Workflow-BS: an integrative workflow for RRBS and WGBS data

Gaëlle Lefort Céline Noirot

INRA

April 8, 2016

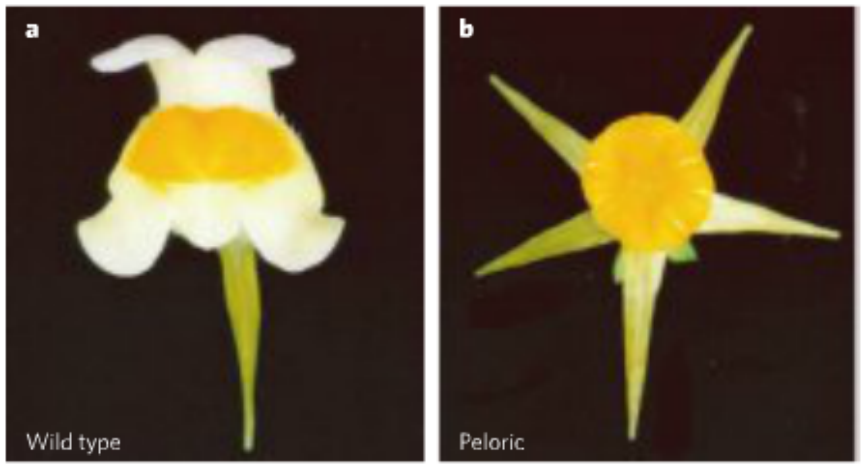
Summary

Summary

Epigenetics

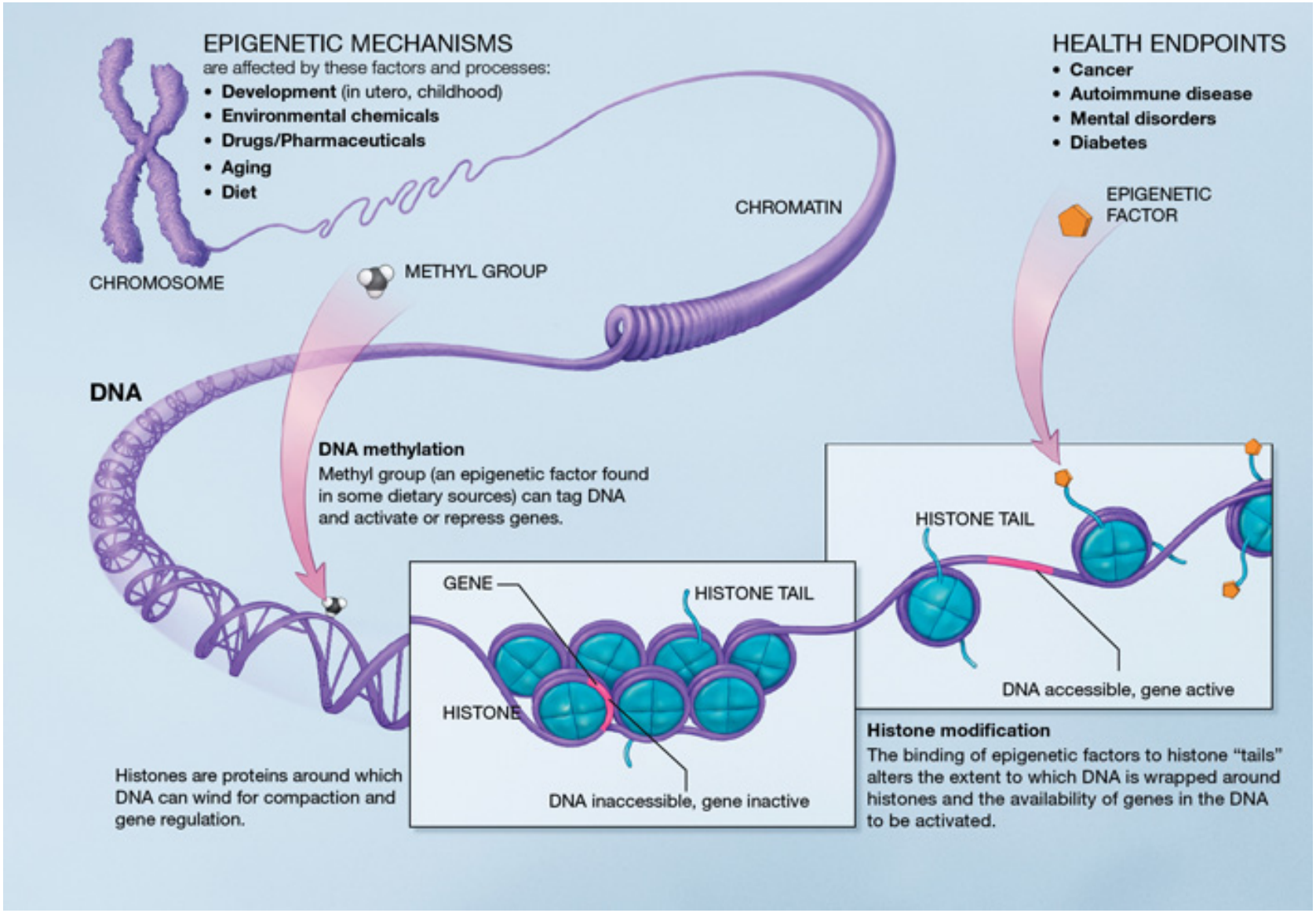
The term epigenetics refers to heritable changes in gene expression (active versus inactive genes) that does not involve changes to the underlying DNA sequence; a change in phenotype without a change in genotype

