anaconda2
OUTDIR=$SCRATCH/methylationMatricesChr22
mkdir -p $OUTDIR
computeMatrix reference-point -R $SCRATCH/Datasets/hg38CpgIslandsFor(
plotProfile -m $OUTDIR/mergedMethylationAroundIslands.tsv.gz -out $OU
plotProfile -m $OUTDIR/mergedMethylationAroundIslands.tsv.gz --perGr(
```