

Workshop 6: DNA Methylation Analysis using Bisulfite Sequencing

Fides D Lay
UCLA
QCB Fellow
lay.fides@gmail.com

Workshop 6 Outline

Day 1:

Introduction to DNA methylation & WGBS

Quick review of linux, Hoffman2 and high-throughput sequencing glossary.

Aligning WGBS reads using bwa-meth

Day 2:

DNA methylation calling using Bis-SNP

Analysis of differentially methylated regions (DMRs) using metilene

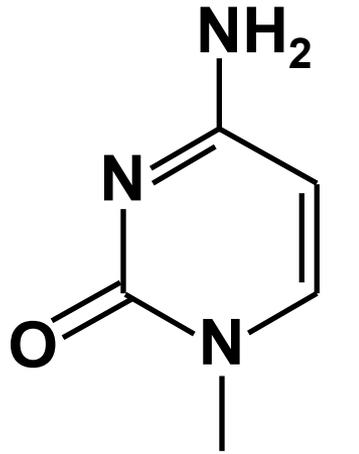
Day 3:

Visualization of DNA methylation data

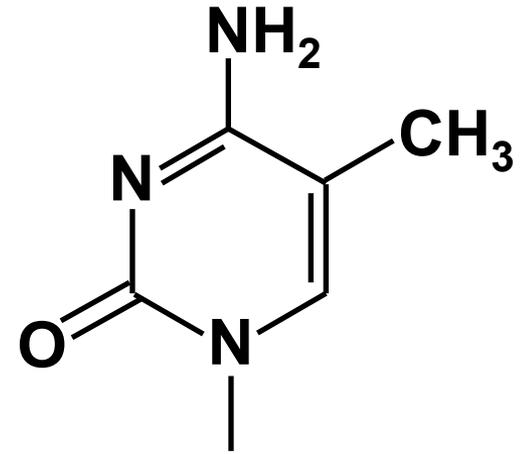
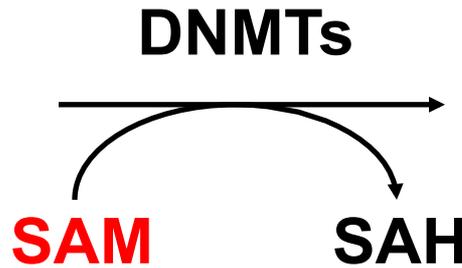
WGBS analysis using BS-Seeker2

Day 1

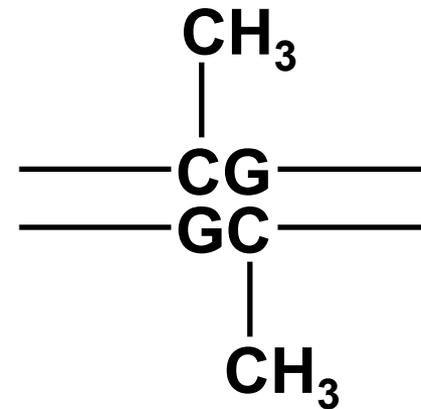
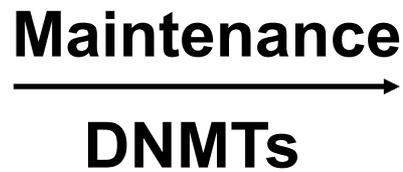
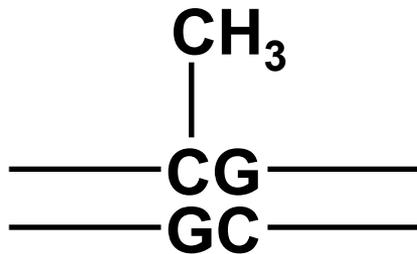
DNA Methylation



**Cytosine
(in DNA)**

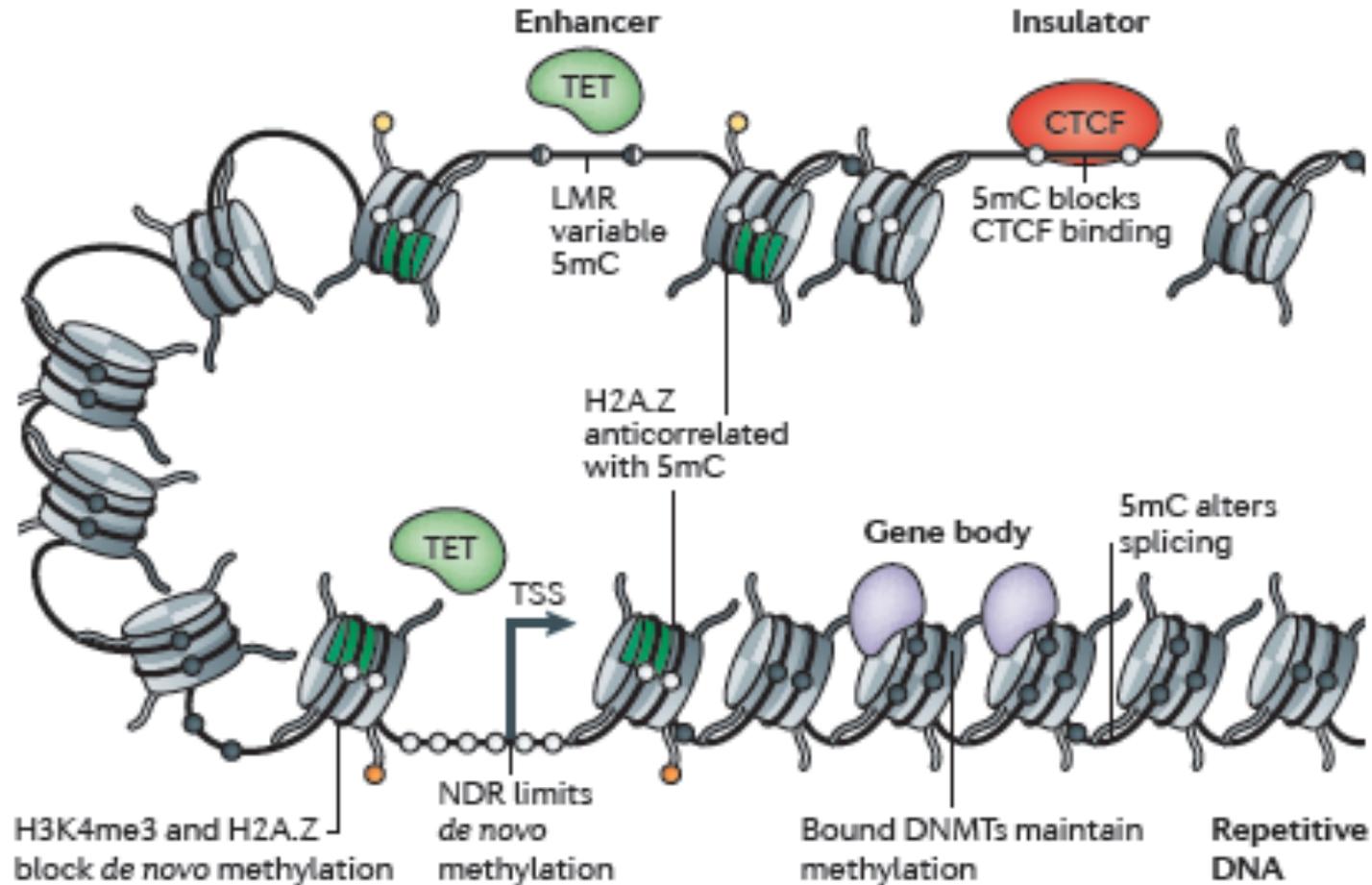


5-Methylcytosine



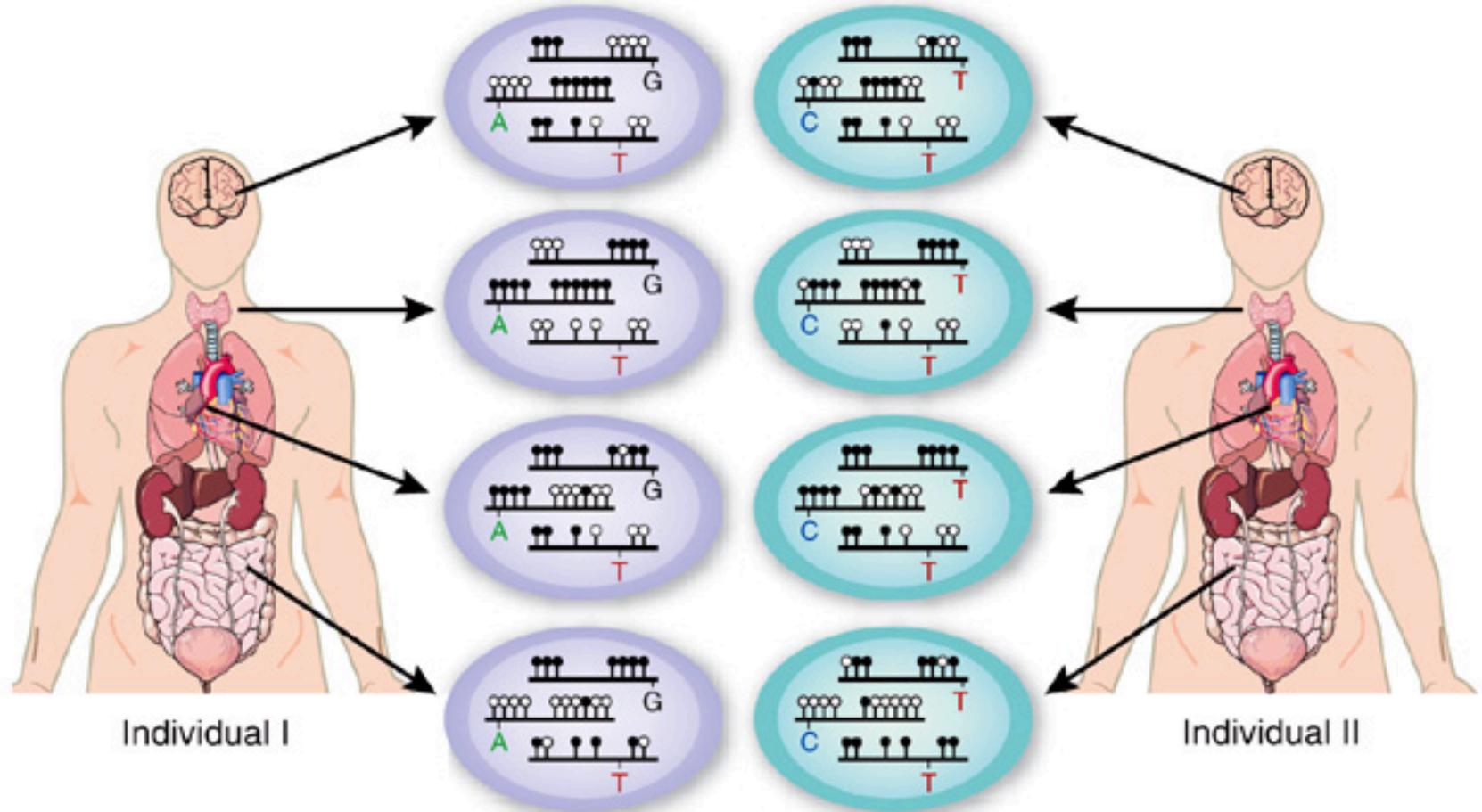
Why Study DNA Methylation?