Morgan et al, 1999

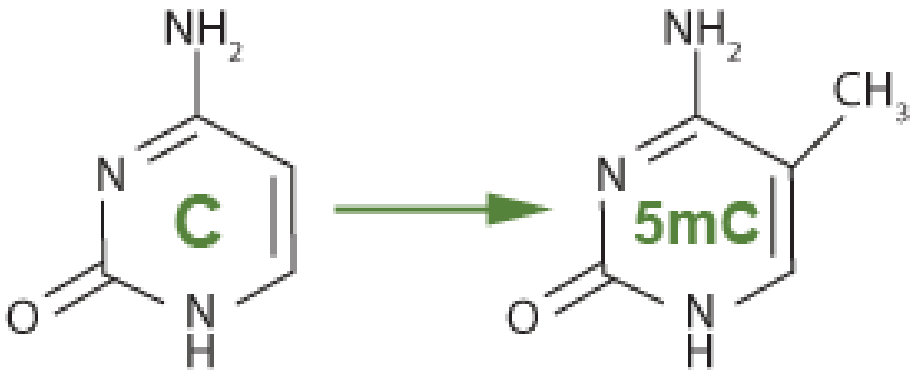


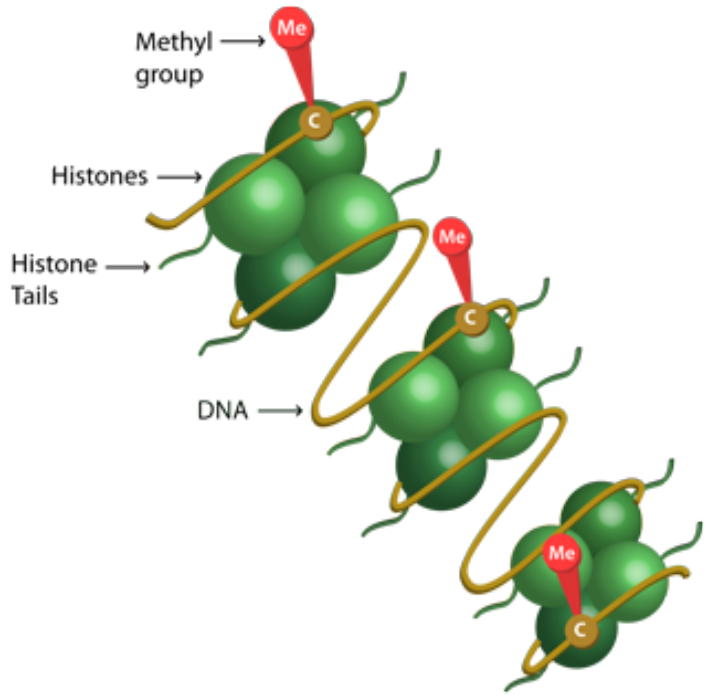
Cubas et al, 1999

Epigenetics



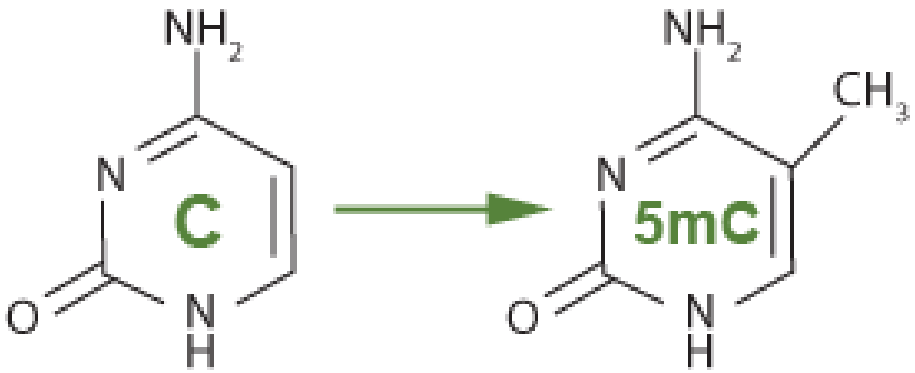
DNA methylation

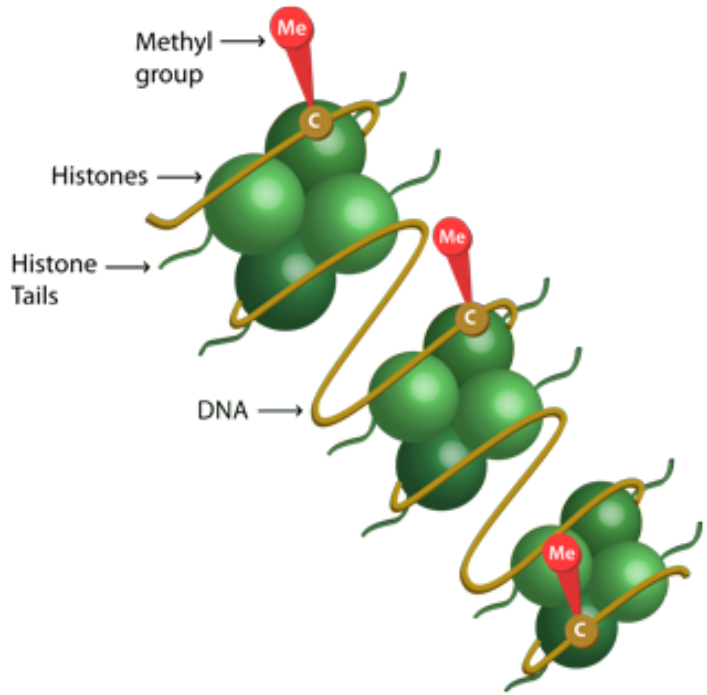




1. The predominant epigenetic modification of DNA in mammalian genomes is methylation of cytosine nucleotides (5-MeC).

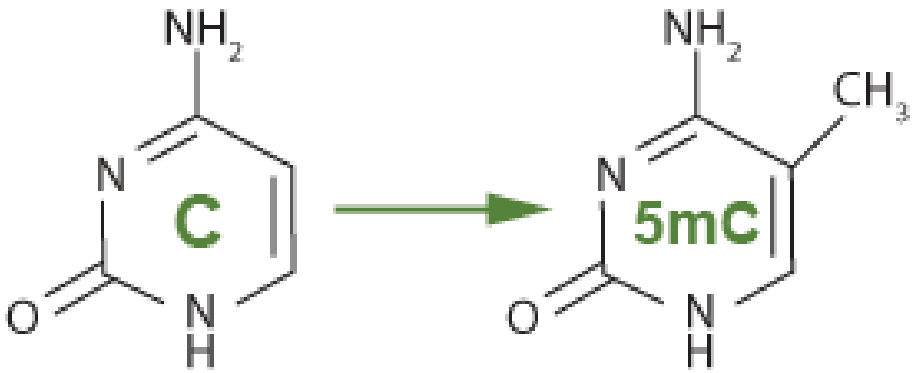
DNA methylation

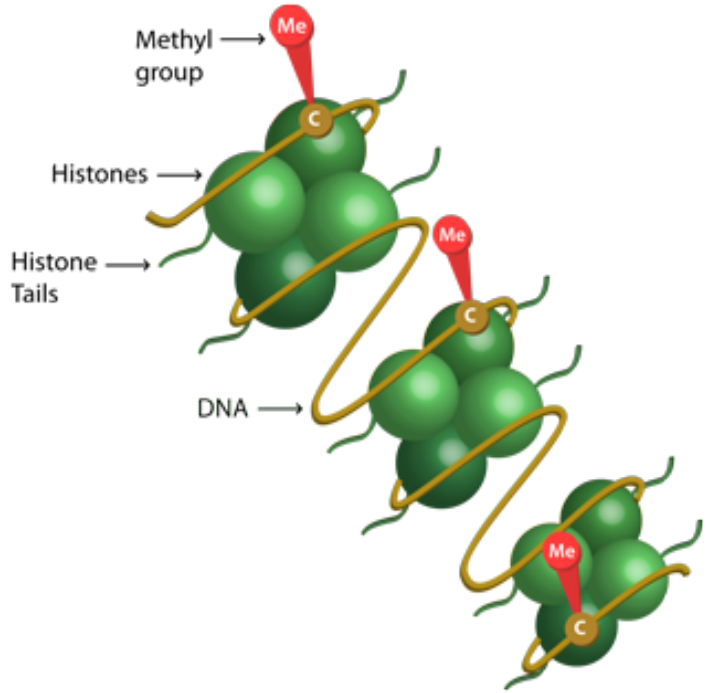




1. The predominant epigenetic modification of DNA in mammalian genomes is methylation of cytosine nucleotides (5-MeC).
2. Mammalian main modification is on 5'-CpG-3' dinucleotides

DNA methylation

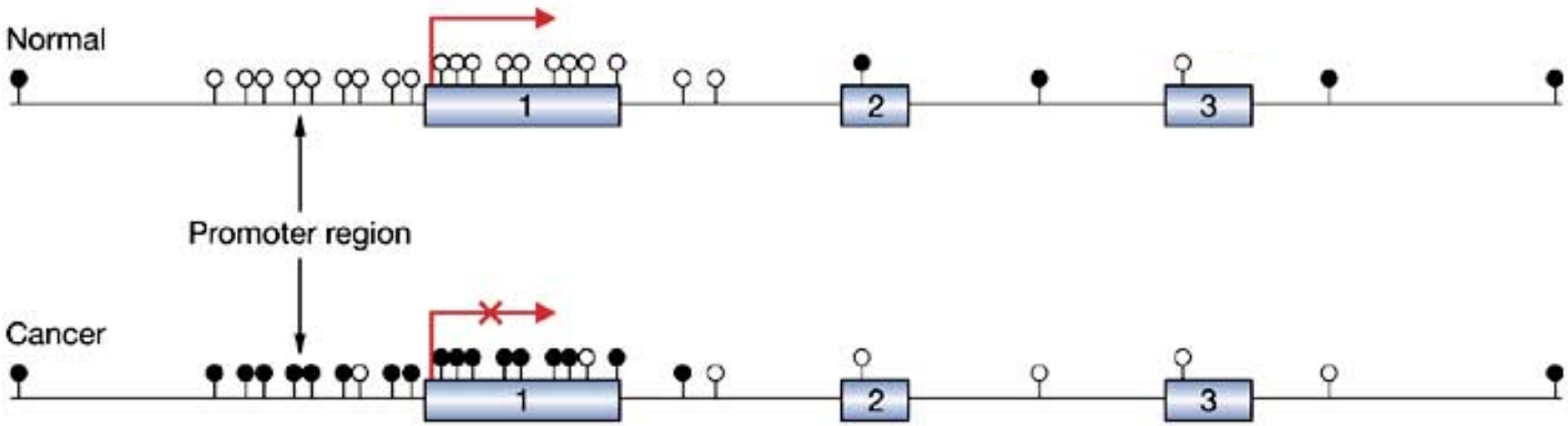




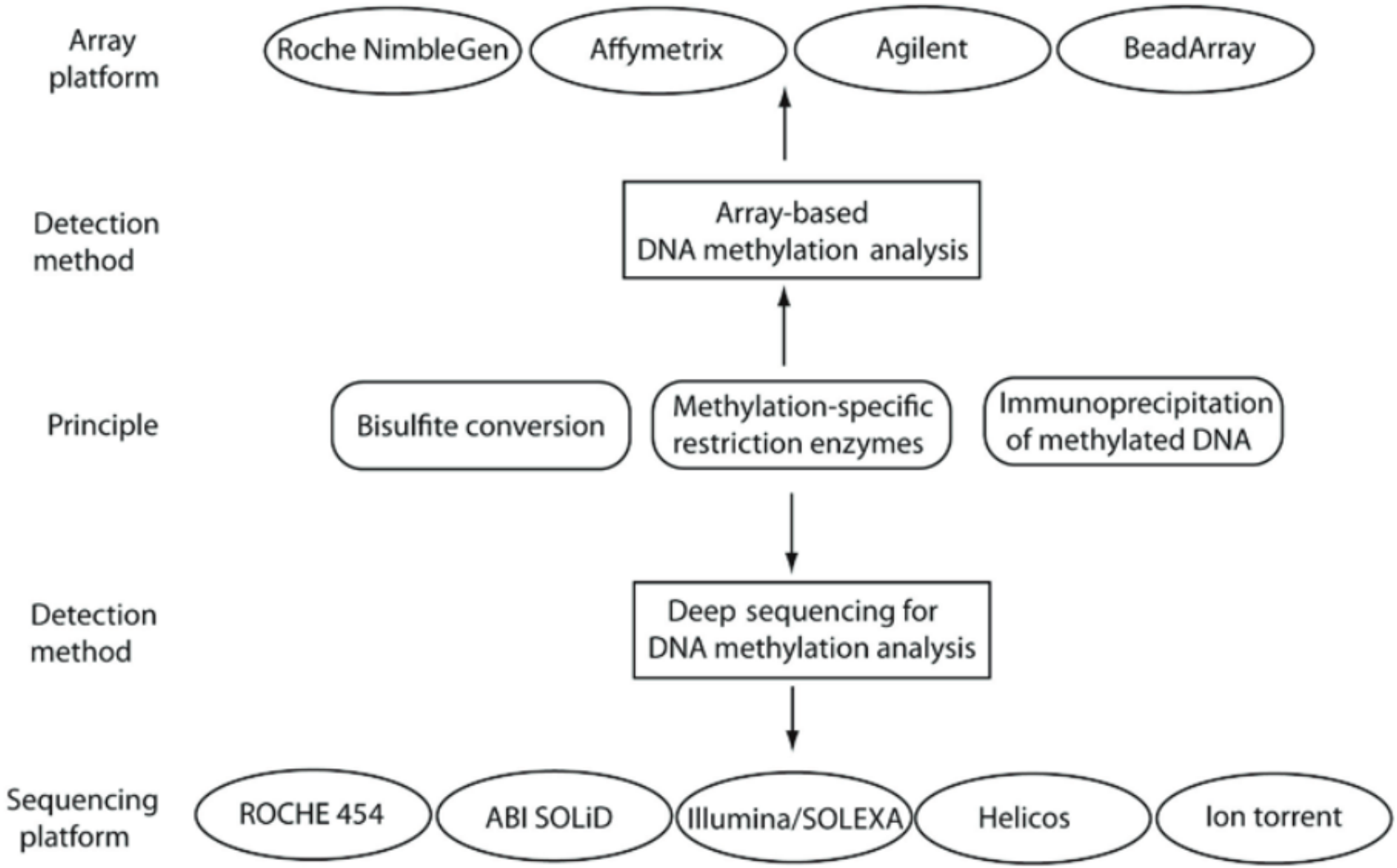
1. The predominant epigenetic modification of DNA in mammalian genomes is methylation of cytosine nucleotides (5-MeC).
2. Mammalian main modification is on 5'-CpG-3' dinucleotides
3. In plant also occurs on CHG and CHH context (where H can be A,C or T)

