SIMPOSIO ANUAL
ÓMICAS

Taller 1
Introducción al análisis de datos epigenómicos

Jenny Johana Gallo Franco
Doctorado en Ingeniería y Ciencias Aplicadas
Pontificia Universidad Javeriana Cali

ÓMICAS

Pontificia Universidad JAVERIANA Cali

Contenido:

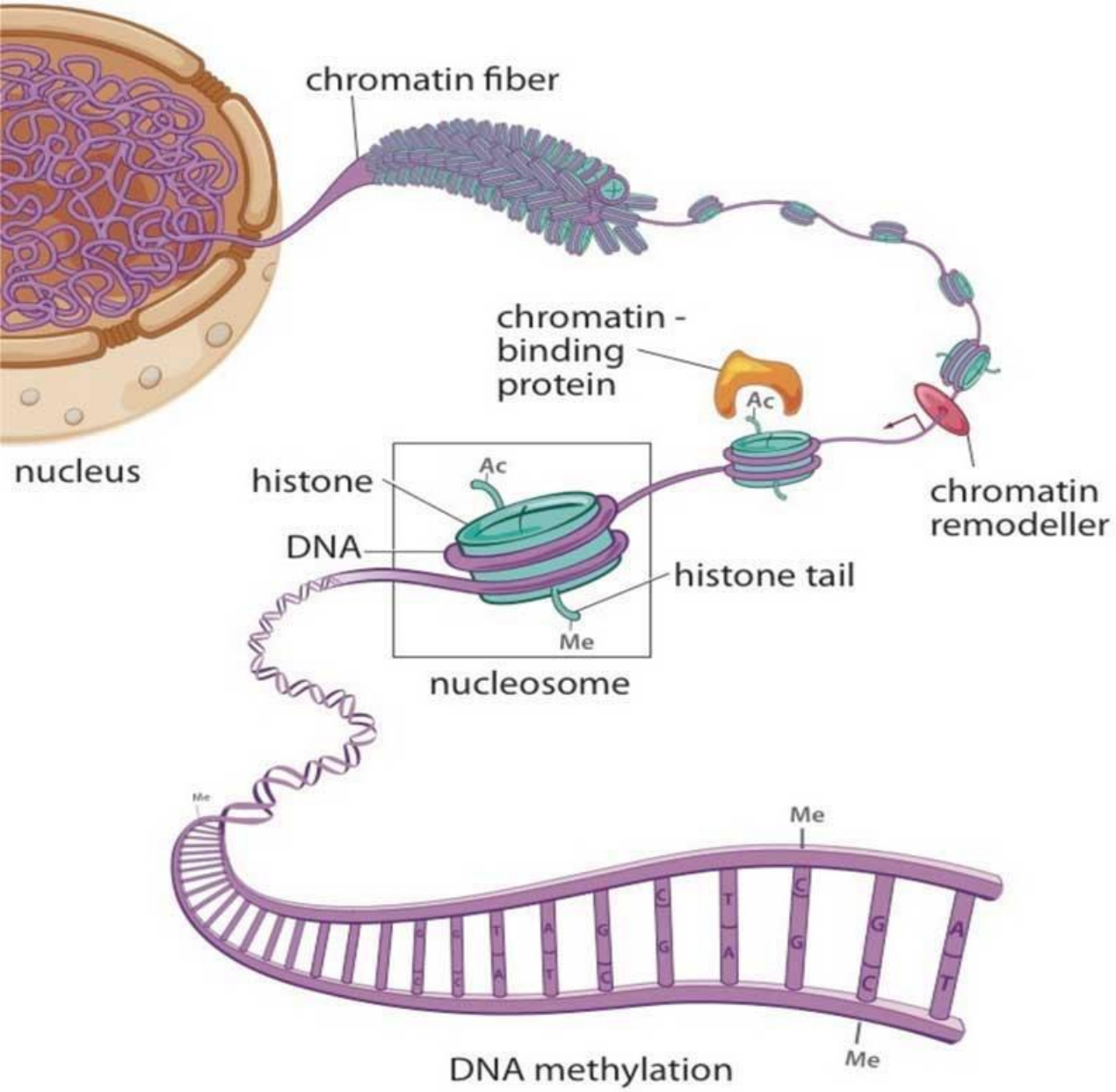
1. Teoría de secuenciación por bisulfito y control de calidad

2. Control de calidad (FastQC), alineamiento y extracción de la metilación – Práctica (Bismark)

3. Visualización y exploración de los niveles de metilación – Práctica (MethylKit)

4. Metilación diferencial, Teoría y práctica (MethylKit)

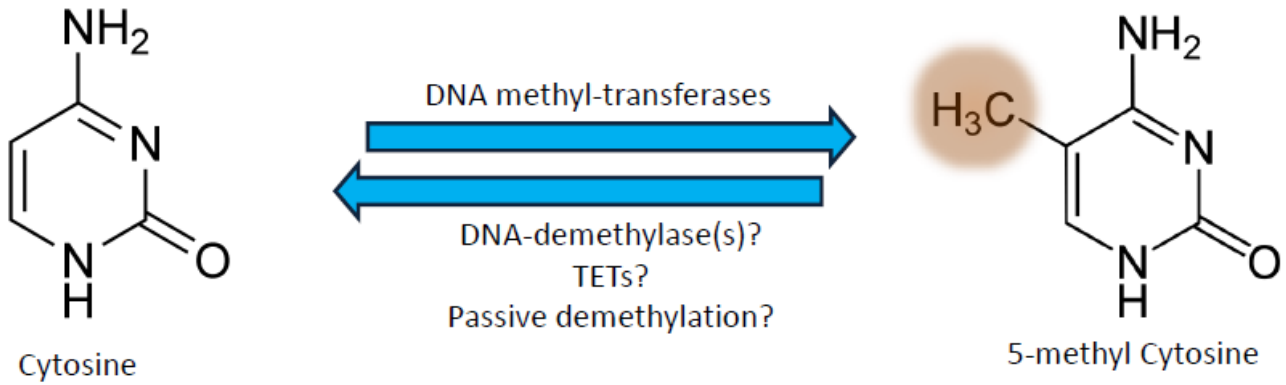
Epigenética



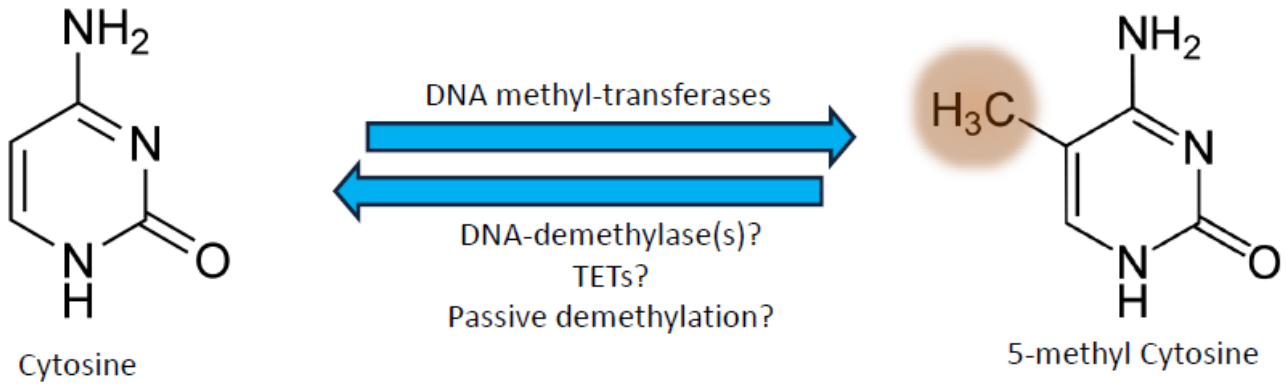
Estudio de los cambios en la expresión de los genes que no están relacionados con la secuencia de ADN