DNA Methylation: A Component of Epigenetic Mechanisms

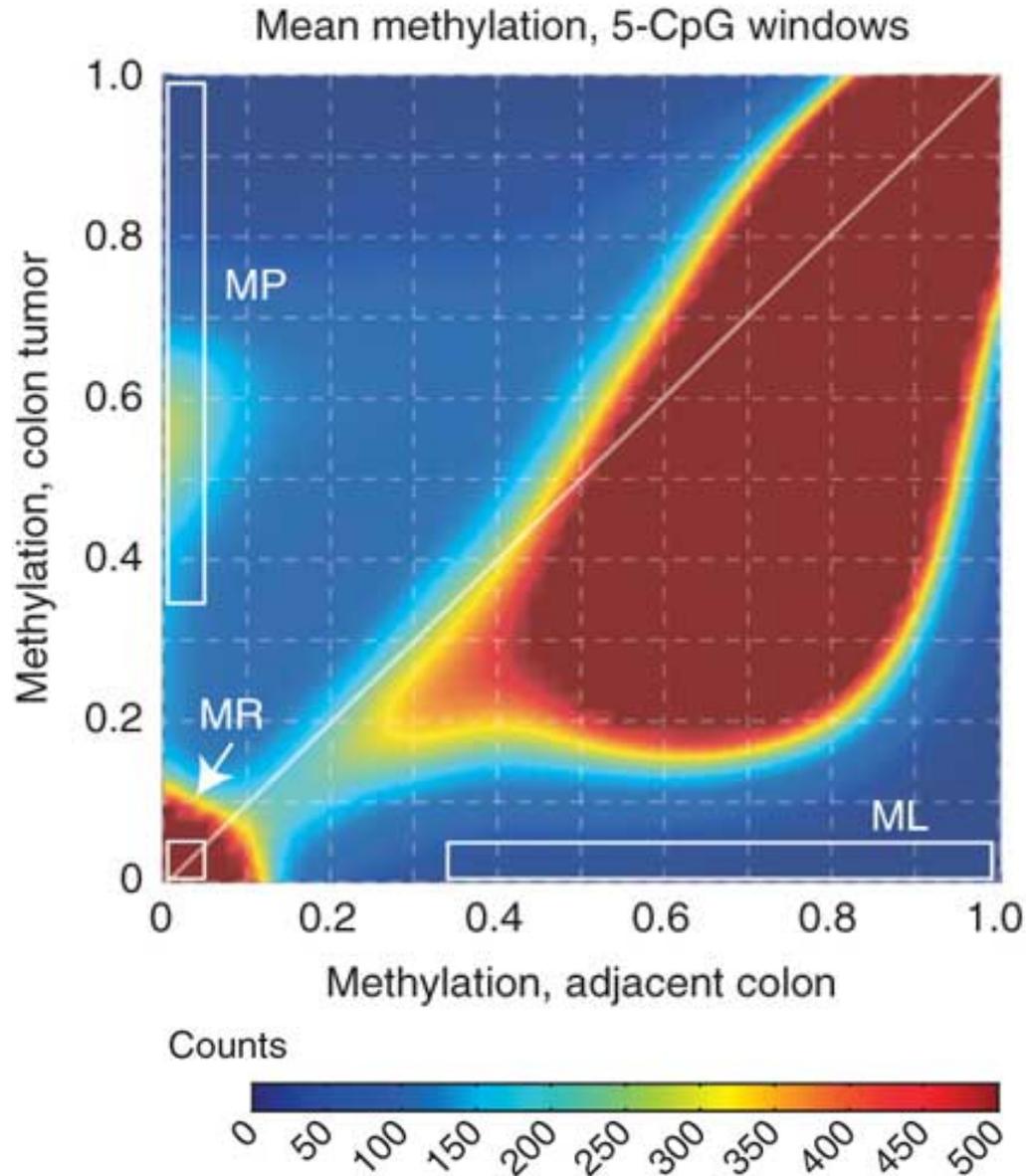


○ Unmethylated CpG	● H3K4me1	■ H2A.Z	● DNMT3A
● Methylated CpG	● H3K4me3		● DNMT3B
● Variable CpG methylation			

DNA Methylation Patterns are Unique to Each Cell Type and May Vary Across Individuals



DNA Methylation Alteration is a Feature of Cancer Cells



What are the functions of DNA Methylation?



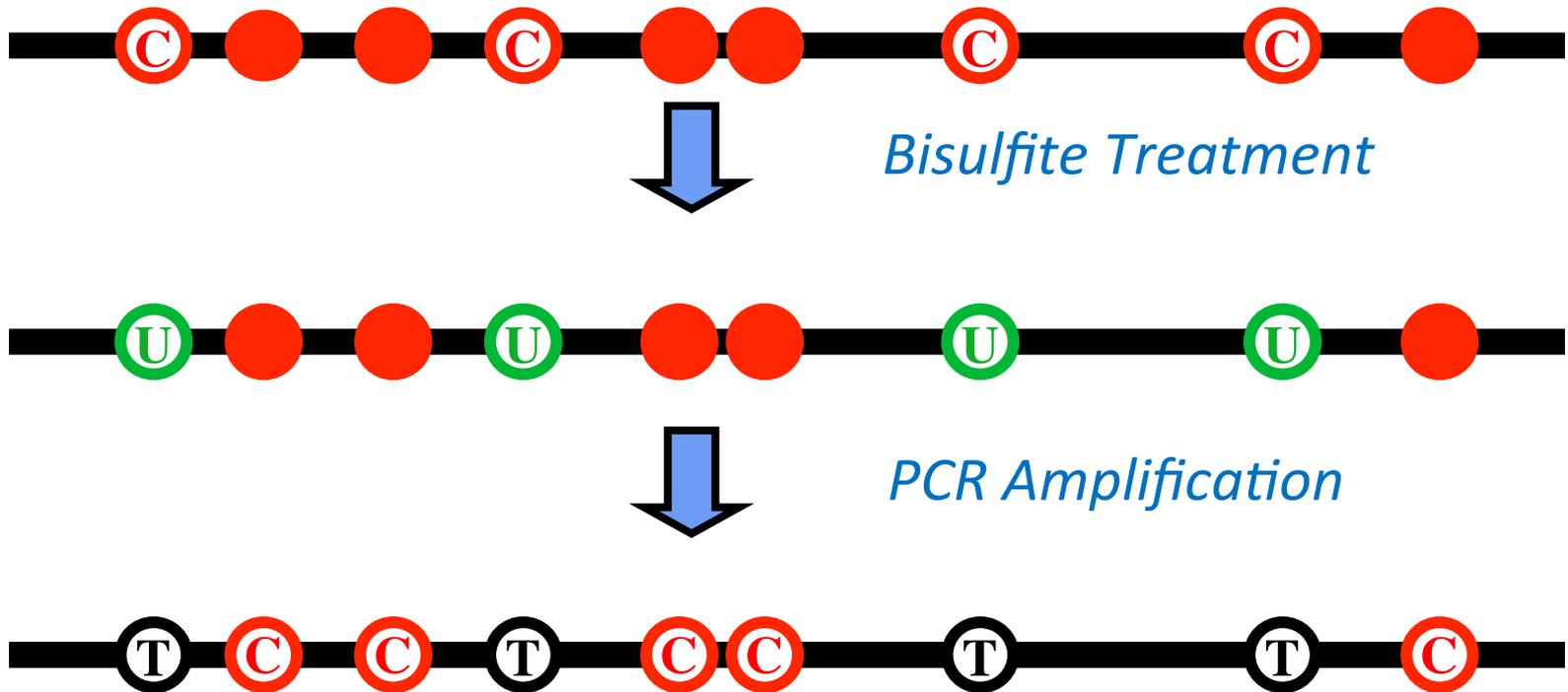
REVIEWS

Functions of DNA methylation: islands, start sites, gene bodies and beyond

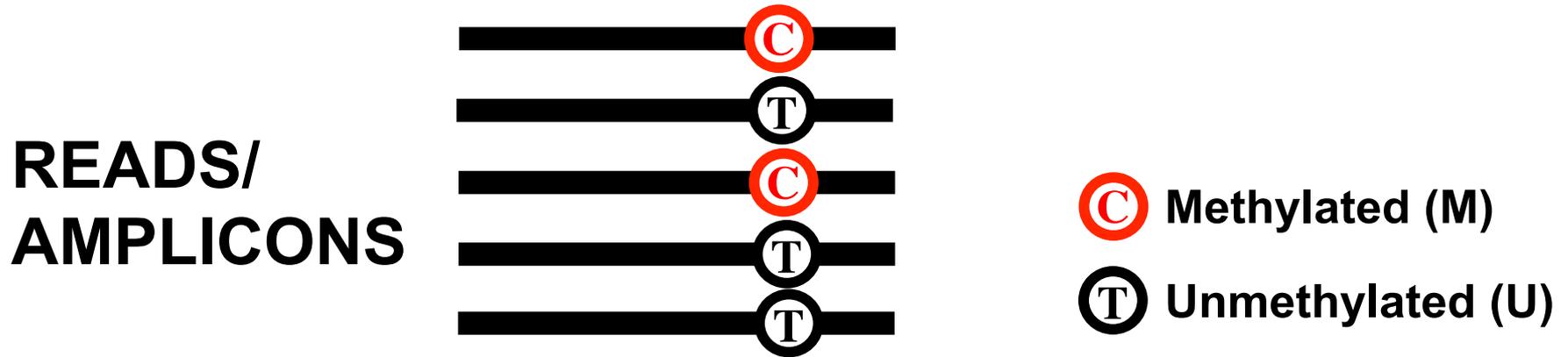
Peter A. Jones

Abstract | DNA methylation is frequently described as a 'silencing' epigenetic mark, and indeed this function of 5-methylcytosine was originally proposed in the 1970s. Now, thanks to improved genome-scale mapping of methylation, we can evaluate DNA methylation in different genomic contexts: transcriptional start sites with or without CpG islands, in gene bodies, at regulatory elements and at repeat sequences. The emerging picture is that the function of DNA methylation seems to vary with context, and the relationship between DNA methylation and transcription is more nuanced than we realized at first. Improving our understanding of the functions of DNA methylation is necessary for interpreting changes in this mark that are observed in diseases such as cancer.

Bisulfite Sequencing



Measuring DNA Methylation



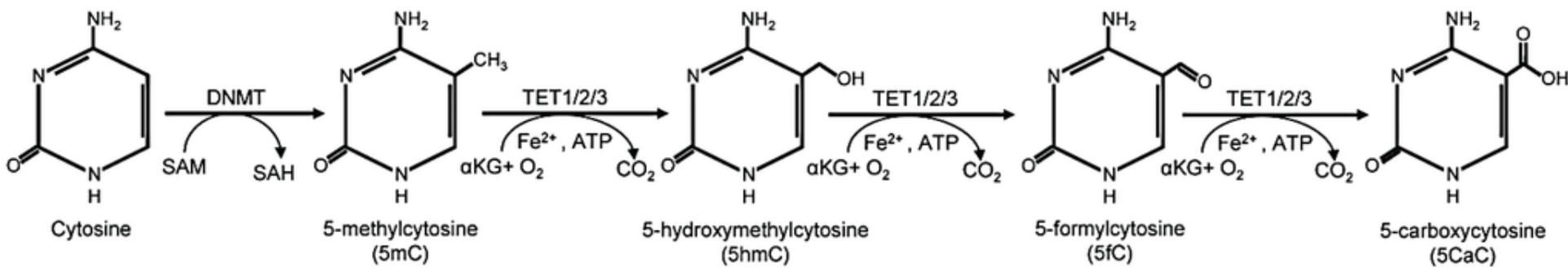
$$\% \text{Methylation} = \frac{M}{M + U} \times 100$$

In the above example, the methylation level of the locus is **40%**

Bisulfite Sequencing-Based Assays

- **Whole-genome Bisulfite Sequencing (WGBS)**
 - ✧ Unbiased (ie. whole genome)
 - ✧ Undigested template
- **Reduced Representation Bisulfite Sequencing (RRBS)**
 - ✧ Enrich for CpG islands area (CGI) which accounts for ~1% of the genome
 - ✧ Digestion with methylation insensitive restriction enzyme (ie. MspI)
- **Nucleosome Occupancy and Methylome Sequencing (NOME-seq)**
 - ✧ Dual measurement of nucleosome occupancy and DNA methylation levels
 - ✧ Treated with GpC Methyltransferase to footprint nucleosome occupancy
- **Chromatin Immunoprecipitation Bisulfite Sequencing (ChIP-BS-seq)**
 - ✧ Measurement of DNA methylation at regions enriched for specific protein or histone marks.