https://www.epigentek.com/catalog/dna-methylation-c-75_21.html

DNA methylation

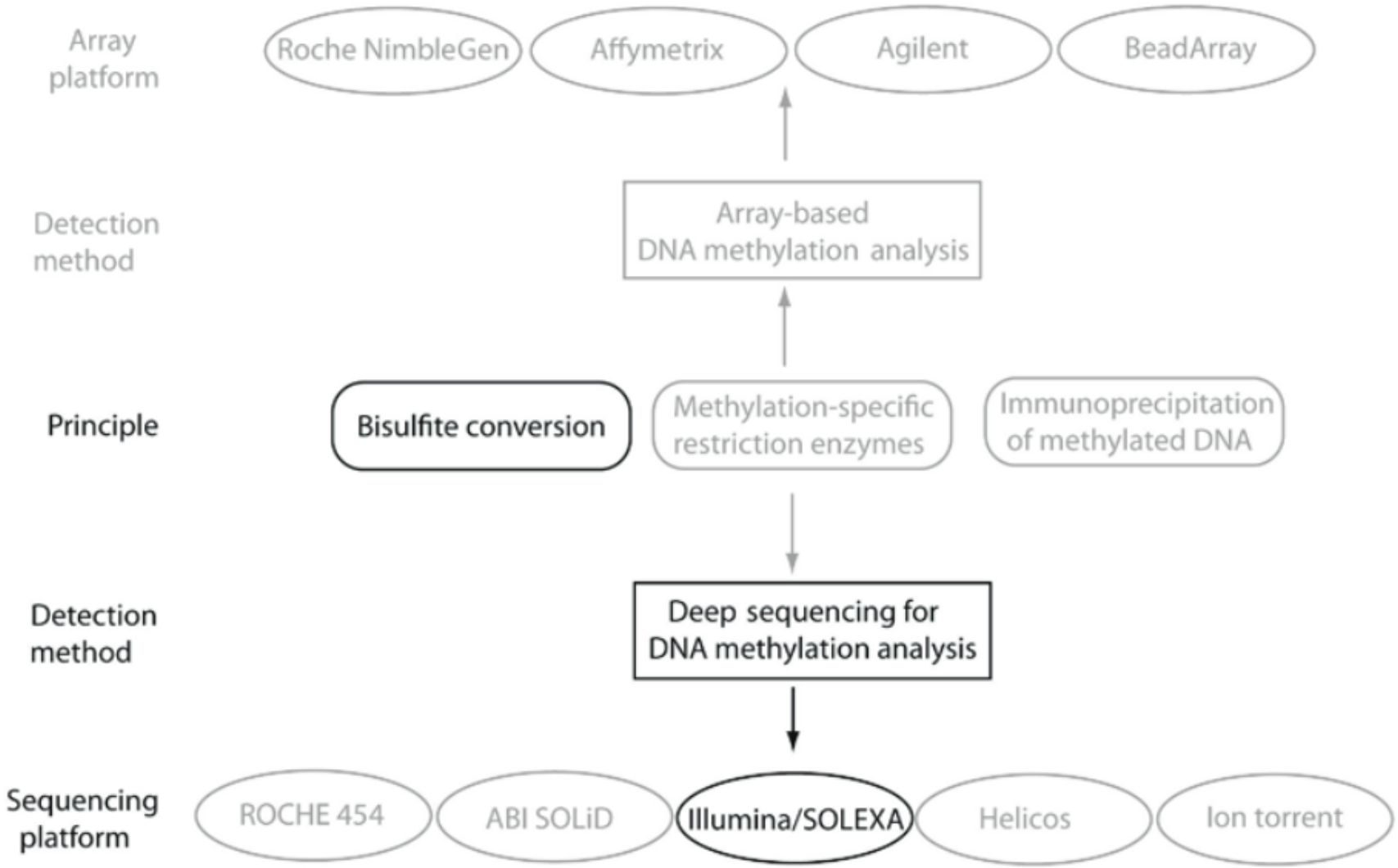


DNA methylation analysis methods



Gupta et al, Review BioTecnique, 2010

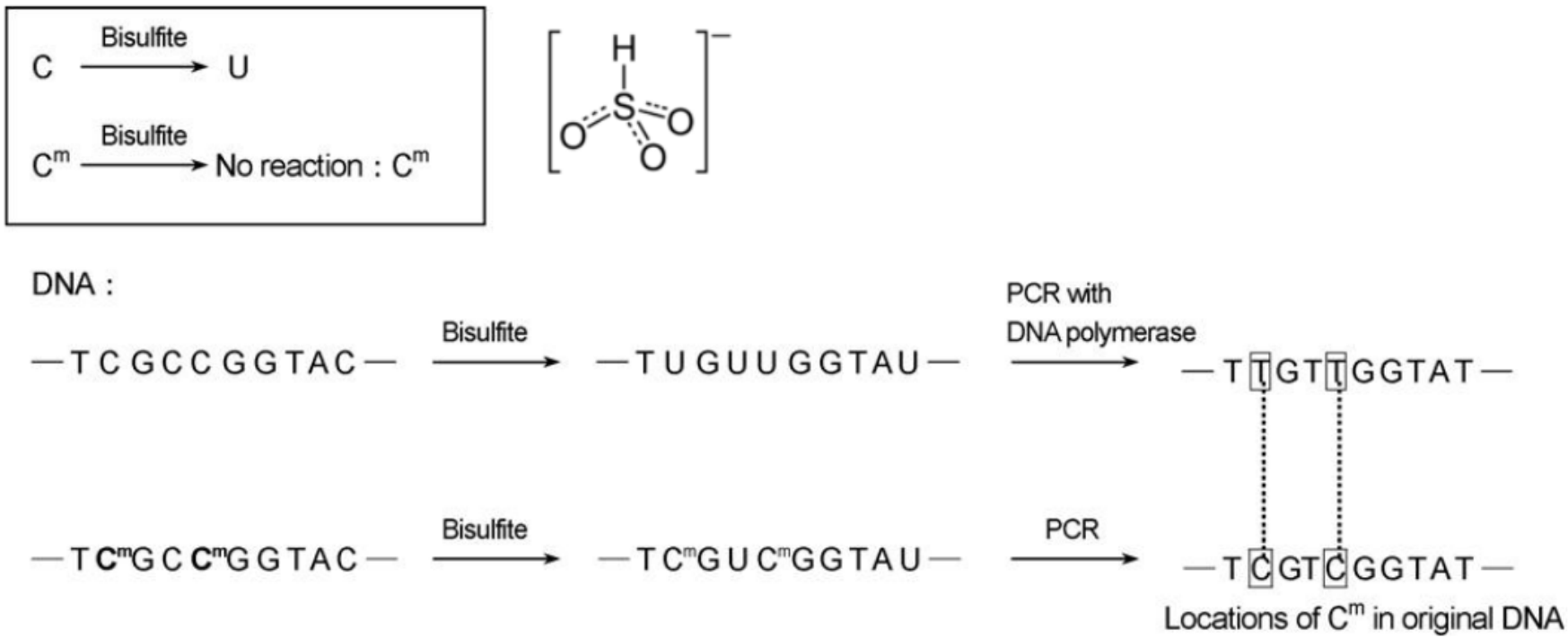
DNA methylation analysis methods



Gupta et al, Review BioTecnique, 2010

Bisulfite treatment

What is bisulfite treatement?

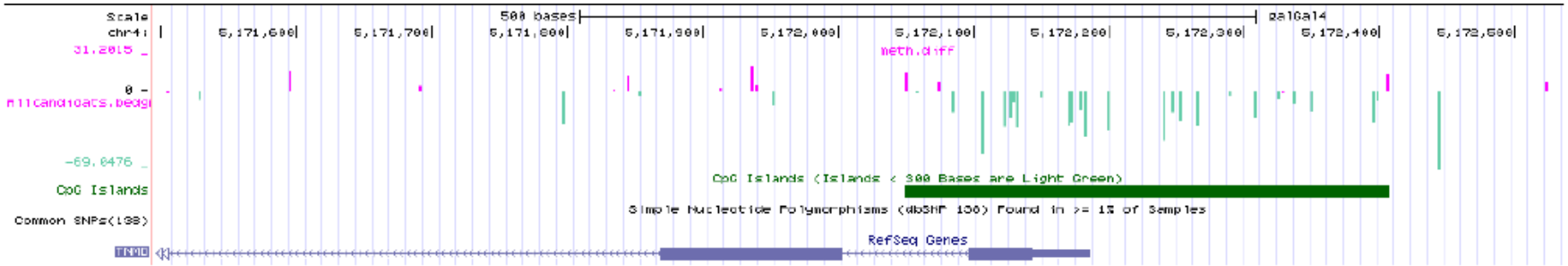


https://www.promega.com/~/media/files/promega%20worldwide/north%20america/promega%20us/webinars%20and%20events/epigeneticwebinarsept2012.pdf

Summary

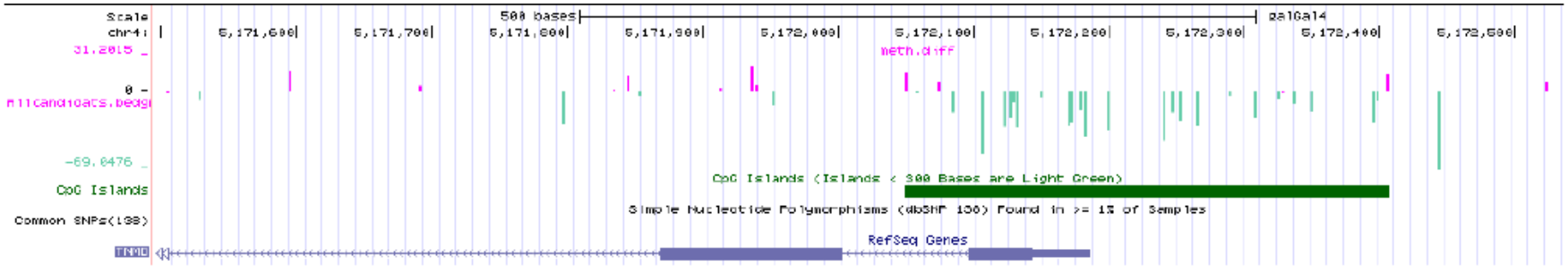
Purpose

Identification of differential methylated regions.



