NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

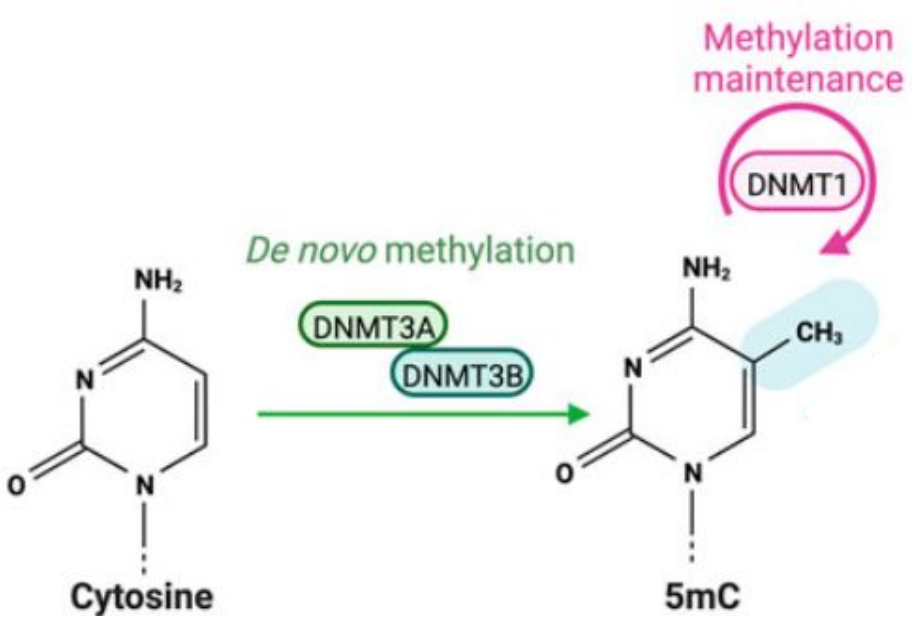
SciLifeLab

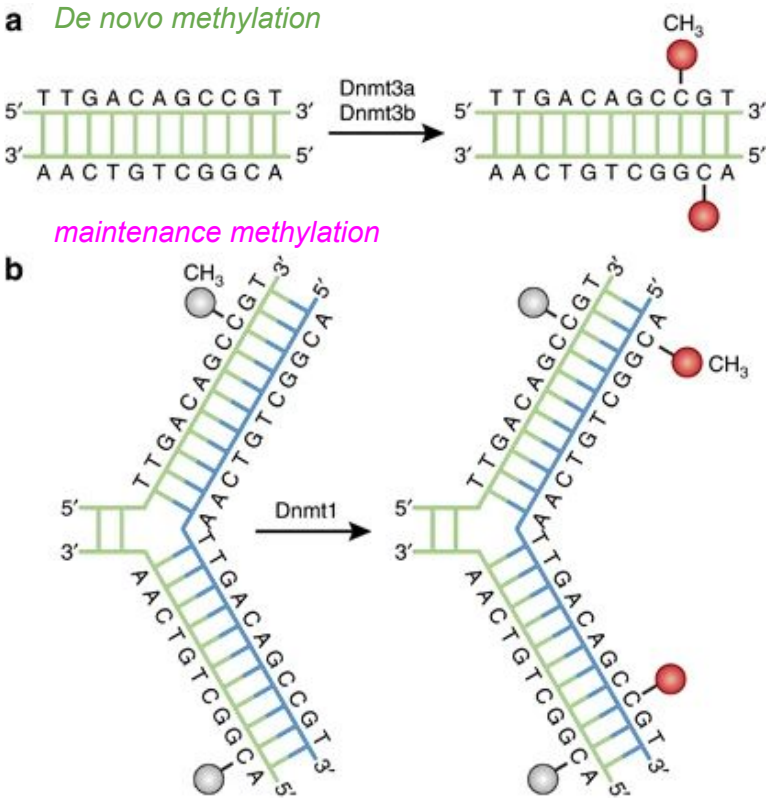
Hands-on Course in Epigenomics: WGBS

Louella Vasquez

25-11-14

DNA methylation is a stable, heritable chemical modification



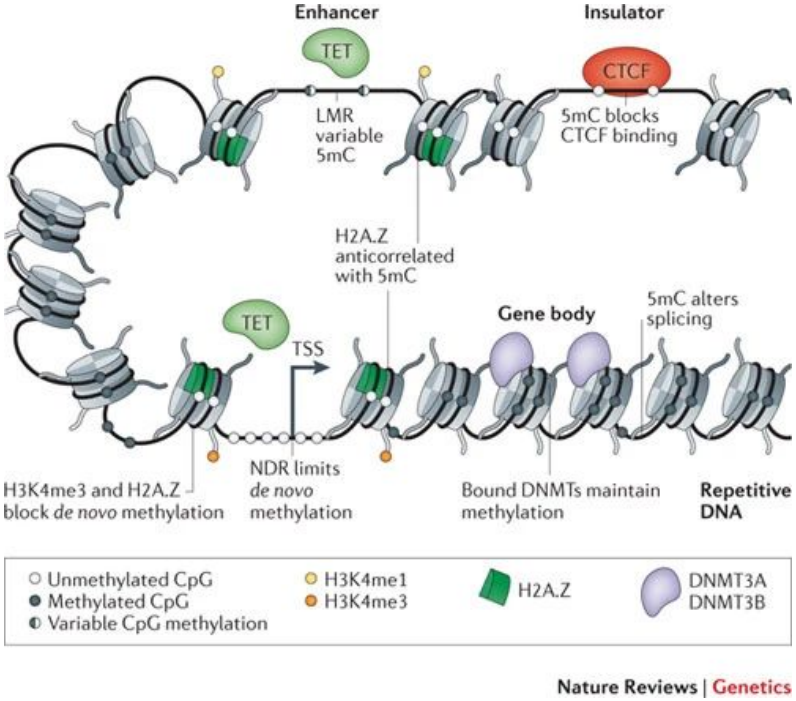


5mC in the genome

- predominantly in CpG (**C**ytosine-**p**hosphate-**G**uanine) dinucleotides in metazoan genomes
  - ~28M CpGs in humans
  - 60–80% methylated in somatic cells

- CpGs in CG-dense regions are CpG islands (CGIs)
  - 200-2000 bp with >50% GC-content
  - CGIs tend to be in promoters, unmethylated for transcribed genes

- non-CpG methylation has been mainly observed in hESCs and neuronal cells in humans
  - CHH, CHG where H = A, C or T



Key functions of DNA methylation

Tissue specific gene regulation

Suppression of transposable elements

Essential for normal development

Genomic imprinting

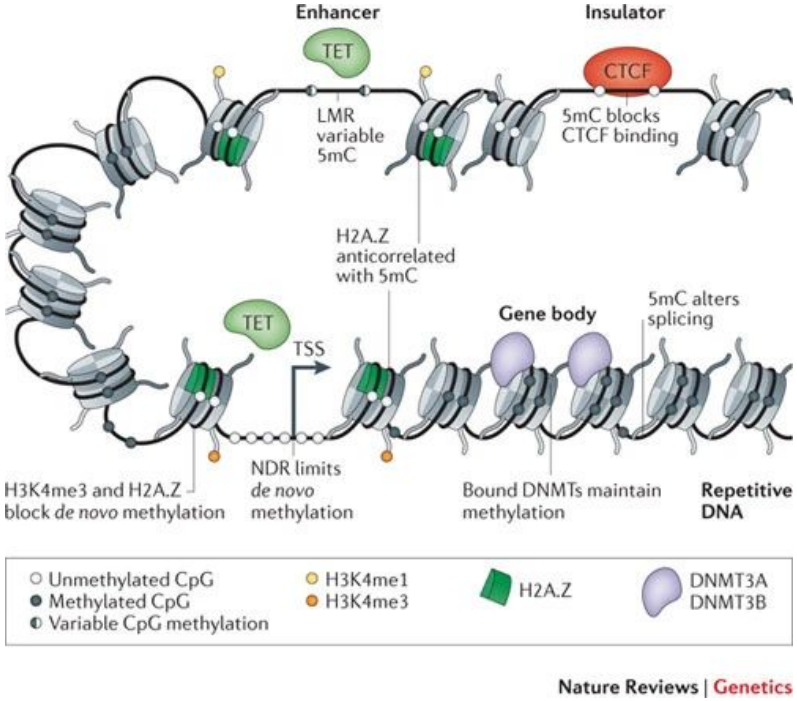
X-chromosome inactivation

Ageing

- Global hypomethylation is proportional to age

Cancer

- Global hypomethylation and locus-specific hypermethylation of CpG islands



DNAM measurement

Differentiate mC from C

C > T Bisulfite conversion

A: 5'-GACCGTTCCAGGTCCAGCAGTGCGCT-3'

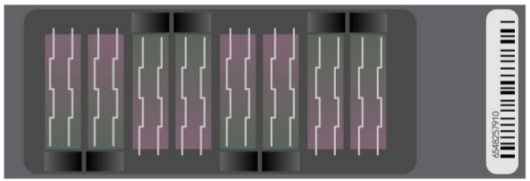
B: 3'-CTGGCAAGGTCCAGGTCGTCACGCGA-5'

A: 5'-GATCGTTTTAGGTTTAGTAGTGCGTT-3'

B: 3'-TTGGCAAGGTTTAGGTTGTTATGCGA-5'