Oxidized Derivatives of 5-MethylCytosine



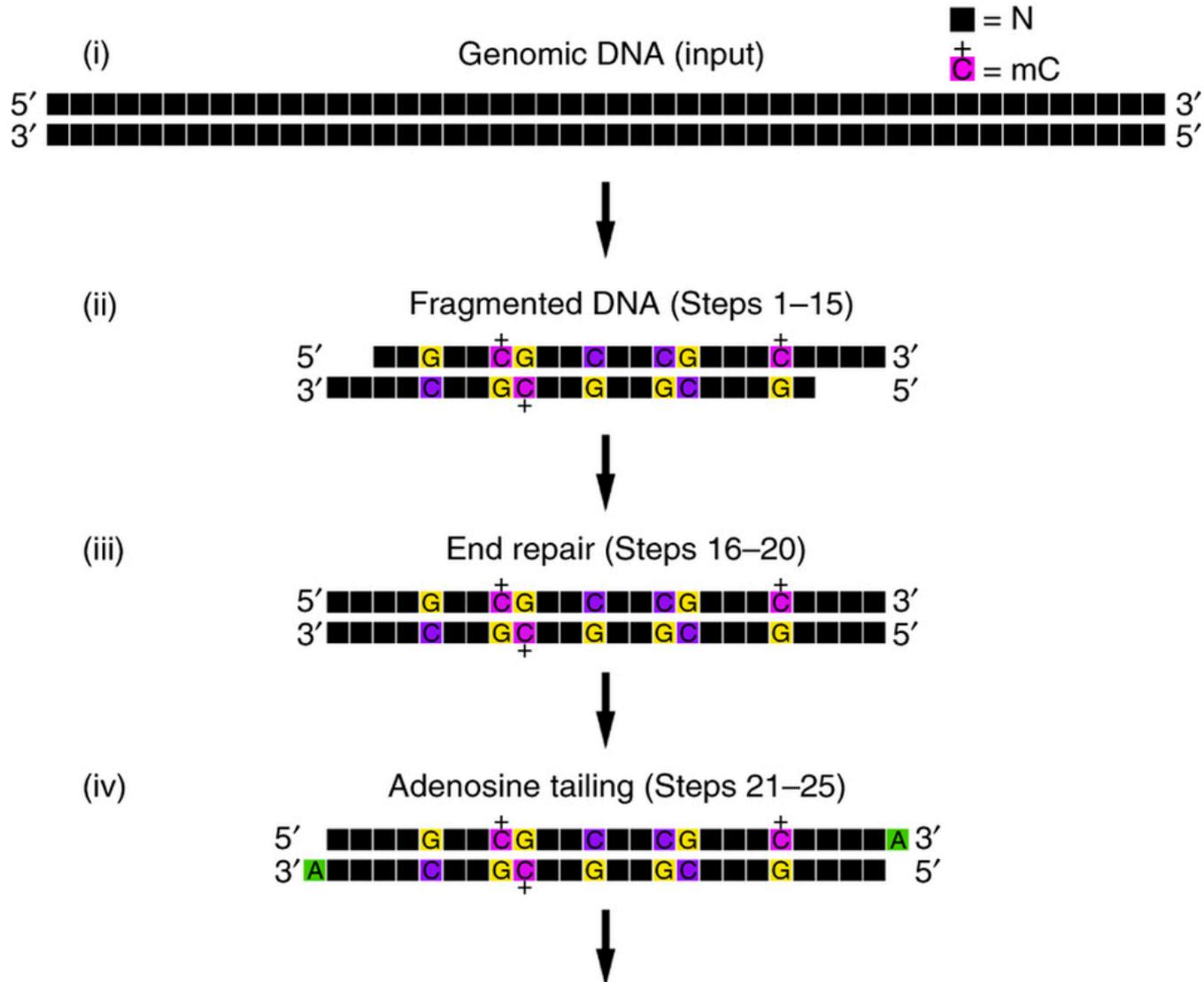
REVIEWS

Advances in the profiling of DNA modifications: cytosine methylation and beyond

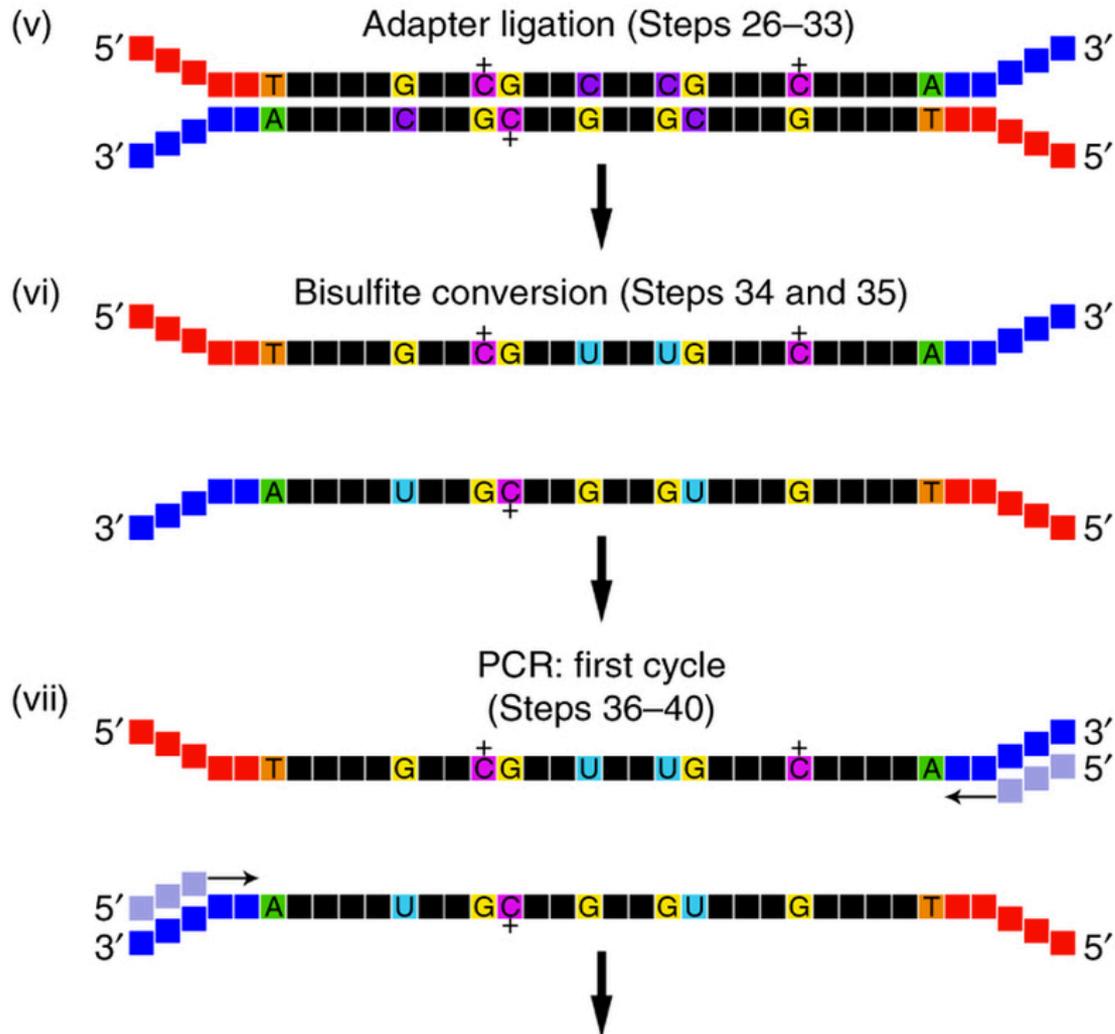
Nongluk Plongthongkum, Dinh H. Diep* and Kun Zhang*

Abstract | Chemical modifications of DNA have been recognized as key epigenetic mechanisms for maintenance of the cellular state and memory. Such DNA modifications include canonical 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC). Recent advances in detection and quantification of DNA modifications have enabled epigenetic variation to be connected to phenotypic consequences on an unprecedented scale. These methods may use chemical or enzymatic DNA treatment, may be targeted or non-targeted and may utilize array-based hybridization or sequencing. Key considerations in the choice of assay are cost, minimum sample input requirements, accuracy and throughput. This Review discusses the principles behind recently developed techniques, compares their respective strengths and limitations and provides general guidelines for selecting appropriate methods for specific experimental contexts.