- Modificaciones de histonas
- RNAs no-codificantes
- **Metilación del ADN**

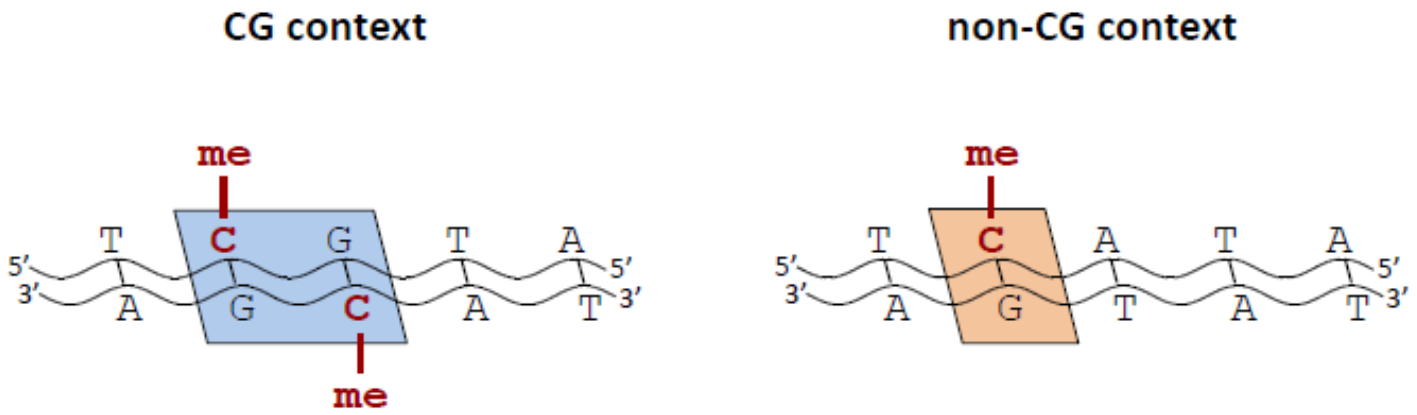
Metilación del ADN



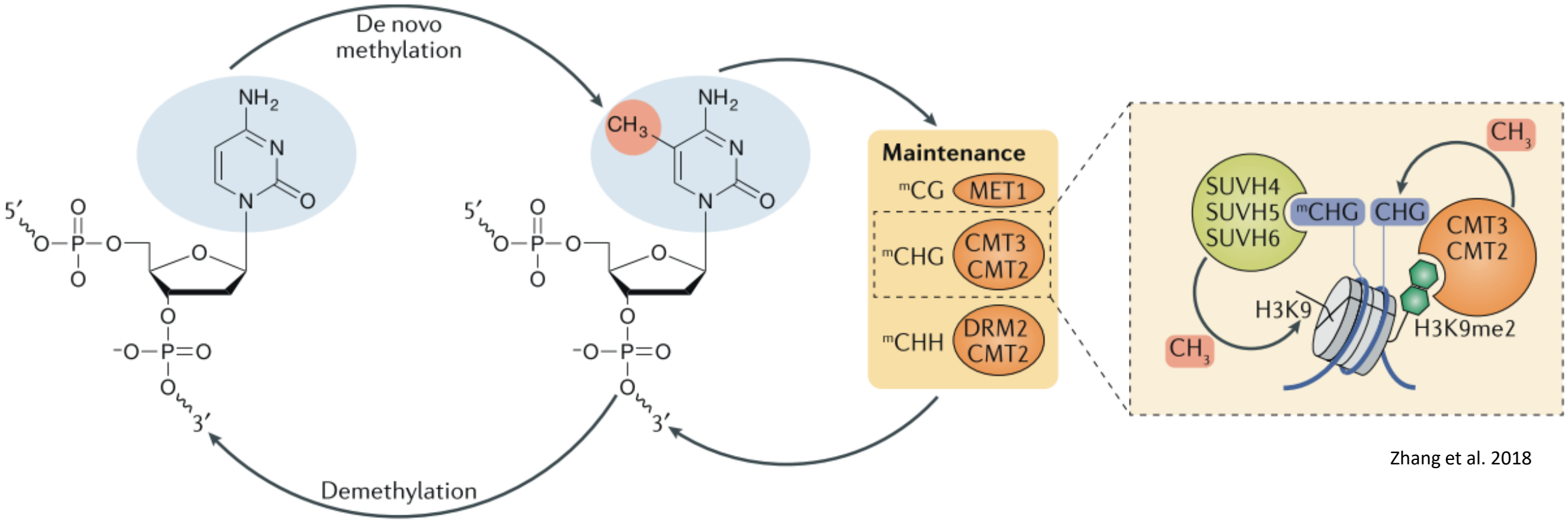
Metilación del ADN



Tipos de metilación en el ADN

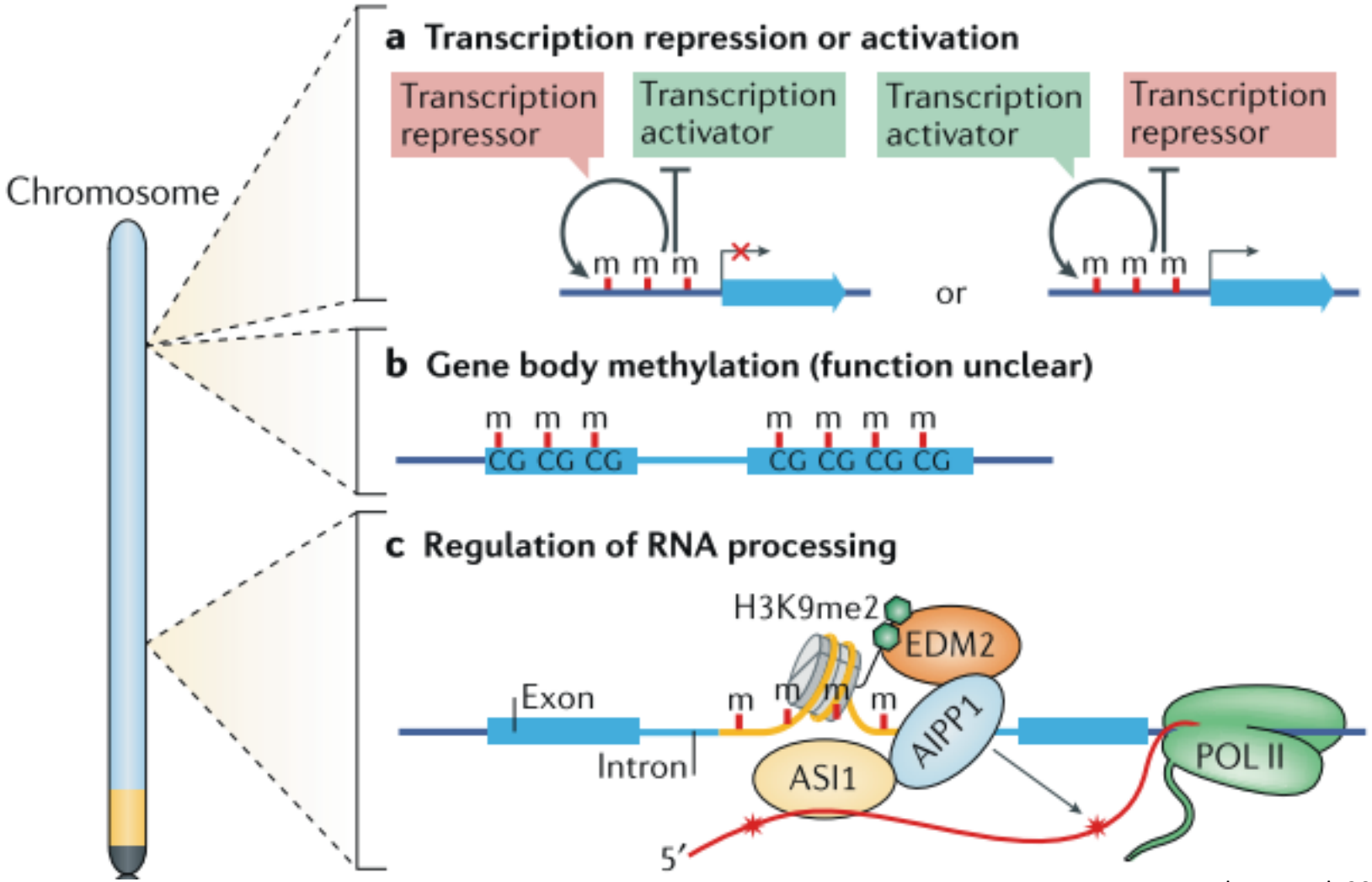


Proceso de metilación



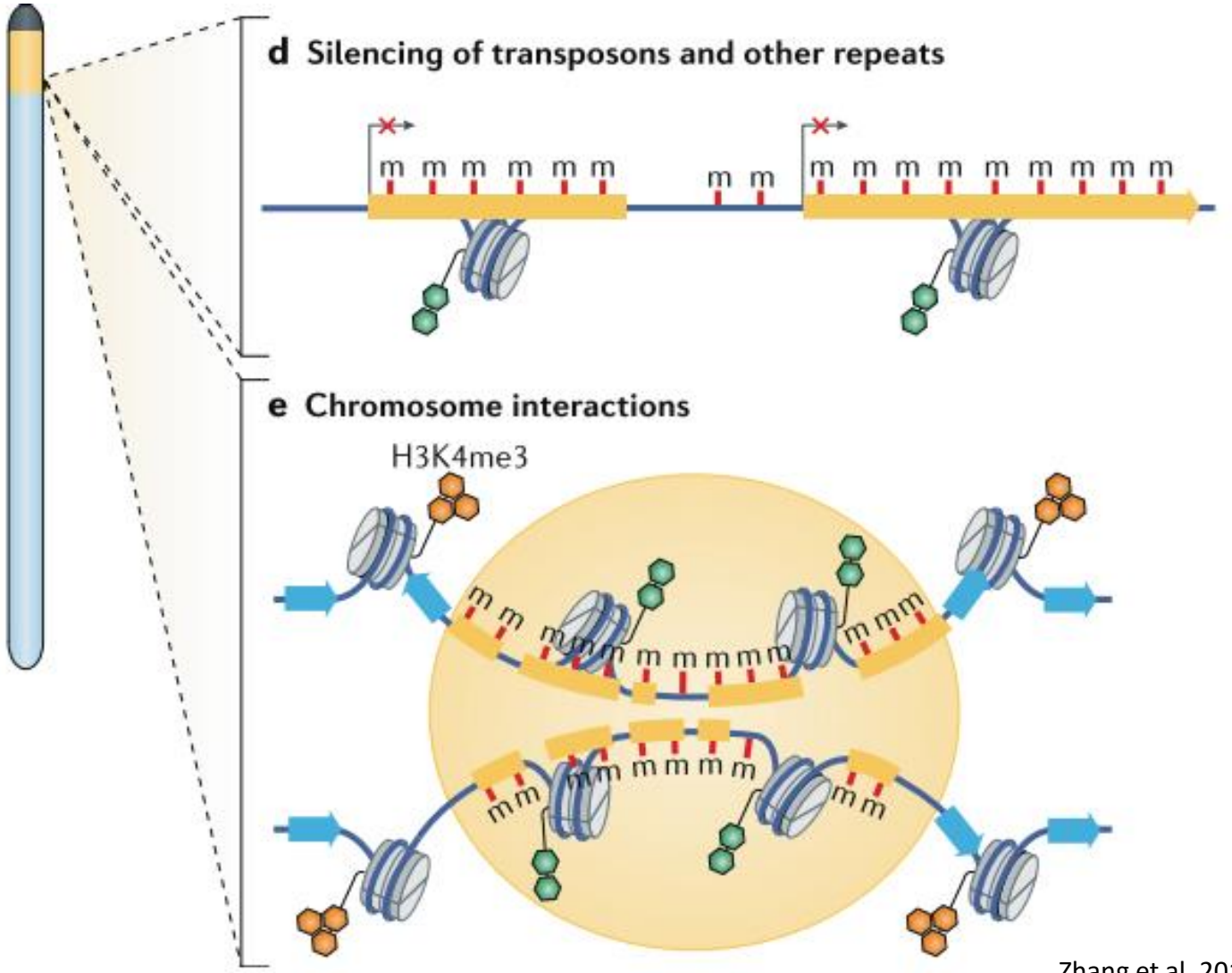
Zhang et al. 2018

Posibles funciones celulares de la metilación en plantas



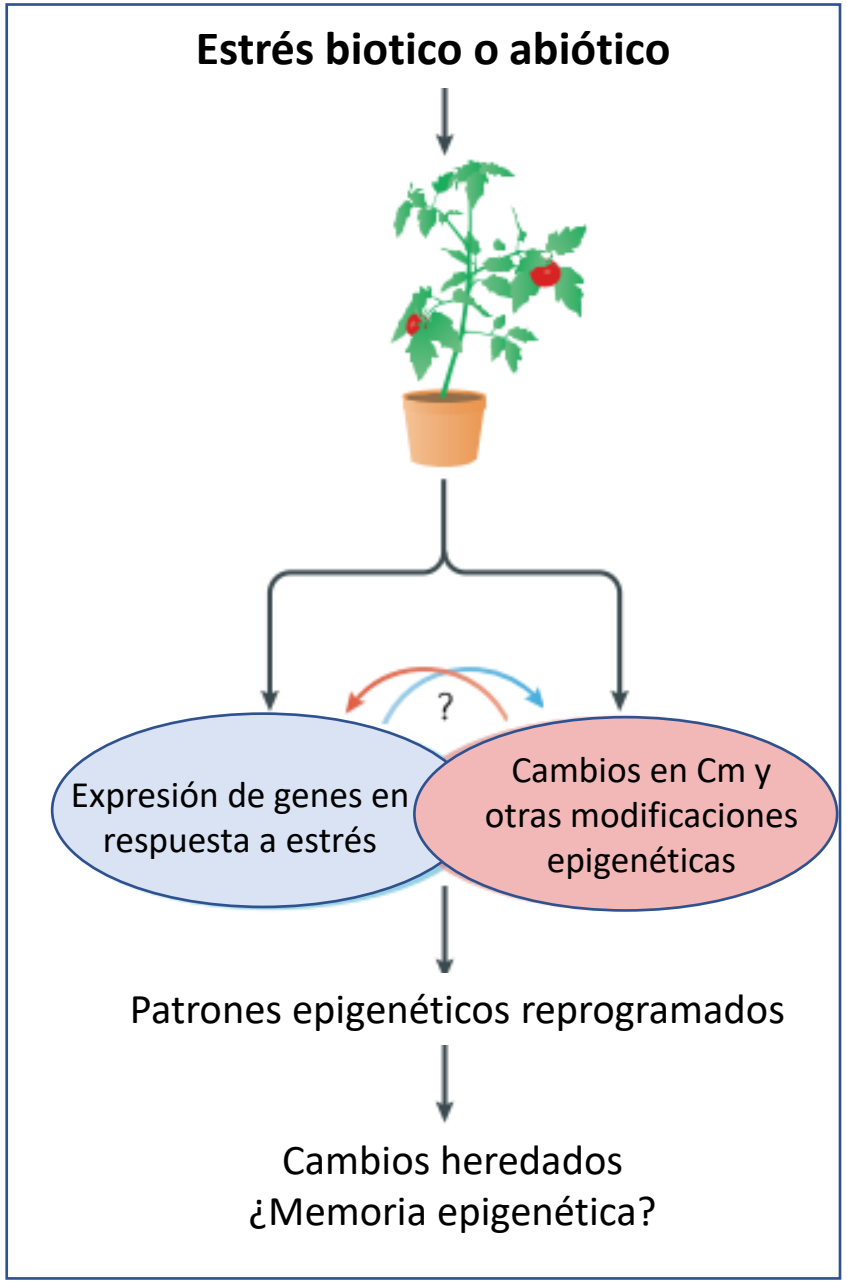
Zhang et al. 2018

Posibles funciones celulares de la metilación en plantas



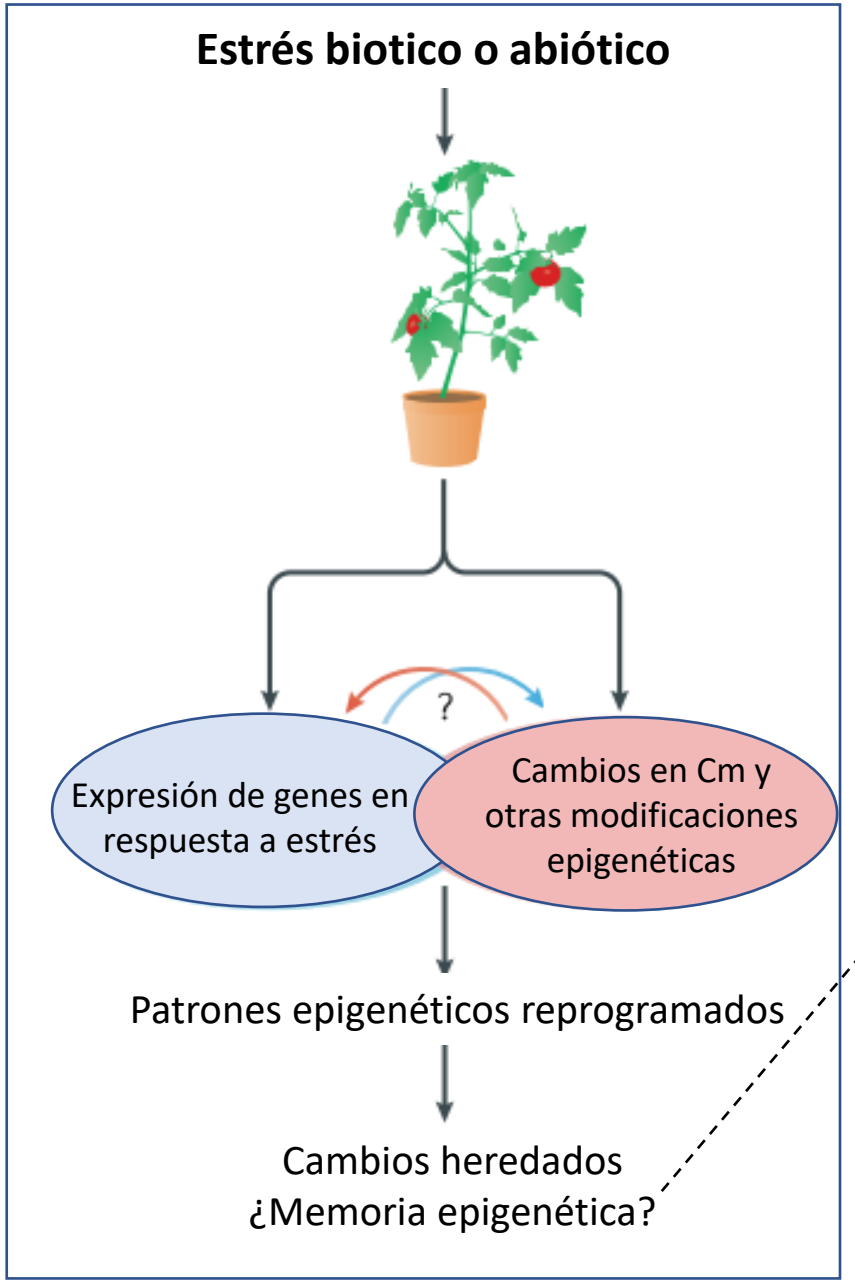
Zhang et al. 2018

Cambios epigenéticos en respuesta a estrés

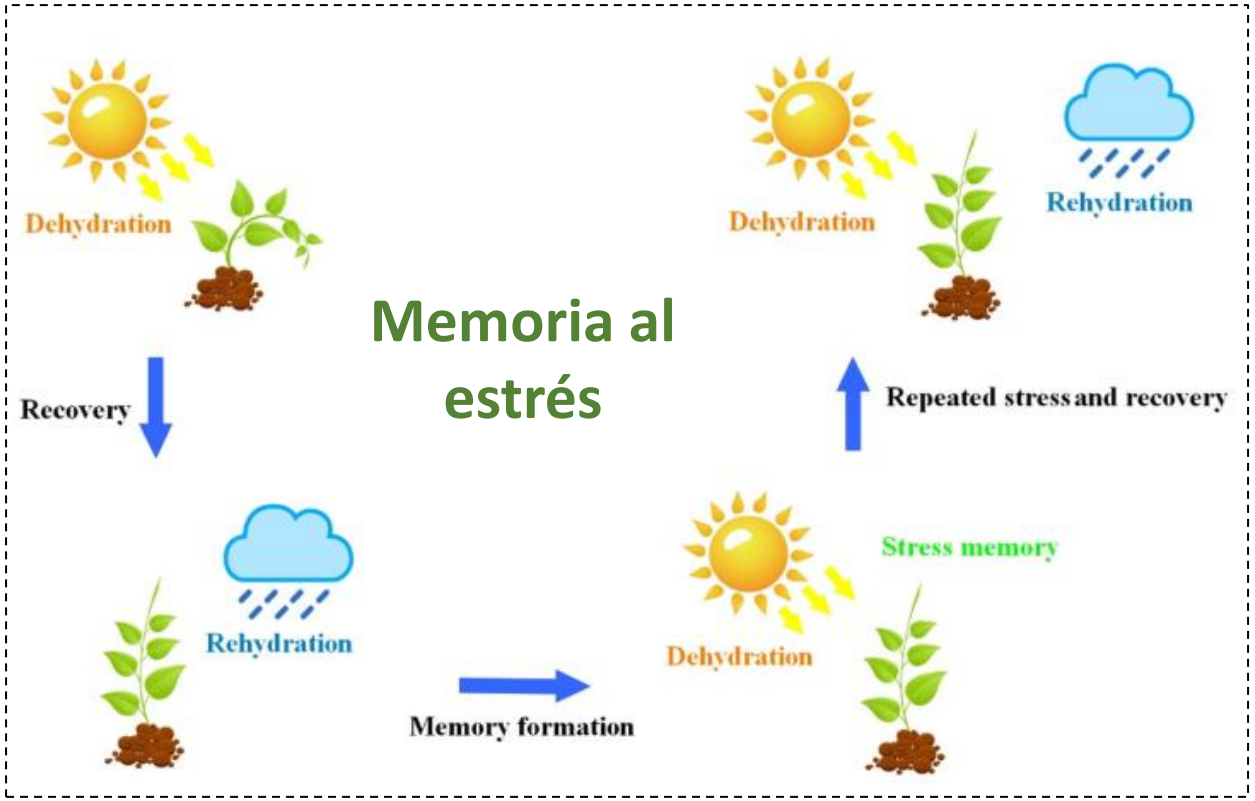


Zhang et al. 2018

Cambios epigenéticos en respuesta a estrés



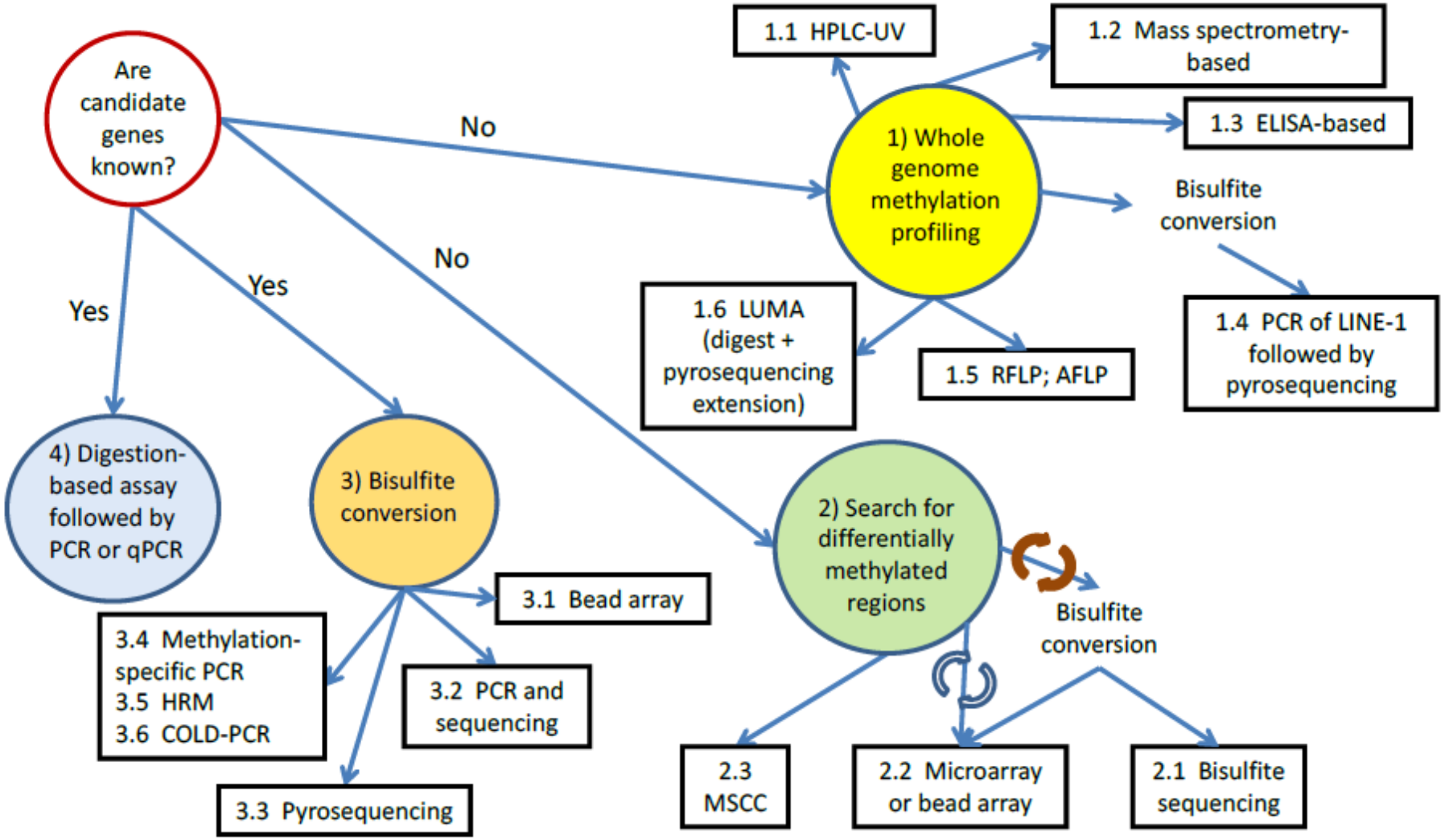
Zhang et al. 2018



Turgut-Kara et al. 2020

¿ Cómo se puede estudiar la metilación del ADN ?

¿ Cómo se puede estudiar la metilación del ADN ?



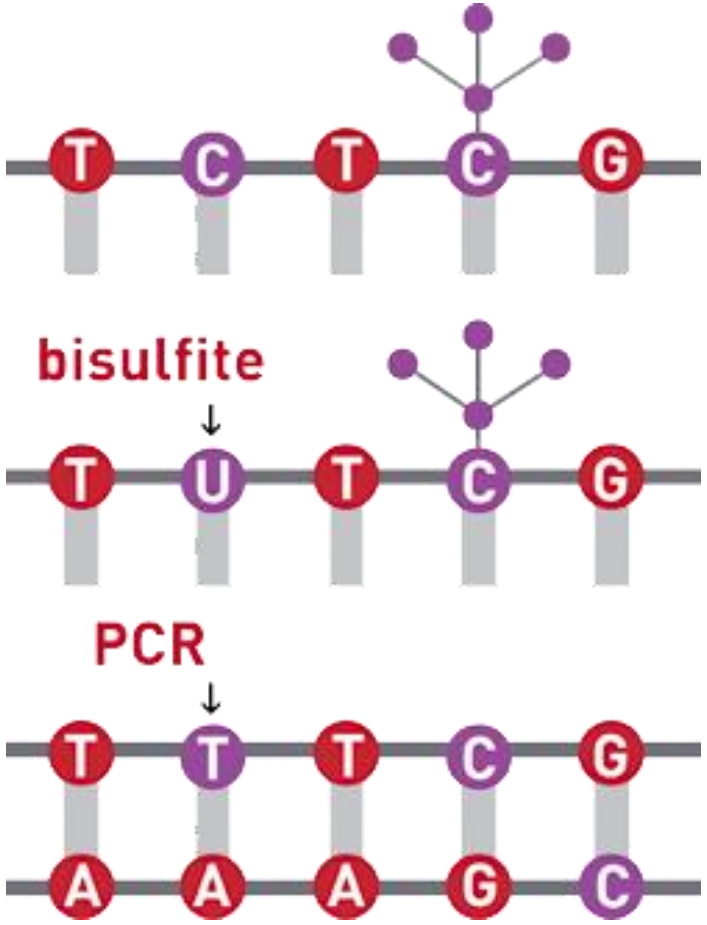
MSCC: Methyl-Sensitive Cut Counting
LUMA: Luminometric Methylation Assay
LINE: Long Interspersed Nuclear Elements
ELISA: Enzyme-Linked Immunosorbent Assay
AFLP: Amplified Fragment Length Polymorphism
RFLP: Restriction Fragment Length Polymorphism
HRM: High Resolution Melting
COLD-PCR: explained in chapter 4.6

Enrichment for methylated regions

Enrichment for CpG rich regions (optional)

Kurdyukov & Bullock 2016

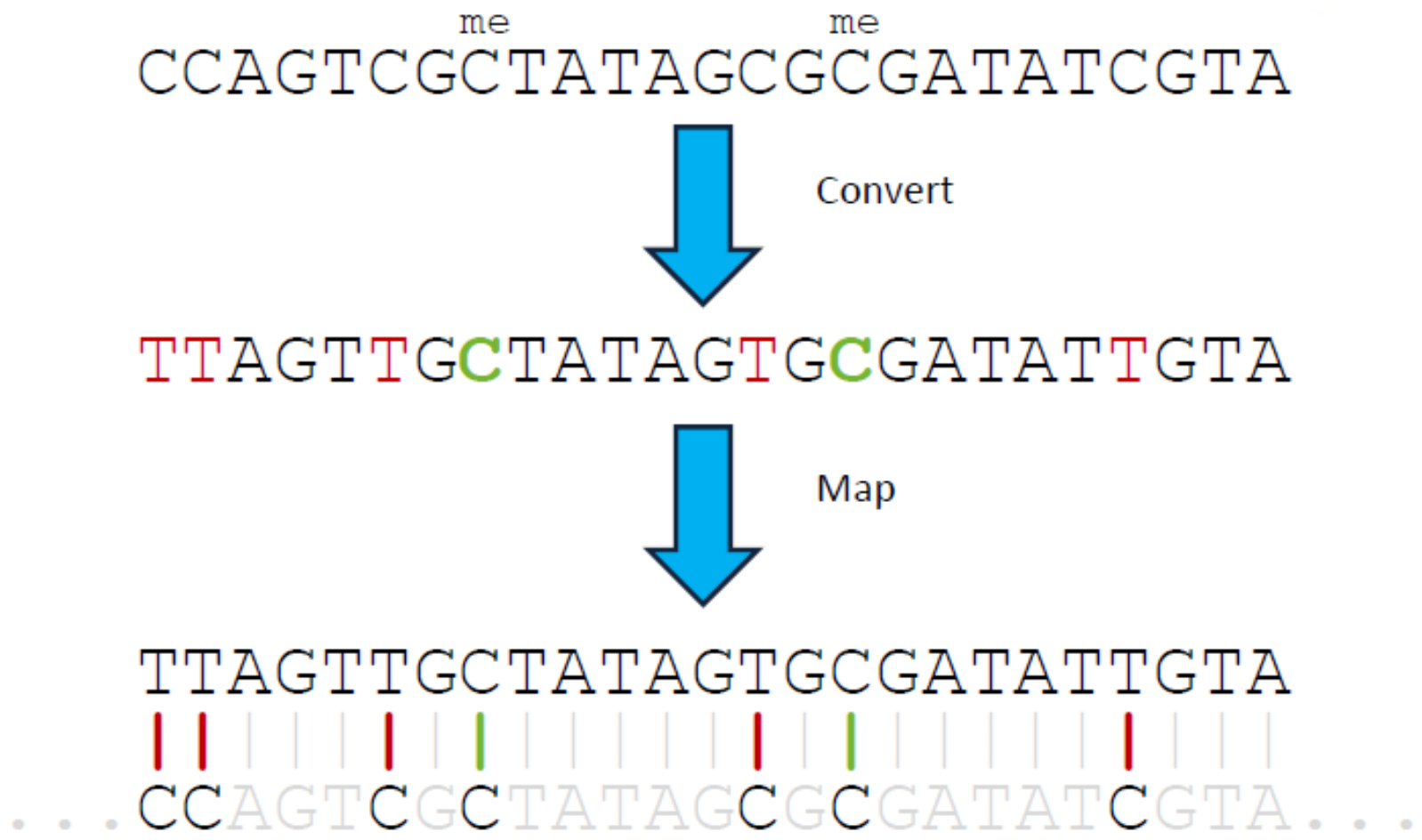
Secuenciación de genoma completo por bisulfito (WGBS)



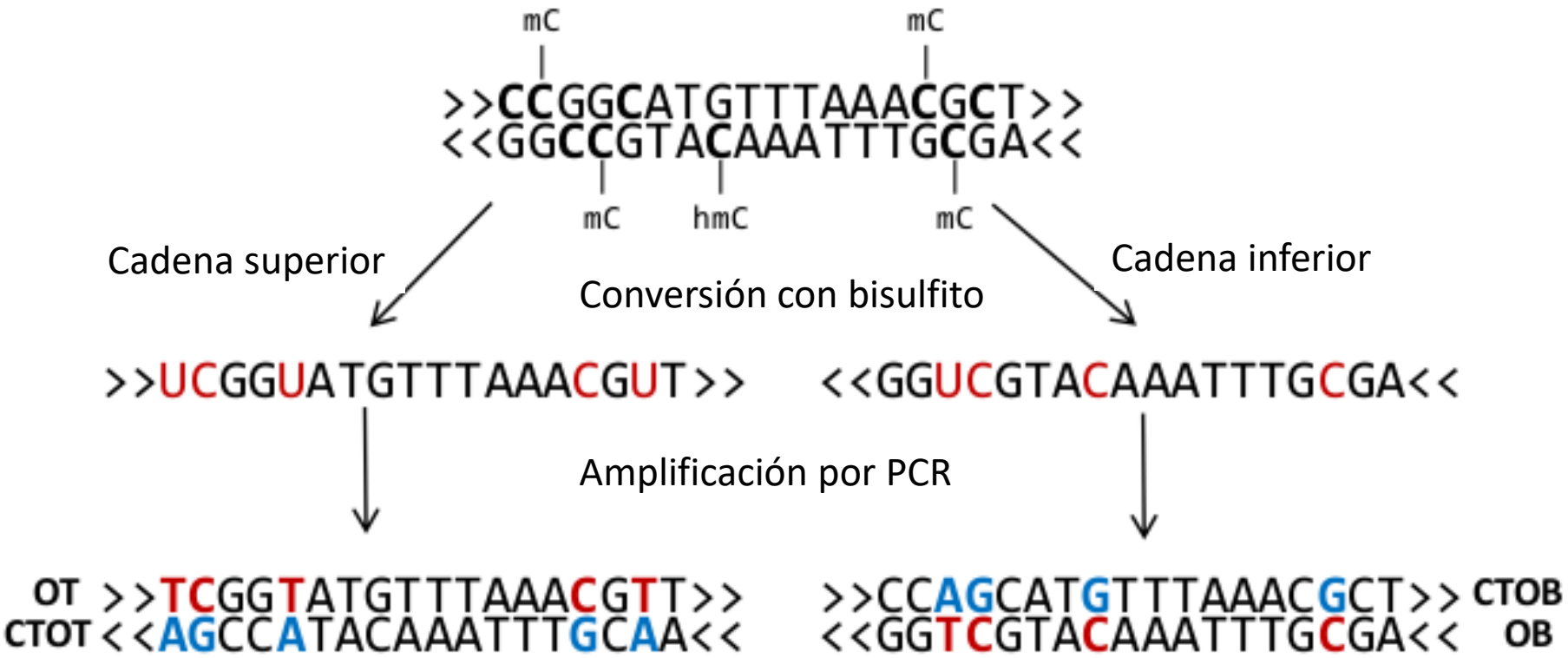
- Los nucleótidos de **citosina no metilados** se convierten en **uracilo,** que se identifican como timinas (T) cuando se secuencia.