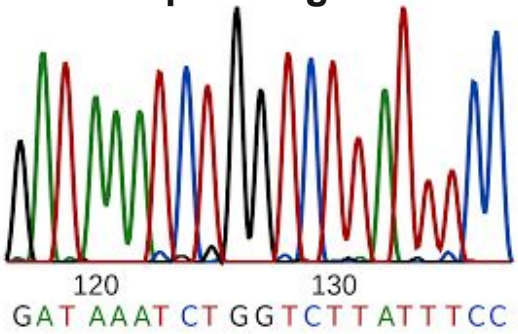
DNA amplification

Methylation Array



DNA Sequencing

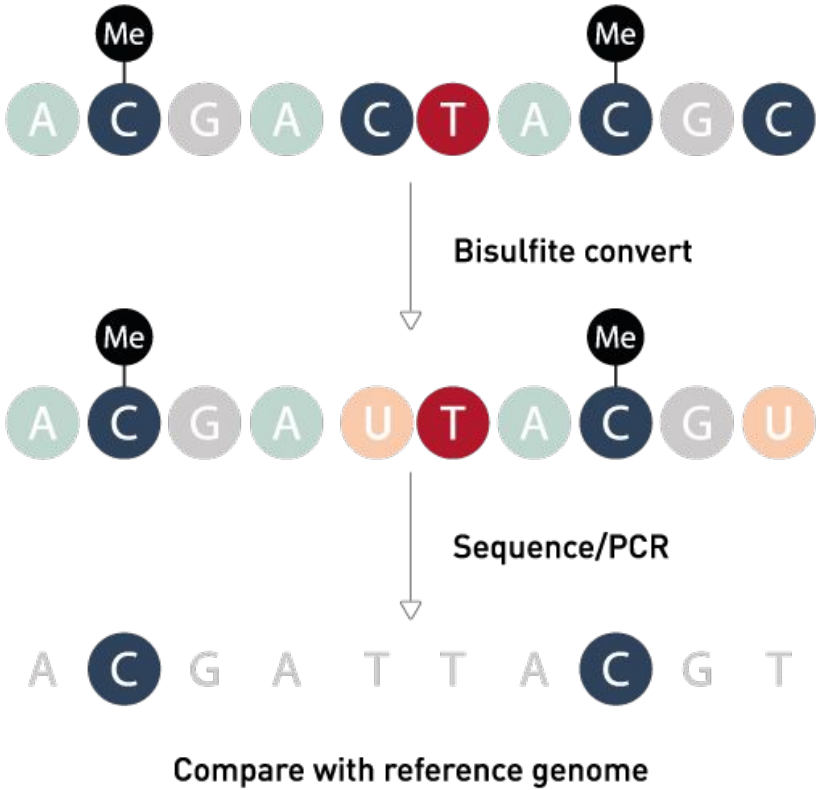
Select regions



Bisulfite conversion

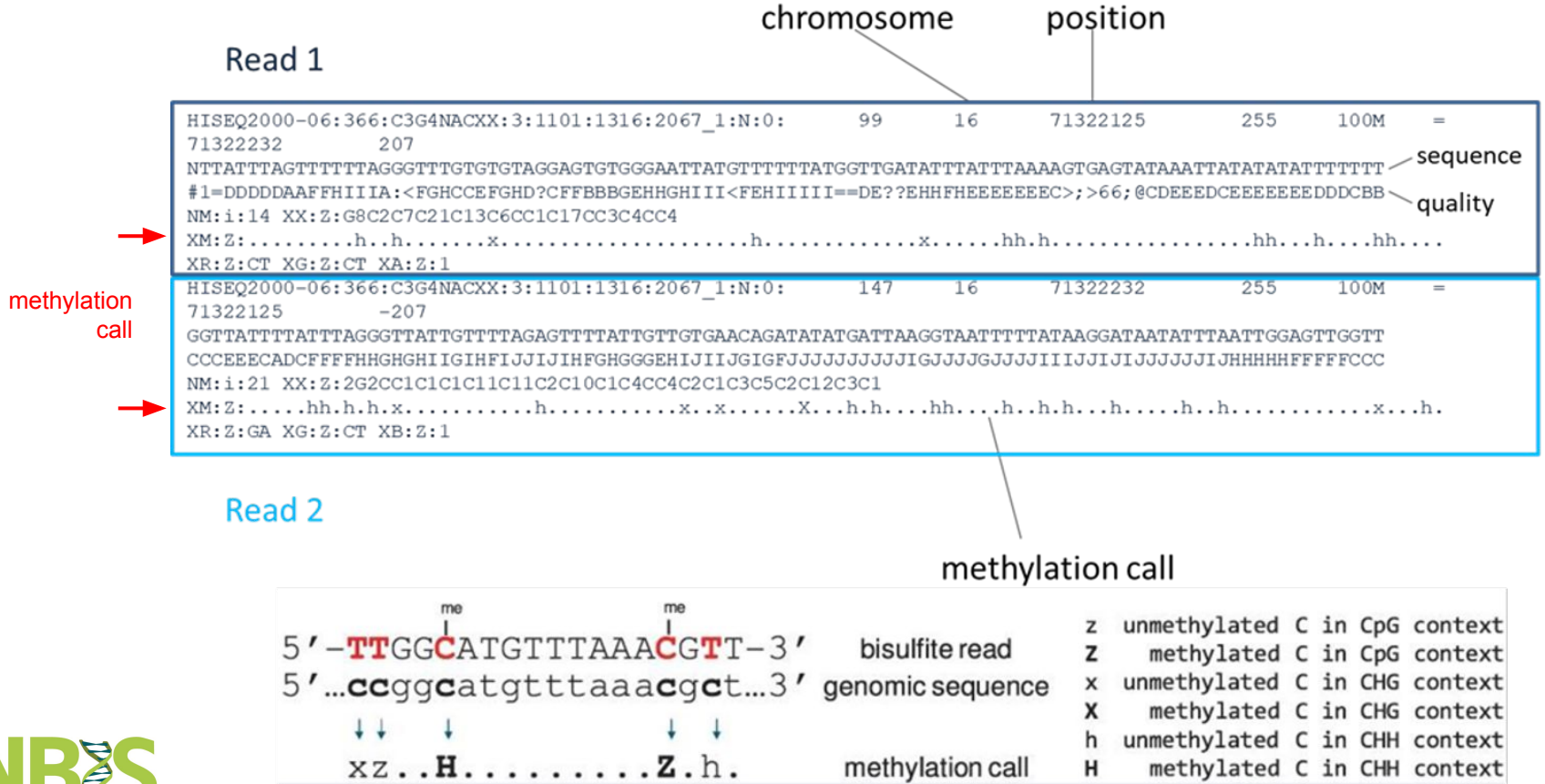
- Used for both array and sequencing

- C → U → (PCR) →T
- mC → C → (PCR) →C



BS-seq output* is a BAM with methylation tags

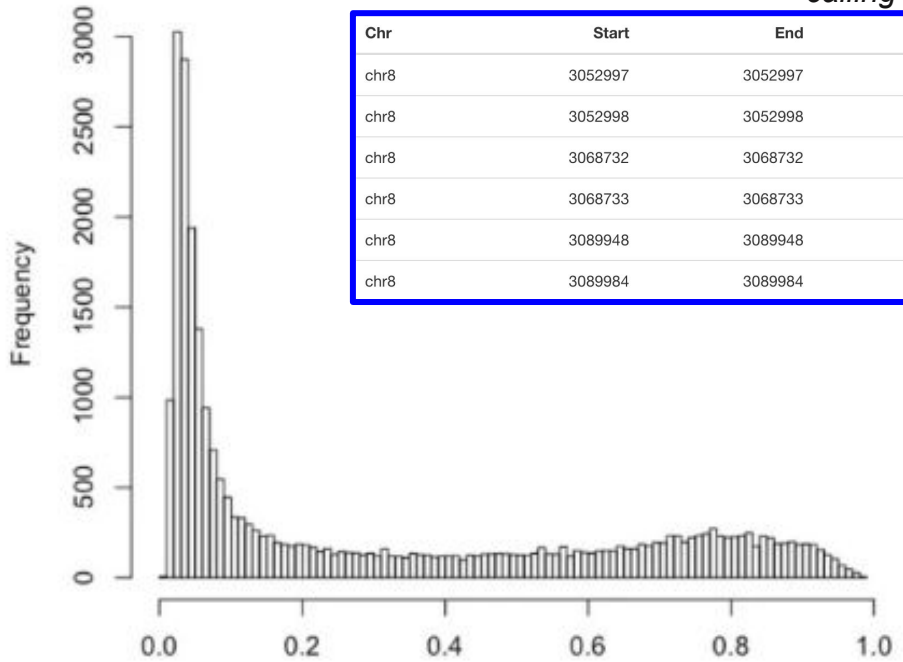
*after alignment



BS-seq output* is %Methylation of a cytosine

**after alignment and methylation calling*

Chr	Start	End	Methylation Prop.	# mC	# C
chr8	3052997	3052997	0.00000	0	1
chr8	3052998	3052998	53.26087	49	43
chr8	3068732	3068732	57.14286	8	6
chr8	3068733	3068733	100.00000	11	0
chr8	3089948	3089948	100.00000	5	0
chr8	3089984	3089984	100.00000	5	0