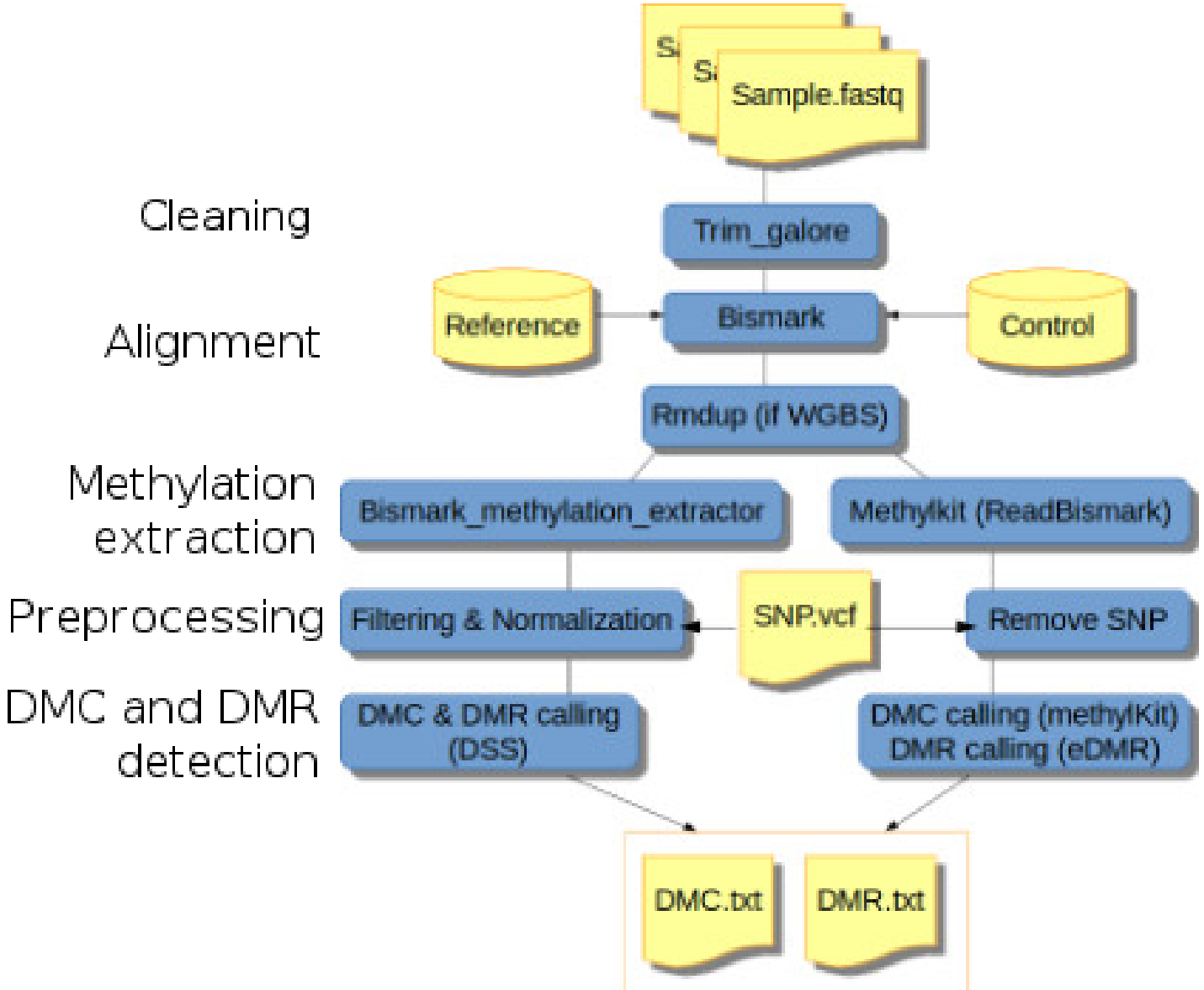
Purpose

Identification of differential methylated regions.



- Projects: Epitherm/Epibird (GenPhySE), Epinod (LIPM), FrAG-ENCODE (SelGen), Mitomics (submitted)
- Kind of data: WGBS, sequence capture, soon RRBS
- Community: France Genomic Methylation WG, FAANG
- Needs: bioinformatics and biostatistics tools

Pipeline overview



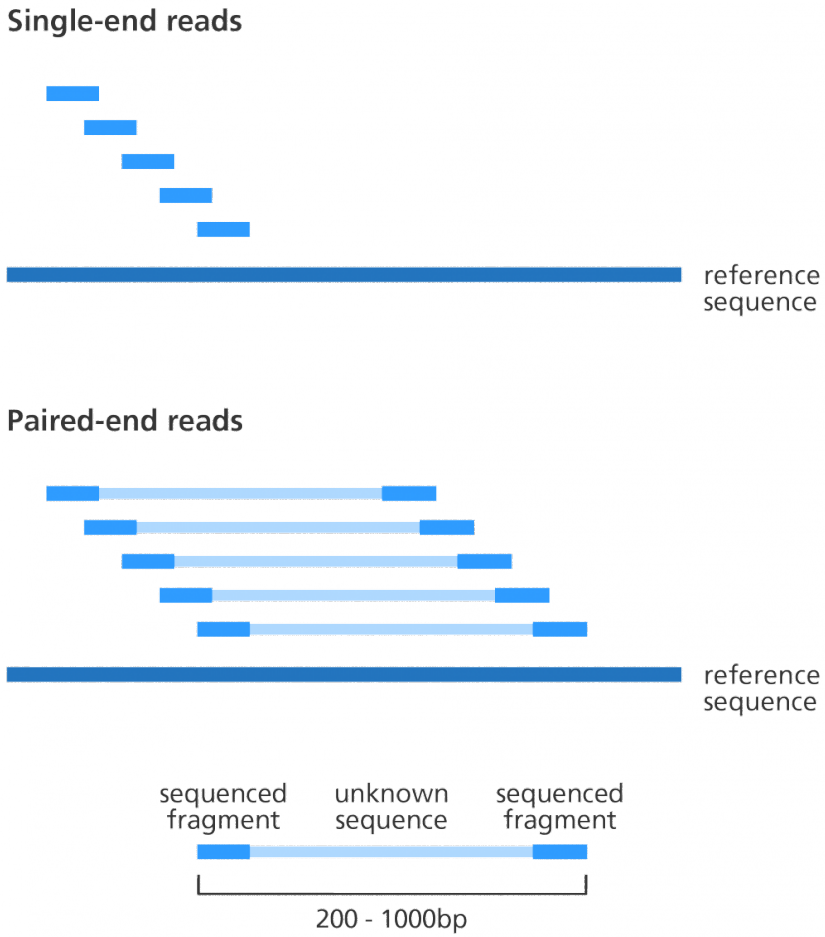
Supported data

- Fastq files from illumina sequencing
  - Bisulfite treated
  - For example a Hiseq 2500 run :
    - ★ up to 2 billion single read or 4 billion paired-end reads
    - ★ read length: from 36bp up to 150bp

```
@HWI-ST314:257:D1WGCACXX:2:1101:1130:51866_1:N:0:TCGAAG
TATGAAAATAAGGTGGATAAGTGTTTTTTGATGTTTATTAGTGATATGTTATTTGAGTTTTGAGTATTTGTATTAGAAAGTTAGATTGATTATTTTAGTT
+
B@@DDFFFHGHHFGHIEHIJJHIIHIJJJGBHIIIJFIJFDFHGIGIGIJJJJJFIJFGIIIGHFHGGIGIJIIGCHHHHCED?@DFCEEEECEEEECCC
@HWI-ST314:257:D1WGCACXX:2:1101:1137:35617_1:N:0:TCGAAG
GGATTTATAGAATTATTTGTTTTCGAAATGCGGGAGGGTATGTTTTTGTTTAATATAGTGAGATAAATAGTAAGTTTGTTGTTTTTTAGGGAGTATATTT
+
BBBDFFFFHHHGHJIJJJHHIJJJEHJJJDHHIGFHIIFCEGHIIJJFHIIIIJHHHEEEEBEDDDEDECEEEEF>CAACACDDDDDD9:?CB?ACCDEE
@HWI-ST314:257:D1WGCACXX:2:1101:1151:27741_1:N:0:TCGAAG
AGGTATAGTGTTTGTTTTGTGTTTTTTTTTTAGTATTGGTTTAGTTGTGTTTTGTGGTTGTATATGTTTTATAGAATATTTTTTTTGTGGTTATGATGGT
+
BBBDDEFFHHHHHGHHIJHGGIIJJJJJJJJJDB=@CGAFHGGECHHAHAHEHFB?C?CC?>CDDD:CCEC>@C>ACCCCFFEDDDD&8<0?B@C>3>@>:
@HWI-ST314:257:D1WGCACXX:2:1101:1160:46348_1:N:0:TCGAAG
AGAAAAAGGTATGAGAGAAAGAATAGGGAAGGAATTATTATTTTAGGTGTTTGGAAATAGATAAGATTAGATTAGTTTTTTAGTGAAGATGATTTTTTTT
+
CCCFFFFFHFFHHGHGIJJJJJJJJJJJJJFCDGHIJJJJIJJJJIGHFHIHHFGIIGIJJJIJJHIHIJIJJHHHHEHHFDDACCDDDDDDDEEEDDDDD
@HWI-ST314:257:D1WGCACXX:2:1101:1186:63430_1:N:0:TCGAAG
ATTTTTTTTTTATTTGTTTGTTTTTTTTTTTTGTGTTAAATTAGGAGGAGAAGGGAAGTTTGAATGAAGTAGTTTATTT
+
CC@FFFFFHHH?FHJ4?FH@GHHIJJJJHFDD',,5?ADA;((,((30((22<A8<88>>@:3:@>C:+:4(4>A4:@C
@HWI-ST314:257:D1WGCACXX:2:1101:1224:90588 1:N:0:TCGAAG
```

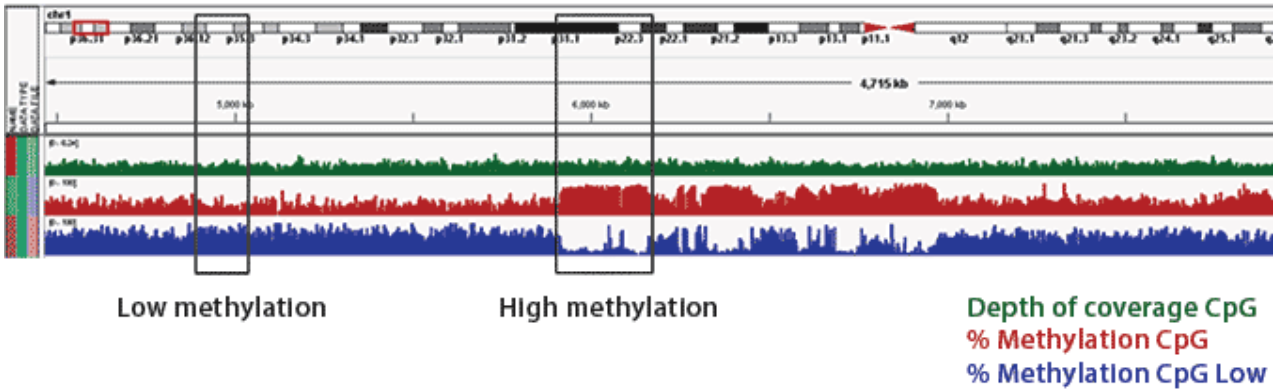
Supported data

- Fastq files from illumina sequencing
- Single or paired reads

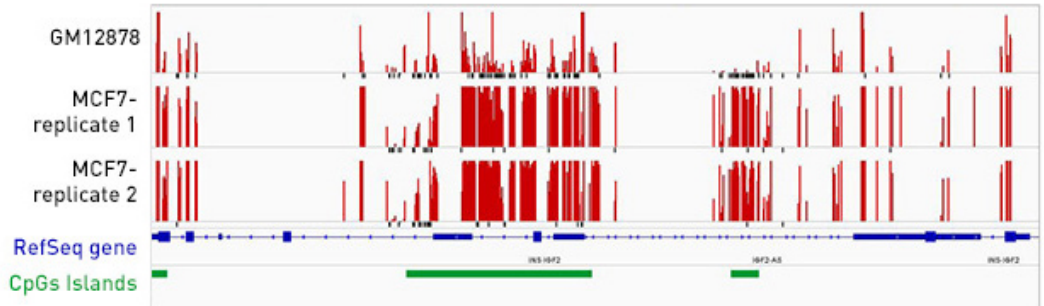


Supported data

- Fastq files from illumina sequencing
- Single or paired reads
- Protocol
  - WGBS: Whole Genome Bisulfite sequencing