- Las **citosinas metiladas** están protegidas de la conversión, por lo que aún se identifican como **citosina (C).**

Análisis de datos de secuenciación bisulfito



Análisis de datos de secuenciación bisulfito
En realidad…



Análisis de datos de secuenciación por bisulfito.

Secuenciación, control de calidad y remoción de adaptadores

Alinear al genoma de referencia Bowtie2

Eliminar sesgo de PCR (Deduplication)

Llamado de metilación

Filtrado por cobertura

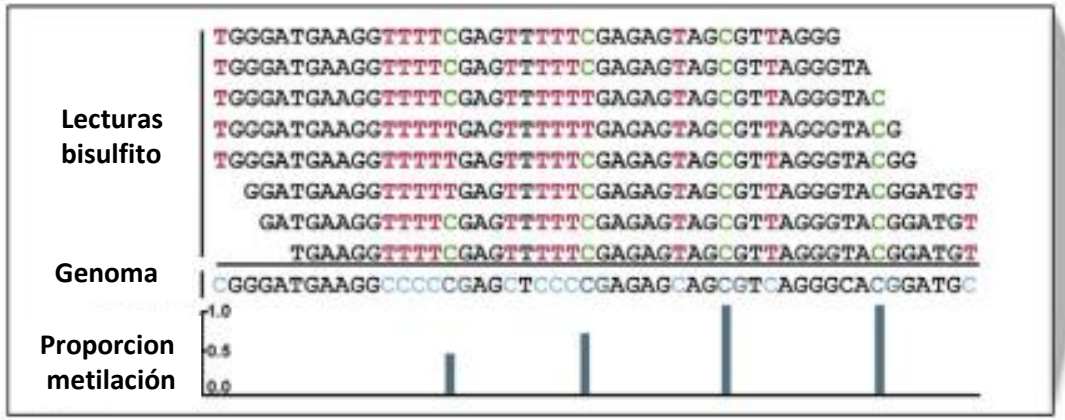
Visualización y agrupamiento

Metilación diferencial

Resultados como tablas, estadísticos de resumen, etc.

Segmentación de la metilación

Anotación de DMR/DMCs



Wreczycka et al. 2017

Formato FASTA

```
>Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
>Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XX
```

¿ Que tipo de secuencias se obtienen de la secuenciación?

FASTA es un formato familiar en datos genómicos

Formato FASTA

```
>Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
>Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XX
```

Formato FASTQ

```
@Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
@Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
```

¿ Que tipo de secuencias se obtienen de la secuenciación?

FASTA es un formato familiar en datos genómicos

Cuatro líneas:

- @ + identificador en la primera línea
- Secuencia
- +
- Valores de calidad

Secuenciación, control de calidad y remoción de adaptadores

Alinear al genoma de referencia Bowtie2

Eliminar sesgo de PCR (Deduplication)

Llamado de metilación

Filtrado por cobertura

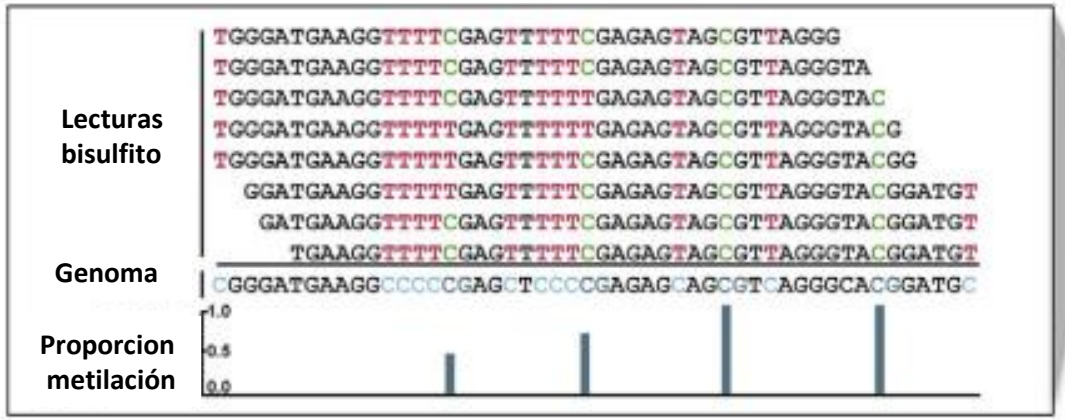
Visualización y agrupamiento

Metilación diferencial

Resultados como tablas, estadísticos de resumen, etc.

Segmentación de la metilación

Anotación de DMR/DMCs



Parte I: Análisis de calidad de las secuencias

FastQC

FastQC Report

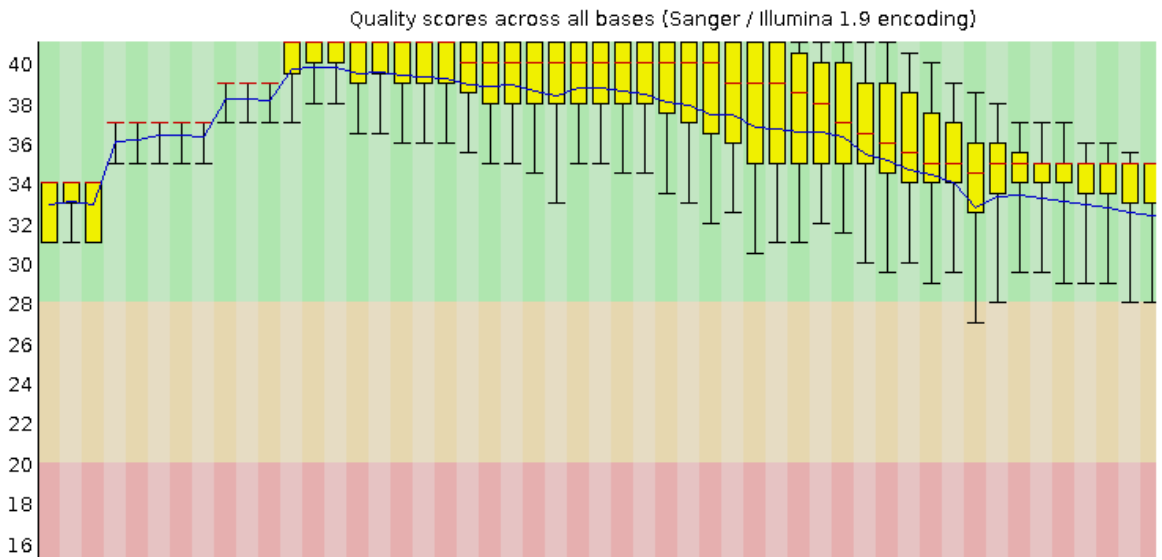
Mon 15 Nov 2021
subset_2_fastq

Summary

Basic Statistics

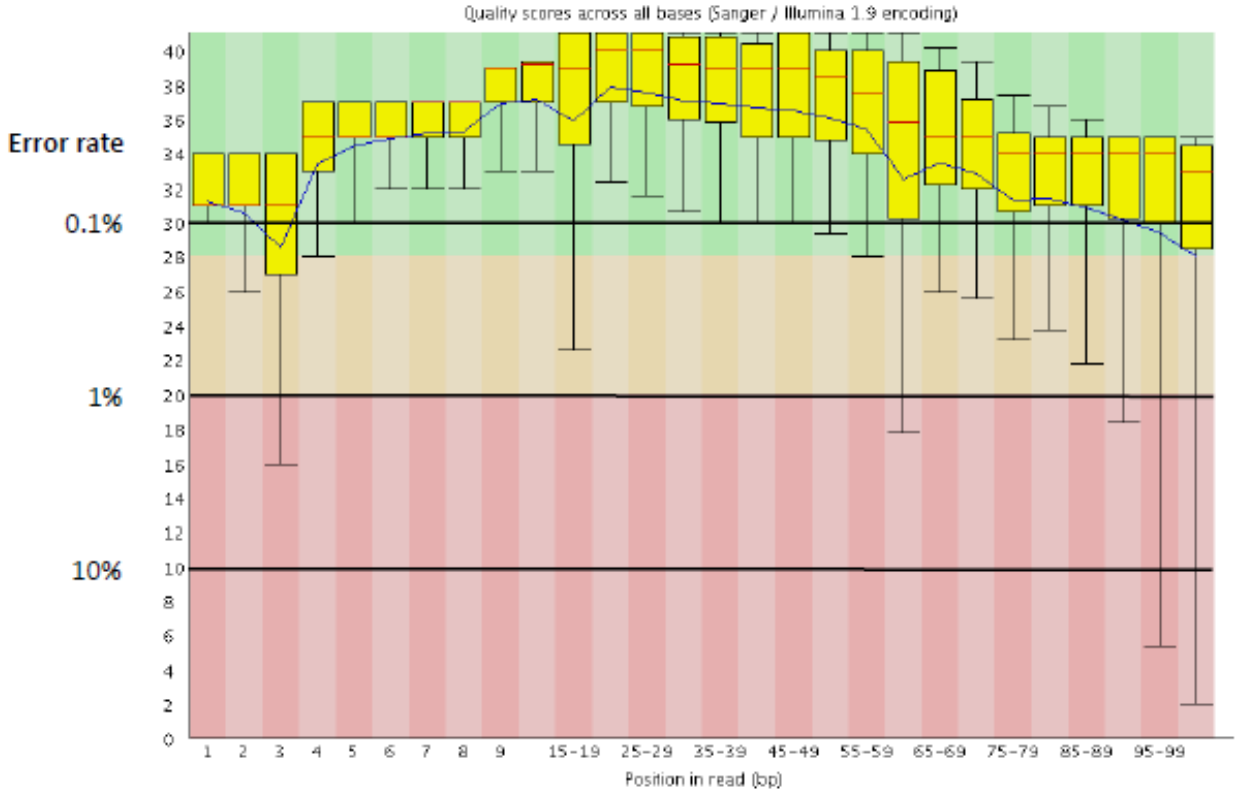
Measure	Value
Filename	MCL1-DK.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	100
%GC	50

Per base sequence quality



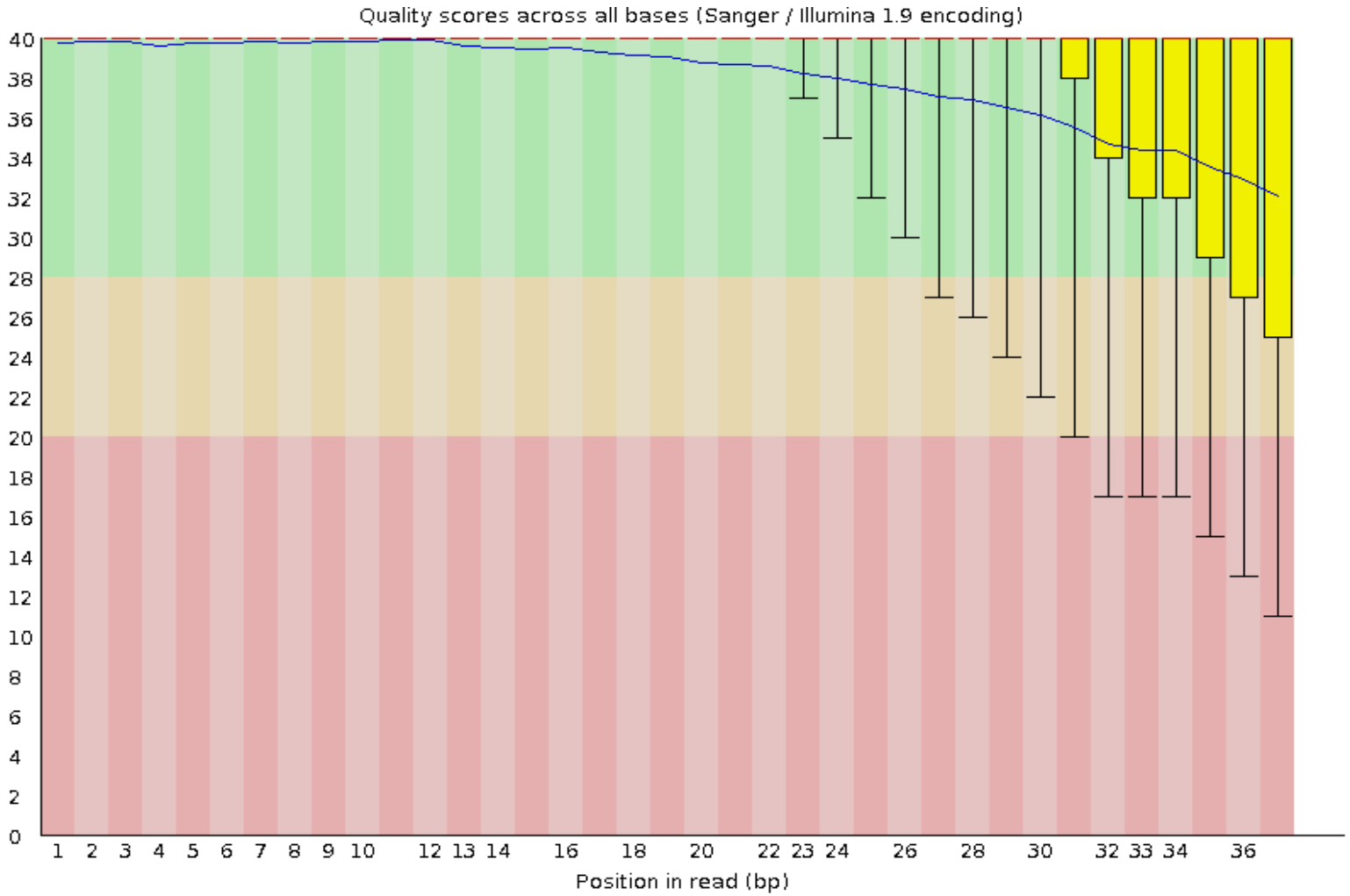
¿ Que nos dice un análisis de FASTQC?

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



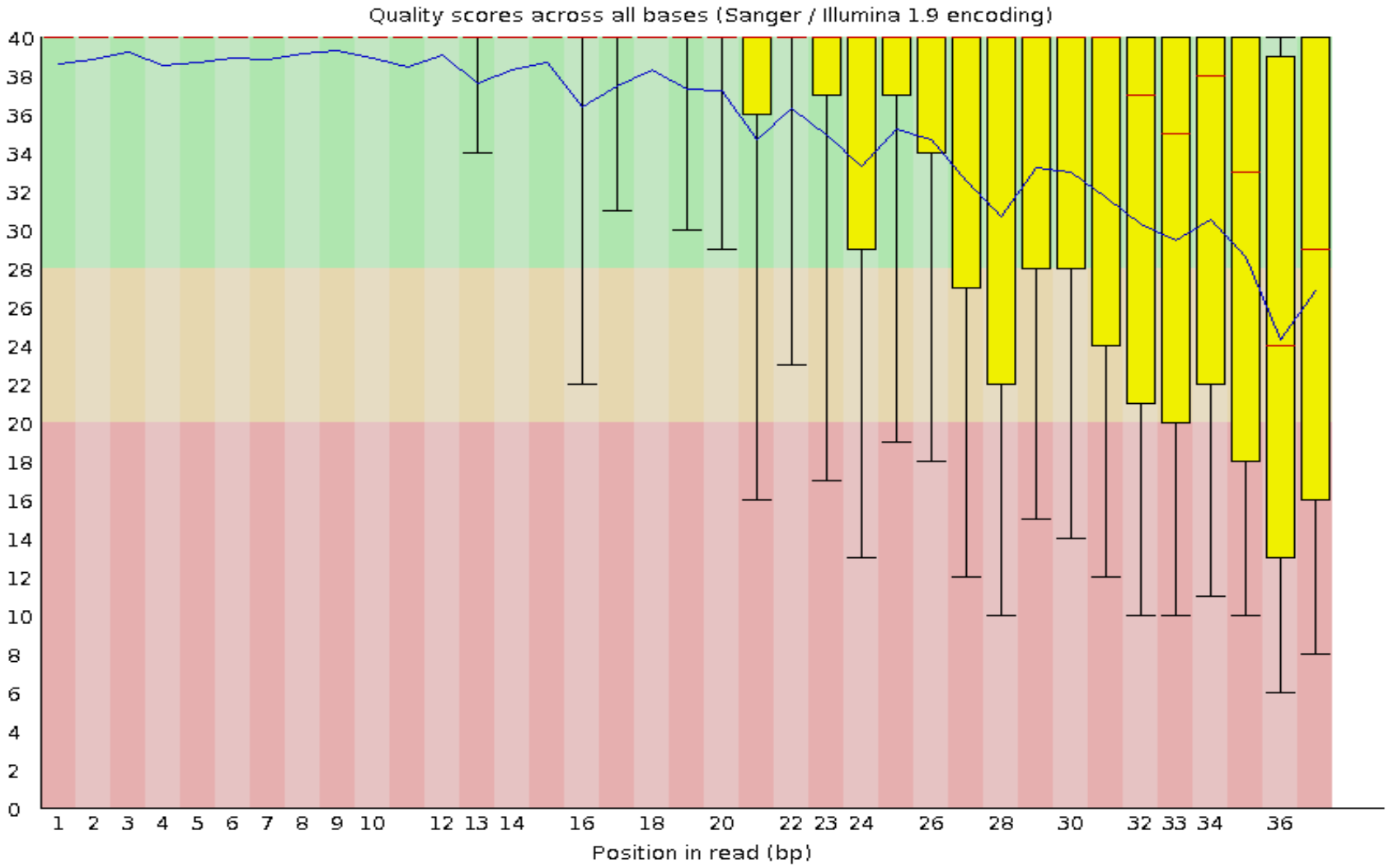
¿Qué es bueno y que es malo de acuerdo al valor de calidad por base?