% Methylation

Pros and Cons of WGBS

✅ Assess 5mC in all context: CpG, CHG, CHH

✅ Covers ~80% of all CpG sites

✅ Lots of data available for reference

❌ BS treatment (⬆️°C, ⬆️pH) degrades DNA

❌ Requires lots of input DNA*
*≥1000ng for conventional BS, ≤100ng for SPLAT

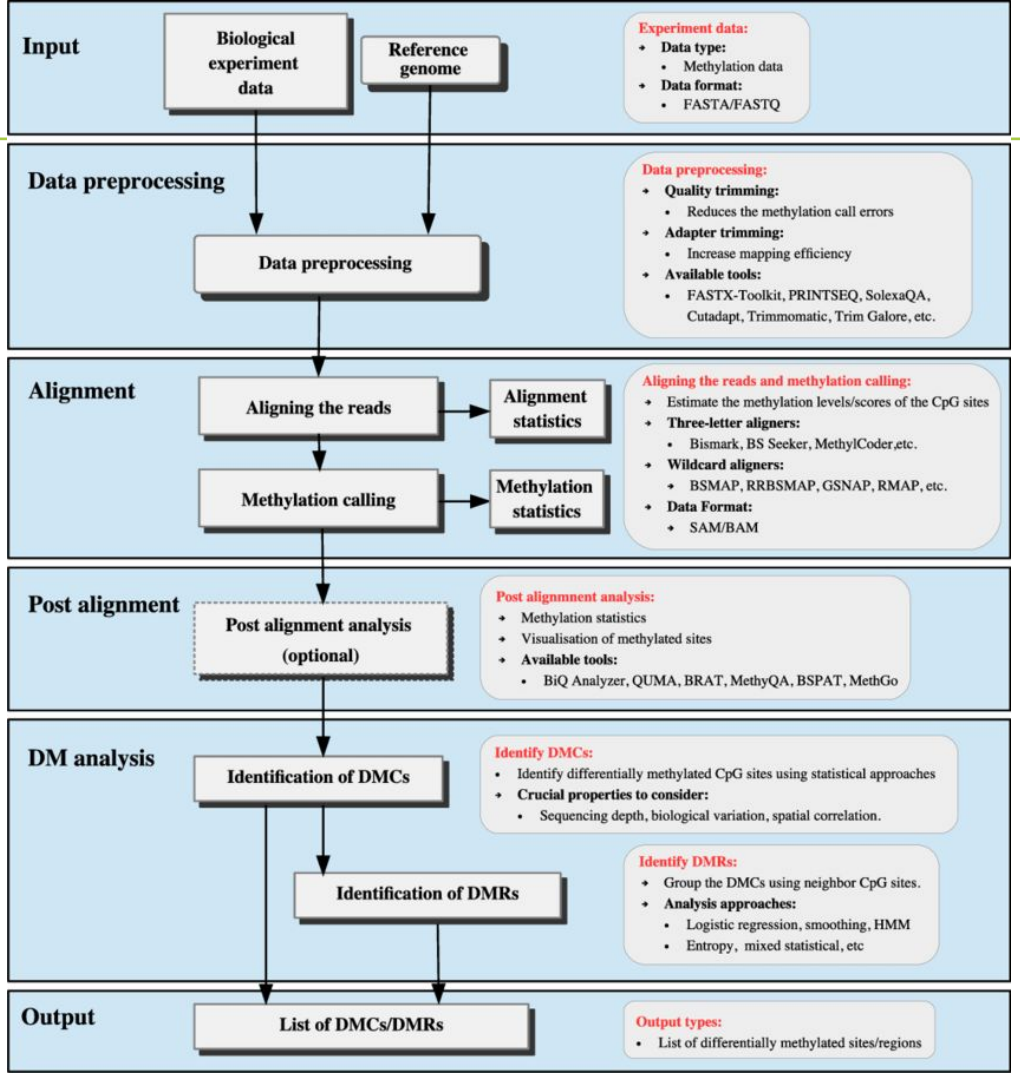
❌ Costly, ~30x coverage recommended for mammalian genome

❌ low coverage in GC-rich region

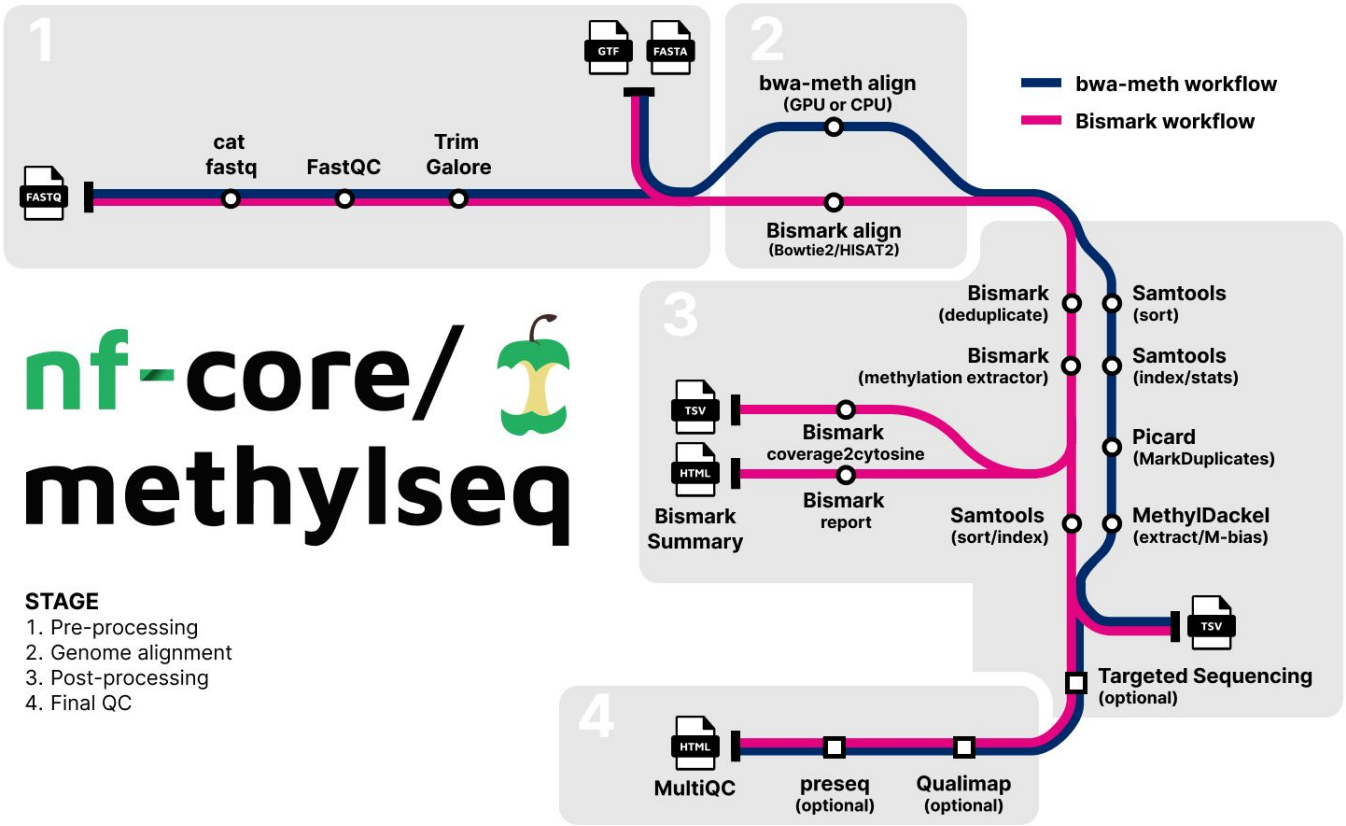
Analysis pipeline

nf-core/methylseq

Exercise for today



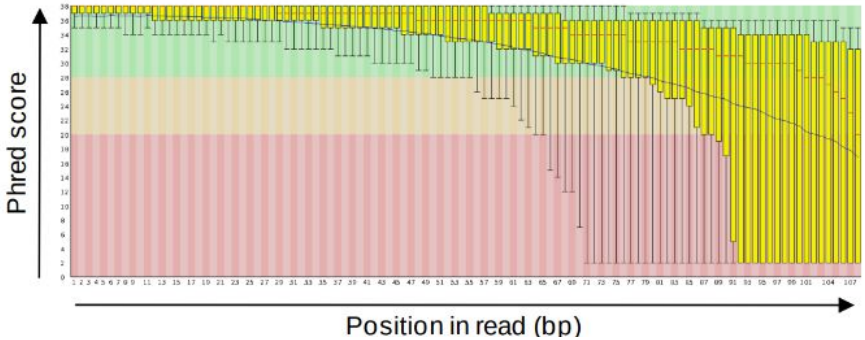
nf-core/methylseq processes BS-seq data to produce alignment and summarised methylation calls



Some notes about WGBS

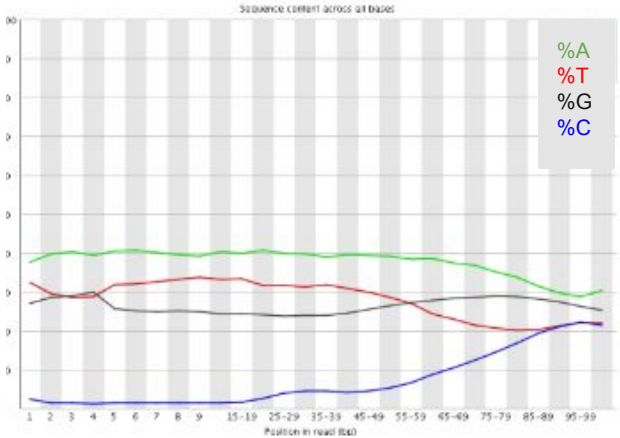
Trim reads of low quality base calls and adapter sequence

- to increase mapping
- to avoid false M calls



Read mapping could be challenging

- BS treatment renders reduced genome complexity
- BS treated reads are very different from the reference sequence
- Up to 4 alignment per locus



Bisulfite conversion efficiency

- non CpG sites in mammalian genome should have > 99.5% conversion in a good experiment
- spike-in controls e.g., phage Lambda