  - RRBS: Reduced Representation Bisulfite sequencing

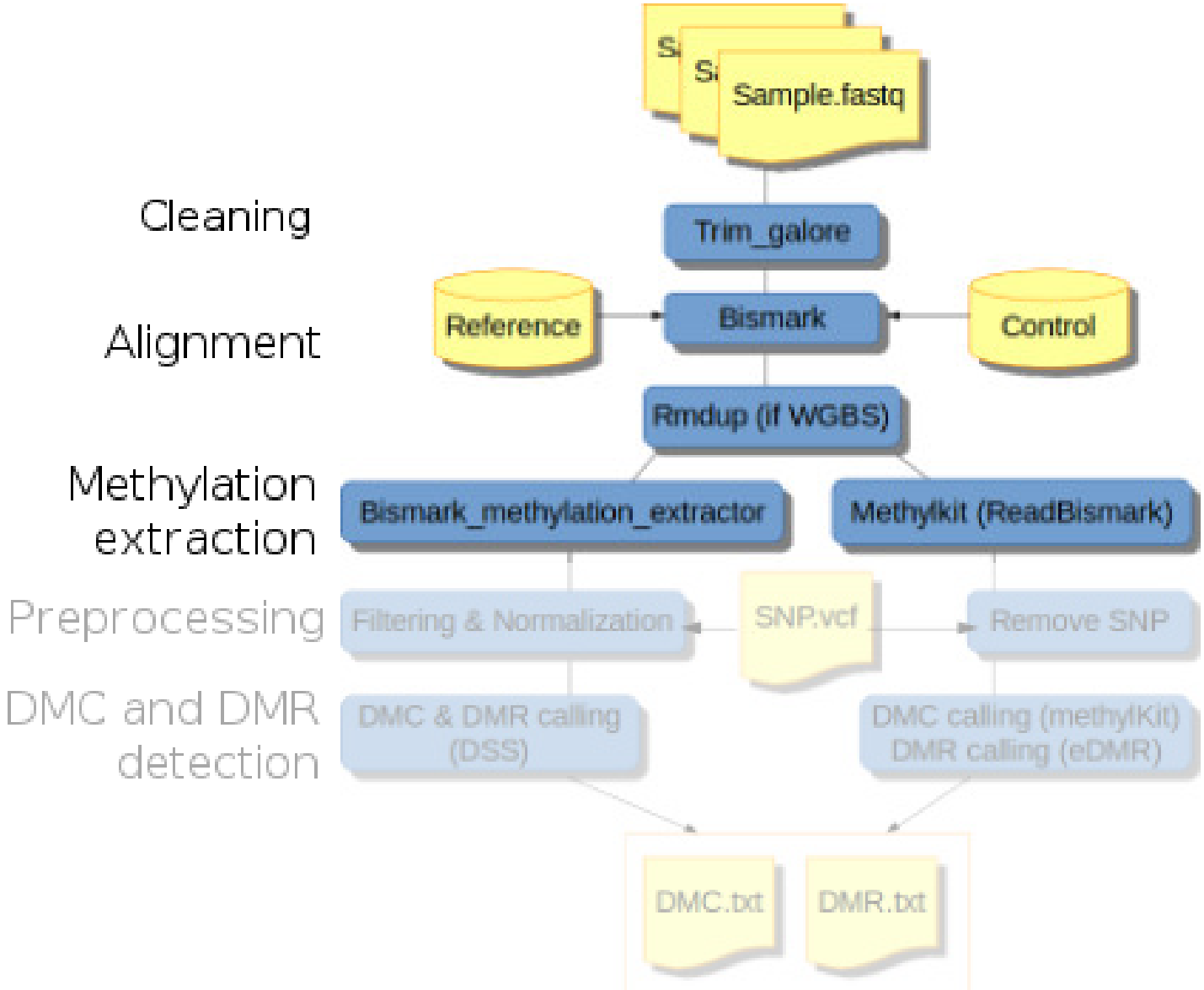


Data from epibird project

- 4 male vs. 4 female chicken embryos
- Sequenced by HiSeq3000
- Whole Genome Bisulfite sequencing

Summary

Bioinformatics steps



Quality control and cleaning

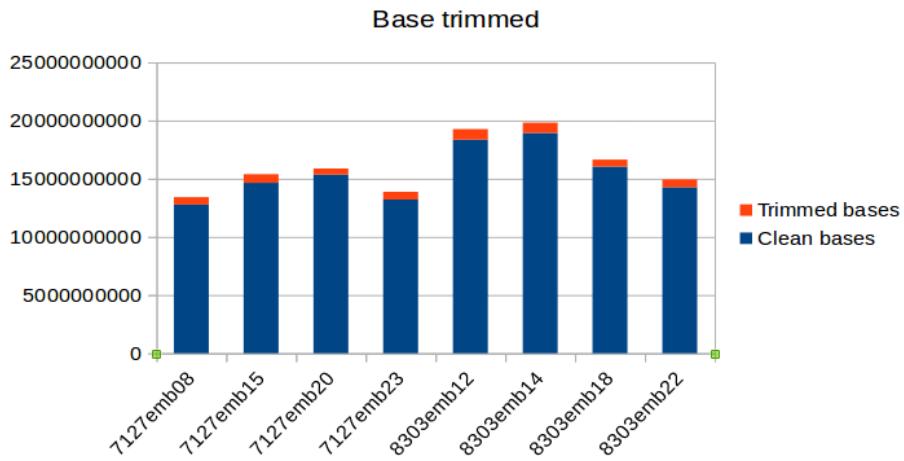
- Trim adapters
- Trim bad quality

Software: Trim_galore

Quality control and cleaning

- Trim adapters
- Trim bad quality

Software: Trim_galore



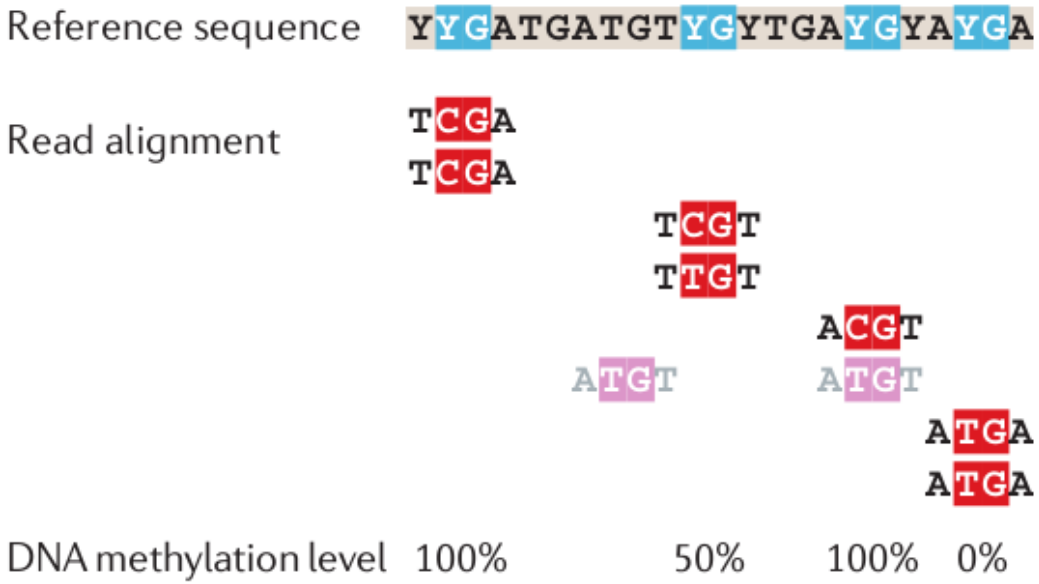
- Reads: about 40% of reads are trimmed
- Bases: about 4% of total bases

Alignment


Genomic DNA sequence
CCGATGATGTCGCTGACGCACGA
DNA methylation level
100%
50%
50%
0%

Alignment

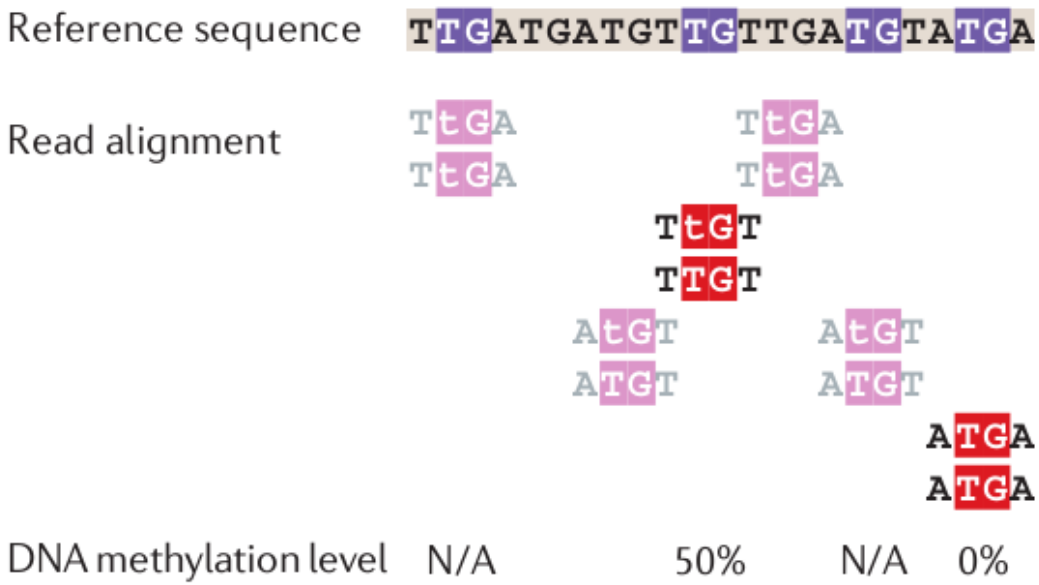
- wild-card alignment



Existing software: BSMAP, GSNAP, Last, Pash, RMAP, RRBSMAP, segemehl...

Alignment

- wild-card alignment
- 3-base encoding



Existing software: **Bismark**, BRAT, BS-Seeker, MethylCoder...