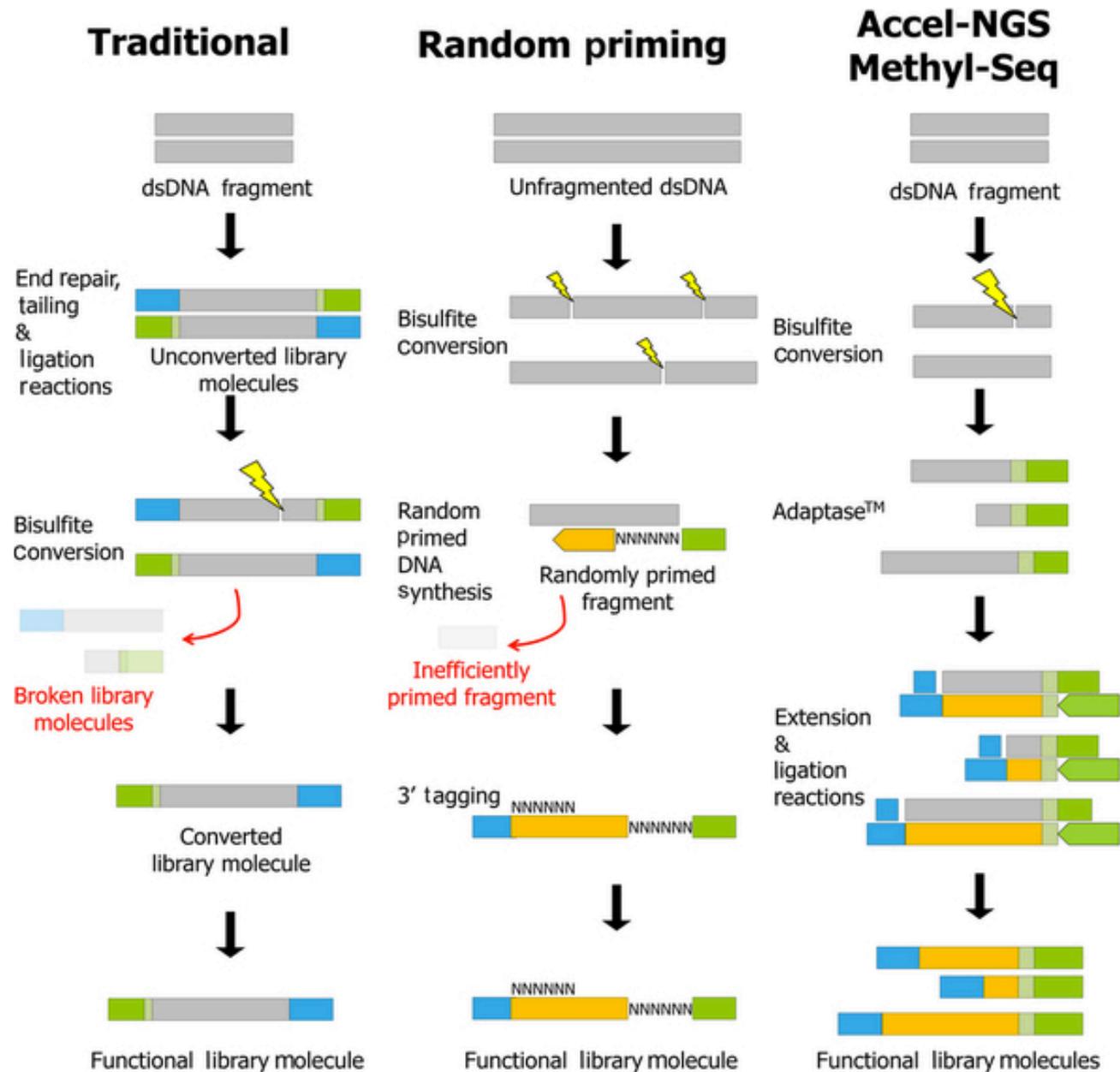
WGBS Library Preparation



WGBS Library Preparation



Post-Bisulfite Adapter Tagging



Study Design

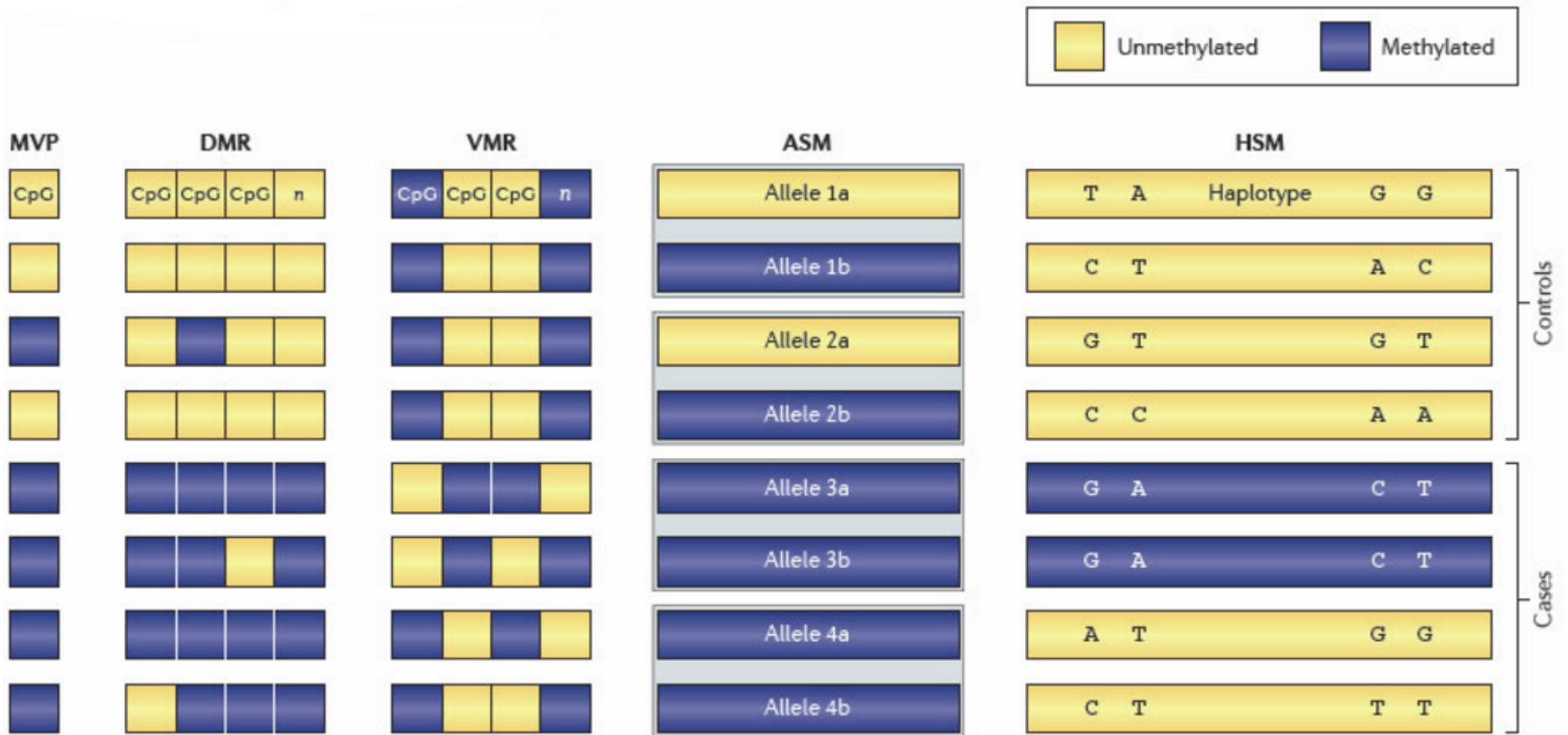
What coverage do you need?

- Generally depends on the magnitude of methylation difference to detect: the larger the difference, the lower coverage needed.
- REMC recommendation: 30x (~800M reads, 100PE)
- Data driven: 5-15x (<400M reads,

How many replicates?

- REMC: 2 biological reps
- Data driven: 3 biological reps (2 acceptable as minimal drop off in sensitivity)

Features and Variation of DNA Methylation



Nature Reviews | Genetics

Features and Variation of DNA Methylation

Methylation variable position (MVP). A CpG site that shows differential methylation — for example, between different disease states, as illustrated in the figure. Given recent findings on non-CpG methylation, potentially all Cs could be MVPs.

Differentially methylated region (DMR). A region of the genome at which multiple adjacent CpG sites show differential methylation. DMRs can occur in many different contexts, such as:

- > iDMR — imprinting-specific differentially methylated region
- > tDMR — tissue-specific differentially methylated region
- > rDMR — reprogramming-specific differentially methylated region
- > cDMR — cancer-specific differentially methylated region
- > aDMR — ageing-specific differentially methylated region.

Variably methylated region (VMR). These regions are defined by increased variability rather than gain or loss of DNAm.

Allele-specific methylation (ASM). These are positions or regions that vary in DNAm depending on the parent-of-origin, the presence of a polymorphism or as a result of a stochastic event.

Haplotype-specific methylation (HSM). This is a differentially methylated region that is defined by a set of co-inherited SNPs (a haplotype).

CpG islands (CGIs). These are regions enriched for CpG sites. Most CGIs are unmethylated in all cell types.

CGI shores. These are regions immediately adjacent to CGIs and display higher variation in DNAm than CGIs despite their lower density of CpG sites.

The figure shows different types of DNAm variation that can be identified with epigenome-wide association studies. The notation n is used to indicate the variable size of the regions shown. For the purpose of this simplified illustration, the cases and controls are assumed to have methylated or unmethylated CpG states only. Real samples will contain populations of different cells and hence display much more heterogeneous methylation levels across the full dynamic range between 0% and 100%.

Various File Format: Covered in Workshop 2

- Fastq: text-based format containing sequencing reads and quality scores

```
[flay@login2 raw]$ head K562_R1.fastq
@DDLZ38V1:356:C33YCACXX:2:1101:1168:2186 1:N:0:GCGCTA
TTTTGTGAATATTGTTATATAATAAGATTAATTTTAAGTTTATTTTTTGATTTTTTATTTTTATAGTTATGAAG
+
@@@DDDAB<D4CF<FH@G>IICI>>@@HHIC<FHEH4:9EFI4<FHIIIGBFHGGEI;CHCGH=G@=DGCE@=CA
@DDLZ38V1:356:C33YCACXX:2:1101:1359:2177 1:N:0:GCGCTA
AAATTTTGTTTTTATTA AAAAATATAAAAATTAGATGGGTTTGGTGGTAAGTGTTTGTAAATTTTAGTTATTTGGGA
+
CC@FFFFFFHHHHHJJJJJJJIIJJJJIIHIJHIHGIJJFGHII: DHFH HH IJJJIJHHHGIIIGIHC GHJIGHE
@DDLZ38V1:356:C33YCACXX:2:1101:1894:2198 1:N:0:GCGCTA
AGGTTTTTTTGAGATATTGTAATAATGATTA AAAATGTTTTTTTTTAAAGGAGGTTTATAATTGATTTATTTA
```

- SAM/BAM: binary format of aligned reads

Various File Format: Covered in Workshop 2

- VCF: Variant Call Format, text file containing meta-information lines, header lines and data lines containing genotype info at each position

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=nyImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:..
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

Various File Format: Covered in Workshop 2

- Bed: tab-delimited text format containing chrNum, chrStart, chrEnd (required), name, score, strand, etc (optional)

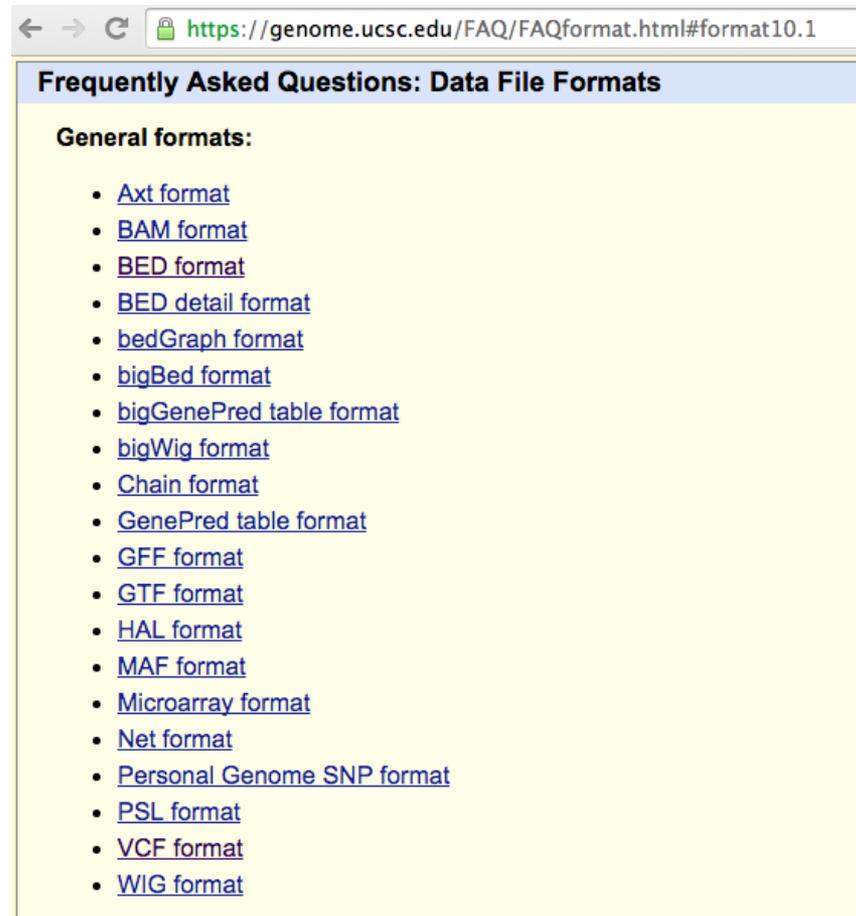
```
[flay@hpc-uec flay]$ head fdft1.bed
chr8      11666017      11666018      cg13963446      0      *
chr8      11665962      11665963      cg11598005      0      *
chr8      11665969      11665970      cg01402994      0      *
```