Deduplication

- recommended for WGBS, not for RRBS

Some notes about WGBS

Trim reads of low quality base calls and adapter sequence

- to increase mapping
- to avoid false M calls

Read mapping could be challenging

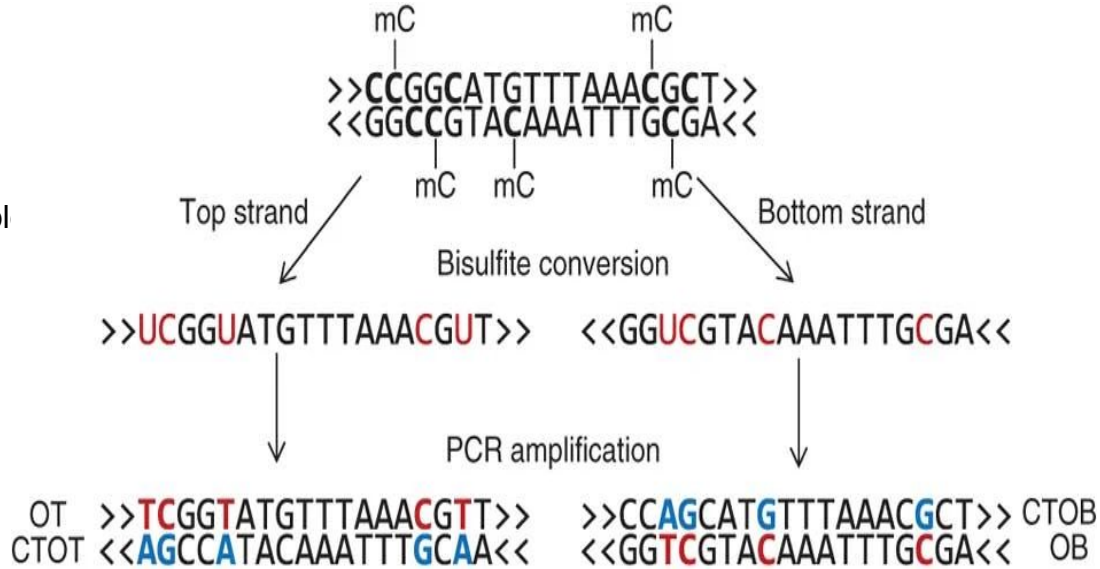
- BS treatment renders reduced genome compl
- BS treated reads are very different from the reference sequence
- Up to 4 read alignment needed per locus

Bisulfite conversion efficiency

- non CpG sites in mammalian genome should 99.5% conversion in a good experiment
- spike-in controls e.g., phage Lambda

Deduplication

- recommended for WGBS, not for RRBS



Some notes about WGBS

Trim reads of low quality base calls and adapter sequence

- to increase mapping
- to avoid false M calls

Read mapping could be challenging

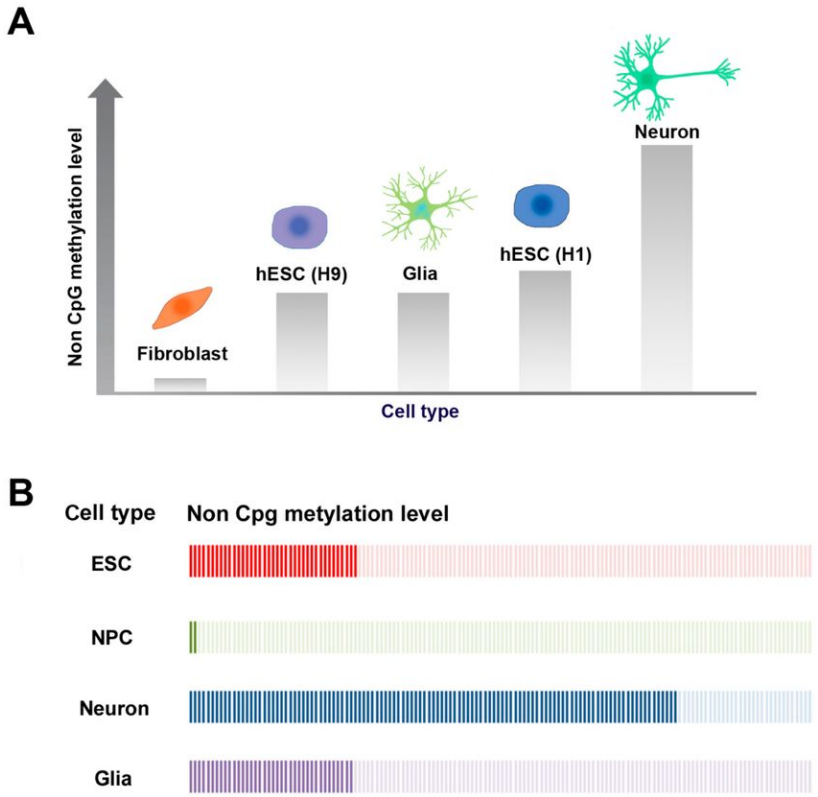
- BS treatment renders reduced genome complexity
- BS treated reads are very different from the reference sequence
- Up to 4 alignment per locus

Bisulfite conversion efficiency

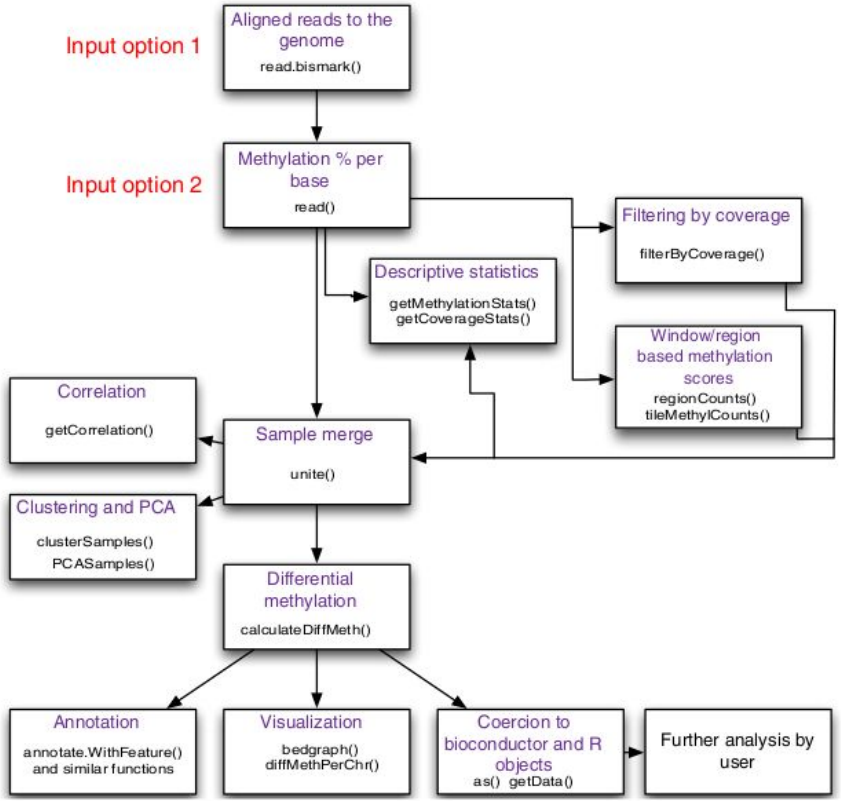
- non CpG sites in somatic tissues (except neuronal cells) of mammalian genome should have > 99.5% conversion in a good experiment
- spike-in controls e.g., phage Lambda