Valor de calidad por base



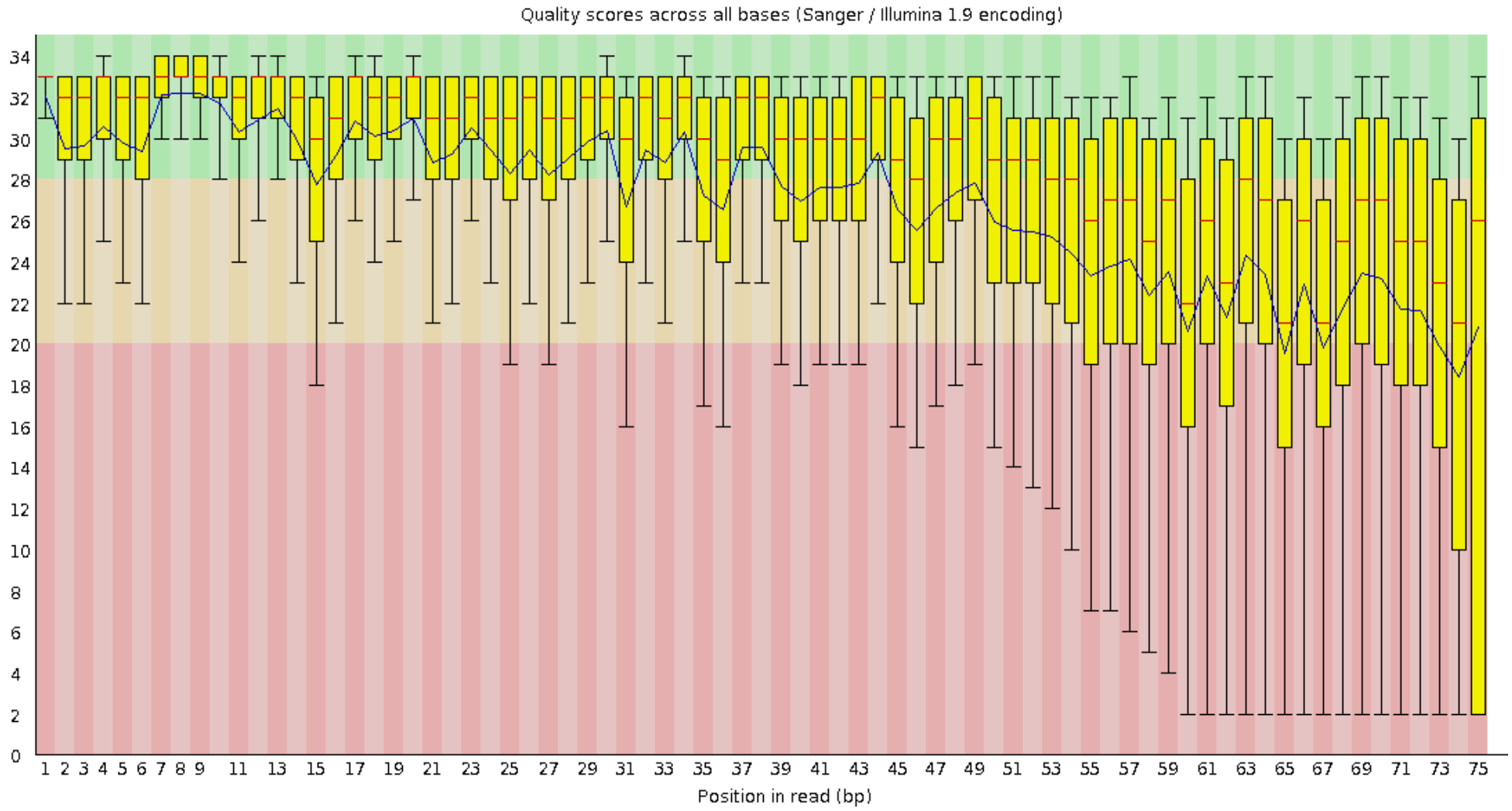
Buena calidad

Valor de calidad por base



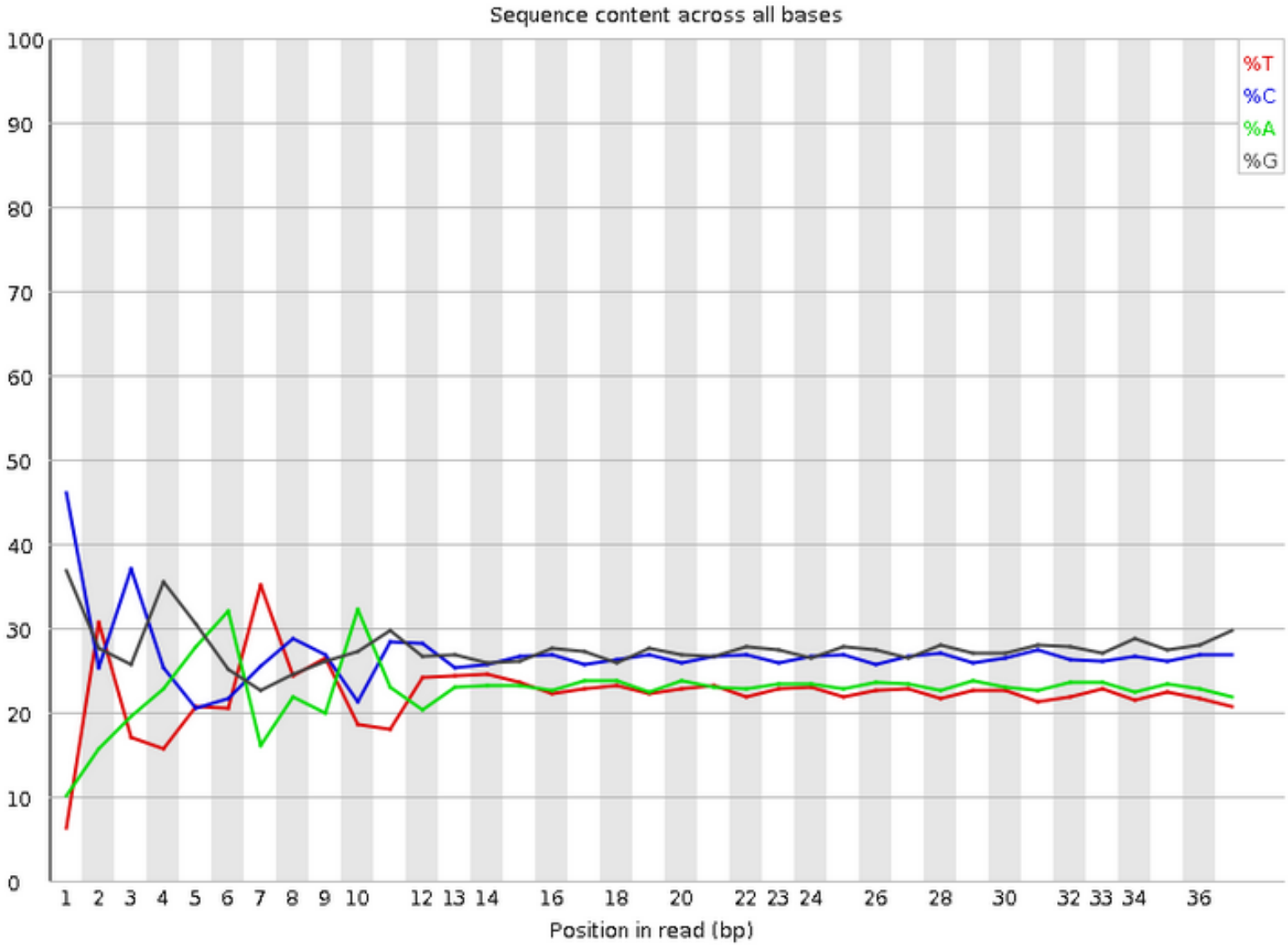
Calidad intermedia

Valor de calidad por base



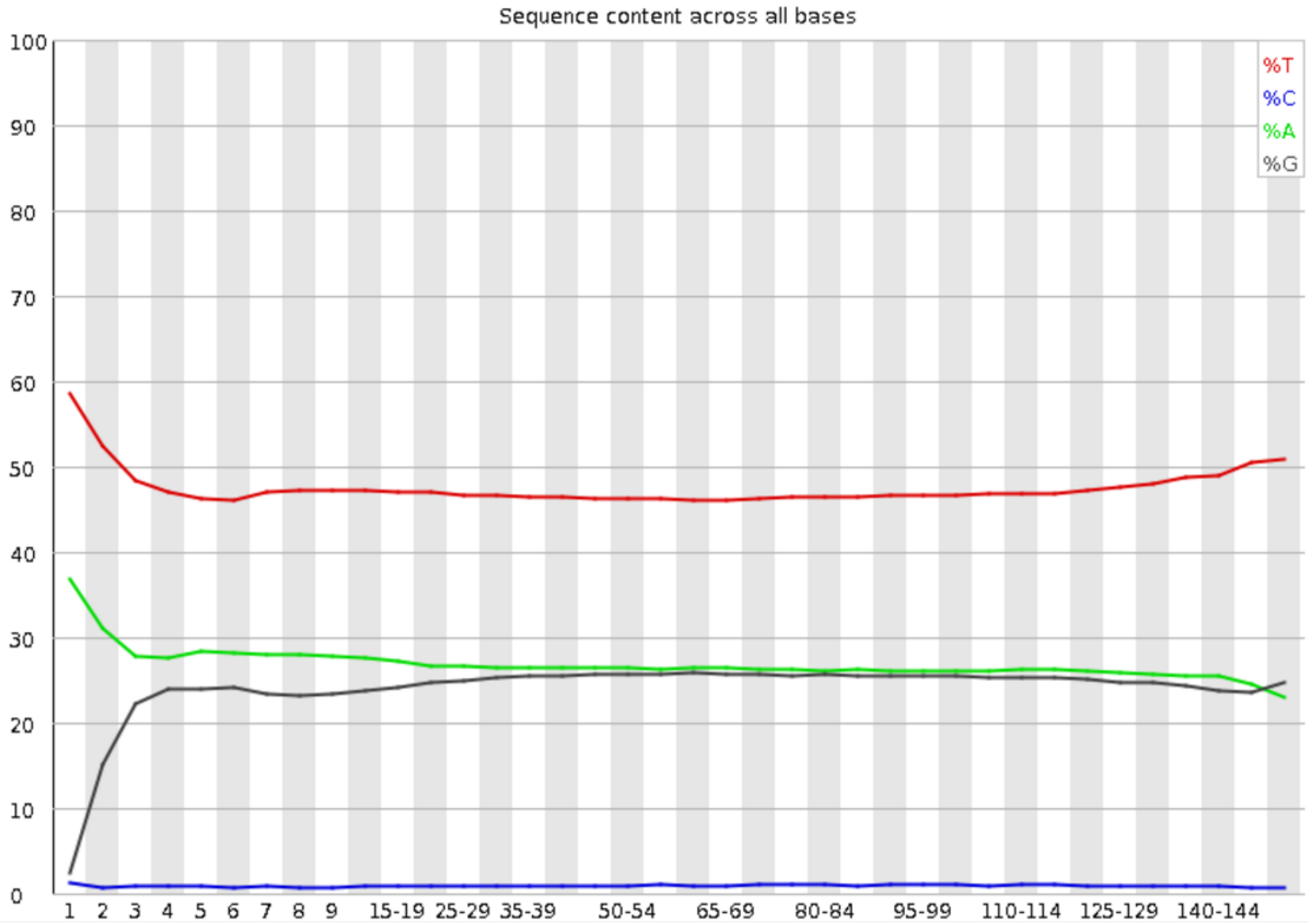
Mala calidad

Contenido de secuencias por base en un set de datos “normal”



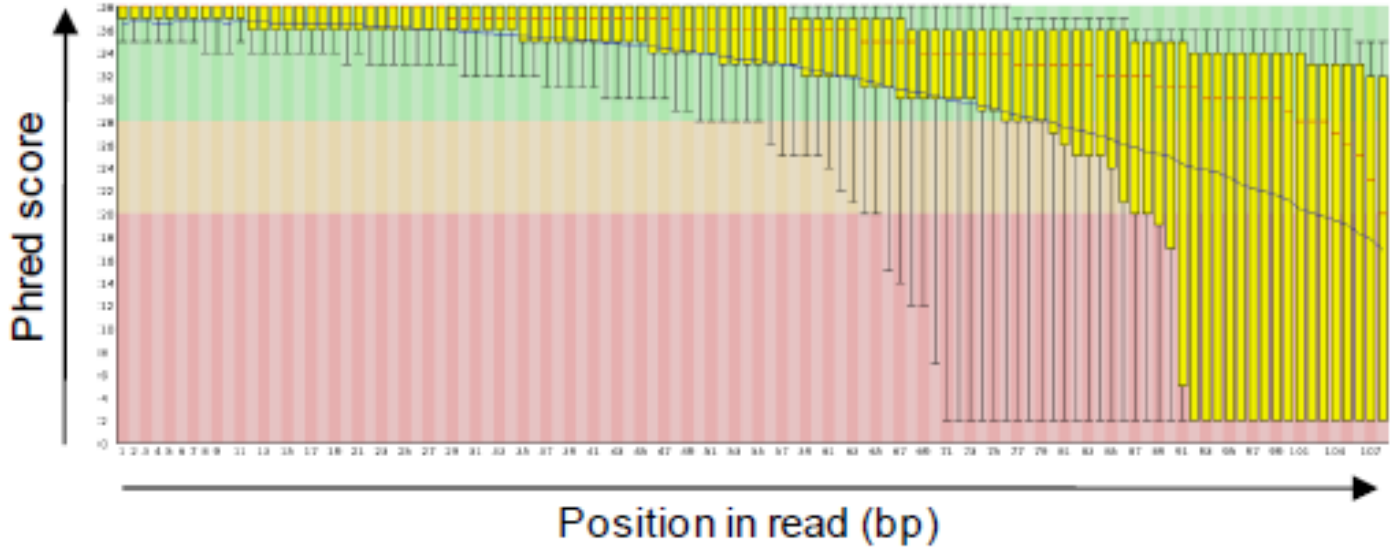
¿Se debería ver igual en secuenciación bisulfito?

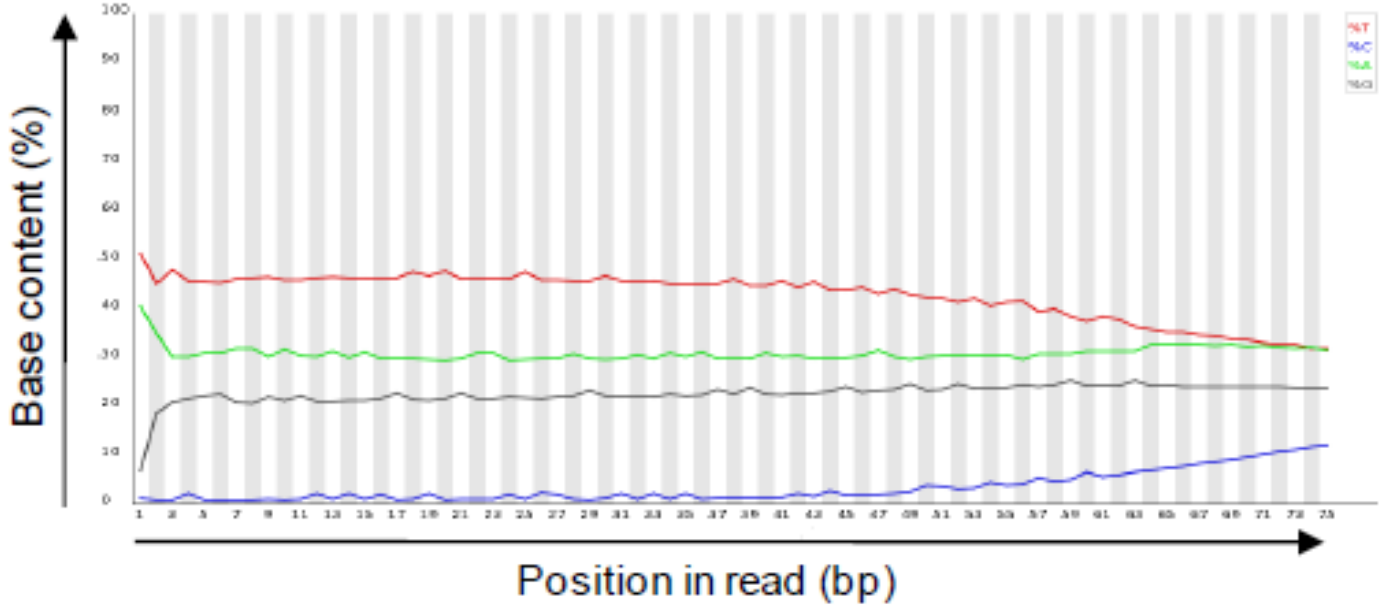
Per base sequence content



Este sería el resultado normal en secuenciación bisulfito

Algunos problemas comunes en secuenciación por bisulfito

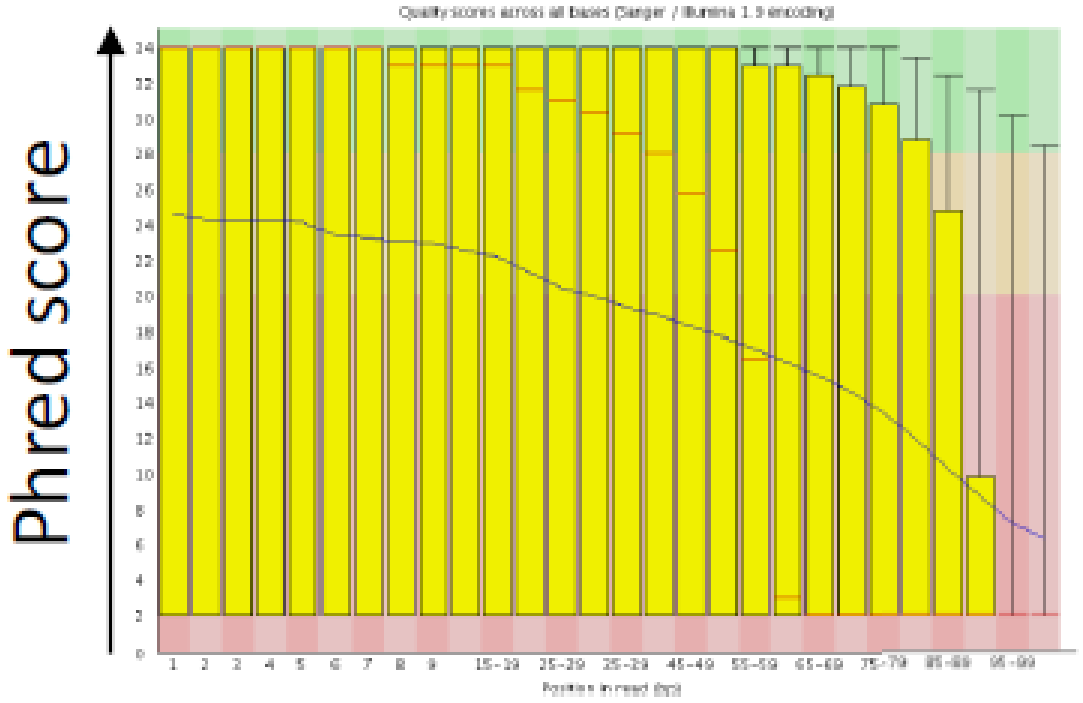




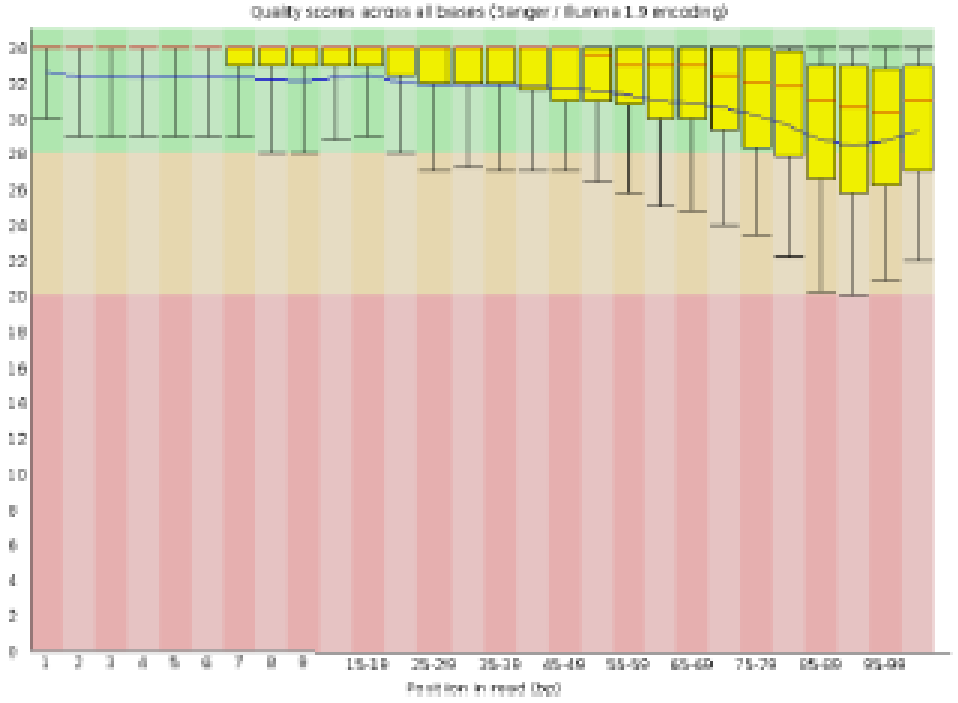
No se observa en librerías normales" como ChIP o RNAseq