- Bedgraph: tab-delimited text format containing chrNum, chrStart, chrEnd and dataValue (required)

```
[flay@login2 bwameth]$ head test.bedgraph
chr1      847784  847784  0.0
chr1      1433528 1433528 100.0
chr1      1433539 1433539 100.0
chr1      1433574 1433574 100.0
chr1      1433638 1433638 100.0
chr1      1730072 1730072 100.0
chr1      1730163 1730163 100.0
chr1      1730204 1730204 100.0
chr1      1433677 1433677 100.0
chr1      3041726 3041726 0.0
```

Various File Format: Covered in Workshop 2

- Wig: a text file format defining continuous data track
- bigWig: condensed data track format for graphing and visualization



The screenshot shows a web browser window with the address bar displaying <https://genome.ucsc.edu/FAQ/FAQformat.html#format10.1>. The page title is "Frequently Asked Questions: Data File Formats". Under the heading "General formats:", there is a list of 20 file formats, each with a blue underlined link:

- [Axt format](#)
- [BAM format](#)
- [BED format](#)
- [BED detail format](#)
- [bedGraph format](#)
- [bigBed format](#)
- [bigGenePred table format](#)
- [bigWig format](#)
- [Chain format](#)
- [GenePred table format](#)
- [GFF format](#)
- [GTF format](#)
- [HAL format](#)
- [MAF format](#)
- [Microarray format](#)
- [Net format](#)
- [Personal Genome SNP format](#)
- [PSL format](#)
- [VCF format](#)
- [WIG format](#)

Useful UNIX Commands: Covered in Workshop 1 & 2

- Where am I? `pwd`
- Change directory `cd` `cd ~/data`
- Move up one level `cd ..`
- List files in folder `ls`
- Look at a file `less fileName`
- Copy a file `cp ~/data/file ~/otherdir/`
- Delete a file `rm fileName`
- Delete a directory `rmdir ~/dirName/`
- Move a file `mv ~/data/file ~/otherdir/file`
- Secure copy `scp user@host1:dir/file user@host2:dir/file`
- Compress a file `gzip -c file > file.gz`
- Uncompress a file `gunzip file.gz`
- Make a new folder `mkdir data2`
- Current directory `./`
- Home directory `~/`
- Count lines in a file `wc -l fileName`

Questions/In Doubt/Lost?

- Unix manual: man functionX
- Google is your best friend!
- Any of your friendly QCB fellows

General WGBS Workflow

Pre-alignment Processing (covered in Workshop 2&4)

- Demultiplexing
- Raw sequencing QC (FastQC)
- Adapter Trimming (TrimGalore)



Alignment

- Align FastQ files to reference genome and output .bam files
- Various tools: BS-Seeker2, BSMAP, Bismark, bwa-meth, etc



Post-alignment

- Methylation Calling (Bis-SNP, Bsmooth, etc)
- DMR identification (metilene, bisseq)
- Visualization (IGV, UCSC, IGB etc)

Publicly Available WGBS Aligners

Program	Version	Aligner type	Aligner description	Language	Alignment engine	Paired end	Color space	Non directional	Multithread	Reference
BatMeth	1.04b	Three letter	FM index of the C-to-T converted genome. Filtering step for low complexity reads with Shannon's entropy.	C, Perl	None	Yes	Yes	Yes	Yes	[11]
Bismark	v0.14.2	Three letter	FM index of the C-to-T converted genome. Mapping step taken into account basecall qualities.	Perl	Bowtie, Bowtie2	Yes	No	Yes	Yes	[12]
Bisulfighter	1.3	Wild-card	Spaced suffix array index of the original (i.e. unconverted) reference genome.	C/C++, Python, R	LAST	Yes	No	Yes	No	[13]
BRAT-BW	2.0.1	Three letter	FM index of the C-to-T converted genome. Multi-seeding starting from different positions within reads.	C++	None	Yes	No	Yes	Yes	[14]
BSMAP	2.74	Wild-card	Hash table of the original genome. Based on SOAP alignment algorithm.	C++, Python	None	Yes	No	Yes	Yes	[15]
BSmooth	0.8.1	Wild-card	FM-index, nucleotide base Y for C and T matches. Support for color-space read mapping.	Perl	Bowtie 2, Merman	Yes (with Bowtie2)	Yes (with Merman)	Yes	Yes	[16]
BS-Seeker	-	Three letter	FM index of the C-to-T converted genome. Only accepts reads in fixed length in FASTQ.	Python	Bowtie	No	No	Yes	No	[17]
BS-Seeker2	v2.0.9	Three letter	FM index of the C-to-T converted genome. Filtering step for the reads with incomplete bisulfite conversion.	Python	Bowtie, Bowtie2, SOAP, RMAP	Yes	No	Yes	Yes	[18]
B-SOLANA	1.0	Color-space	FM-index of the C-to-T converted genome. Support for color-space read mapping.	Python	Bowtie	No	Yes	No	Yes	[19]
GSNAP	2014-01-21	Wild-card	Hash table of the C-to-T converted genome. Uses wild-card letter Y for Cs and Ts in reads.	C, Perl	None	Yes	No	Yes	Yes	[20]
ERNE-BSS	2.1	Wild-card	Hashing the reference genome with 5-letter, A, T, G, Cn, Cn.	C++	None	Yes	Yes	No	Yes	[21]
LAST	548	Wild-card	Spaced suffix array index of the original genome. Modified score matrix for C-to-T and G-to-A matches.	C/C++, Python	None	Yes	No	Yes	Yes (requires GNU parallel)	[22]
MAQ	v0.6.6	Wild-card	Multiple hash tables of reads. Nonunique reads assigned randomly to one of the best-matching positions.	C/C++, Perl	None	Yes	Yes	No	No	[23]

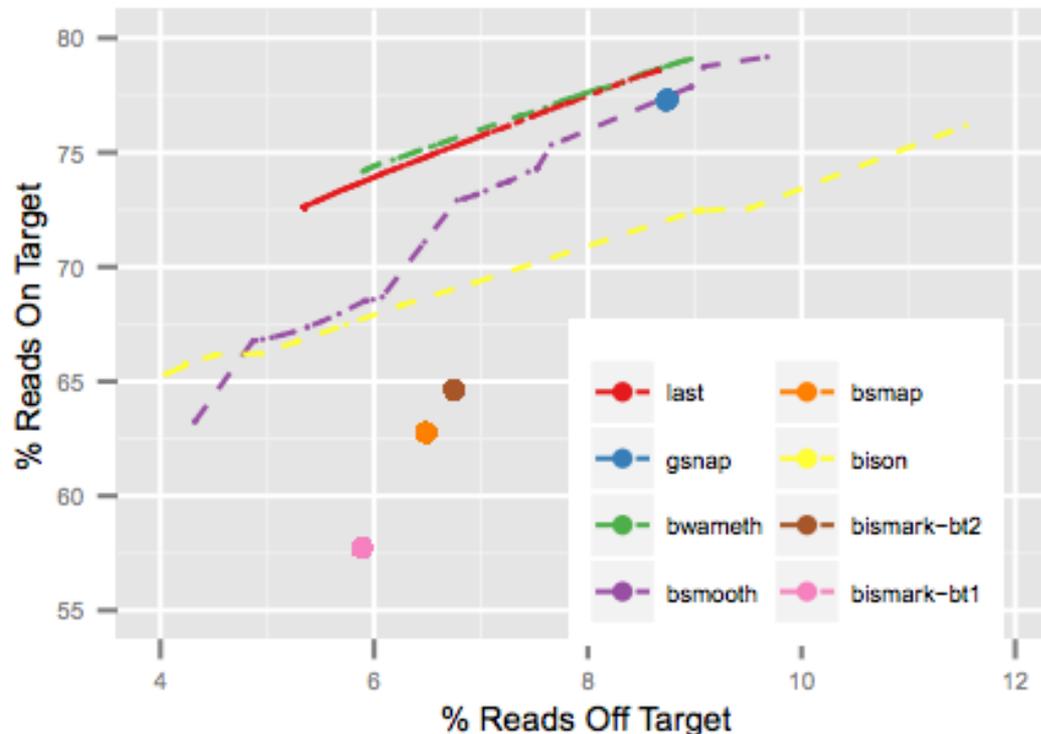
bwa-meth

Fast and accurate alignment of long bisulfite-seq reads

Brent S. Pedersen^{1,*}, Kenneth Eyring¹, Subhajyoti De^{1,2}, Ivana V. Yang¹ and David A. Schwartz¹

¹Department of Medicine, University of Colorado Denver, School of Medicine, Denver, Colorado, USA. 80045

²University of Colorado Cancer Center, Molecular Oncology Program, Aurora, Colorado, USA



bwa-meth

Advantages:

- Simplicity and ease of use: a single script for fastq alignment and methylation calling; wraps bwa mem option
- Speed: efficient bwa parallelization
- Memory usage: compressed or uncompressed input files, reads streamed directly into aligner and not written to disk, so lower memory requirement
- Useful output: sorted & indexed .bam files containing map-q score, read-group & alignment flags, compatible with picard tools and gatk for downstream processing
- Strand-specificity: increase accuracy
- Accuracy without trimming: due to improved bwa local alignment

Disadvantages:

- Only works for directional library preparation protocol

Logging into Hoffman2

ssh flay@hoffman2.idre.ucla.edu

A screenshot of a macOS Terminal window. The title bar reads "Terminal — ssh — 139x38". The terminal content shows the command "ssh flay@hoffman2.idre.ucla.edu" being entered. The prompt is "Fides-Lays-MacBook-Pro:~ fideslay\$". Below the command, the prompt "flay@hoffman2.idre.ucla.edu's password:" is visible with a cursor. Two red circles are drawn around the "ssh" and "flay" parts of the command. Two red arrows originate from these circles and point towards the text "change to your user ID" below the terminal.

```
Fides-Lays-MacBook-Pro:~ fideslay$ ssh flay@hoffman2.idre.ucla.edu
flay@hoffman2.idre.ucla.edu's password: █
```

change to your user ID

Logging into Hoffman2

flay@login3:~ — ssh — 139x38

```
Fides-Lays-MacBook-Pro:~ fideslay$ ssh flay@hoffman2.idre.ucla.edu
```

```
flay@hoffman2.idre.ucla.edu's password:
```

```
Last login: Wed Mar  2 10:03:31 2016 from 164.67.160.102
```

```
Welcome to the Hoffman2 Cluster!
```

```
Hoffman2 Home Page:      http://www.hoffman2.idre.ucla.edu
```

```
Consulting:              https://support.idre.ucla.edu/helpdesk
```

```
All login nodes should be accessed via "hoffman2.idre.ucla.edu".
```

```
Please do NOT compute on the login nodes.
```

```
Processes running on the login nodes which seriously degrade others' use of the system may be terminated without warning. Use qssh to obtain an interactive shell on a compute node for CPU or I/O intensive tasks.
```

```
The following news items are currently posted:
```

```
  Globus access changes February 13, 2016
```

```
  Mathematica upgraded to 10.3
```

```
  IDRE Winter 2016 HPC Classes
```

```
  News Archive On Web Site
```

```
Enter shownews to read the full text of a news item.
```

```
[flay@login3 ~]$ █
```