Deduplication

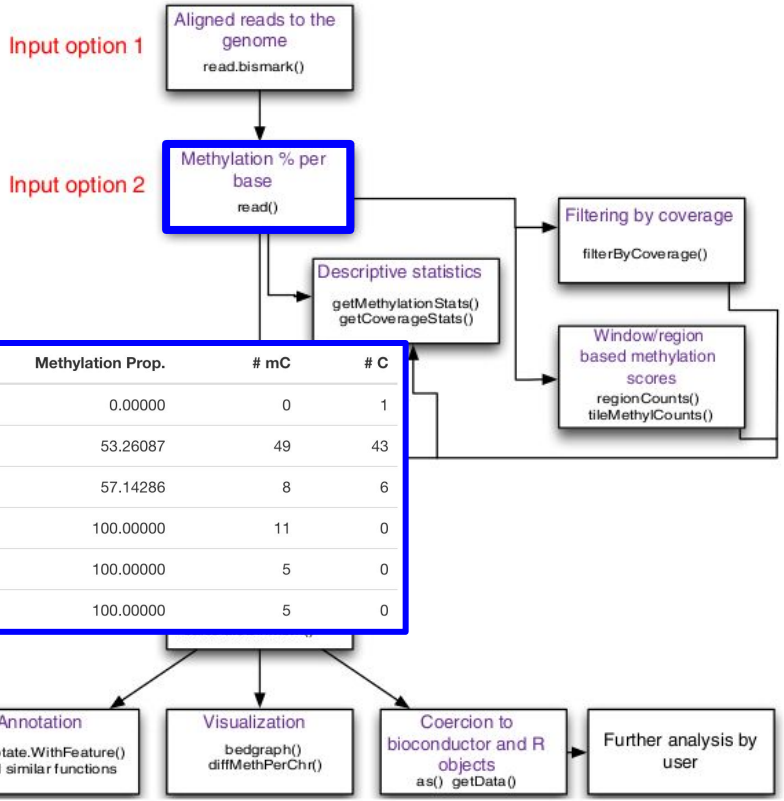
- recommended for WGBS, not for RRBS



Analysis workflow for WGBS data

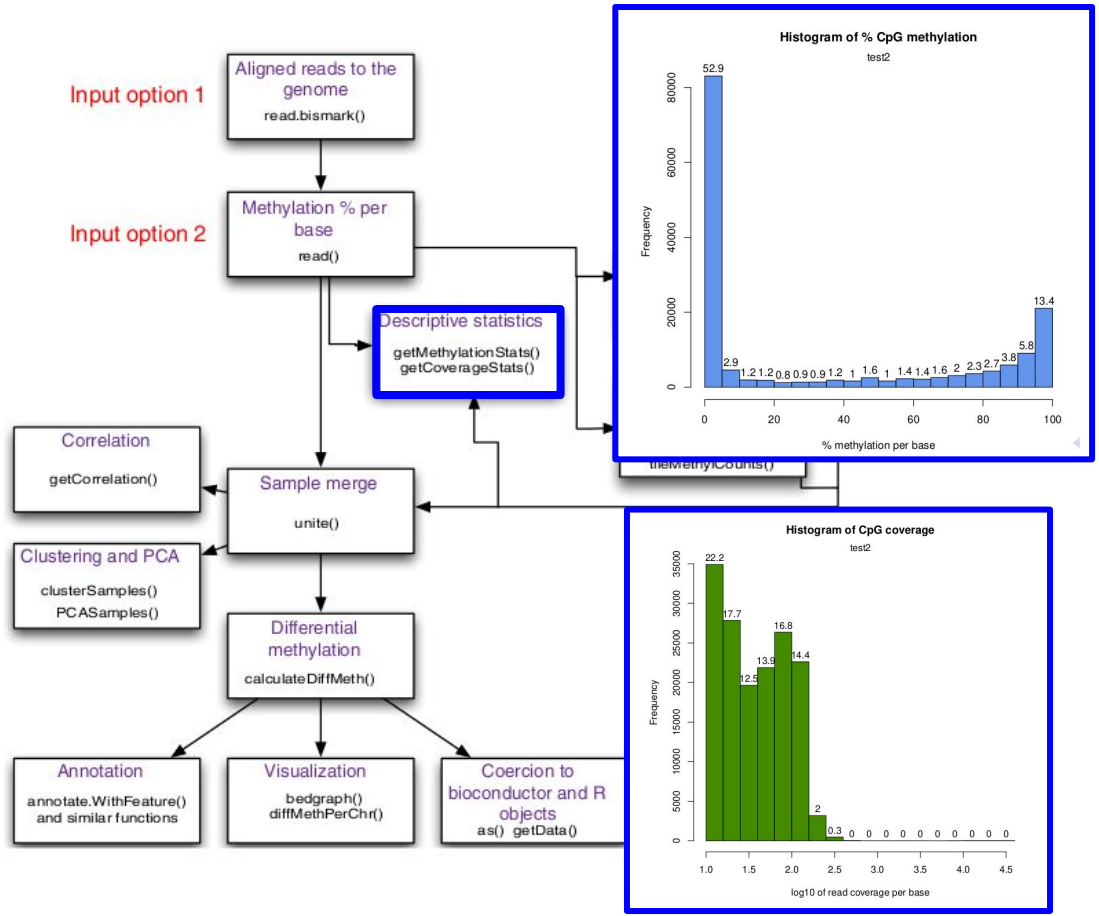


Analysis workflow for WGBS data



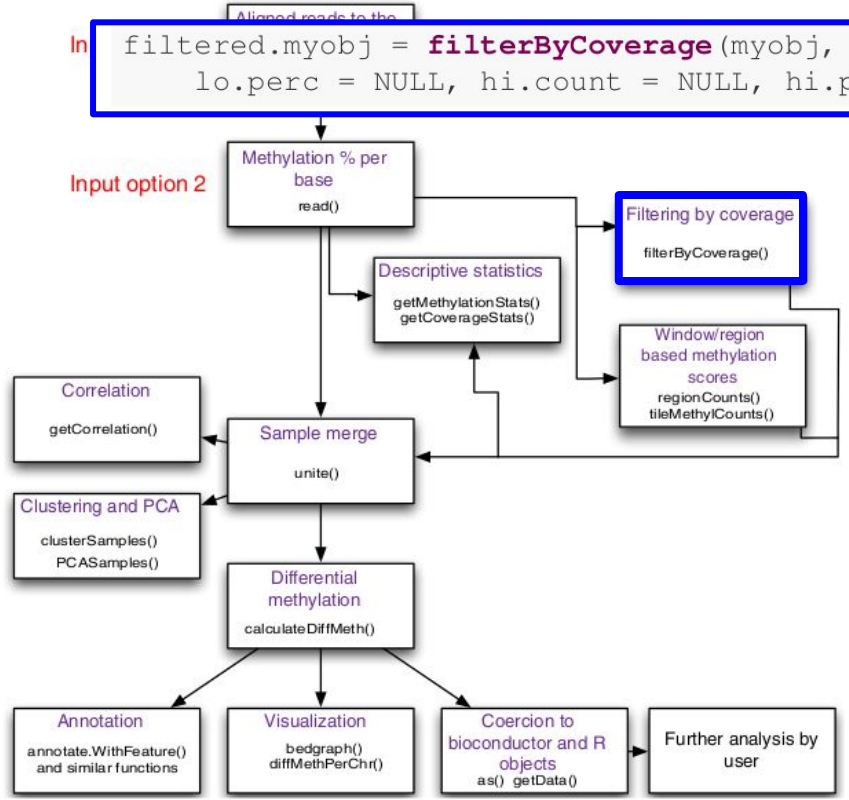
Chr	Start	End	Methylation Prop.	# mC	# C
chr8	3052997	3052997	0.00000	0	1
chr8	3052998	3052998	53.26087	49	43
chr8	3068732	3068732	57.14286	8	6
chr8	3068733	3068733	100.00000	11	0
chr8	3089948	3089948	100.00000	5	0
chr8	3089984	3089984	100.00000	5	0

Analysis workflow for WGBS data

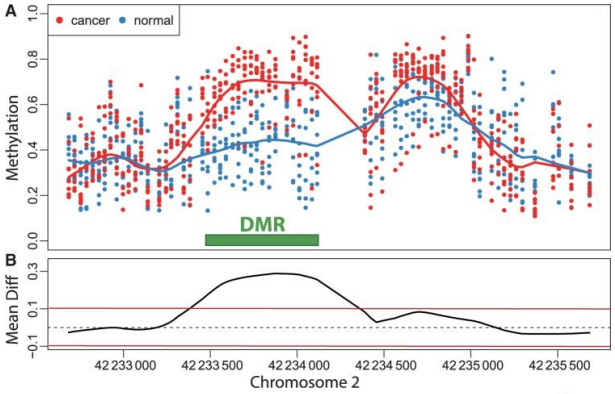


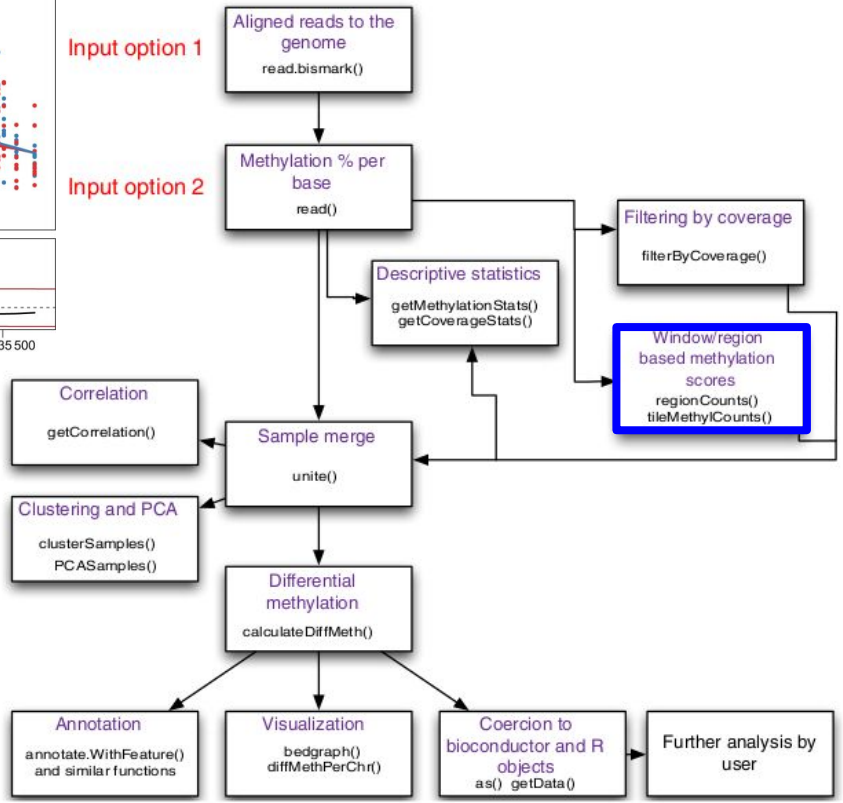
Analysis workflow for WGBS data

```
filtered.myobj = filterByCoverage(myobj, lo.count = 10,
    lo.perc = NULL, hi.count = NULL, hi.perc = 99.9)
```