FASTQC - Eliminación de bases o lecturas de baja calidad

Antes

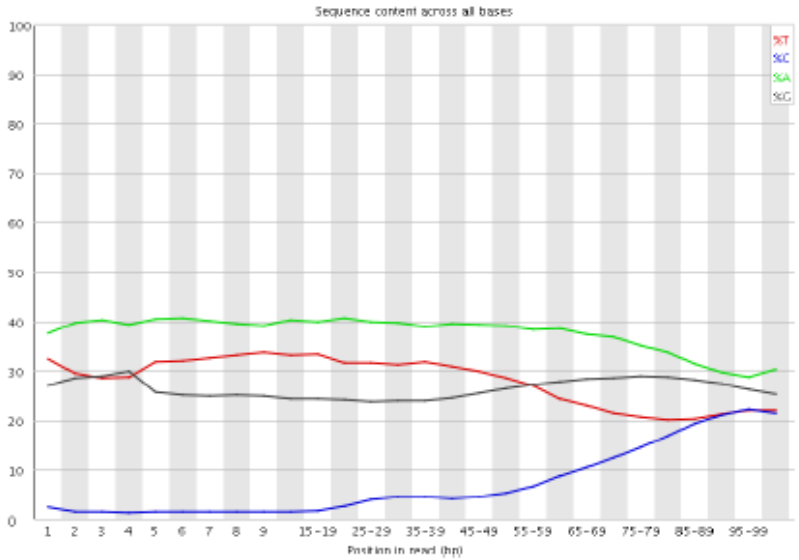


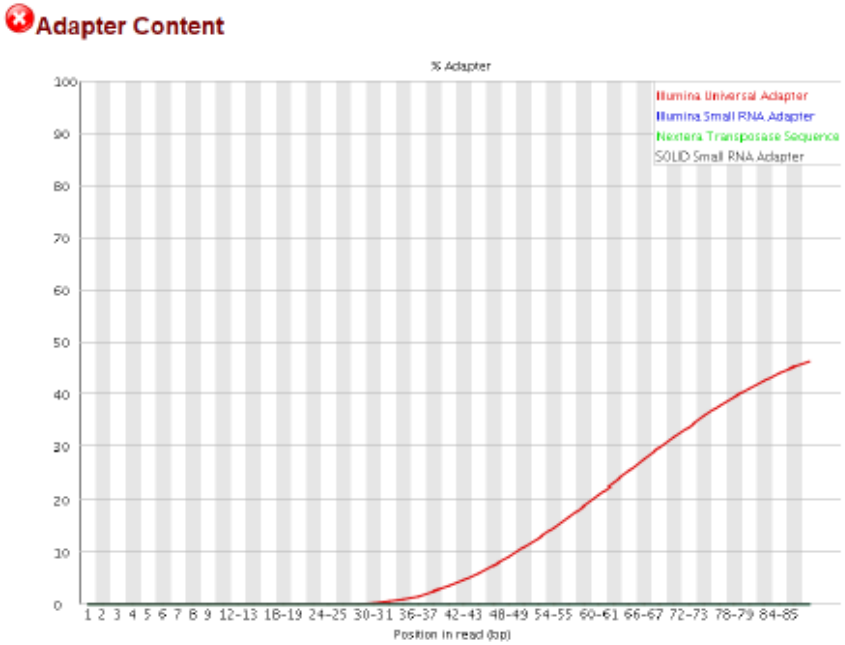
Después



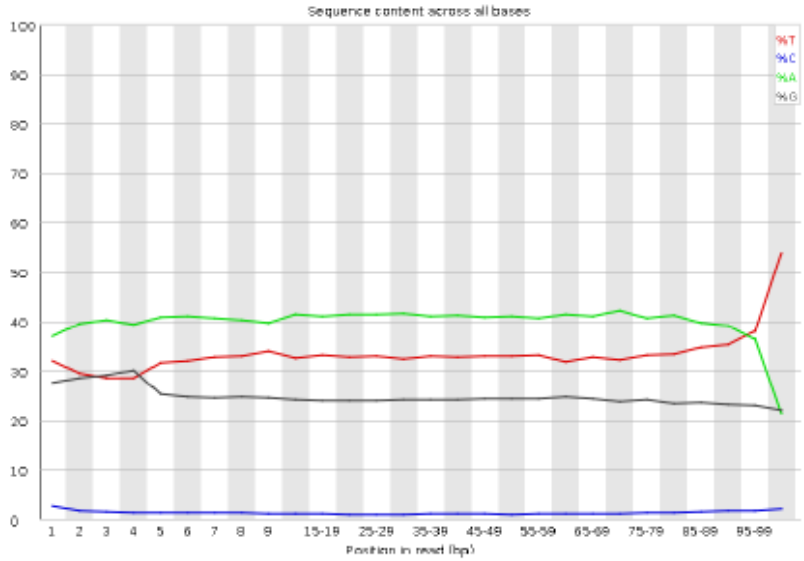
FASTQC - Remoción de la contaminación por adaptadores

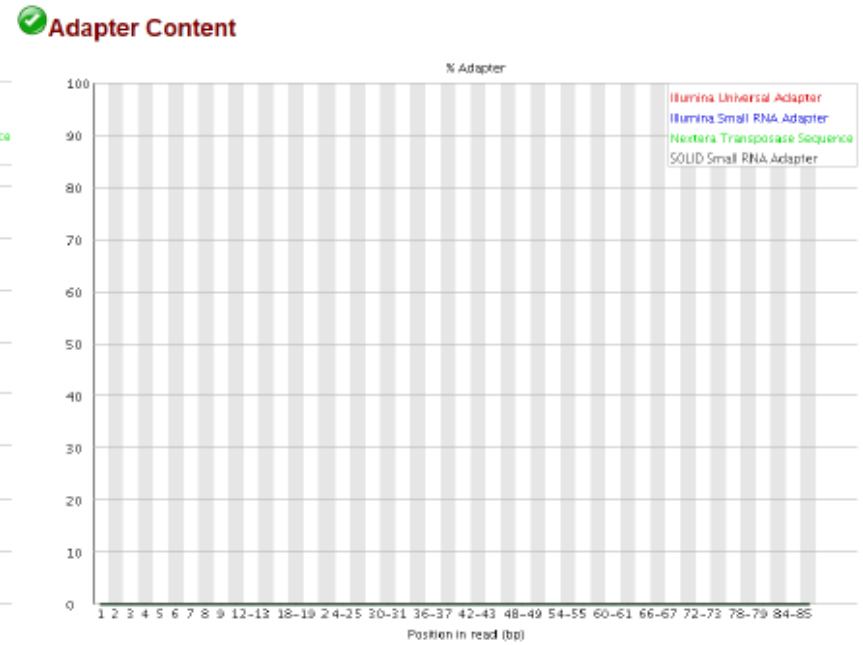
Antes





Después





Parte II: Alineamiento de secuencias y llamado de la metilación

Secuenciación, control de calidad y remoción de adaptadores

Alinear al genoma de referencia Bowtie2

Eliminar sesgo de PCR (Deduplication)

Llamado de metilación

Filtrado por cobertura

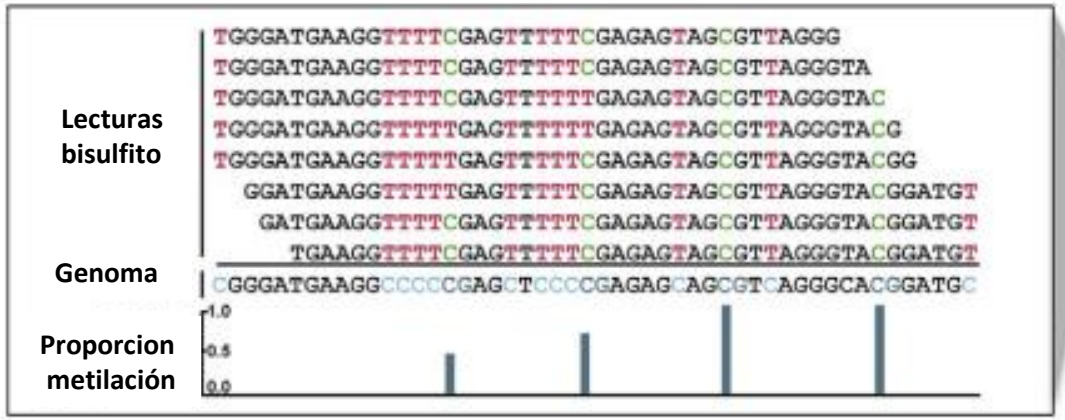
Visualización y agrupamiento

Metilación diferencial

Resultados como tablas, estadísticos de resumen, etc.

Segmentación de la metilación

Anotación de DMR/DMCs



Bismark v0.22.3

Babraham Bioinformatics

(Krueger & Andrews, 2011)

ÓMICAS

Bismark v0.23.0

Babraham Bioinformatics

https://github.com/FelixKrueger/Bismark/tree/master/Docs

(Krueger & Andrews, 2011)

Bismark es un programa para mapear lecturas de secuenciación tratadas con bisulfito a un genoma de interés y realizar el llamado de la metilación en un solo paso.

El archivo de salida se puede importar fácilmente a un visor de genoma, como SeqMonk y permite al investigador analizar los niveles de metilación de sus muestras de inmediato.

ÓMICAS

Bismark v0.23.0

Babraham Bioinformatics

https://github.com/FelixKrueger/Bismark/tree/master/Docs

(Krueger & Andrews, 2011)

Bismark es un programa para mapear lecturas de secuenciación tratadas con bisulfito a un genoma de interés y realizar el llamado de la metilación en un solo paso.

El archivo de salida se puede importar fácilmente a un visor de genoma, como SeqMonk y permite al investigador analizar los niveles de metilación de sus muestras de inmediato.

Sus principales características son:

- Mapeo de bisulfito y llamado de la metilación en un solo paso - Admite alineaciones de lectura single-end y paired-end
- Admite alineaciones sin gaps y con gaps
- Tiene parámetros de alineamiento ajustables
- Los archivos de salida se discriminan entre metilación de citosina en los contextos CpG, CHG y CHH

ÓMICAS

Bismark v0.23.0

Babraham Bioinformatics

https://github.com/FelixKrueger/Bismark/tree/master/Docs

(Krueger & Andrews, 2011)

Bismark es un programa para mapear lecturas de secuenciación tratadas con bisulfito a un genoma de interés y realizar el llamado de la metilación en un solo paso.

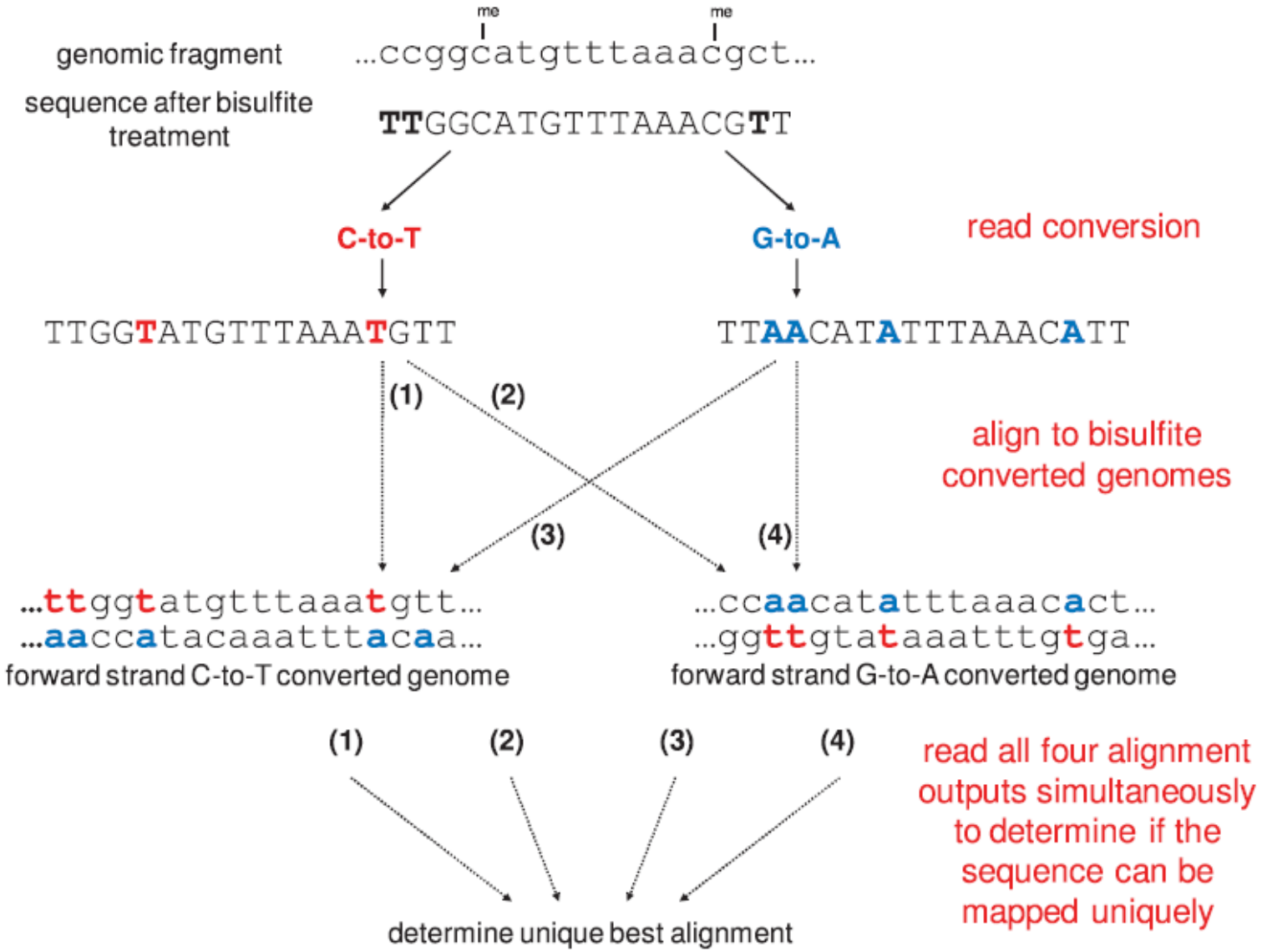
El archivo de salida se puede importar fácilmente a un visor de genoma, como SeqMonk y permite al investigador analizar los niveles de metilación de sus muestras de inmediato.

El mapeo de secuencias tratadas por bisulfito a un genoma de referencia constituye un reto computacional debido a:

- La reducida complejidad del código de ADN
- Hasta cuatro cadenas de ADN tienen que ser analizadas
- Cada lectura puede existir en todos los estados posibles de metilación

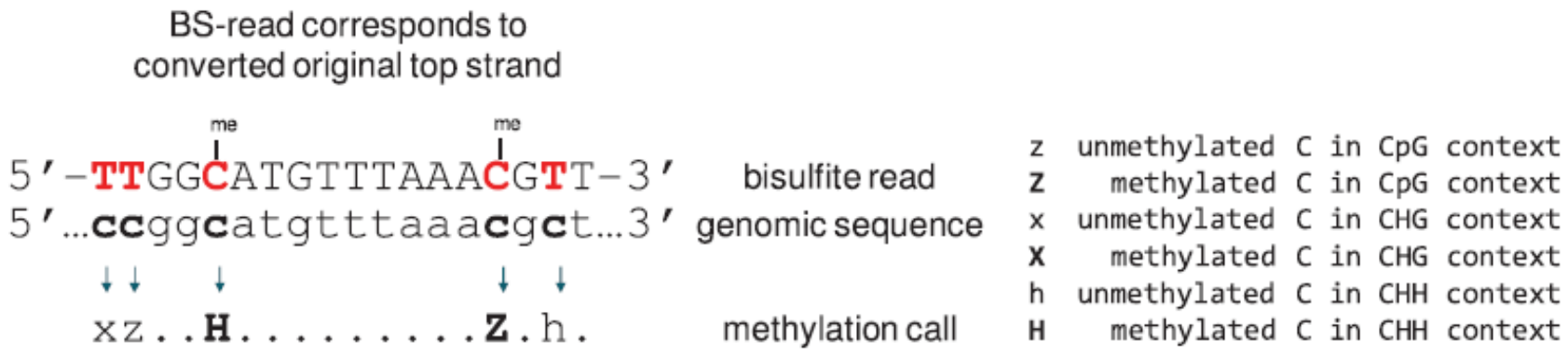
ÓMICAS

Enfoque de Bismark para el mapeo de secuencias bisulfito



Krueger & Andrews 2011

Determinación del estado de metilación de cada posición



Dependiendo de la hebra, una lectura mapeada contra una secuencia dada puede implicar buscar sustituciones de C a T (como se muestra aquí) o G a A.

Krueger & Andrews 2011

Corriendo Bismark…

El genoma de interés debe ser convertido a bisulfito e indexado para permitir el alineamiento por Bowtie

(I) Corriendo bismark_genome_preparation

USAGE:

```
insert your code bismark_genome_preparation [options] <path_to_genome_folder>
```

Un comando estandar para correr la preparación del genoma se vería asi:

```
bismark_genome_preparation --path_to_aligner /usr/bin/bowtie2 --verbose /home/jenny/Anal
ysis-I/genome/
```

or

```
bismark_genome_preparation --bowtie2 --verbose ./genome/
```

Corriendo Bismark…

El genoma de interés debe ser convertido a bisulfito e indexado para permitir el alineamiento por Bowtie

(I) Corriendo bismark_genome_preparation

USAGE:

```
insert your code bismark_genome_preparation [options] <path_to_genome_folder>
```

Un comando estandar para correr la preparación del genoma se vería asi:

```
bismark_genome_preparation --path_to_aligner /usr/bin/bowtie2 --verbose /home/jenny/Anal
ysis-I/genome/
```

or

```
bismark_genome_preparation --bowtie2 --verbose ./genome/
```

Solo se debe correr una vez para el genoma de interés

(II) Corriendo `bismark`

Alineamiento y llamado de la metilación

USAGE:

```
bismark [options] --genome <genome_folder> {-1 <mates1> -2 <mates2> | <singles>}
```

Ejemplo típico de un alineamiento:

```
bismark --multicore 4 --genome ./genome/ -1 WABQ1_L2_1.fq.gz,WABQ2_L1_1.fq.gz
-2 WABQ1_L2_2.fq.gz,WABQ2_L1_2.fq.gz --output_dir ./out/
```

Esto va a producir dos archivos de salida:

```
test_dataset_bismark_bt2.bam  (Contiene todos los alineamientos y el llamado de metilación)
test_dataset_bismark_SE_report.txt  (contiene el resumen del alineamiento y la metilación)
```

(II) Corriendo `bismark`

Alineamiento y llamado de la metilación

USAGE:

```
bismark [options] --genome <genome_folder> {-1 <mates1> -2 <mates2> | <singles>}
```

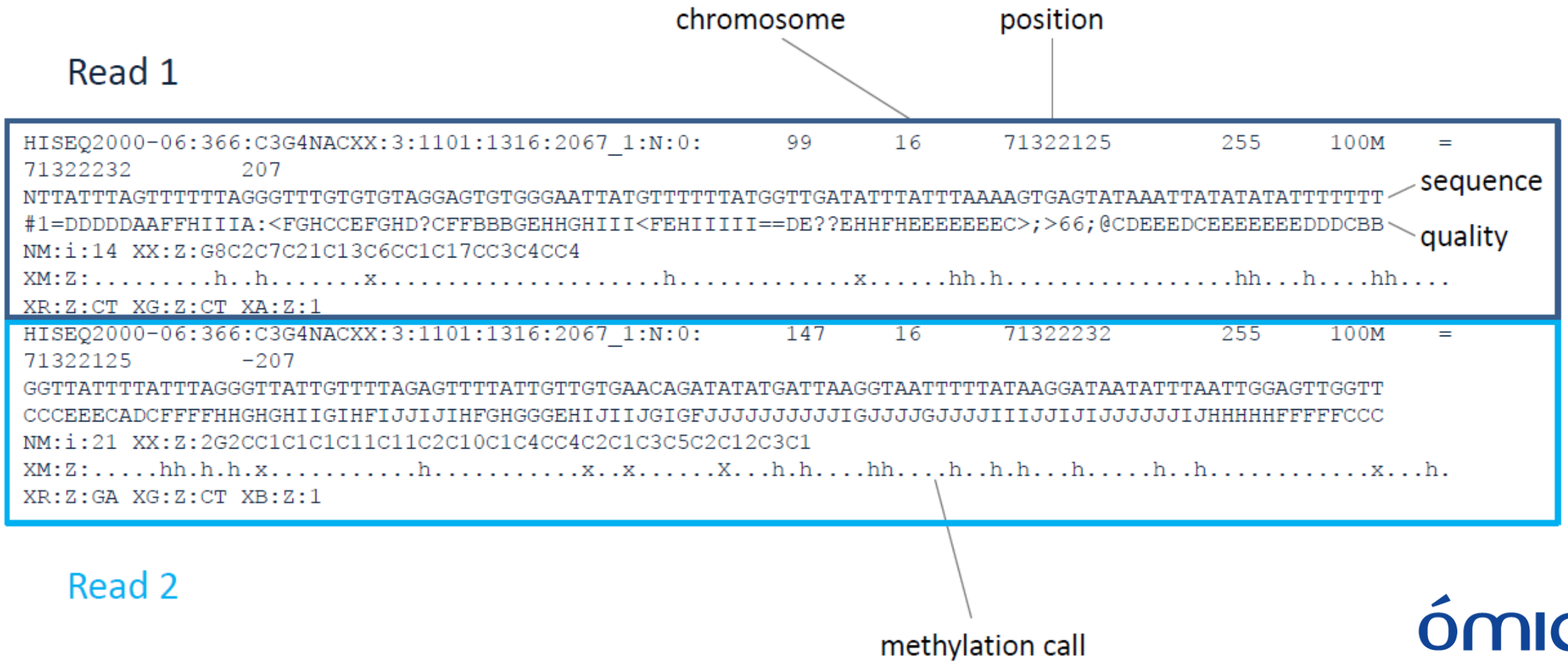
Ejemplo típico de un alineamiento:

```
bismark --multicore 4 --genome ./genome/ -1 WABQ1_L2_1.fq.gz,WABQ2_L1_1.fq.gz
-2 WABQ1_L2_2.fq.gz,WABQ2_L1_2.fq.gz --output_dir ./out/
```

Esto va a producir dos archivos de salida:

```
test_dataset_bismark_bt2.bam  (Contiene todos los alineamientos y el llamado de metilación)
test_dataset_bismark_SE_report.txt  (contiene el resumen del alineamiento y la metilación)
```

Alineamiento de secuencias
Archivo de salida del primer alineamiento en Bismark (BAM file)



(III) Corriendo `deduplicate_bismark`

USAGE:

```
deduplicate_bismark --bam [options] <filenames>
```

Este comando deduplicará el archivo BAM de alineación Bismark y eliminará todas las lecturas excepto una que se alinee en la misma posición y en la misma orientación. Este paso se recomienda para muestras de bisulfito de genoma completo, pero no debe usarse para bibliotecas de representación reducida como RRBS, amplicones o bibliotecas de enriquecimiento.