Useful Tools on Hoffman2

module available

```
[flay@login3 ~]$ module available
```

```
----- /u/local/Modules/modulefiles -----
ATS
R/2.12.0
R/2.12.1
R/2.12.2
R/2.13.2
R/2.15.1
R/2.9.1
R/3.0.1(default)
R/3.1.1
R/3.2.1
R/3.2.3
Rstudio/0.98(default)
Rstudio/R-2.15.1
abaqus/6.11
abaqus/6.12
abaqus/6.13
abaqus/6.14(default)
activeperl/5.16(default)
affymetrix/apt-1.14.3(default)
amber/14
amber-intel11.1-mvapich1.9a2/12
annovar/v2811Jun18
annovar/v2811Oct02
annovar/v2814Jul14(default)
ansys/14.0
armadillo/4.600.4
bcftools/1.2
bedtools/2.13.3
bedtools/2.17.0(default)
bedtools/2.23.0
blat/35
boost/1_55_0(default)
boost/1_59_0
bowtie/0.12.7
bowtie/0.12.8(default)
bowtie2/2.1.0
bowtie2/2.2.5(default)
bwa/0.6.2
bwa/0.7.7(default)
castep/5.5(default)
cern_root/5.26.00
cern_root/5.30.00(default)
cern_root/5.32.01
cern_root/5.34.18
cernlib/2006(default)
clhep/2.1.0.1(default)
clhep/2.1.1.0
clhep/2.1.3.1
cmake/2.8.7
cmake/3.0.2(default)
consol/4.3a
consol/4.3b
consol/4.4
consol/5.0
consol/5.2(default)
cuda/4.0(default)
cuda/4.1
cuda/4.2
cuda/5.0
cuda/5.5
cuda/7.0
cufflinks/2.0.2
cufflinks/2.1.1(default)
cufflinks/2.2.1
ddd/3.3.12(default)
ddplot/2.5
ddplot/4.0(default)
eclipse/3.6
eclipse/4.3(default)
espresso/5.0.3
fastx_toolkit/0.0.13.2
ffmpeg/0.11.1
ffmpeg/1.1.1(default)
freebayes/17Feb12
freetds/0.91
fribidi/0.19.6
gatk/1.0.4
gatk/2.7.2
gatk/3.1.1
gatk/3.3.0(default)
gcc/4.3.5
gcc/4.4(default)
gcc/4.7.2
gcc/4.9.3
gcta/0.93.9
gdal/1.9.2
gdc/0.30
gdc/4.9.3(default)
geant4/10.00.p02_wG4ML
geant4/4.0.3.p02(default)
geant4/4.0.4.p02
geant4/4.0.5
geant4/9.6.2
geant4/9.6.p02_wG4ML
geos/3.4.2(default)
gftp/2.0.19
ghc/7.6.3
gmp/4.3.2
gnuplot/4.2.3
gnuplot/4.4.4(default)
gotoblas2/1.13_multi_threaded
gotoblas2/1.13_single_threaded
gpac/0.5.1
grads/2.0.1(default)
graphviz/2.28.0(default)
gromacs/4.6.5
handbrake/0.10.0(default)
handbrake/0.9.8
haskell/2013.2.0.0
hdfs/1.8.11_intel13.cs
hdfs/1.8.11_intel13.cs_intelmpi4.1.1
hdfs/1.8.14_intel-13.1.1_intelmpi-4.1.1
homer/4.7
hyperworks/11.0
hyperworks/12.0(default)
idl/6.3
idl/6.11(default)
idr/2010-10
impute/2.3.0(default)
intel/11.1
intel/12.0
intel/12.1
intel/13.0
intel/13.cs(default)
intel/14.cs
intelmpi/4.1.1
intelmpi/4.1.3
intelmpi/5.0.0
jags/3.3.0
jags/3.4.0(default)
java/1.6.0_23(default)
java/1.7.0_45
julia/0.3.11
lammps/10Feb15-intel14.0-imp15.0(default)
lammps/22Feb13-intel11.1-openmpi.4.5
lammps/28Jun14-intel13.1-imp14.1
lynx/2.8.7(default)
maple/16
maple/17(default)
maq/0.7.1(default)
mathematica/10.0
mathematica/10.3(default)
mathematica/8.0.4
mathematica/9.0
matlab/7.11
matlab/7.14
matlab/7.7
matlab/8.2
matlab/8.4
matlab/8.6(default)
mats/3.0.7
mecab/0.996
nemo/2.0(default)
nopac/2012-cpu
mpc/0.8.1
mpfr/2.4.2
mplayer
nag/5.3
ncl/5.2.1(default)
nco/4.0.6(default)
nco/4.4.4
netbeans/7.4
netbeans/8.0.2(default)
netcdf/4.1.3(default)
netcdf/4.1.3-shared
netcdf/4.1.3_i
netcdf/4.2.3-c
netcdf/4.4.2-fortran
ngsplot/2.47
nchen/6.5(default)
octave/3.6.1
octave/3.6.4(default)
openbabel/2.3.1(default)
openmpi/1.6_gcc-4.4
opensees/2.4.1
opensees/2.4.3
opensees/2.4.4(default)
opensees/2.4.4_parallel
paraview/3.6.2(default)
perl/5.10.1(default)
pgi_compiler/15.3
pgsl/9.1
plink/1.08
pop-c++/1.3(default)
preseq/1.0.2
proj/4.8.0
pypy/1.9
pypy3/2.4.0
python/2.6(default)
python/2.7
python/2.7.3
python/3.1
python/3.4
qcachegrind/0.7.4
qchem/3201s
qchem/3202s
qchem/4.3
qchem/4.3mpich(default)
qchem/4002mpich
qchem/4002s
qchem/4101
qchem/4101mpich
qiime/1.8.0
qiime/1.9.0(default)
relion/1.3
ruby/1.9.2(default)
samtools/0.1.17
samtools/0.1.18
samtools/0.1.19(default)
samtools/1.2
seqlogo/2.8.2
shapet/1
solar/4.3.1(default)
soxr/0.1.0
splicetrap/0.90.5
stata/11
stata/13
stata/14(default)
svn/1.6(default)
tcltk/8.4_thread-disabled(default)
tecplot360/2014
tecplot360/2015(default)
texlive/2012(default)
tms/1.3(default)
tophat/1.3.3
tophat/2.0.14(default)
tophat/2.0.4
tophat/2.0.9
treemix/1.12
trinity/2013-08-14
vcftools/0.1.12a
vcftools/0.1.14(default)
vcftools/0.1.9
vegas/0.8.27
vmd/1.8.7(default)
vtk/5.8(default)
x264/000(default)
x264/20141209-2245
yasm/1.2.0(default)
zlib/1.2.8
```

Useful Tools on Hoffman2

ls /u/local/apps/

```
[flay@login3 ~]$  
[flay@login3 ~]$ ls /u/local/apps/  
abaqus          blcr             emacs            grads            libgtextutils   mumps            picard-tools     scilab           tvmet  
abaqusdocs      blitz           espresso         graphicsmagick  libtool         muscle           plink            scones           uclust  
accelrys        bmapuclatools  fastphase       graphviz        libxc           mygroup         pop-c++         sentaurus       udunits  
Accelrys        boost          fasttree        gromacs         lmbd            mysql           povray          shapeit         usearch  
activeperl      boost-jam      fastx_toolkit   gsl             loni_pipeline   namd            pplcer          shapelib        usr_lib_GOverride  
ActiveTcl       bowtie         fbat            hadoop          lumeral         ncl             preseq          skampi          valgrind  
adina           bowtie2        fe-safe         handbrake       lynx            nco             proj.4          slots           vasp  
Adobe           bwa            ffmpeg          harminv         mach            netbeans        protobuf        snappy          vcftools  
affymetrix      caffe          fftw2           haskell         mafft          netcdf          pypy           solar           vegas  
AFNI            casava         fftw3           hdf             manorm         newbler         pypy3          sortmerna      ViennaRNA  
aida            cdbtools       flex            hdf5            maple           ngsplot        python          soxr           vim  
amber           cd-hit         fltk            homer           maq            nlopt          qcachegrind     splicetrapp    visit  
ampliconnoise  cernlib        freebayes       hyperworks     maqview        numpy           qchem           spm             vmd  
anaconda        cern_root      freesurfer      iaida           mathematica     nwchem         qgfe            stata           votca  
annovar         clearcut       freetds         idl             matio           ocaml          qhull           stressapptest  vtk  
ansys_inc       clhep         fribidi         idr             matlab          octave         qiime           structure       wannier90  
ant             cln            fsl             igraph         mats           omssa          qiime_data      subversion     weblogo  
antlr           cmake          gamess          ilog            mecab          oommf         qrupdate        suitesparse    wolfram  
armadillo       common         gatk            IM-IMA         meep           openbabel      qsub            sundials       x264  
arpack         consol         gatk-queue     impute         meld           openfoam       [redacted]       superlu_dist   x86_open64  
atlas           conan          gaussian        imsl            merlin         openfoam       quantiSNP       swarm           xerces-c  
atomeye        condor         gaussview       infernal        metis          openmotif      queue.logs      sysstat        xfdtd  
atompaw        consed         gcta           inspect        mfold          opensees      R               szip            xfig_OLD  
autoconf       cp2k          gdal            installation   microbioemutil osiris         raxml           t500            xilinx-vivado  
automake       cpmd          gdc             intel           migrate        osu-micro-benchmarks rdp_classifier  tau            xmd  
bamtools       ctffind        geant4          isight         minirosetta    papi           relion          tcL             xmcdcon  
bcftools       cufflinks     genetorrent     jags           mira           paraview       remcom          tcsh            xpdf_OLD  
bcl2qfastq     cytoscape     geos            jam             molcas         parmetis       repeatmasker    tecplot360     yaml  
beagle         ddd           gflags         java            molder         parsec         RepeatModeler   texlive        yasm  
beaglecall     ddpilot       gftp           jaxodraw       molpro         parsinsert     resmap          tmux           zlib  
beagle_utilities dejagnu       ginac          julia           mono           pcre           rosetta         tophat         trilinos  
bedtools       dirac         git_OLD        kepler         mopac          pdt            rstudio         toscastucture  trinity  
bfast          diskusage     globalarrays    kepler         mothur         penncnv        rsync           totalview      trans-ABYSS  
bioscope       dl_poly      glog           lammps         mpc            perl_modules   rtax            treemix  
blas           dx            gmp            lapack         mpfr           pgsq           ruby            trilmix  
blast          eclipse       gnuplot        lapack++       mpiblast       phantompeakqualtools sage             trillinos  
blast+         eigen         gpac           ldope          mplayer        phase          samtools        trinity  
blat           eigensoft     grace          leveldb        mrt            phase          scalapack       tt
```

Alignment Using bwa-meth: On Hoffman2 Interactively

#grab a Hoffman2 session:

qrsh

or

qrsh -now n -l i,mem=5G,time=02:00:00,exclusive=TRUE -pe shared 6

Request exact configuration

time limit of the session
(default is 2 hours,
24hrs max)