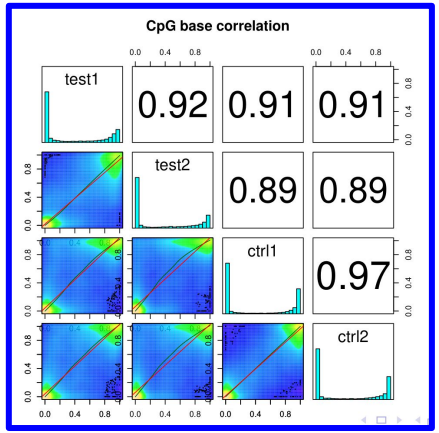


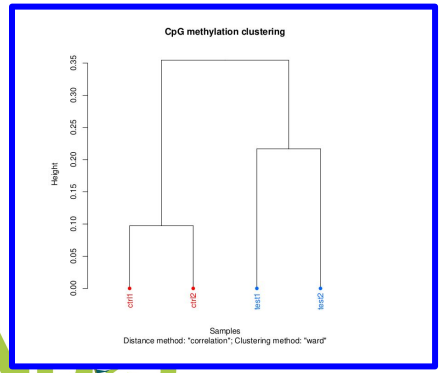
Analysis workflow for WGBS data

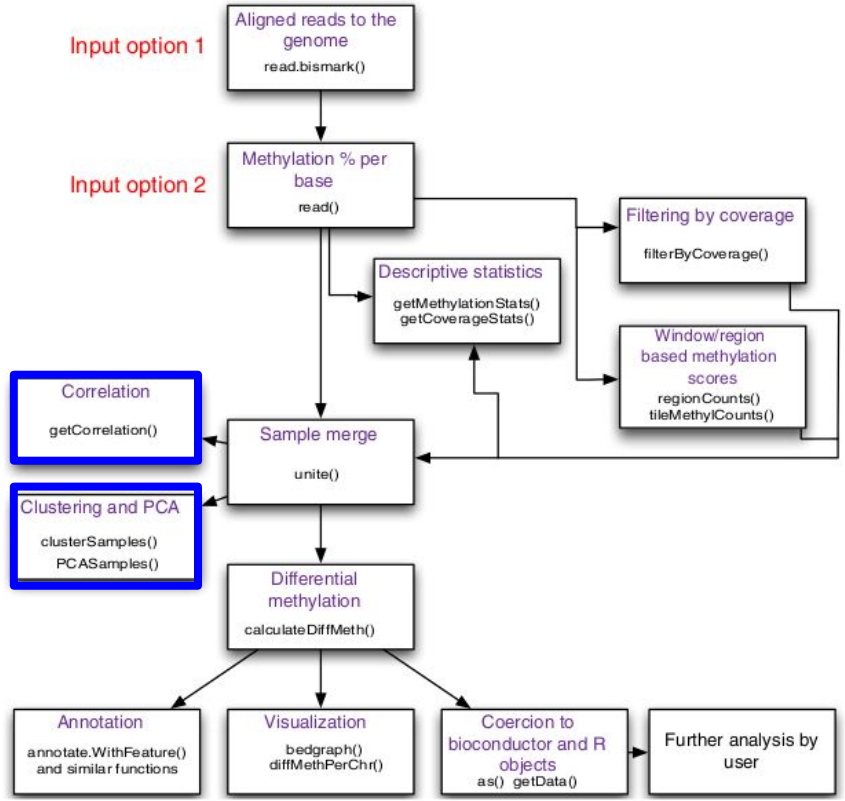


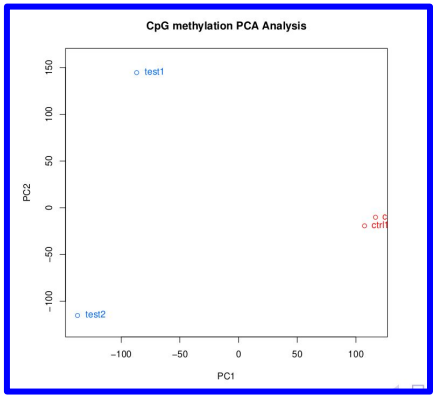


Analysis workflow for WGBS data









Differential Methylation

Remove CpGs with little variation

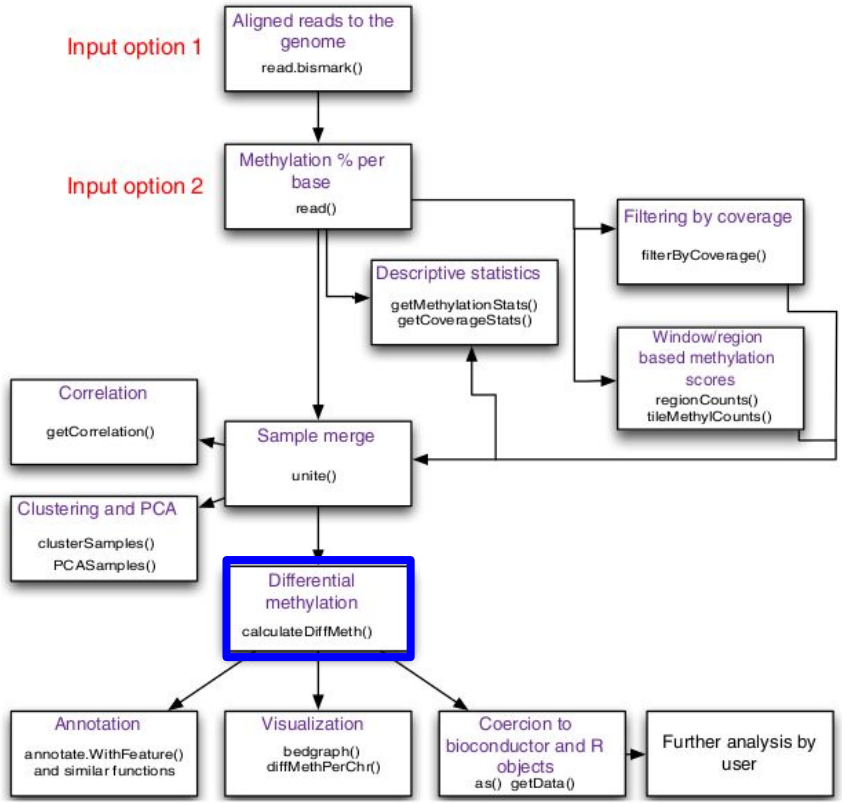
Remove CpGs that overlap C>T SNPs

No replicates: Fisher's exact test

With replicates:

- Logistic regression
- Beta Binomial

Covariates can be included in the model



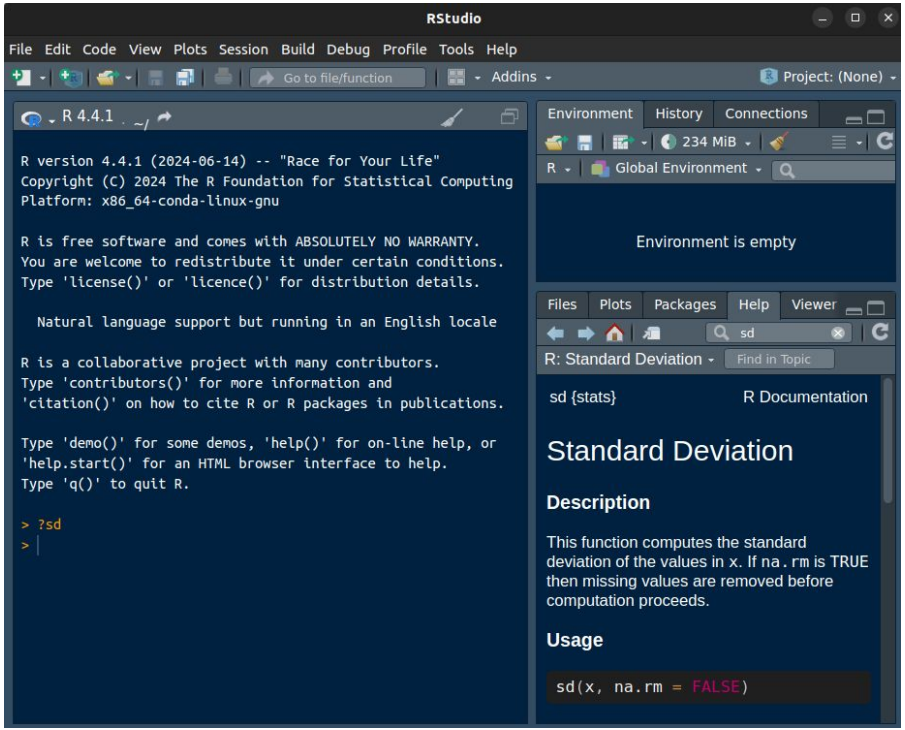
Coding skills from the tutorial

- R coding
- RStudio

- Linux commands

- use resources of an HPC
  - start an interactive compute session and specify its requirements