(IV) Corriendo `bismark_methylation_extractor`

USAGE:

```
bismark_methylation_extractor [options] <filenames>
```

Un camando típico para extraer la metilación contexto-dependiente (CpG/CHG/CHH) sería asi:

```
bismark_methylation_extractor --gzip --bedGraph test_dataset_bismark_bt2.bam
```

or

```
bismark_methylation_extractor --gzip --bedGraph --buffer_size 10G --comprehensive ./out/
WABQ1_L2_1_bismark_bt2_pe.deduplicated.bam --output ./out/met_extractor
```

(IV) Corriendo `bismark_methylation_extractor`

USAGE:

```
bismark_methylation_extractor [options] <filenames>
```

Un camando típico para extraer la metilación contexto-dependiente (CpG/CHG/CHH) sería asi:

```
bismark_methylation_extractor --gzip --bedGraph test_dataset_bismark_bt2.bam
```

or

```
bismark_methylation_extractor --gzip --bedGraph --buffer_size 10G --comprehensive ./out/
WABQ1_L2_1_bismark_bt2_pe.deduplicated.bam --output ./out/met_extractor
```

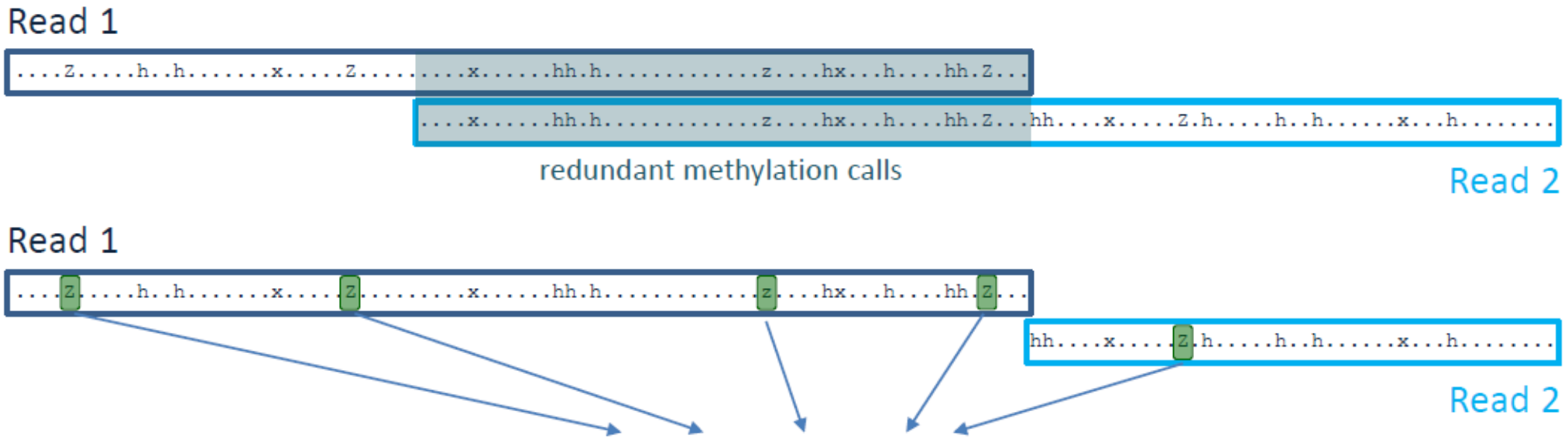
Esto va a producir tres archivos de salida:

```
CpG_context_test_dataset_bismark_bt2.txt.gz
CHG_context_test_dataset_bismark_bt2.txt.gz
CHH_context_test_dataset_bismark_bt2.txt.gz
```

Formato de salida →

1. seq-ID
2. methylation state
3. chromosome
4. start position (= end position)
5. methylation call

Extracción de la metilación



CpG methylation output

```
Bismark methylation extractor version v0.10.1
HS9_11915:8:2311:4022:38651#13/1        +       1       3029229 Z
HS9_11915:8:1208:13025:95413#13/1       +       1       3079409 Z
HS9_11915:8:1301:11752:81850#13/1       -       1       3104640 z
HS9_11915:8:2112:15483:84166#13/1       +       1       3104862 Z
HS9_11915:8:2110:8777:33683#13/1        -       1       3104862 z
HS9_11915:8:2208:16561:25806#13/1       +       1       3104862 Z
HS9_11915:8:2308:15290:100335#13/1      -       1       3124392 z
HS9_11915:8:2308:15290:100335#13/1      +       1       3124416 Z
HS9_11915:8:2212:13818:79056#13/1       +       1       3124416 Z
HS9_11915:8:2105:9522:91783#13/1        +       1       3124392 Z
HS9_11915:8:2105:9522:91783#13/1        +       1       3124416 Z
```

read ID | meth state | chr | pos | context

Extracción de la metilación I

```
Bismark methylation extractor version v0.10.1
HS9_11915:8:2311:4022:38651#13/1        +       1       3029229 Z
HS9_11915:8:1208:13025:95413#13/1       +       1       3079409 Z
HS9_11915:8:1301:11752:81850#13/1       -       1       3104640 z
HS9_11915:8:2112:15483:84166#13/1       +       1       3104862 Z
HS9_11915:8:2110:8777:33683#13/1        -       1       3104862 z
HS9_11915:8:2208:16561:25806#13/1       +       1       3104862 Z
```

CpG methylation output

bismark2bedGraph

```
1       5705370 5705370 100     1       0
1       5706335 5706335 60      3       2
1       5706336 5706336 100     3       0
1       5706453 5706453 75      3       1
1       5706454 5706454 0       0       2
1       5706845 5706845 71.4285714285714        5       2
1       5706846 5706846 66.6666666666667        2       1
1       5707925 5707925 0       0       1
1       5707926 5707926 66.6666666666667        2       1
1       5709177 5709177 100     2       0
1       5709178 5709178 0       0       1
1       5710030 5710030 66.6666666666667        4       2
```

chr pos methylation percentage meth unmeth

bedGraph/coverage output

Extracción de la metilación II

1	10525	10525	66.6666666666667		2	1
1	10542	10542	100	3	0	
1	10563	10563	66.6666666666667		2	1
1	10571	10571	100	3	0	
1	10577	10577	66.6666666666667		2	1
1	10579	10579	100	3	0	
1	10589	10589	50	2	2	
1	10609	10609	0	0	1	
1	10617	10617	0	0	1	
1	10620	10620	0	0	1	

coverage output

coverage2cytosine

1	10525	+	2	1	CG	CGC
1	10526	-	0	0	CG	CGG
1	10542	+	3	0	CG	CGA
1	10543	-	0	0	CG	CGG
1	10563	+	2	1	CG	CGC
1	10564	-	0	0	CG	CGT
1	10571	+	3	0	CG	CGC
1	10572	-	0	0	CG	CGG
1	10577	+	2	1	CG	CGC
1	10578	-	0	0	CG	CGA
1	10579	+	3	0	CG	CGG
1	10580	-	0	0	CG	CGC
1	10589	+	2	2	CG	CGG

chr pos strand meth unmeth di-nuc tri-nuc

optional: merge into
CpG dinucleotide entities

Genome wide CpG report

(IV) Corriendo `bismark_methylation_extractor`

USAGE:

```
bismark_methylation_extractor [options] <filenames>
```

Un camando típico para extraer la metilación contexto-dependiente (CpG/CHG/CHH) incluyendo el archivo .bed y el reporte por pares de nucleotidos sería asi:

```
bismark_methylation_extractor --gzip --bedGraph --buffer_size 10G --comprehensive --cyto
sine_report --genome_folder /path_to_genome_folder/ WABQ1_L2_1_bismark_bt2_pe.deduplicat
ed.bam --output ./out
```

(V) Corriendo `bismark2report and bismark2summary`

USAGE:

```
bismark2report [options]

bismark2summary [options]
```

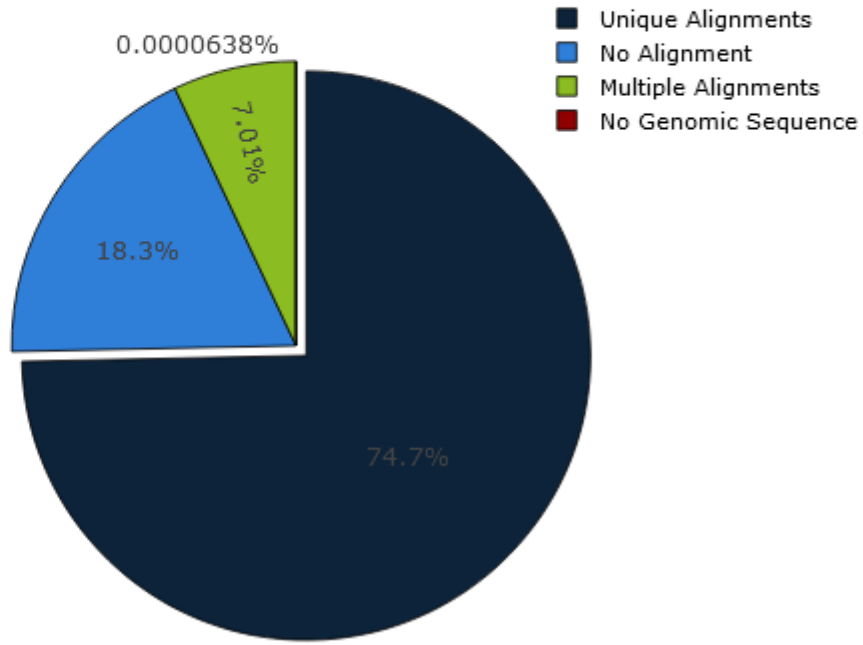
Bismark Processing Report

SRR5008109_WGBS_Seq_E14_d11_1_control_1_val_1.fq.gz and SRR5008109_WGBS_Seq_E14_d11_1_control_2_val_2.fq.gz

Data processed at 11:25 on 2018-08-16

Alignment Stats

Sequence pairs analysed in total	9399055
Paired-end alignments with a unique best hit	7018098
Pairs without alignments under any condition	1722286
Pairs that did not map uniquely	658671
Genomic sequence context not extractable (edges of chromosomes)	6

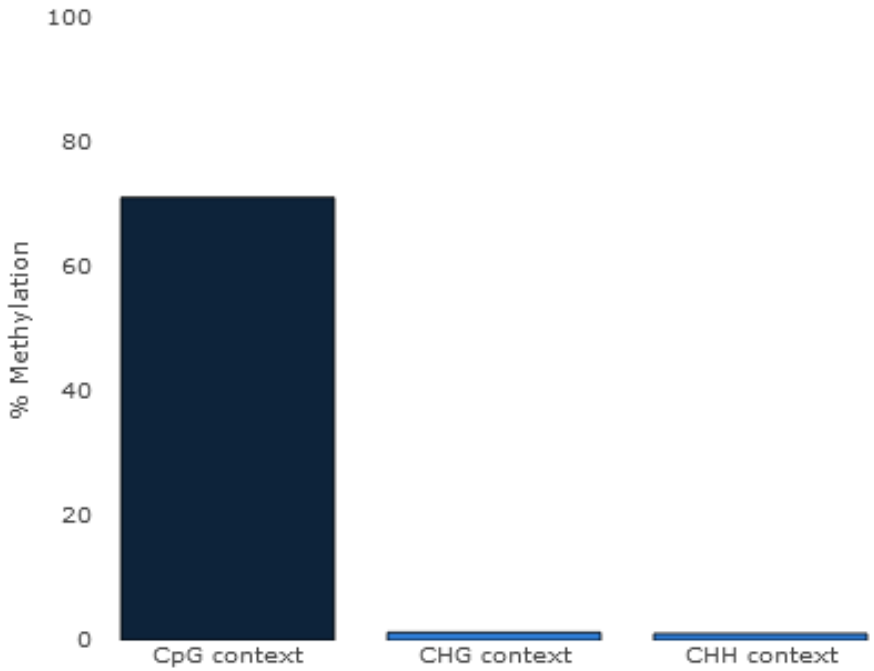


https://www.bioinformatics.babraham.ac.uk/projects/bismark/PE_report.html

Bismark Processing Report

Cytosine Methylation

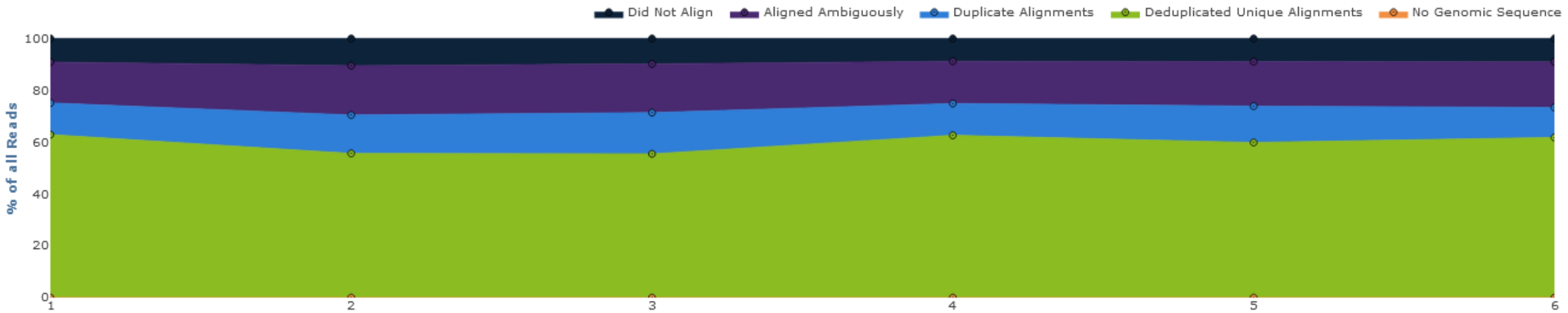
Total C's analysed	264469464
Methylated C's in CpG context	8364359
Methylated C's in CHG context	755431
Methylated C's in CHH context	2166832
Methylated C's in Unknown context	535
Unmethylated C's in CpG context	3380818
Unmethylated C's in CHG context	57592010
Unmethylated C's in CHH context	192210014
Unmethylated C's in Unknown context	56967
Percentage methylation (CpG context)	71.2%
Percentage methylation (CHG context)	1.3%
Percentage methylation (CHH context)	1.1%
Methylated C's in Unknown context	N/A%



https://www.bioinformatics.babraham.ac.uk/projects/bismark/PE_report.html

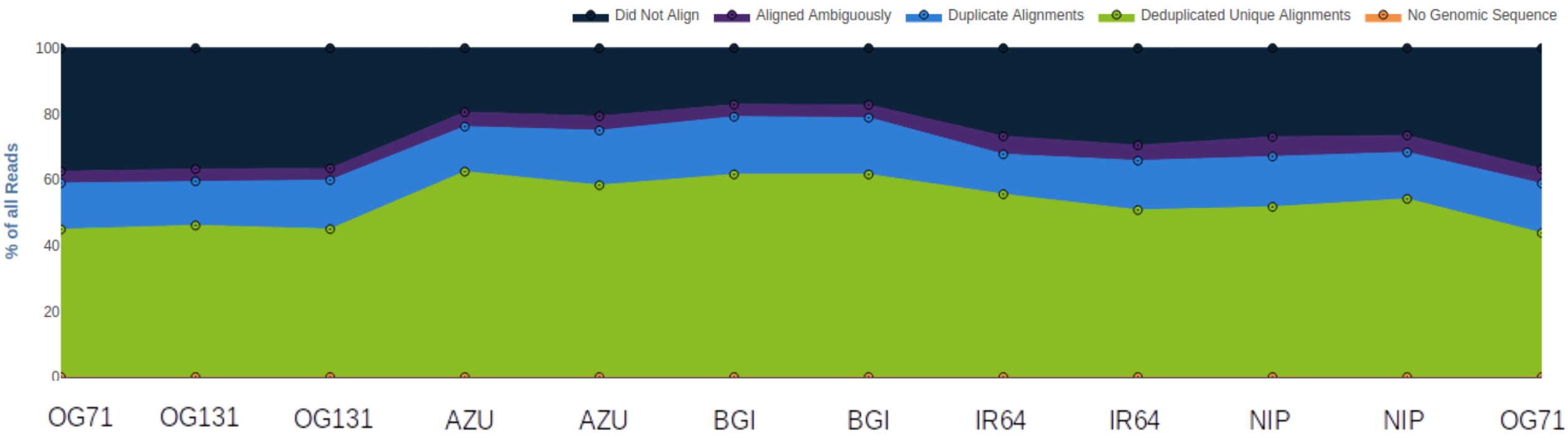
Bismark Project Overall Summary

Alignment Statistics



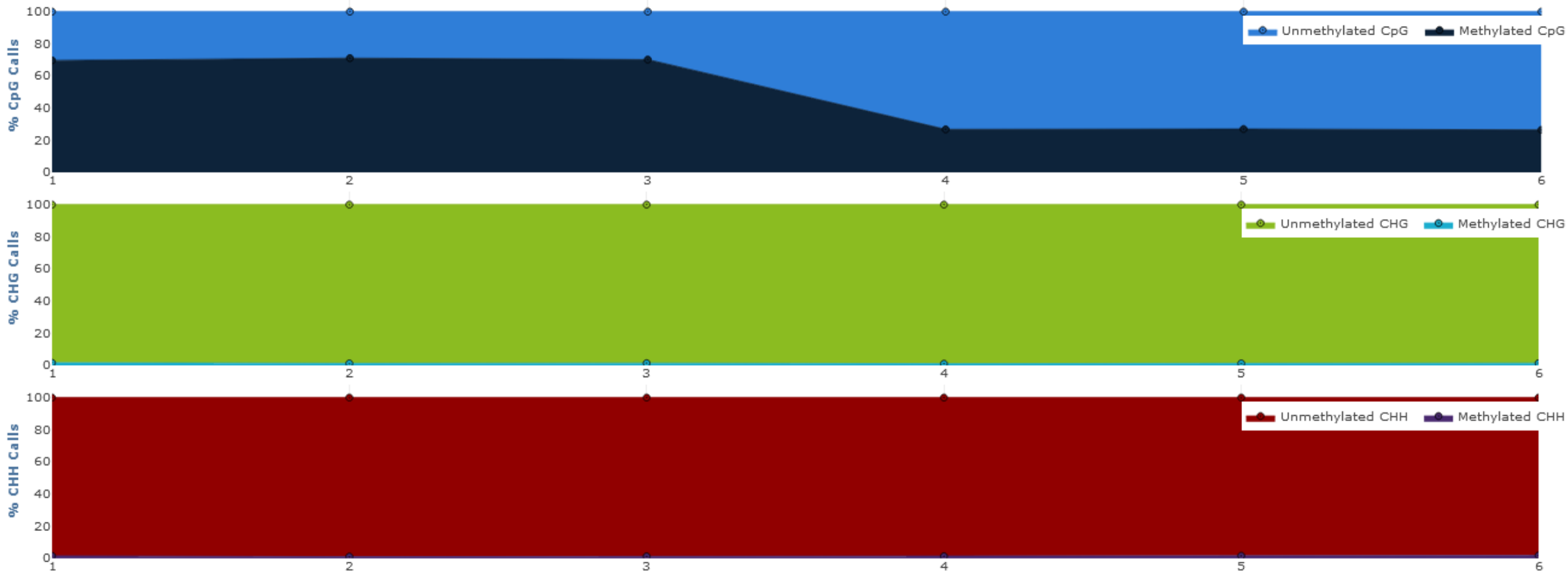
https://www.bioinformatics.babraham.ac.uk/projects/bismark/bismark_summary_WGBS.html

