Bisulfite Mapping

Download Example Data

Because our tutorial is only lasting a few hours, it makes sense to use an example that wont take very long to run. So for this example we are going to use the model plant genome *Arabidopsis_thaliana*, which co-incidently was one of the first organisms to be run using whole-genome bisulfite sequencing (BS-seq/MethylC-seq) in Ryan Lister's 2008 *Cell* paper. *Arabidopsis_thaliana* is roughly 125Mb in size, making it a easy organism to map to, and there have been some great papers published which sequence the methylomes of a range of DNA methyltransferase (DNMT) mutants.

Lets firstly download some example data and convert them to FASTQ files and get some basic QC stats. For simpliticity, we will only use two samples, with no replicates. But to really assess the variation within each sample, its generally a good idea to sequence biological and/or technical replcates.

```
# Download the data off NCBI SRA
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR534/SRR534239/SRR534239.sra
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR534/SRR534177/SRR534177.sra

# Convert SRA to compressed FASTQ files
fastq-dump --gzip -Z SRR534239.sra > GSM981031_met1.fastq.gz
fastq-dump --gzip -Z SRR534177.sra > GSM980986_colWT.fastq.gz

# Run fastQC
fastqc -t 2 GSM981031_met1.fastq.gz GSM980986_colWT.fastq.gz
```

The two samples are the Columbia wild-type reference and *met1*, a mutation of the CG methyltransferase MET1 resulting in elimination of CG methylation throughout the genome.

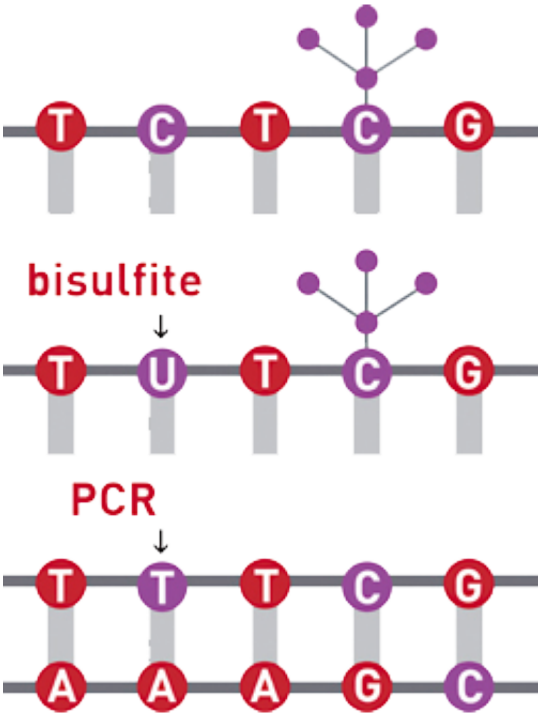
The reference genome of *Arabidopsis_thaliana* is in the /data directory, so lets format that for bisulfite sequence mapping. We need to make sure that bismark recognises that you want to use bowtie2 to do the mapping.

```
# Create directory for bismark BS-seq genome
mkdir Athal

# Move data file in there
mv data/TAIR10.fa Athal/

# Run bismark to format our Athaliana genome (TAIR10)
bismark_genome_preparation --bowtie2 Athal
```

We may need to trim quality and adapters if the fastQC report indicates that there is adapter contamination and/or poor quality at the end of sequence reads.

```
# Trim adapters and quality
AdapterRemoval --file1 GSM981031_met1.fastq.gz --basename met1_trim --trimns
--trimqualities --gzip  --minlength 25
AdapterRemoval --file1 GSM980986_colWT.fastq.gz --basename colWT_trim --
trimns --trimqualities --gzip --minlength 25
```

Because we're dealing with DNA that has been treated with Sodium bisulfite, all cytosines (C) will be converted to a Uracil (U), while 5' methyl cytosine's will be left unconverted as a cytosine (C). During amplification the converted Uracil's (U) are read as Thymine's (T)



www.diagenode.com

```
bisnark -p 2 --bam --bowtie2 Athal met1_trim.truncated.gz
bismark -p 2 --bam --bowtie2 Athal colWT_trim.truncated.gz
```

Data cleanup and processing

After we're done with mapping our data, we need to process our BAMs. We ned to do the following tasks:

1. Sort our bam file
2. Deduplicate the data to remove clonal deuplicates
3. Make methylation call files

Firstly to sort and deduplicate. One of the most widely-used Bioinformatics tools used today is samtools. And it can be frustrating to use because it lacks multi-threading options which would allow scalability to larger systems.

However now we can use sambamba which fills those needs. Lets use it to sort our bismark-made BAMs and mark and remove duplicates.

```
sambamba sort -t 2 met1.TAIR10.bam
sambamba sort -t 2 colWT.TAIR10.bam

sambamba markdup -p -r -t 2 met1.TAIR10.sorted.bam
met1.TAIR10.sorted.markdup.bam
sambamba markdup -p -r -t 2 colWT.TAIR10.sorted.bam
colWT.TAIR10.sorted.markdup.bam
```

Now we have two cleaned bam files that

Methylation calls

Now we have an alignment for each of our samples, we need to extract the respective 5mC contexts from the data. Generally most mammalian BSseq sequencing project can get away with only identifying the proportion of methylated cytosines in CpG conetxts, but with whole-genome data it is always recommended to look at the other contexts to get a wide picture of the methylation profile.

Bismark does come with methylation call extraction scripts, however I have found that the program PileOMeth is very quick and easy to use.

```
for i in *.sorted.markdup.bam
 do
    PileOMeth extract --mergeContext -o ${i%%.*}.CpG Athal/TAIR10.fasta $i
    PileOMeth extract --noCpG -CHG --mergeContext -o ${i%%.*}.CHG
Athal/TAIR10.fasta $i
    PileOMeth extract --noCpG -CHH --mergeContext -o ${i%%.*}.CHH
Athal/TAIR10.fasta $i
done
```

Now......Differential Methylation!