- launch a Nextflow process



Connect to Rackham

Follow the instruction in Setup

Option A.1

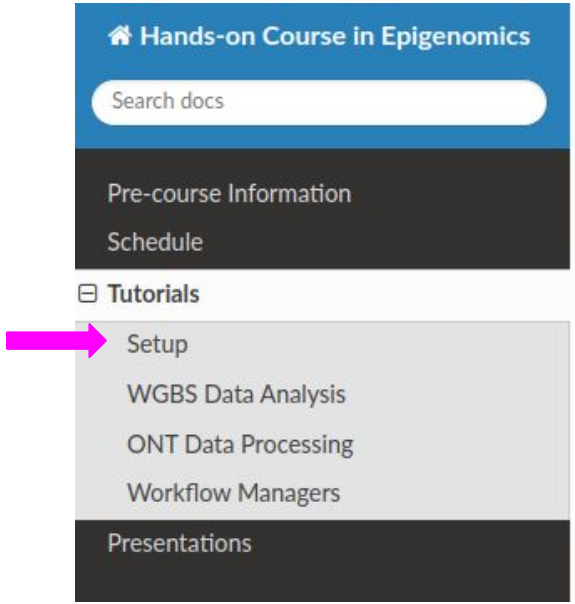
Use module system

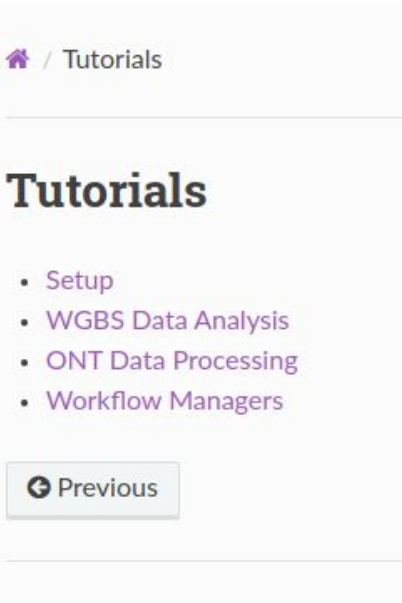
Start RStudio

Option A.2

Use module system

Start R session





Option B

Launch the RStudio server from a container

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

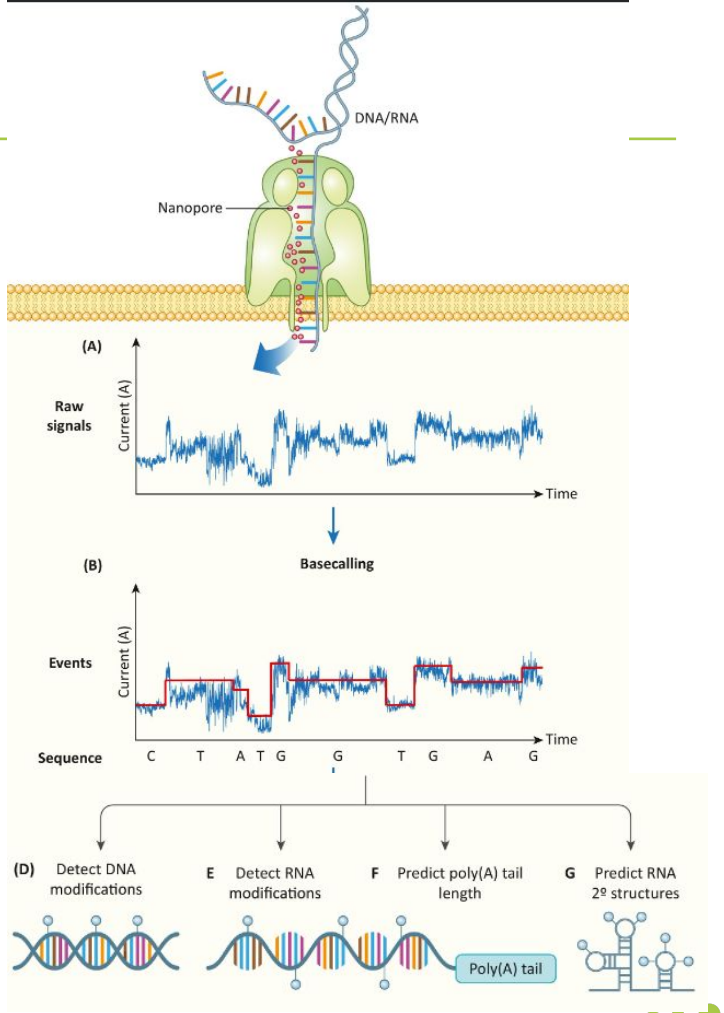
Hands-on Course in Epigenomics: ONT

Louella Vasquez

25-11-14

Oxford Nanopore Technology

- ✅ Direct sequencing of native DNA / RNA
  no PCR amplification
- ✅ Detect DNAM without chemical or enzymatic treatments
- ✅ Distinguish between different types of base modifications
- ✅ Long read sequencing can resolve highly dense GC regions;
  Allele-specific methylation from phasing
- ❌ Requires more input DNA than WGBS to accurately measure DNAM
- ❌ High computational cost and data storage
- ❌ Long turnaround time



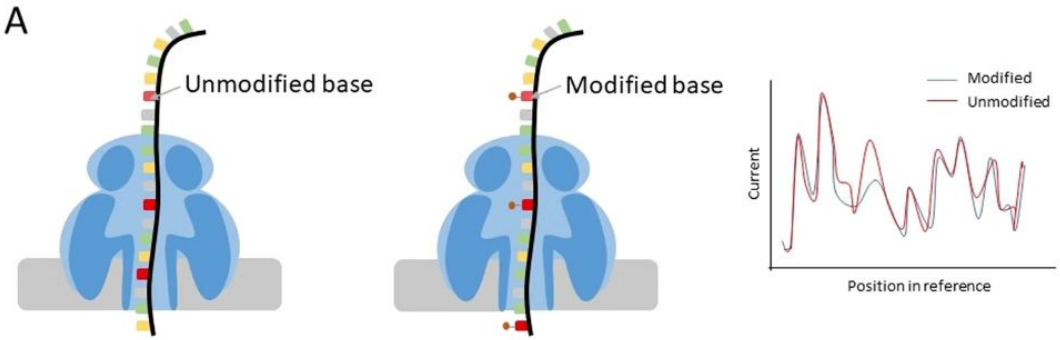
Reference: https://www.cell.com/trends/genetics/fulltext/S0168-9525%2821%2900257-2

Comparison of current methods for DNAM profiling

	Whole genome			Targeted
	ONT	WGBS	EM-seq	EPIC
DNA Input	1–5 µg	1 – 500 ng	10–200 ng	250 ng–500 ng
Single-base Resolution	Yes	Yes	Yes	No
Approximate Run Time	80–84 h	20–24 h	20–24 h	30 min
Yield [Gb]	139	163	137	NA
Sequencing Coverage (x)	34	46	41	NA
Total Reads (M)	7.5	1132.5	986	NA
Number of QC-Passed Reads (M)	NA*	1041.7	976	NA
Percentage of Mapped Reads	90.8%	99.87%	99.99%	NA
Percentage of Mapped Duplicates	0	9.5%	7.0%	NA
Mean Read Length (bp)	16,922	150	150	NA
Longest Read (bp)	856,100	150	151	NA
Number of Called CpGs	56,715,299	53,912,145	54,178,937	865,596
Computational Run Time	Very high	High	High	Low
Complexity of Analytic Pipeline**	High	Medium	Medium	Low
Generated Data Size (GB)	~ 1200	~ 120	~ 70	~ 150mb
Turnaround Time (TAT)	7–12 days	6–10 days	6–10 days	3–4 days

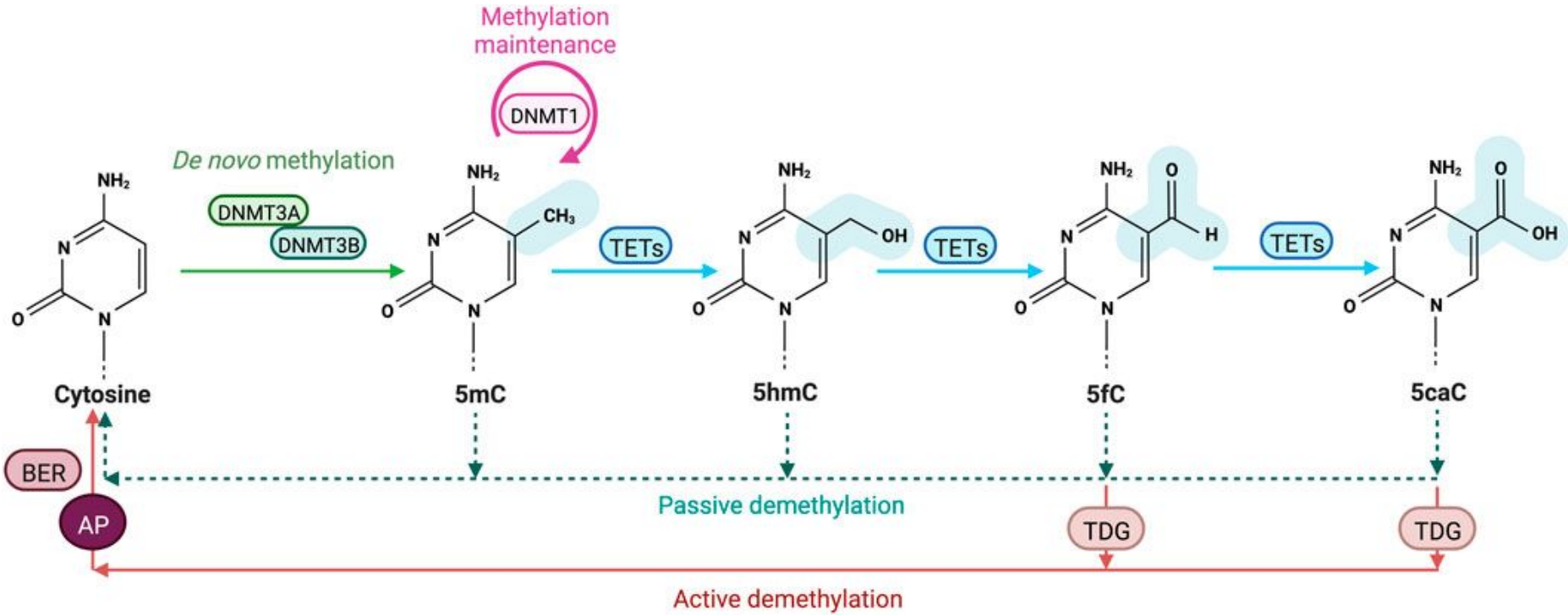
Modified Basecalling

- **5mC**: 5-methylcytosine
- **5hmC**: 5-hydroxymethylcytosir
- **4mC**: N(4)-methylcytosine
- **m6A**: N(6)-methyladenosine



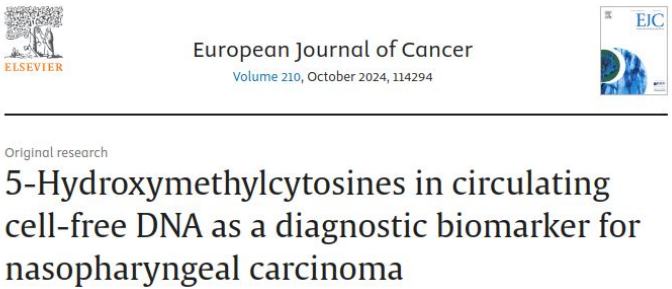
Bacteria exhibit three main types of DNA methylation: 4mC, 5mC, and 6 mA, which regulate crucial processes including DNA replication and repair, virulence factor expression, environmental adaptation, and host-pathogen interactions

5hmC arises during demethylation but it is a stable mark



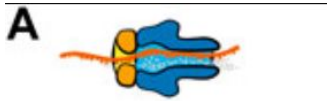
Functions of 5hmC

- Associated with active gene expression and open chromatin
- Cellular differentiation and identity
- role in maintaining genome stability
- 5hmC is abundant in the brain; aberrant 5hmC pattern has been linked to Alzheimer's disease
- Global reduction of 5hmC in many cancers; silencing of tumour suppressor genes