Bismark Project Overall Summary



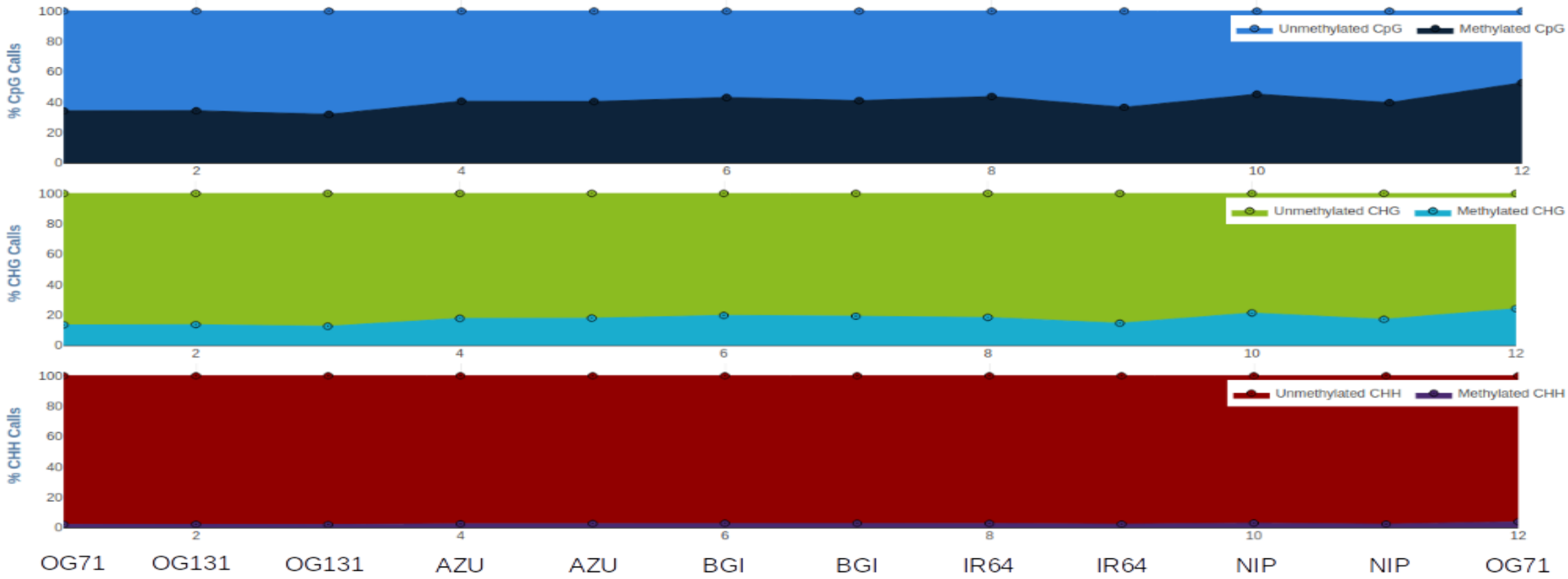
Bismark Project Overall Summary

Cytosine Methylation

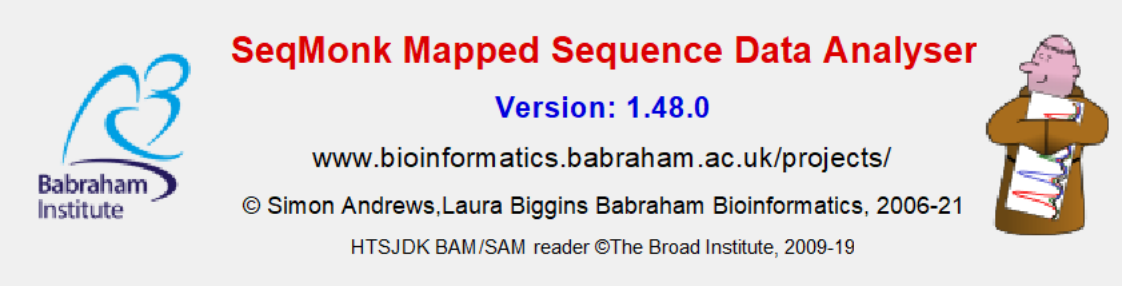


Bismark Project Overall Summary

Cytosine Methylation



Visualización de datos utilizando Seq-Monk



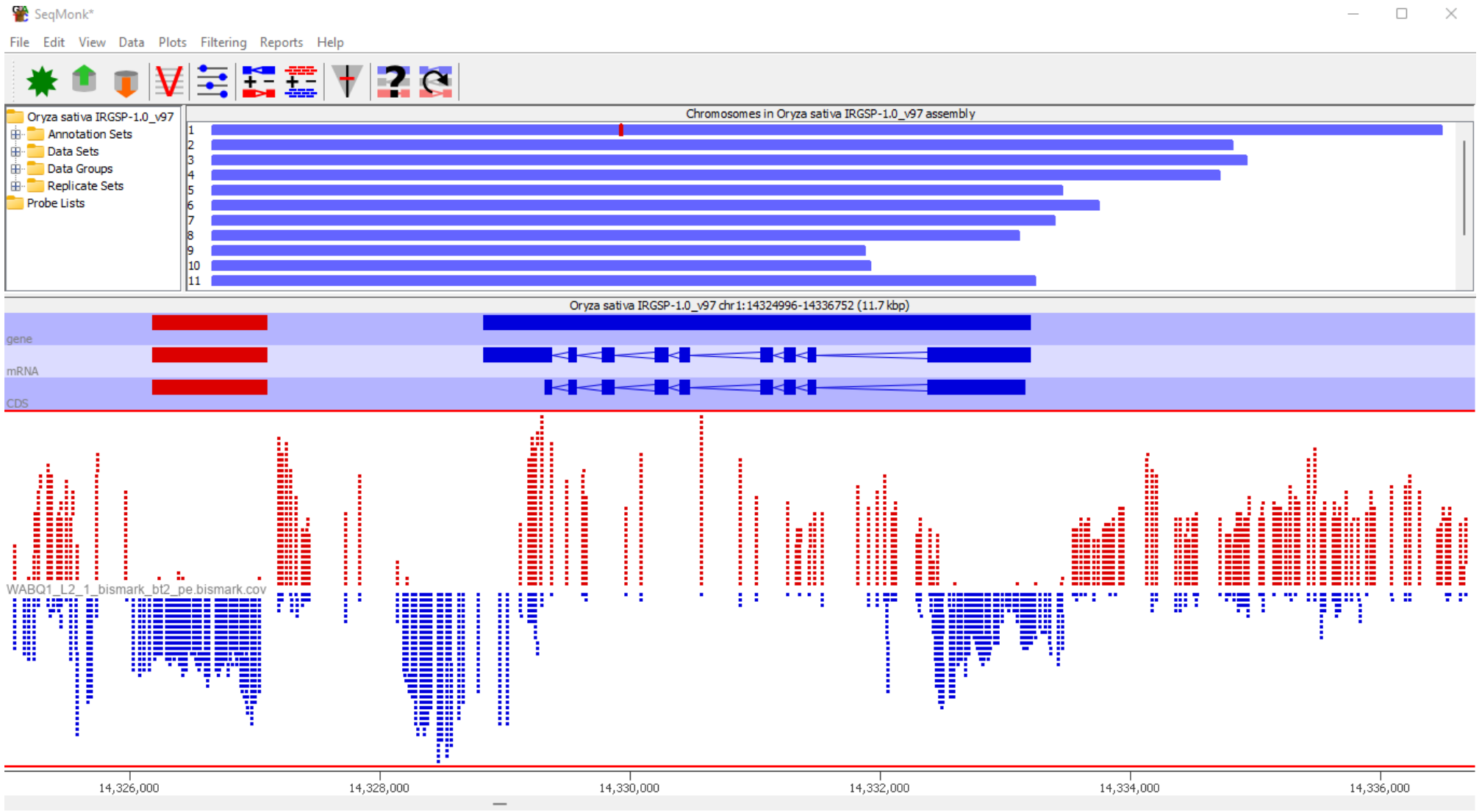
Datos de entrada:

-**BAM/SAM** provenientes de Bismark
-Archivos de **cobertura** provenientes de Bismark
-**BedGraph** archivos provenientes de Bismark

- Archivos de llamado de la metilación generados en methylkit (.txt)

Archivos de **cobertura**

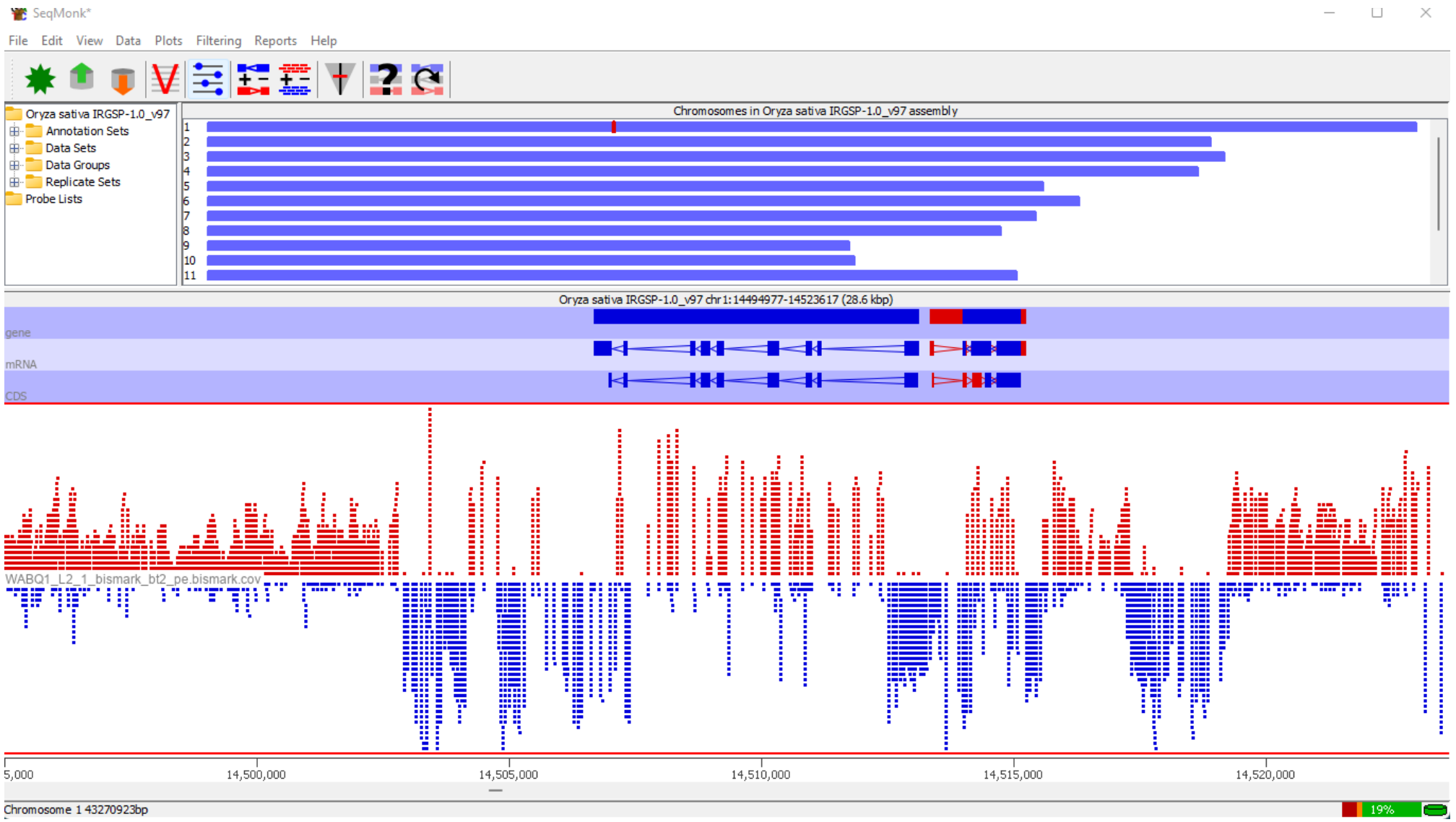
Visualización de datos utilizando Seq-Monk



Rojo = Metilado

Azul = No-Metilado

Visualización de datos utilizando Seq-Monk



Rojo = Metilado

Azul = No-Metilado

Análisis de datos para datos de secuenciación por bisulfito.

Secuenciación, control de calidad y remoción de adaptadores

Alinear al genoma de referencia Bowtie2

Eliminar sesgo de PCR (Deduplication)

Llamado de metilación

Filtrado por cobertura

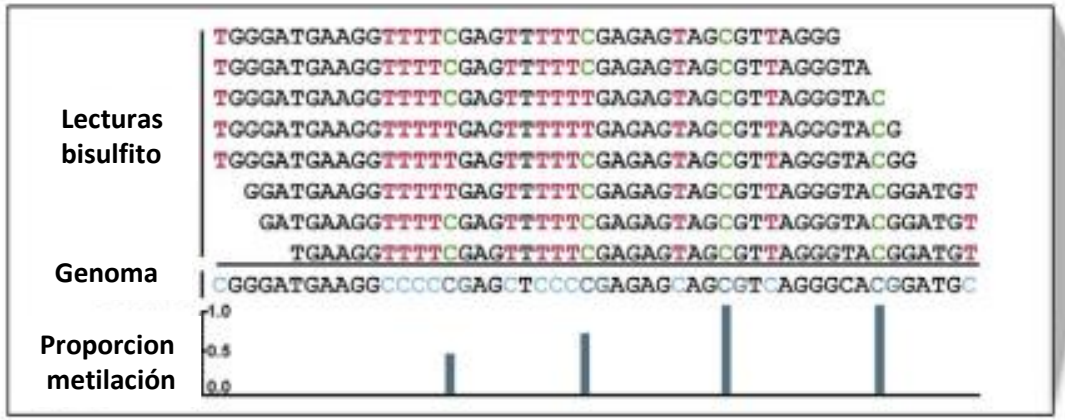
Visualización y agrupamiento

Metilación diferencial

Resultados como tablas, estadisticos de resumen, etc.

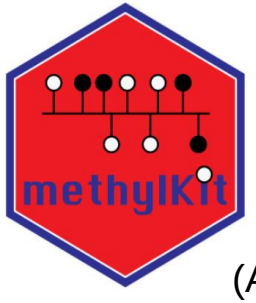
Segmentación de la metilación

Anotación de DMR/DMCs



Methylkit v.1.15

(Akalin et al., 2012)



Methylkit v.1.15

(Akalin et al., 2012)

https://bioconductor.org/packages/devel/bioc/vignettes/methylKit/inst/doc/methylKit.html

MethylKit es un paquete de R para el análisis y anotación de datos de metilación del ADN obtenidos por tecnologías de secuenciación masiva.

Descarga de archivos para la parte practica

https://drive.google.com/drive/folders/11l8hNlqHVw0OBo-arh3MtBxM5rykD1__?usp=sharing

Corriendo MethylKit…

1.1 Crear el objeto methylkit a partir de los archivos BAM provenientes de Bismark

```
my.methRaw = processBismarkAln(location = Bam_files, sample.id=samples, assembly="rice",
                               treatment = c(0,1), read.context="CpG",
                               mincov=10, save.folder=outDir)
```

	chrBase	chr	base	strand	coverage	freqC	freqT
2	Chr1.10145	Chr1	10145	F	24	0.00	100.00
3	Chr1.10149	Chr1	10149	F	24	0.00	100.00
4	Chr1.10290	Chr1	10290	F	29	0.00	100.00
5	Chr1.10327	Chr1	10327	F	29	0.00	100.00
6	Chr1.10366	Chr1	10366	F	25	0.00	100.00
7	Chr1.10419	Chr1	10419	F	19	0.00	100.00
8	Chr1.10429	Chr1	10429	F	18	0.00	100.00
9	Chr1.10476	Chr1	10476	F	19	0.00	100.00
10	Chr1.10544	Chr1	10544	F	16	0.00	100.00
11	Chr1.10549	Chr1	10549	F	16	0.00	100.00
12	Chr1.10573	Chr1	10573	F	15	0.00	100.00
13	Chr1.10592	Chr1	10592	F	16	0.00	100.00
14	Chr1.10596	Chr1	10596	F	15	0.00	100.00
15	Chr1.10646	Chr1	10646	F	11	0.00	100.00
16	Chr1.1192	Chr1	1192	F	10	100.00	0.00
17	Chr1.1218	Chr1	1218	F	13	100.00	0.00
18	Chr1.1222	Chr1	1222	F	14	100.00	0.00
19	Chr1.1238	Chr1	1238	F	15	100.00	0.00

Archivo generado en Methylkit

sativaCpG.txt

Corriendo MethylKit…

1.1 Crear el objeto methylkit a partir de los archivos BAM provenientes de Bismark

```
my.methRaw = processBismarkAln(location = Bam_files, sample.id=samples, assembly="rice",
                               treatment = c(0,1), read.context="CpG",
                               mincov=10, save.folder=outDir)
```

1.2 Crear el objeto methylkit a partir de los ***archivos de texto*** obtenidos en el paso anterior y los guarda como archivos "flat file database" en la carpeta methylDB_CpG

```
my.methRaw2 = methRead(location = files, sample.id=samples, assembly="rice",
                       treatment = c(0,1), context="CpG",
                       dbtype = "tabix", dbdir = "methylDB_CpG")
```

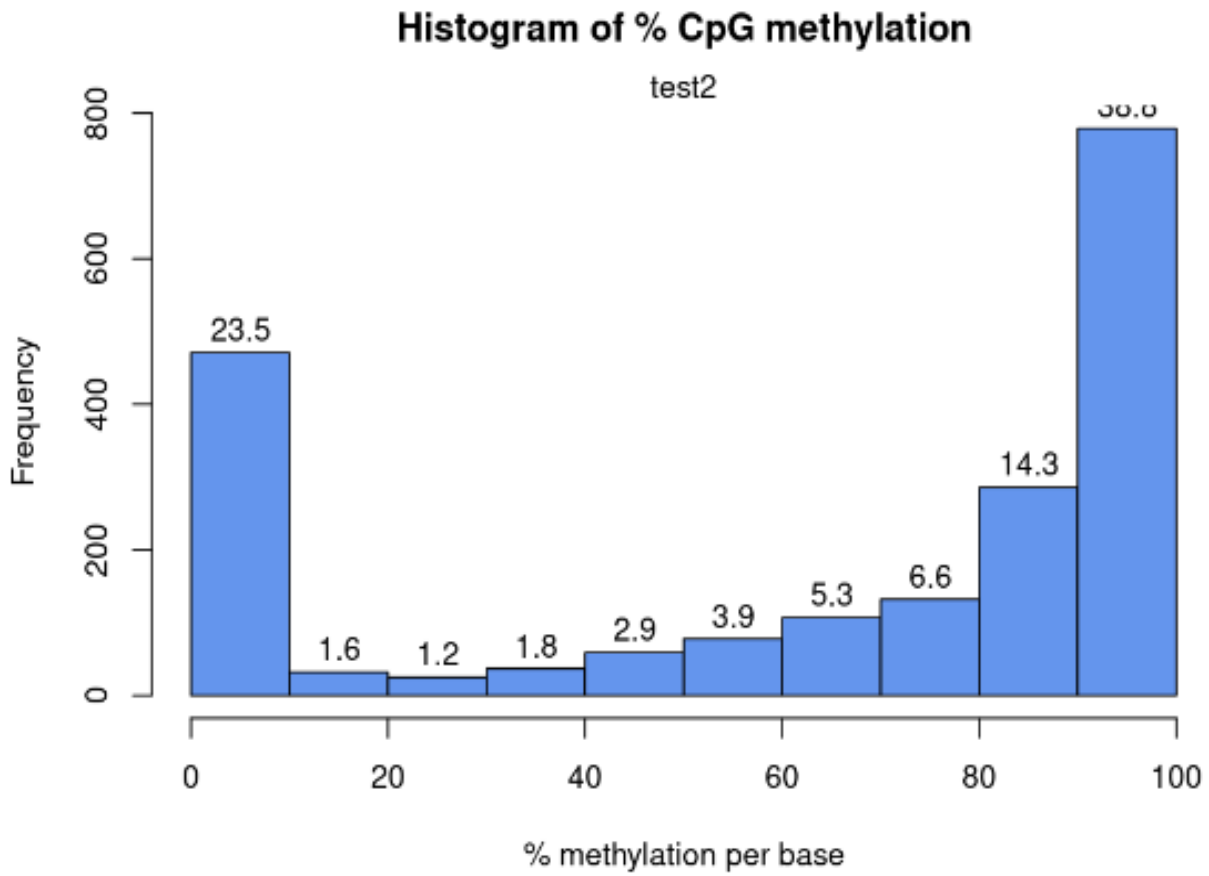
2. Estadísticos descriptivos de las muestras

```
getMethylationStats(my.methRaw2[[1]],plot=FALSE,both.strands=FALSE)
```

```
methylation statistics per base
summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00   76.67   51.66   95.24  100.00
percentiles:
       0%       10%       20%       30%       40%       50%       60%       70%
  0.00000   0.00000   0.00000   0.00000   6.25000  76.66667  87.50000  92.85714
      80%       90%       95%       99%     99.5%     99.9%      100%
100.00000 100.00000 100.00000 100.00000 100.00000 100.00000 100.00000
```

2. Estadísticos descriptivos de las muestras

```
getMethylationStats(my.methRaw2[[1]],plot=TRUE,both.strands=FALSE)
```



En cualquier célula dada, se esperaría que cualquier base esté metilada o no.