Reserve the whole node

Request 6 cores

Memory requested per
core (default is 1G)

bwa-meth

ls /u/home/galaxy/collaboratory/apps

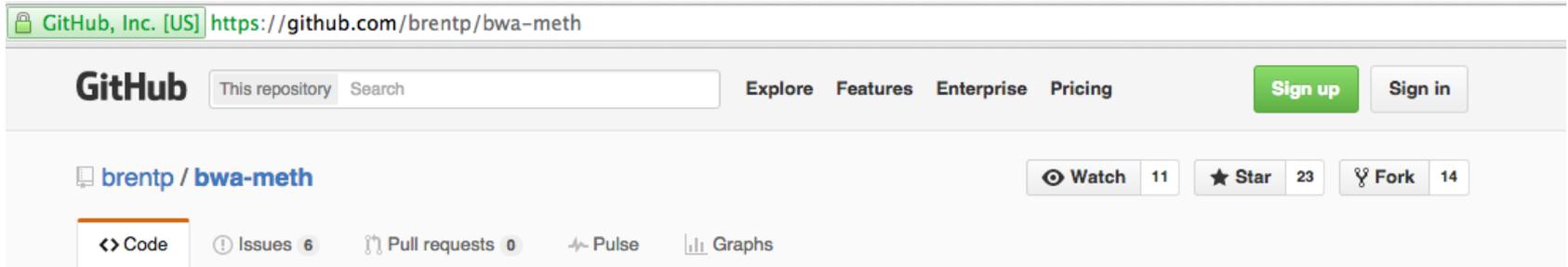
```
[flay@login3 apps]$ ls /u/home/galaxy/collaboratory/apps
bedops_v2.4.20.v2  FastQC      HiC-Pro      samtools-1.1
bin               fit-hi-c    hicup_v0.5.8 sratoolkit.2.7.0-centos_linux64
biscuit-release   gatk        MACS          v0.10.tar.gz
bwa-meth          HiCPlotter picard.jar
bwa-meth-0.10     hic-pro     release.zip
```

ls /u/home/galaxy/collaboratory/apps/bwa-meth-0.10

```
[flay@login3 apps]$ ls /u/home/galaxy/collaboratory/apps/bwa-meth-0.10/
bias-plot.py  compare  ez_setup.py  paper  requirements.txt  setup.py
bwameth.py   example  LICENSE      README.md  scripts
```

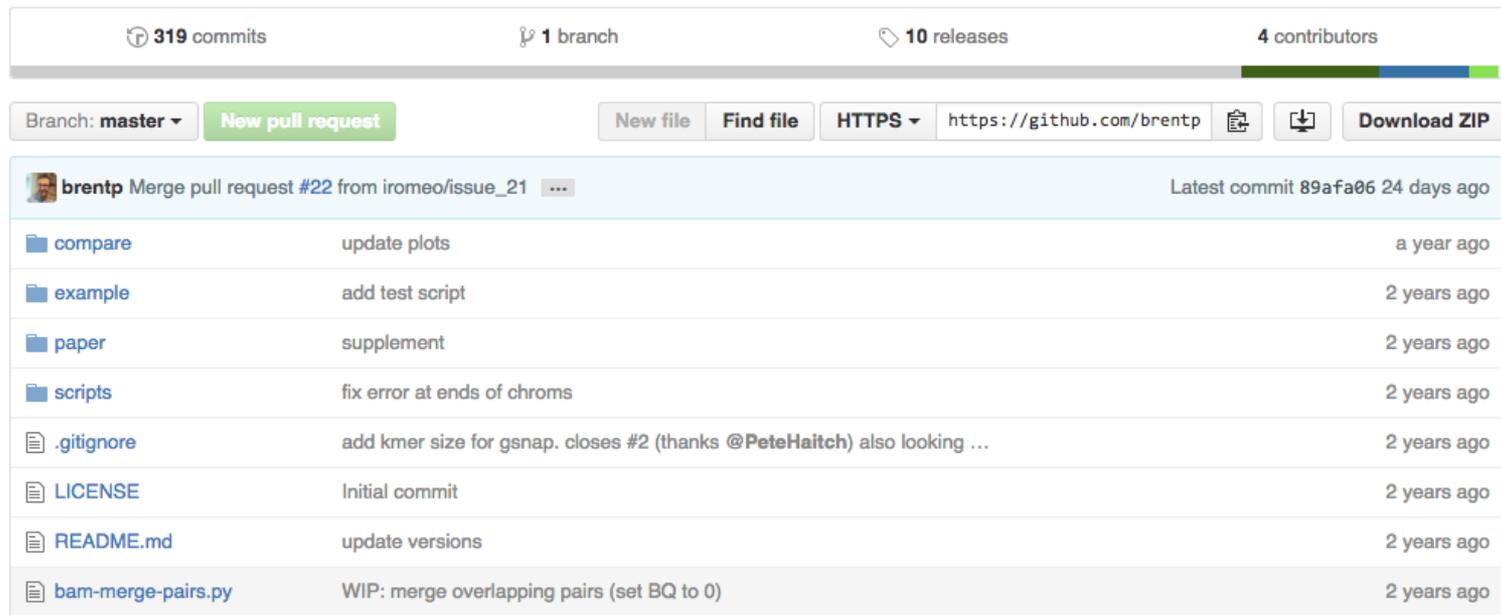
bwa-meth

#For documentation, go to <https://github.com/brentp/bwa-meth>



The screenshot shows the GitHub repository page for `brentp / bwa-meth`. At the top, there's a navigation bar with the GitHub logo, a search box for "This repository", and links for "Explore", "Features", "Enterprise", and "Pricing". There are "Sign up" and "Sign in" buttons. Below the navigation bar, the repository name "brentp / bwa-meth" is displayed, along with statistics: 11 Watchers, 23 Stars, and 14 Forks. A secondary navigation bar includes "Code", "Issues 6", "Pull requests 0", "Pulse", and "Graphs".

align BS-Seq reads and extract methylation without intermediate temp files



This section shows the repository's activity and file structure. It includes statistics: 319 commits, 1 branch, 10 releases, and 4 contributors. Below these are buttons for "New pull request", "New file", "Find file", "HTTPS", and "Download ZIP". The main content is a list of commits, with the most recent being a merge pull request #22 from `iromeo/issue_21` by `brentp`, committed 24 days ago. Below this is a list of files and folders with their commit messages and dates.

File/Folder	Commit Message	Time
compare	update plots	a year ago
example	add test script	2 years ago
paper	supplement	2 years ago
scripts	fix error at ends of chroms	2 years ago
.gitignore	add kmer size for gsnap. closes #2 (thanks @PeteHaitch) also looking ...	2 years ago
LICENSE	Initial commit	2 years ago
README.md	update versions	2 years ago
bam-merge-pairs.py	WIP: merge overlapping pairs (set BQ to 0)	2 years ago

bwa-meth: Dependencies

#load all the dependencies prior to running your program

```
module load bwa
```

```
module load python/2.7.3
```

```
module load samtools/1.2
```

#install toolshed and seaborn for python

```
pip install toolshed --user
```

```
[flay@login2 bwa-meth]$ module load python/2.7.3
```

```
[flay@login2 bwa-meth]$ pip install toolshed --user
```

```
Downloading/unpacking toolshed
```

```
  Downloading toolshed-0.4.5.tar.gz
```

```
  Running setup.py (path:/work/tmp/pip_build_flay/toolshed/setup.py) egg_info for package toolshed
```

```
Installing collected packages: toolshed
```

```
  Running setup.py install for toolshed
```

```
    Installing toolshed script to /u/home/f/flay/.local/bin
```

```
Successfully installed toolshed
```

```
Cleaning up...
```

```
[flay@login2 bwa-meth]$
```

```
pip install seaborn --user
```

```
##this is optional
```

```
[flay@login2 bwa-meth]$ pip install seaborn --user
```

```
Downloading/unpacking seaborn
```

```
  Downloading seaborn-0.7.0.tar.gz (154kB): 154kB downloaded
```

```
  Running setup.py (path:/work/tmp/pip_build_flay/seaborn/setup.py) egg_info for package seaborn
```

```
Downloading/unpacking pandas (from seaborn)
```

```
  Downloading pandas-0.17.1.tar.gz (6.7MB): 6.7MB downloaded
```

```
  Running setup.py (path:/work/tmp/pip_build_flay/pandas/setup.py) egg_info for package pandas
```

bwa-meth

python /u/home/galaxy/collaboratory/apps/bwa-meth-0.10/bwameth.py

```
[flay@n2066 ~]$ python /u/home/galaxy/collaboratory/apps/bwa-meth-0.10/bwameth.py --h
usage:
map bisulfite converted reads to an insilico converted genome using bwa mem.
A command to this program like:
```

```
python bwameth.py --reference ref.fa A.fq B.fq
```

Gets converted to:

```
bwa mem -pCMR ref.fa.bwameth.c2t '<python bwameth.py c2t A.fq B.fq'
```

So that A.fq has C's converted to T's and B.fq has G's converted to A's and both are streamed directly to the aligner without a temporary file. The output is a corrected, sorted, indexed BAM.

```
[-h] --reference REFERENCE [-t THREADS] [-p PREFIX] [--calmd]
[--read-group READ_GROUP] [--set-as-failed {f,r}]
fastqs [fastqs ...]
```

positional arguments:

```
fastqs          bs-seq fastqs to align. Run multiple sets separated by
                 commas, e.g. ... a_R1.fastq,b_R1.fastq
                 a_R2.fastq,b_R2.fastq note that the order must be
                 maintained.
```

optional arguments:

```
-h, --help          show this help message and exit
--reference REFERENCE
                   reference fasta
-t THREADS, --threads THREADS
-p PREFIX, --prefix PREFIX
--calmd
--read-group READ_GROUP
                   read-group to add to bam in same format as to bwa:
                   '@RG\tID:foo\tSM:bar'
--set-as-failed {f,r}
                   flag alignments to this strand as not passing QC
                   (0x200). Targetted BS-Seq libraries are often to a
                   single strand, so we can flag them as QC failures.
                   Note f == OT, r == OB. Likely, this will be 'f' as we
                   will expect reads to align to the original-bottom (OB)
                   strand and will flag as failed those aligning to the
                   forward, or original top (OT).
```

Alignment Using bwa-meth: Copy Files

#copy all the files you need to scratch folder

cd /u/scratch/f/flay/ change to your ID

cp /u/scratch/f/flay/workshop6.tar.gz .