ONT data processing pipeline

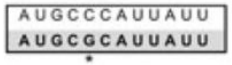
pod5 from ONT run



Basecalling using **`Dorado`**

Quality control using **`pycoQC`**

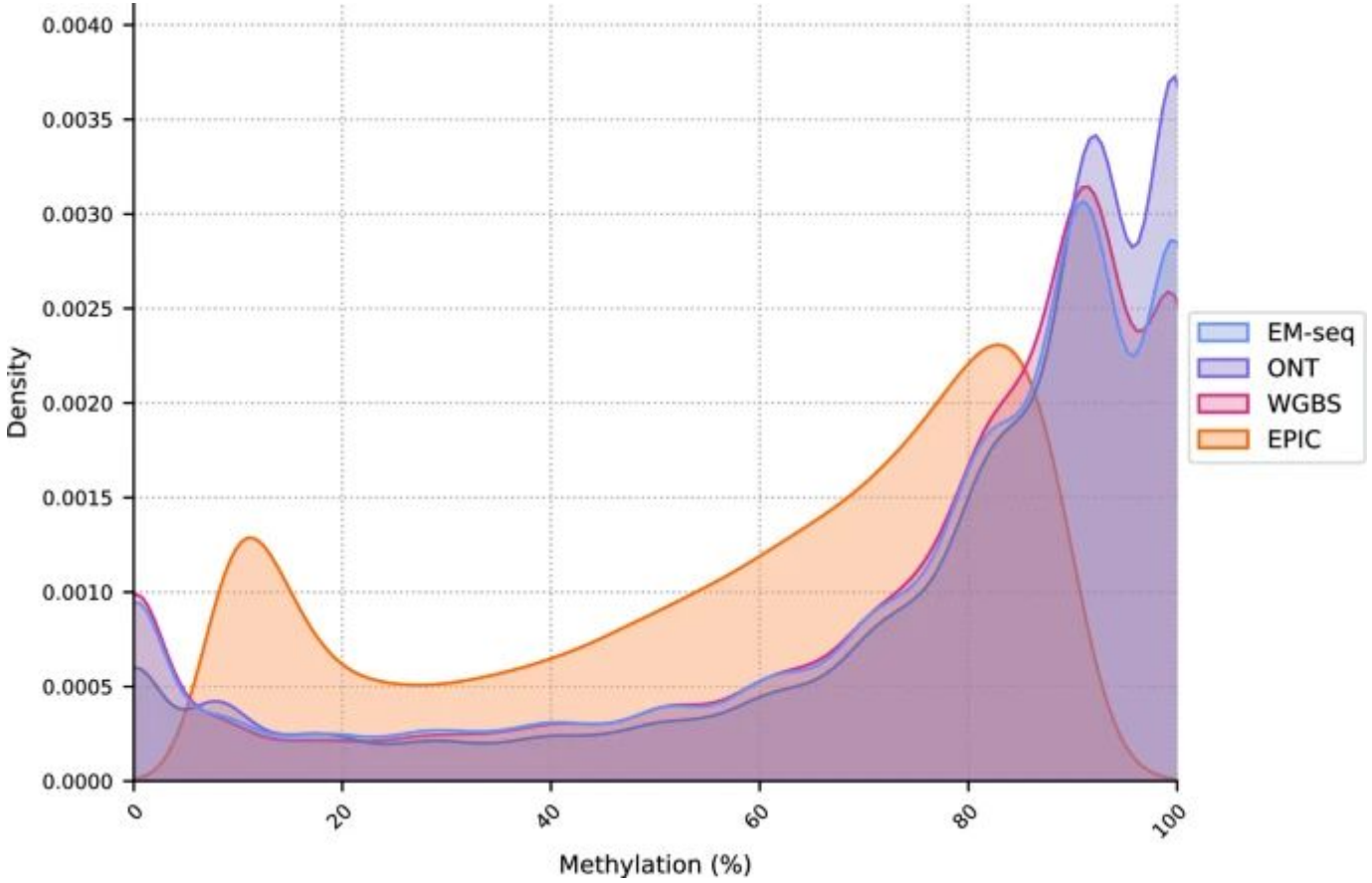
Align reads using **`minimap2`**



Pileup reads for M calling **`modkit`**

Mod base	strand	Nvalid_cov	%mod	Nmod	Ncanonical	Nother_mod
a	.	22	13.64	3	19	0
h	.	4851	0.41	20	22	4809
m	.	4851	99.13	4809	22	20

Modkit pileup output is %Methylation of a cytosine

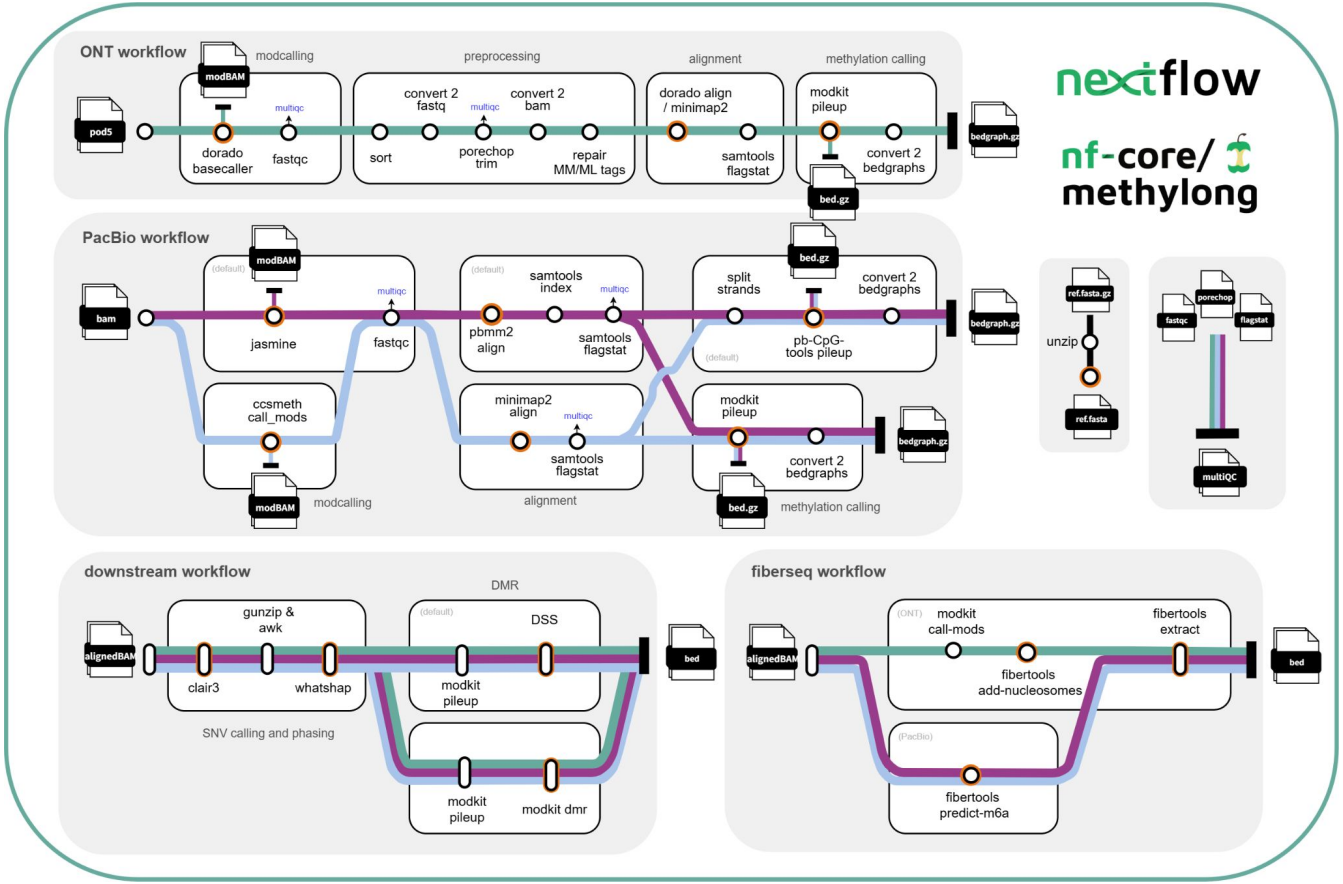


Alternative workflow: epi2-me wf-basecalling

Nextflow workflow can be used to perform:

- Basecalling of a directory of pod5 or fast5 signal data
- Basecalling in Duplex mode
- Modified basecalling
- Basecalling in real time
- Output basecalled sequences in various formats: FASTQ, CRAM or Unaligned BAM
- If a reference is provided a sorted and indexed BAM or CRAM will be output for basecalling a directory of pod5 or fast5 signal data with dorado and aligning it with minimap2 to produce a sorted, indexed CRAM.

Alternative workflow: nf-core/methylong



Coding skills from the tutorial

- Setup a project directory
- Linux commands
- bash command
- use resources of an HPC
  - Submit a computing job to slurm with GPU core requirement
- IGV for visualisation
- launch a Nextflow process

```
#!/bin/bash -l
#SBATCH -A uppmax2025-2-309            # Replace with your NAISS project name
#SBATCH -p gpu                         # Request a GPU partition or node
#SBATCH --gres=gpu:1                   # Request generic resources  of 1 gpu
#SBATCH -t 24:00:00                    # Set a limit of the total run time, format is days-hours:minutes:seconds
#SBATCH -J DORADO                      # Specifies name for the job
#SBATCH -e DORADO_%j_error.txt         # output file for the bash script standard error
#SBATCH -o DORADO_%j_out.txt           # output file for the bash script standard output


# location of a precompiled dorado binary
dorado="/proj/uppmax2025-2-309/nobackup/ngi-epigenomics/tools/dorado-1.1.0-linux-x64/bin/dorado"
#
# location of a precompiled modkit binary
modkit="/proj/uppmax2025-2-309/nobackup/ngi-epigenomics/tools/dist_modkit_v0.5.1_8fa79e3/modkit"
#
# location of a precompiled pycoQC binary
pycoQC="/home/louel/.conda/envs/pycoQC/bin/pycoQC"
#
# load samtools - latest version
module load SAMtools



# input raw POD5 file. CHANGE to your project folder!
inpod5="/proj/uppmax2025-2-309/nobackup/ngi-epigenomics/students/louella/data/modbase-validation_2024.10/subset/5mC_rep1.pod5"
#
# reference genome in fasta format.  CHANGE to your project folder!
reffasta="/proj/uppmax2025-2-309/nobackup/ngi-epigenomics/students/louella/data/modbase-validation_2024.10/references/all_5mers.fa"


# specify the output directory to store the output files.  CHANGE to your project folder!
outputdir=/proj/uppmax2025-2-309/nobackup/ngi-epigenomics/students/louella/output
#
# specify the output filename
outputbam=$(echo $inpod5 | xargs basename -s .pod5)
outputbam="hac.$outputbam"
echo "Saving output file to .... $outputbam"
echo
sleep 3s



#
#
# 1.
# run dorado basecaller command
# output is unaligned BAM file
$dorado basecaller hac,5mC_5hmC $inpod5 > $outputdir/$outputbam.unaligned.bam
```

Pelle

Open a terminal and log in to Pelle by ssh.