Alignment

Mapping efficiency (percent of mapped reads) depends

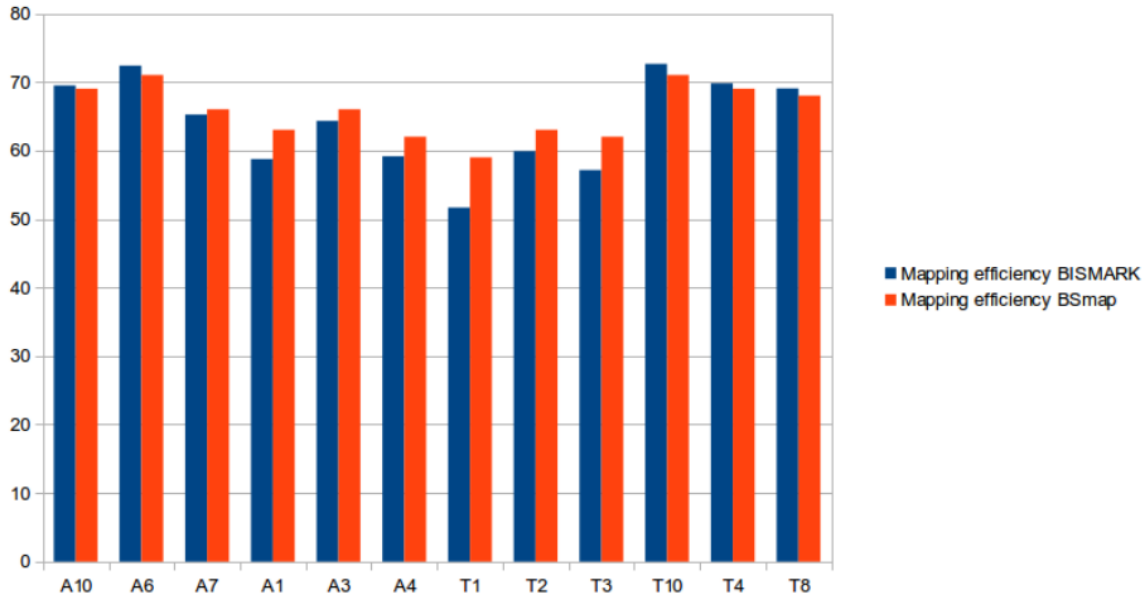
- read length
- single or paired end library
- genome composition and variability

Alignment

Mapping efficiency (percent of mapped reads) depends

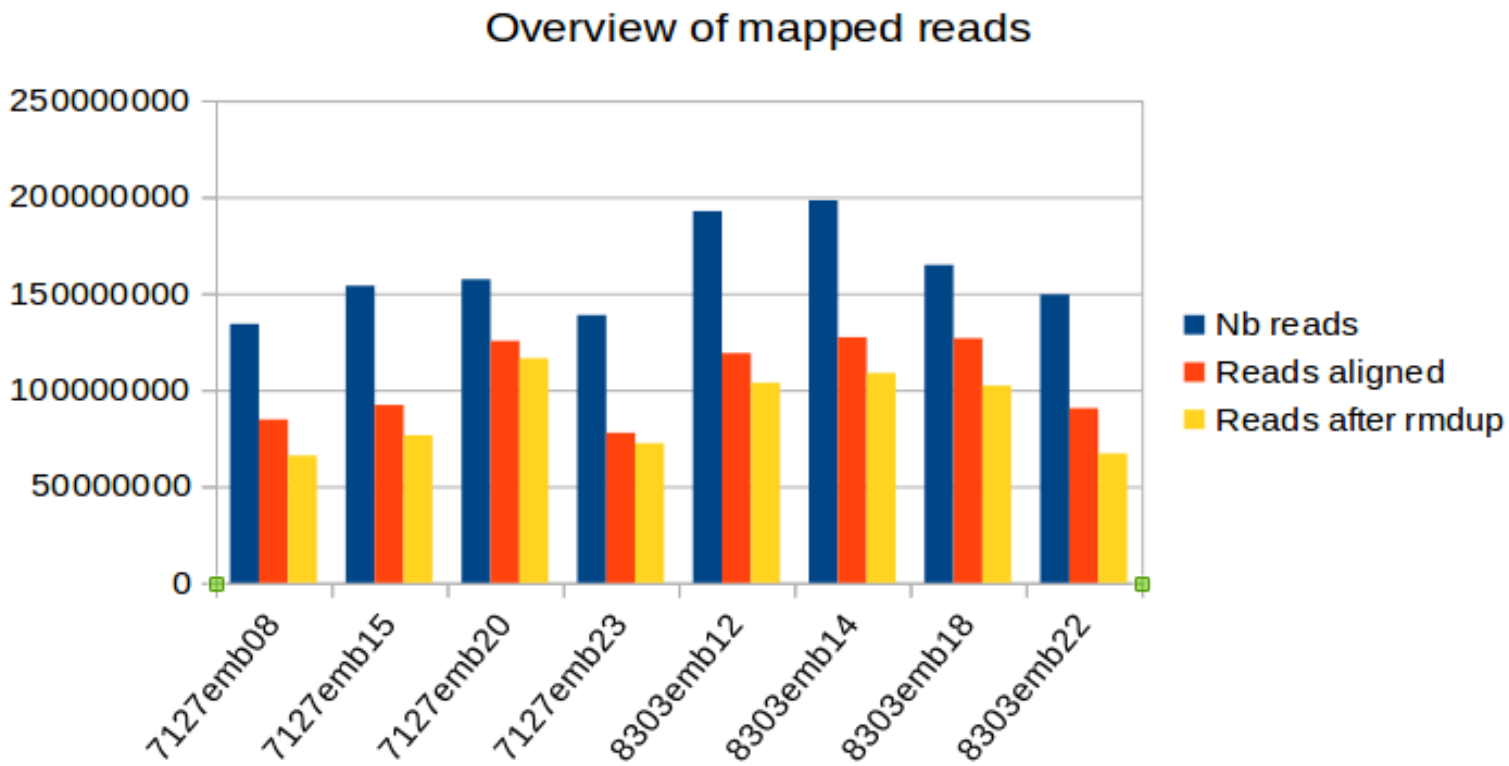
- read length
- single or paired end library
- genome composition and variability

Comparison of 2 strategies :



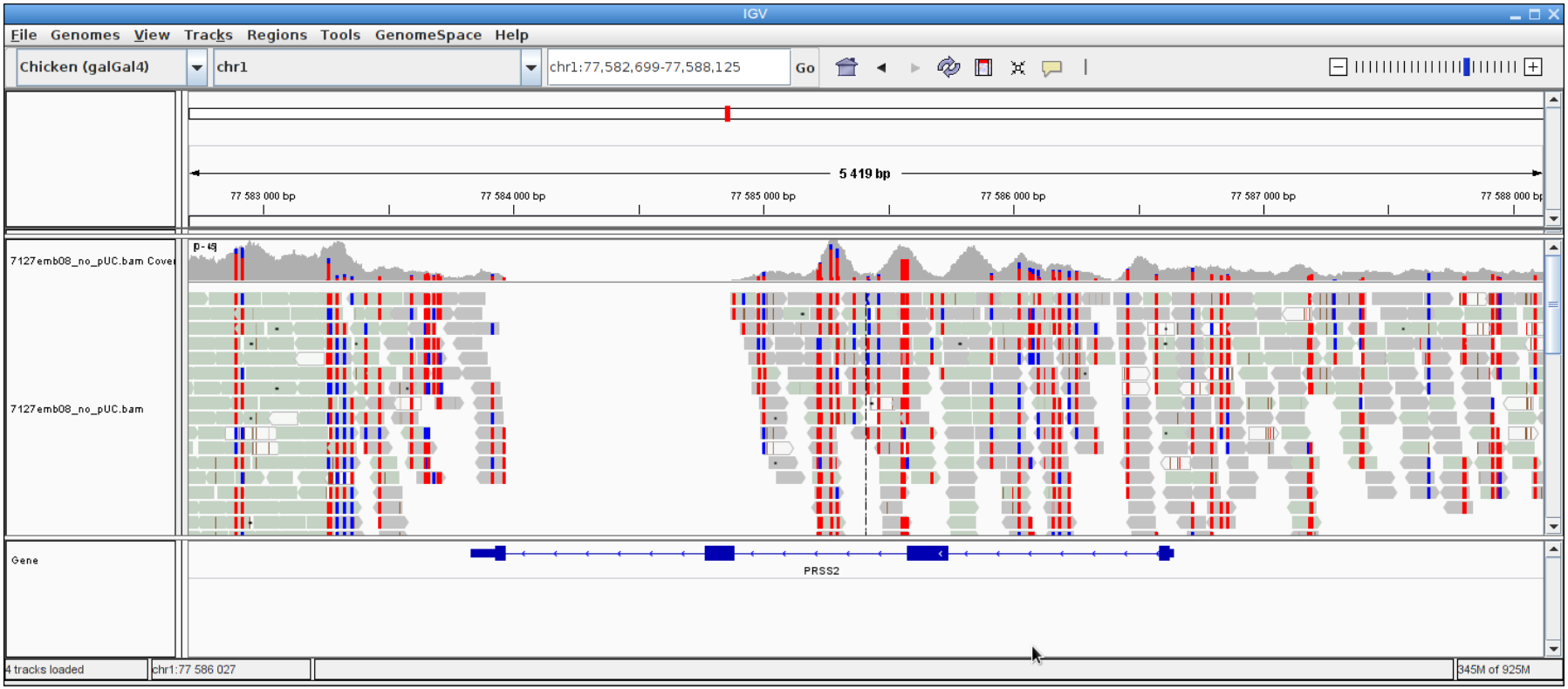
M2 Bioinfo 2014 - Julien Plennecassagne - Projet Epiterm

Alignment: Epibird results



- Bismark: 61 to 81% of mapping efficiency
- Rmdup: 73 to 93% of reads kept after rmdup

Per base methylation extraction



Per base methylation extraction

chrBase	chr	base	strand	coverage	freqC	freqT
chr1.913	chr1	913	R	1	100.00	0.00
chr1.417	chr1	417	R	3	100.00	0.00
chr1.258	chr1	258	F	1	100.00	0.00
chr1.699	chr1	699	F	3	100.00	0.00
chr1.589	chr1	589	R	6	83.33	16.67
chr1.718	chr1	718	R	6	0.00	100.00
chr1.573	chr1	573	F	8	87.50	12.50
chr1.832	chr1	832	R	3	100.00	0.00
chr1.755	chr1	755	R	7	85.71	14.29
chr1.233	chr1	233	F	1	100.00	0.00
chr1.403	chr1	403	R	3	100.00	0.00
chr1.608	chr1	608	F	5	40.00	60.00
chr1.684	chr1	684	R	4	100.00	0.00
chr1.700	chr1	700	R	3	100.00	0.00
chr1.831	chr1	831	F	5	100.00	0.00
chr1.931	chr1	931	F	1	100.00	0.00
chr1.739	chr1	739	F	6	83.33	16.67
chr1.252	chr1	252	F	1	100.00	0.00
chr1.633	chr1	633	R	3	33.33	66.67
chr1.717	chr1	717	F	4	0.00	100.00