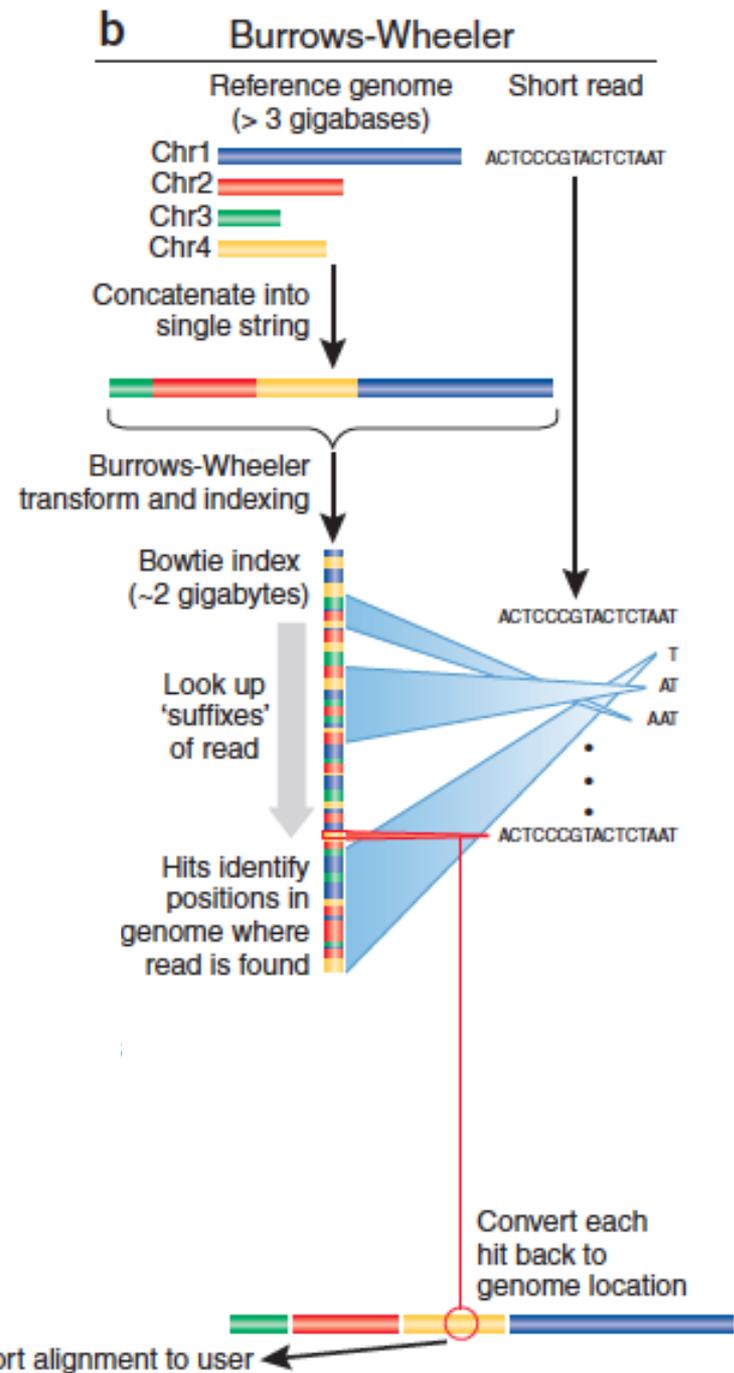
Indicates
current
directory

tar -xvzf workshop6.tar.gz

cd workshop6

ls

Bwa Alignment and Indexing



Alignment Using bwa-meth: Indexing the Reference Genome

#Use chr1 as an example

#Change directory to where reference genome fasta is

```
cd genome/chr1
```

```
python /u/home/galaxy/collaboratory/apps/bwa-meth-0.10/
```

```
bwameth.py index chr1.fa
```

Alignment Using bwa-meth: Indexing the Reference Genome

```
[flay@n2066 chr1]$ python /u/home/galaxy/collaboratory/apps/bwa-meth-0.10/bwameth.py index chr1.fa
converting c2t in chr1.fa to chr1.fa.bwameth.c2t
indexing: chr1.fa.bwameth.c2t
[bwa_index] Pack FASTA... 5.88 sec
[bwa_index] Construct BWT for the packed sequence...
[BWTIncCreate] textLength=788781728, availableWord=67501160
[BWTIncConstructFromPacked] 10 iterations done. 100000000 characters processed.
[BWTIncConstructFromPacked] 20 iterations done. 195555488 characters processed.
[BWTIncConstructFromPacked] 30 iterations done. 280551552 characters processed.
[BWTIncConstructFromPacked] 40 iterations done. 356092224 characters processed.
[BWTIncConstructFromPacked] 50 iterations done. 423228976 characters processed.
[BWTIncConstructFromPacked] 60 iterations done. 482896288 characters processed.
[BWTIncConstructFromPacked] 70 iterations done. 535924720 characters processed.
[BWTIncConstructFromPacked] 80 iterations done. 583052480 characters processed.
[BWTIncConstructFromPacked] 90 iterations done. 624935760 characters processed.
[BWTIncConstructFromPacked] 100 iterations done. 662157696 characters processed.
[BWTIncConstructFromPacked] 110 iterations done. 695236640 characters processed.
[BWTIncConstructFromPacked] 120 iterations done. 724633216 characters processed.
[BWTIncConstructFromPacked] 130 iterations done. 750756992 characters processed.
[BWTIncConstructFromPacked] 140 iterations done. 773971824 characters processed.
[bwa_index] 348.12 seconds elapse.
[bwa_index] Update BWT... 3.95 sec
[bwa_index] Pack forward-only FASTA... 4.10 sec
[bwa_index] Construct SA from BWT and Occ... 117.57 sec
[main] Version: 0.7.7-r441
[main] CMD: bwa index -a bwtsv chr1.fa.bwameth.c2t
[main] Real time: 491.483 sec; CPU: 479.627 sec
```

Alignment Using bwa-meth: Indexing the Reference Genome

#Output files after indexing

```
[flay@n2192 mm9]$ ls -lh *chr1*  
-rw-r--r-- 1 flay matteop 192M Mar  1 17:42 chr1.fa  
-rw-r--r-- 1 flay matteop 380M Mar  4 22:17 chr1.fa.bwameth.c2t  
-rw-r--r-- 1 flay matteop 1.2K Mar  4 22:24 chr1.fa.bwameth.c2t.amb  
-rw-r--r-- 1 flay matteop  83 Mar  4 22:24 chr1.fa.bwameth.c2t.ann  
-rw-r--r-- 1 flay matteop 377M Mar  4 22:24 chr1.fa.bwameth.c2t.bwt  
-rw-r--r-- 1 flay matteop  95M Mar  4 22:24 chr1.fa.bwameth.c2t.pac  
-rw-r--r-- 1 flay matteop 189M Mar  4 22:26 chr1.fa.bwameth.c2t.sa
```

Alignment Using bwa-meth: Understanding Commands and Options

```
[flay@n2066 chr1]$ python /u/home/galaxy/collaboratory/apps/bwa-meth-0.10/bwameth.py
usage:
map bisulfite converted reads to an insilico converted genome using bwa mem.
A command to this program like:
```

```
python bwameth.py --reference ref.fa A.fq B.fq
```

Gets converted to:

```
bwa mem -pCMR ref.fa.bwameth.c2t '<python bwameth.py c2t A.fq B.fq'
```

So that A.fq has C's converted to T's and B.fq has G's converted to A's and both are streamed directly to the aligner without a temporary file. The output is a corrected, sorted, indexed BAM.

```
[-h] --reference REFERENCE [-t THREADS] [-p PREFIX] [--calmd]
[--read-group READ_GROUP] [--set-as-failed {f,r}]
fastqs [fastqs ...]
```

```
map bisulfite converted reads to an insilico converted genome using bwa mem.
A command to this program like:
```

```
python bwameth.py --reference ref.fa A.fq B.fq
```

Gets converted to:

```
bwa mem -pCMR ref.fa.bwameth.c2t '<python bwameth.py c2t A.fq B.fq'
```

So that A.fq has C's converted to T's and B.fq has G's converted to A's and both are streamed directly to the aligner without a temporary file. The output is a corrected, sorted, indexed BAM.

Alignment Using bwa-meth: Align Fastq Files to the Reference Genome

#aligned paired-end reads to the reference genome (hg19)

```
cd /u/scratch/f/flay/workshop6/data/raw
```

```
python /u/home/galaxy/collaboratory/apps/bwa-meth-0.10/  
bwameth.py --threads 5 --prefix N25 --reference /u/scratch/  
f/flay/workshop6/genome/hg19_rCRSChrm.fa N25_R1.fastq  
N25_R2.fastq
```

Alignment Using bwa-meth: Align Fastq Files to the Reference Genome

```
[flay@n2066 raw]$ python /u/home/galaxy/collaboratory/apps/bwa-meth-0.10/bwameth.py --threads 5 --prefix N25 --reference /u/scratch/f/flay/workshop6/genome/hg19_rCRSchrn.fa N25_R1.fastq N25_R2.fastq
running: bwa mem -T 40 -B 2 -L 10 -CM -U 100 -p -R '@RG ID:N25_R SM:N25_R' -t 5 /u/scratch/f/flay/workshop6/genome/hg19_rCRSchrn.fa.bwameth.c2t '</u/local/apps/python/2.7.3/bin/python /u/home/galaxy/collaboratory/apps/bwa-meth-0.10/bwameth.py c2t N25_R1.fastq N25_R2.fastq'
writing to:
samtools view -bS - | samtools sort - N25
converting reads in N25_R1.fastq,N25_R2.fastq
WARNING: 12500 reads with length < 80
      : this program is designed for long reads
[M::main_mem] read 12500 sequences (950000 bp)...
[M::mem_pestat] # candidate unique pairs for (FF, FR, RF, RR): (0, 3755, 0, 0)
[M::mem_pestat] skip orientation FF as there are not enough pairs
[M::mem_pestat] analyzing insert size distribution for orientation FR...
[M::mem_pestat] (25, 50, 75) percentile: (138, 152, 172)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (70, 240)
[M::mem_pestat] mean and std.dev: (155.65, 25.45)
[M::mem_pestat] low and high boundaries for proper pairs: (36, 274)
[M::mem_pestat] skip orientation RF as there are not enough pairs
[M::mem_pestat] skip orientation RR as there are not enough pairs
[M::mem_process_seqs] Processed 12500 reads in 18.257 CPU sec, 5.553 real sec
[main] Version: 0.7.7-r441
[main] CMD: bwa mem -T 40 -B 2 -L 10 -CM -U 100 -p -R @RG ID:N25_R SM:N25_R -t 5 /u/scratch/f/flay/workshop6/genome/hg19_rCRSchrn.fa.bwameth.c2t </u/local/apps/python/2.7.3/bin/python /u/home/galaxy/collaboratory/apps/bwa-meth-0.10/bwameth.py c2t N25_R1.fastq N25_R2.fastq
[main] Real time: 157.097 sec; CPU: 74.820 sec
running: samtools index N25.bam
```

Alignment Using bwa-meth: Output Files

```
[flay@n2176 raw]$ ls -lh
total 6.3M
-rw-r--r-- 1 flay matteop 906K Jun 16 22:17 N25.bam
-rw-r--r-- 1 flay matteop 1.6M Jun 16 22:17 N25.bam.bai
-rw-r--r-- 1 flay matteop 1.3M Jun 16 22:04 N25_R1.fastq
-rw-r--r-- 1 flay matteop 1.3M Jun 16 22:04 N25_R2.fastq
```

Assessing Library Quality: Alignment and Pairing Rate

#use samtools calculated statistics to look at alignment rate
samtools flagstat fileName.bam

```
[flay@n2176 raw]$ samtools flagstat N25.bam
12197 + 306 in total (QC-passed reads + QC-failed reads)
3 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
11878 + 306 mapped (97.38%:100.00%)
12194 + 306 paired in sequencing
6097 + 153 read1
6097 + 153 read2
11180 + 0 properly paired (91.68%:0.00%)
11830 + 306 with itself and mate mapped
45 + 0 singletons (0.37%:0.00%)
474 + 0 with mate mapped to a different chr
212 + 0 with mate mapped to a different chr (mapQ>=5)
```

Submitting Alignment Jobs on Hoffman2

#make a bash script containing bwameth commands.

#For example:

```
vi bwameth_align.sh  
i
```

```
#!/bin/sh
```

```
bwameth.py --threads 6 --prefix fileName --reference /u/scratch/f/  
flay/genome/hg19_rCRSchrm.fa /u/scratch/f/flay/test/  
fileName_R1.fastq /u/scratch/f/flay/test/fileName_R2.fastq
```

```
qsub -cwd -V -N bwameth -l h_data=4G,exclusive -l h_rt=20:00:00  
-pe shared 8 bwameth.sh
```

Change to current Job name
working directory

Export the environment
you are in