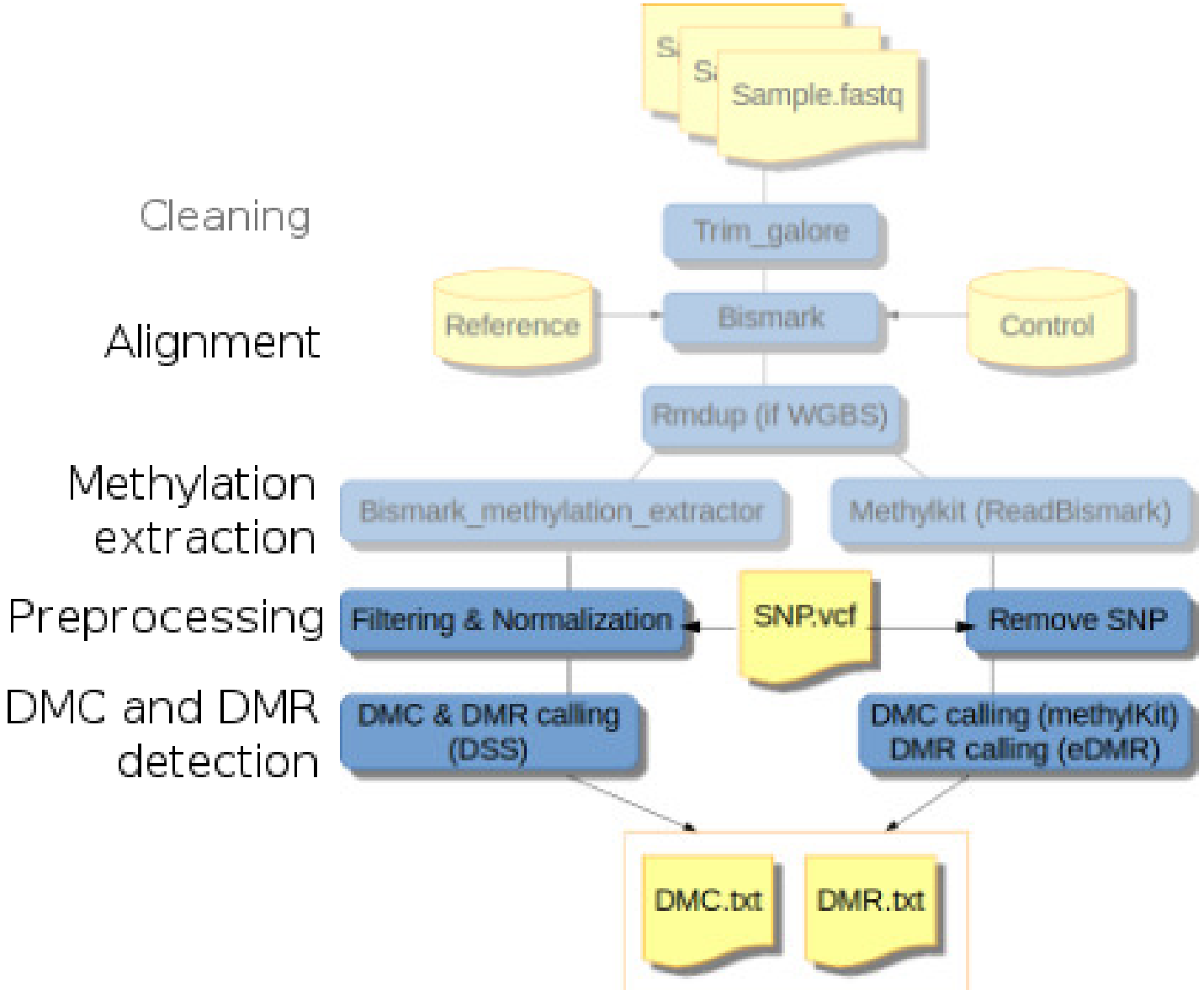
Per sample extract :

1. C in specific context (CpG, CHG, CHH)
2. choose coverage threshold

Existing software: **methylKit**, bismark_methylation_extraction ...

Summary

Biostatistics steps



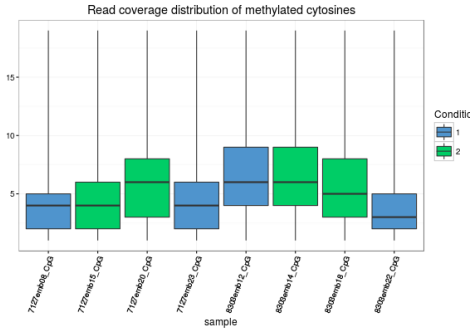
Steps

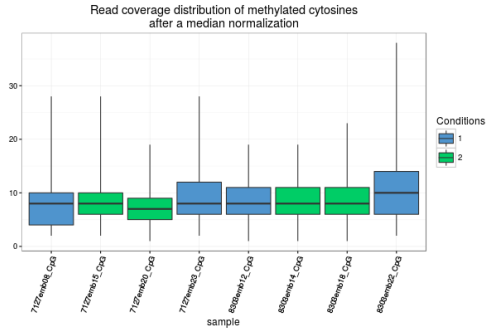
1. Normalization and filter on coverage
2. Identification of differentially methylated cytosines (DMCs)
3. Identification of differentially methylated regions (DMRs)

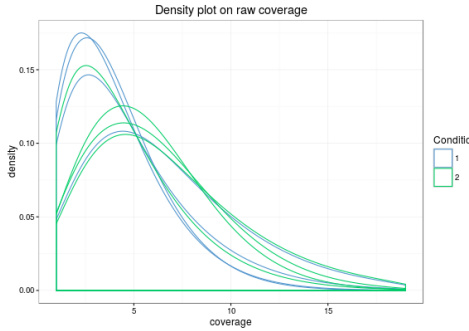
Step 1: Preprocessing on methylation data

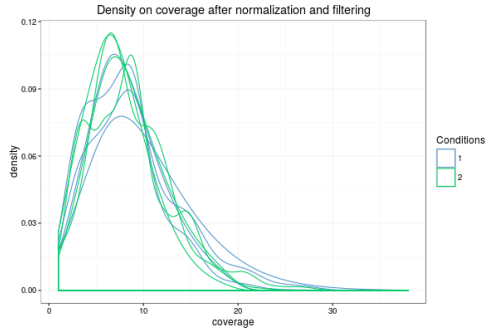
1. Remove known SNPs
2. Remove bases with a very high read coverage
3. Normalize read coverage of each cytosine (5 methods: libsize, median, upper-quartile, RLE and LR)
4. Remove bases with a very low read coverage (a minimum coverage of 5x is recommended)

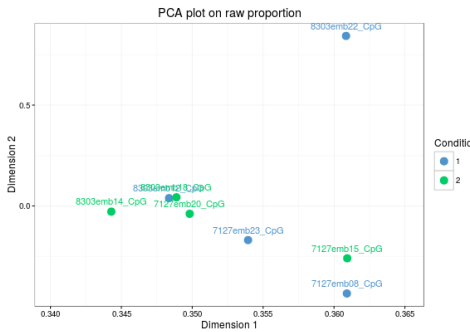
Normalization diagnostics plots

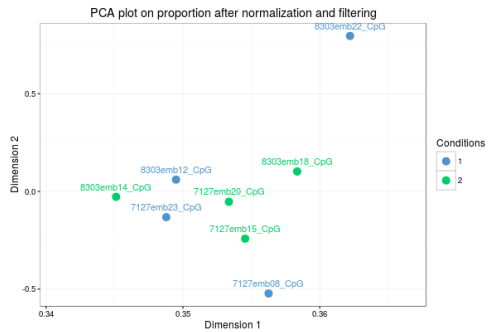


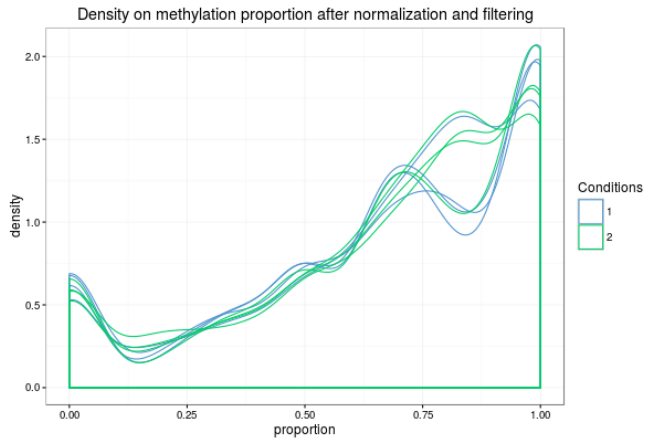












Step 2: identification of DMCs

Used methods

☹ Fisher exact test: all replicates are pooled (methylKit without replicates)

😐 Logistic regression: the hypothesis is that all data is from the same distribution (methylKit)

☺ Beta-binomial model: take into account of the biological variability between samples (DSS)
Case without replicats: nearby cytosines can be used to estimate variability (DSS)

☺ *Hidden Markov model: take into account of the spatial correlation between nearby cytosines*

Beta-binomial model (from DSS package)

Model

$$X_{ijk}|p_{ijk}, N_{ijk} \sim \text{Binomial}(N_{ijk}, p_{ijk})$$

with

$$p_{ijk} \sim \text{Beta}(\mu_{ij}, \phi_{ij})$$

and

$$\phi_{ij} \sim Log - Normal(m_{0j}, r_{0j}^2)$$

Estimation Empirical Bayes approach

Test Wald test on mean methylation levels μ_{ij}

Comparison of normalization methods

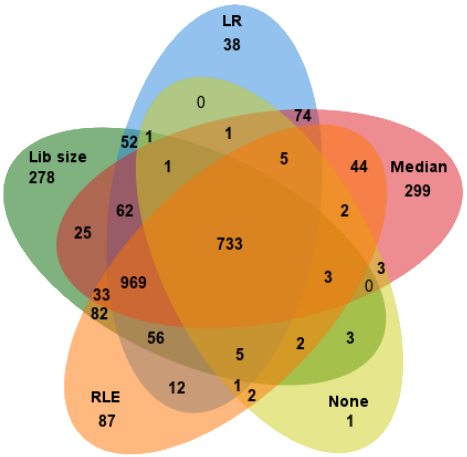


Figure : DMCs identified with DSS depending on the normalization method

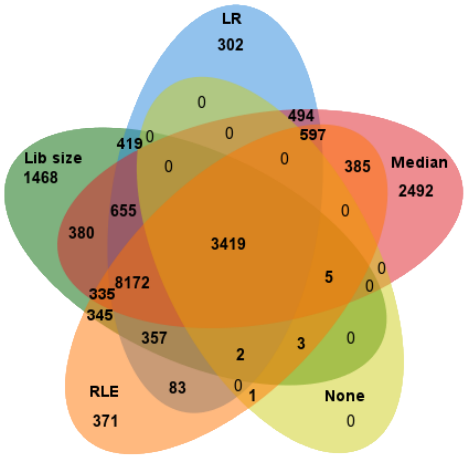


Figure : DMCs identified with methylKit depending on the normalization method

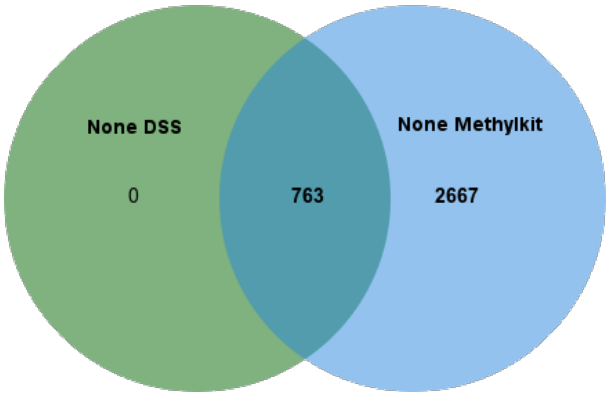


Figure : DMCs identified with methylKit and DSS without normalization

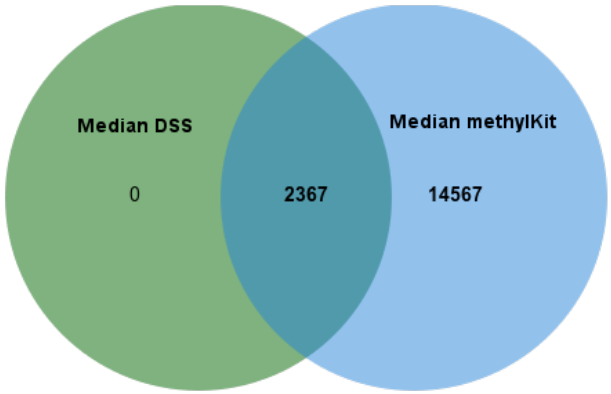


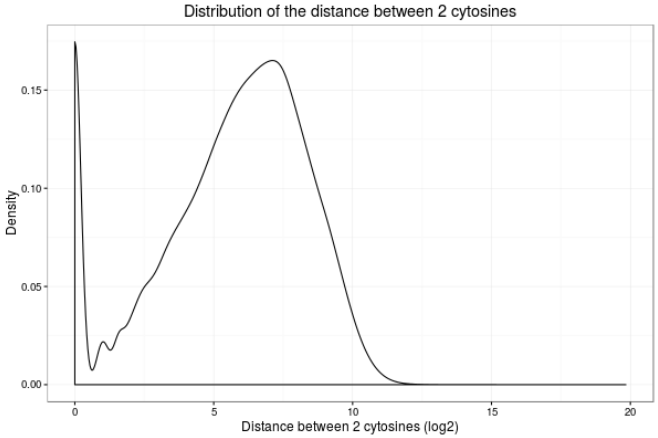
Figure : DMCs identified with methylKit and DSS with median normalization

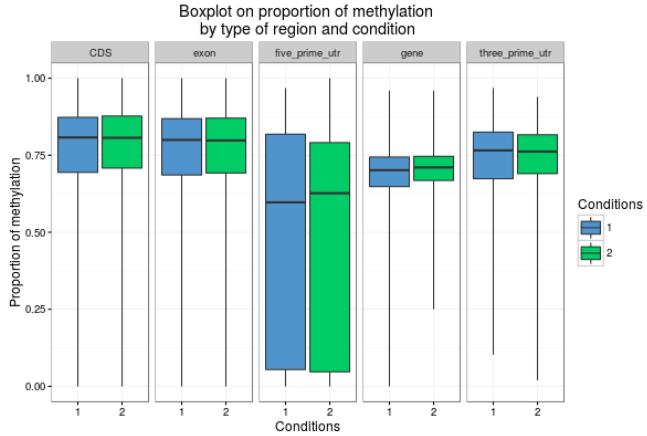
Step 3: identification of DMRs

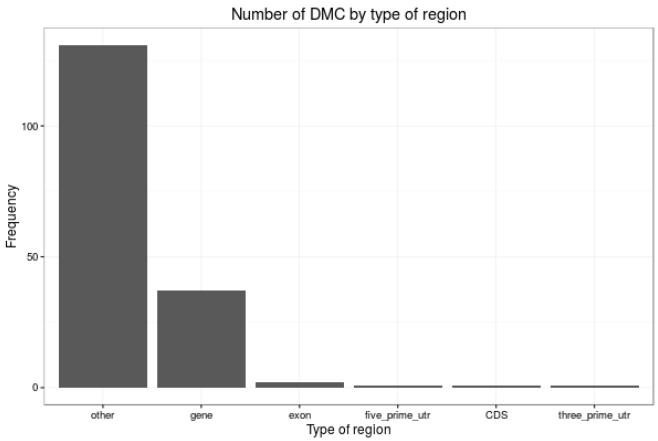
Used methods

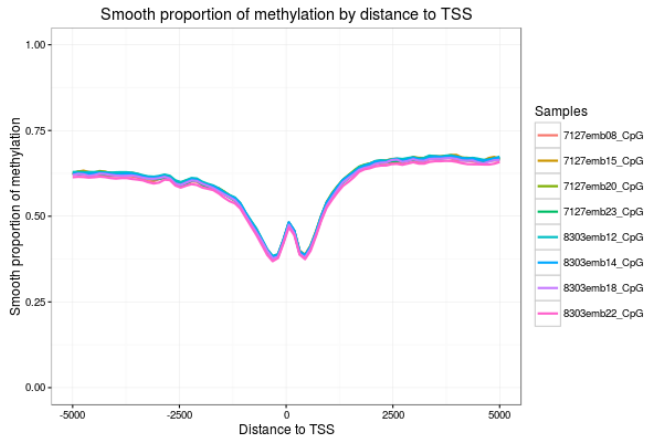
- Sliding windows or predefined regions (MethylKit, DMRcaller...)
- From results on DMCs (DSS, eDMR...)
- With a hidden Markov model (Bisulfighter)

Differential analysis results and other plots









Summary

Pipeline conclusion

- based on Jlow: error recovery, use most HPC (SGE, Condor, ...), extensible
- configuration with one config file
- include all steps (bioinfo and biostats) within a single command line
- re-runable after main step : alignment and methylation extraction

Available :

- github FAANG consortium
- mulcyber: `https://mulcyber.toulouse.inra.fr/plugins/mediawiki/wiki/jflow-toolshed/index.php/Accueil`

Coming soon ...

- Normalization and DSS
- New aligners
- A web server

Coming soon ...

- Normalization and DSS
- New aligners
- A web server

Next step :

- Link methylation regions with transcriptomics data ...

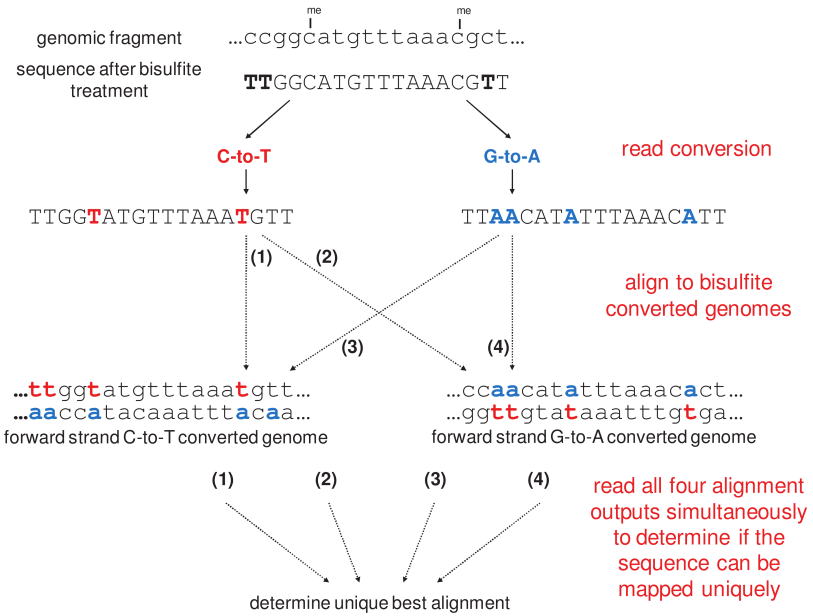
Acknowledgement

- Nathalie Villa-Vialaneix (MIAT)
- Frédérique Pitel, Gérald Salin, Sylvain Foissac, Marjorie Merch, Julien Plennecassagne, Diane Esquerre (GenPhySE)
- Erika Sallet, Pascal Gamas (LIPM)
- Monique Falières (Genotoul)

Thank you for your attention

Questions ?

Bismark


me
me
genomic fragment
...ccggcatgtttaaacgct...
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
(3)
(4)
align to bisulfite converted genomes
...ttggtatgtttaaatgtt...
...aaccatacaaatttacaa...
forward strand C-to-T converted genome
...ccaacatatttaaacact...
...ggttgtataaatttgtga...
forward strand G-to-A converted genome
(1)
(2)
(3)
(4)
read all four alignment outputs simultaneously to determine if the sequence can be mapped uniquely
determine unique best alignment

Config file

```
########################################################################
#Input files

--reference-genome
workflows/methylseq/data/Gallus_gallus.Galgal4.dna.chromosome.7.fa

#--control-genome
#/path/to/control.fa

--snp-reference
workflows/methylseq/data/snp.vcf

########################################################################
#Sample description

--sample
sample-name=sample1
read1=workflows/methylseq/data/sample1.R1.fastq.gz
read2=workflows/methylseq/data/sample1.R2.fastq.gz
#bam=/nosave/cnoirot/work/methylseq/wf000083/Bismark_paired/7127emb20.bam
#methylkit=

--sample
sample-name=sample2
read1=workflows/methylseq/data/sample2.R1.fastq.gz
read2=workflows/methylseq/data/sample2.R2.fastq.gz

########################################################################
#Protocol parameters :
#IF BAM PROVIDED : Set true if you provide alignment of single-end library
#--is-single
#
#Type of methylation context to extract and analyze :
#--rrbs
#
#To set if the libraries are non directional (Default : False)
#--non-directional
#
```

Config file

```
########################################################################
#Bismark alignment parameters
#Sets the number of mismatches to allowed in a seed alignment during multi-seed alignment
#--alignment-mismatch
#10
#
#The maximum insert size for valid paired-end alignments.
#--max-insert-size
#800
#
#Use bowtie1 instead of bowtie2 (longer, better for reads < 50bp) - default False
#--bowtie1
#
#Force to not perform rmdup :
#--no-rmdup
#
########################################################################
# Comment following lines if you do not want to perform statistical analysis
#Methylation extraction parameters (methylkit)
#For extraction of C define min coverage
--coverage
5

#Type of methylation context to extract and analyze [CpG|CHG|CHH]
#specifie several times if you want to analyze several contect
--context
CpG
#--context
#CHH

########################################################################
#DMC parameters will be perform on each context (you can specify several --test)
--test
test-name=sex
pool1=sample1,sample4,sample5,sample8
pool2=sample2,sample3,sample6,sample7
#perform methylKit logical normalization 0/1
normalization=1
#filter position with coverage less than 5 and with coverage above 99% quantile.
filter=1
#method to adjust p-values for multiple testing [BH|bonferroni]
correct=BH
#significance level of the tests (i.e. acceptable rate of false-positive in the list of DMC)
alpha=0.05
```