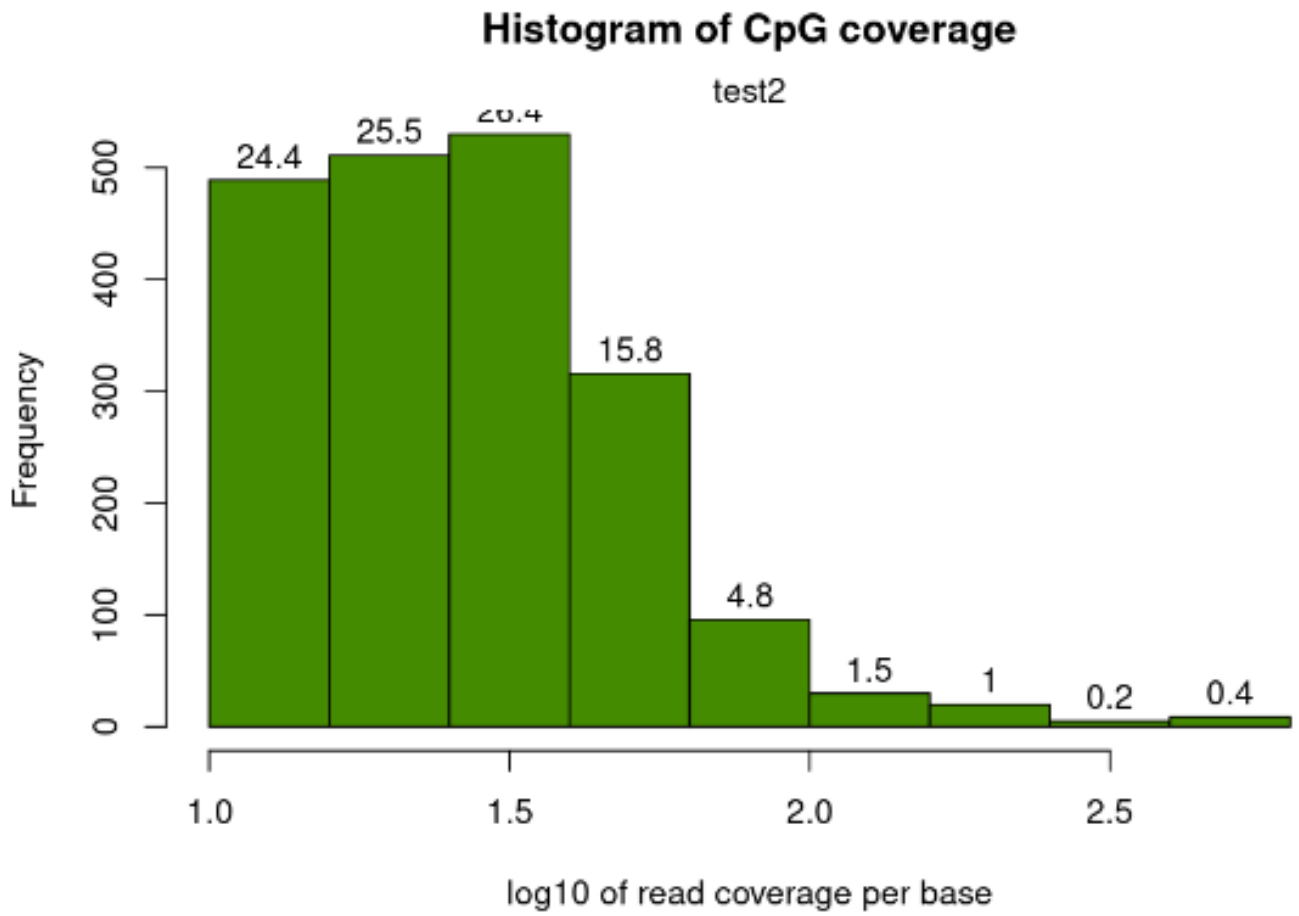
2. Estadísticos descriptivos de las muestras – Estadísticas de cobertura

```
getCoverageStats(my.methRaw2[[1]],plot=FALSE,both.strands=FALSE)
```

```
read coverage statistics per base
summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   12.00   16.00   19.39   22.00  298.00
percentiles:
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%   95%   99% 99.5% 99.9%
   10    11    12    13    14    16    18    20    24    30    37    70   101   196
 100%
  298
```

2. Estadísticos descriptivos de las muestras – Estadísticas de cobertura

```
getCoverageStats(my.methRaw2[[1]],plot=TRUE,both.strands=FALSE)
```



Experimentos con un alto sesgo de duplicación por PCR tendrán un pico secundario hacia el lado derecho del histograma.

Análisis comparativo

3. Filtrado de secuencias de acuerdo a la cobertura de lecturas

```
filtered.obj=filterByCoverage(my.methRaw2_p, lo.count=10, lo.perc=NULL,
                              hi.count=NULL, hi.perc=99.9)
```

Puede resultar útil filtrar muestras según la cobertura…

- Si las muestras sufren un sesgo de PCR, sería útil descartar bases con una cobertura de lectura muy alta.

- Es necesario descartar las bases que tienen una cobertura de lectura baja para aumentar la potencia de las pruebas estadísticas.

Análisis comparativo

4. Unión de las muestras

```
meth = unite(my.methRaw2, destrand=TRUE)
```

La función ‘unite’ unirá todas las muestras en un objeto para la ubicación de las pares de bases que están cubiertas en todas las muestras

Análisis comparativo

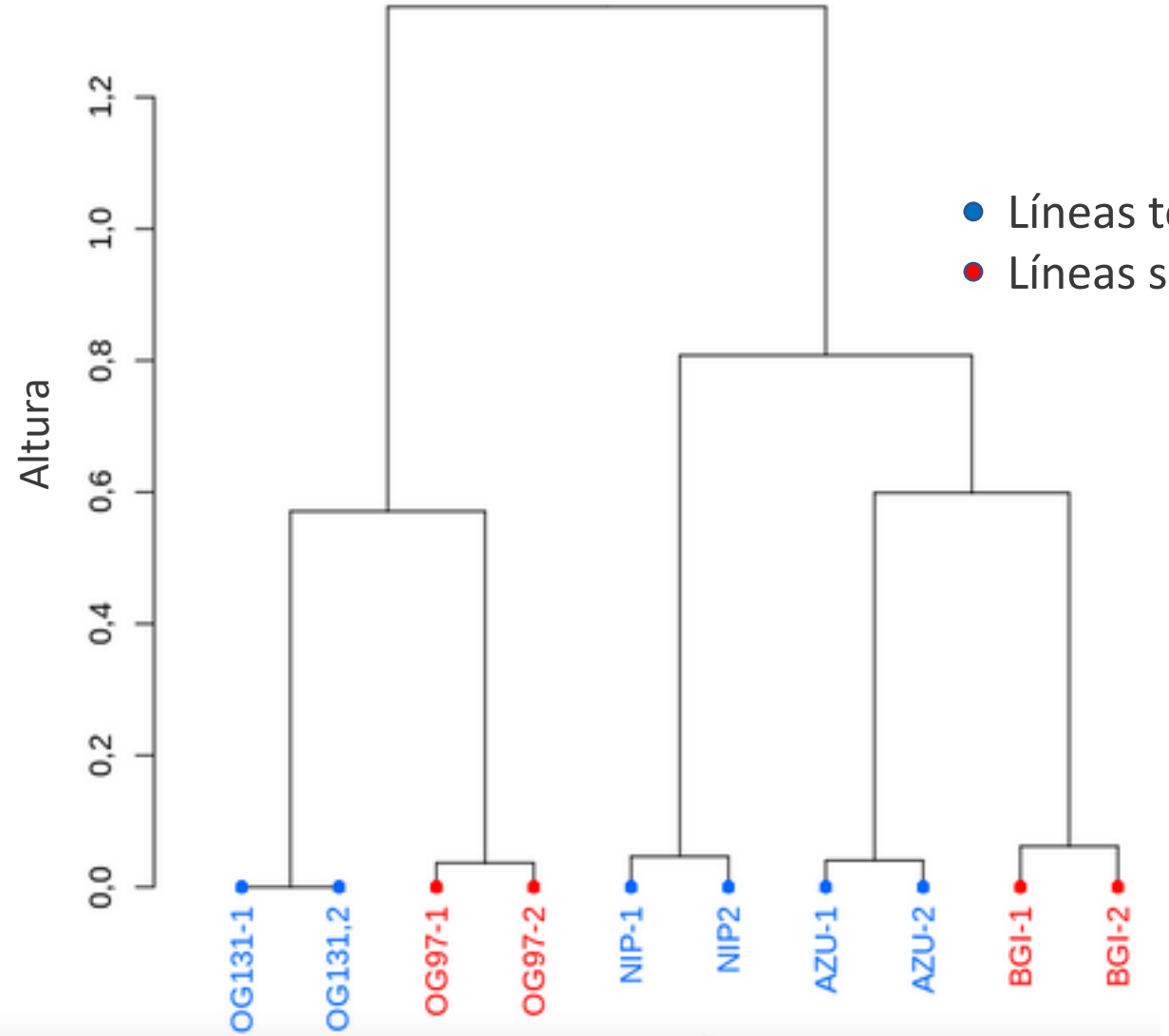
4. Análisis de agrupamiento

```
clusterSamples(meth, dist="correlation", method="ward", plot=TRUE)
```

5. Análisis de componentes principales (PCA)

```
PCASamples(meth, adj.lim=c(0.5,0.5))
```

Correlación de las muestras según el contexto CpG



- Líneas tolerantes
- Líneas susceptibles

Mismo patrón con metilación CHG y CHH

Método de distancia: 1 - Pearson
Método de agrupamiento: Ward

Análisis de datos para datos de secuenciación por bisulfito.

Secuenciación, control de calidad y remoción de adaptadores

Alinear al genoma de referencia Bowtie2

Eliminar sesgo de PCR (Deduplication)

Llamado de metilación

Filtrado por cobertura

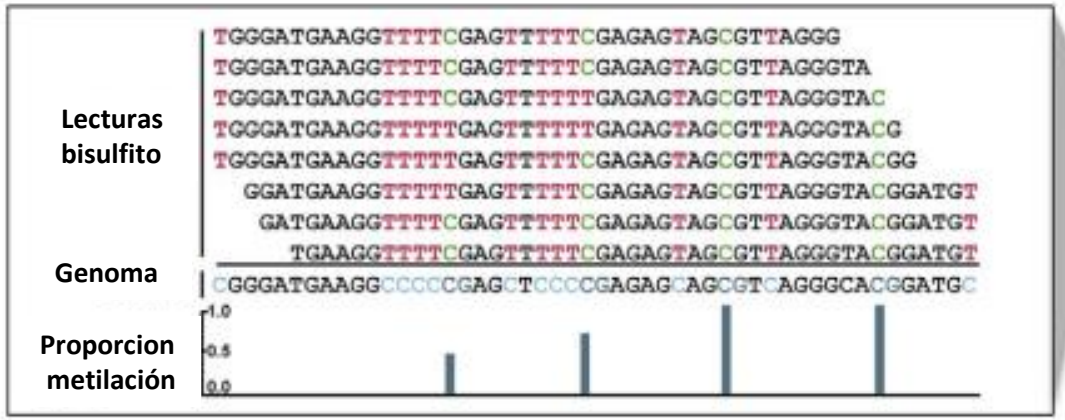
Visualización y agrupamiento

Metilación diferencial

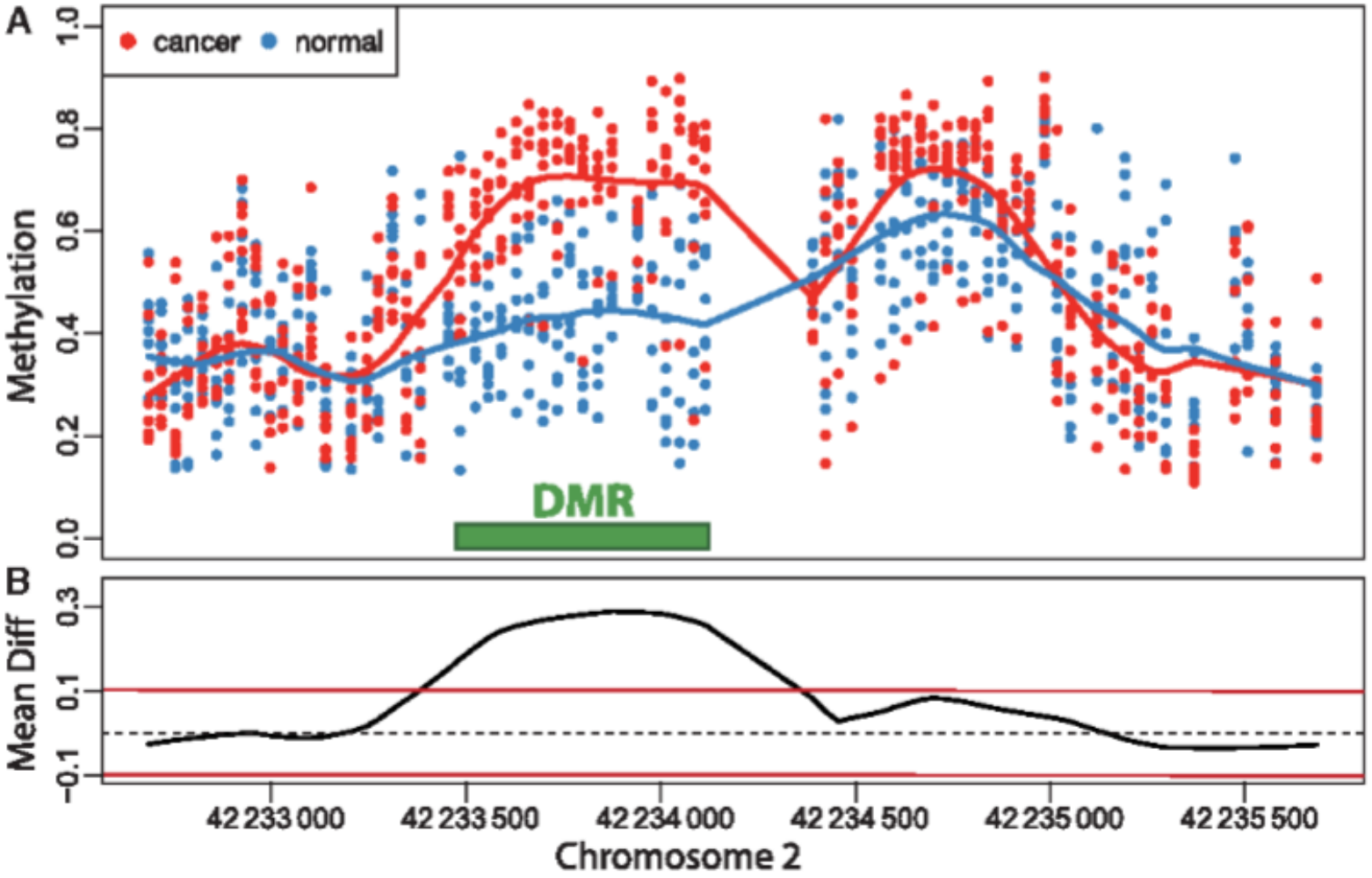
Resultados como tablas, estadísticos de resumen, etc.

Segmentación de la metilación

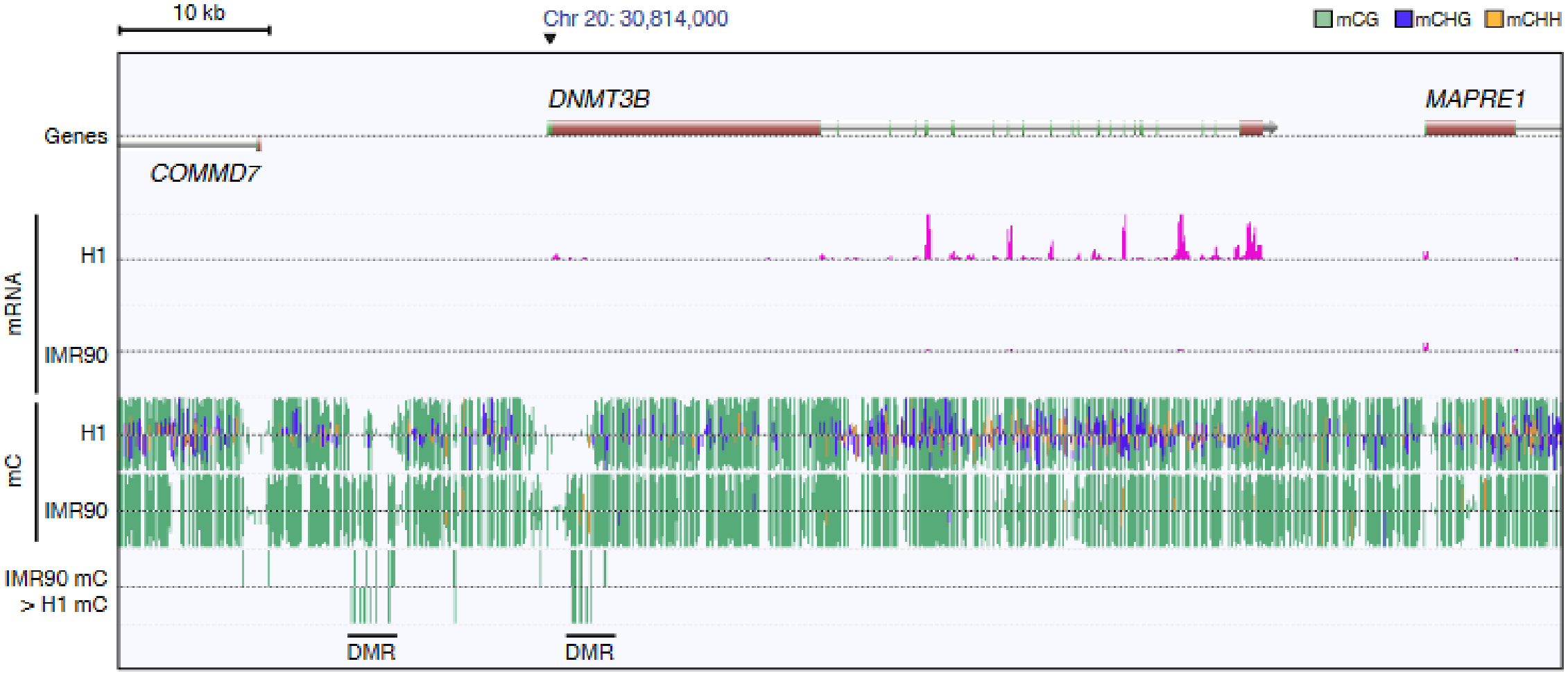
Anotación de DMRs



Regiones diferencialmente metiladas (DMRs)

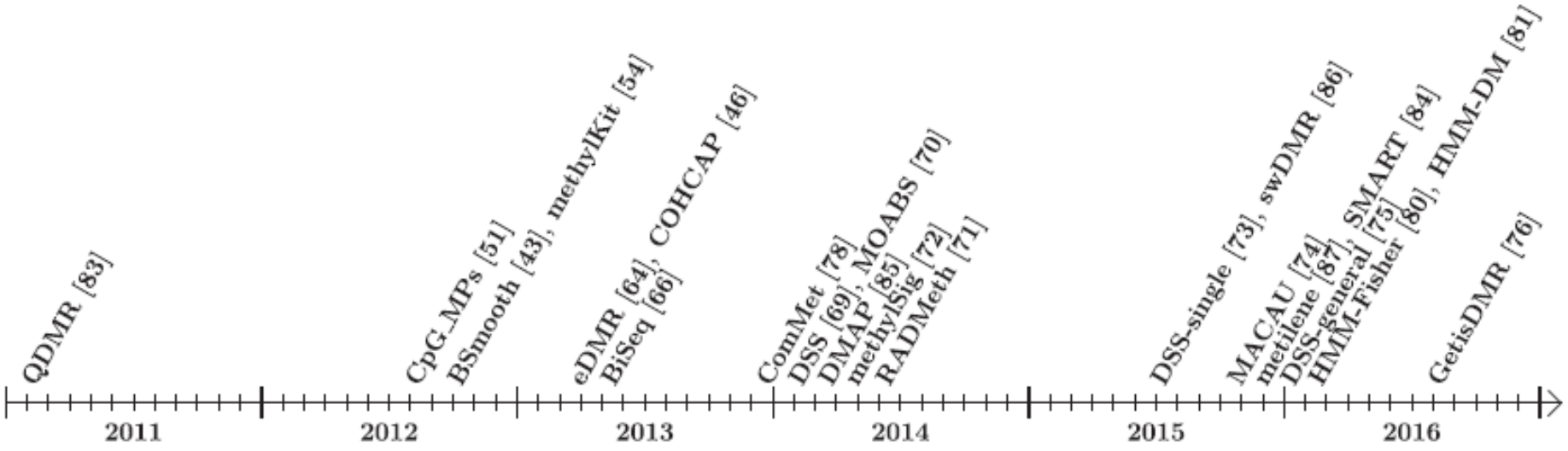


Regiones diferencialmente metiladas (DMRs)



Metilación diferencial

La principal ventaja de MethylKit es que puede tener en cuenta la cobertura de secuenciación, puede incorporar covariables adicionales en el modelo y trabajar con la metilación de CHG o CHH.
También proporciona funcionalidades como compilación de la metilación de muestras, agrupación de muestras, anotación y visualización de DM, etc.



Shafi et al. 2018

Identificación de DMRs utilizando el método de sliding windows

La función *calculateDiffMeth ()* es la función principal para calcular la metilación diferencial.

Dependiendo del tamaño de la muestra por cada conjunto, se utilizará la regresión logística o exacta de Fisher para calcular los valores P.

Los valores P se ajustarán a los valores Q utilizando el método SLIM (Wang et al. 2011) https://pubmed.ncbi.nlm.nih.gov/21098430/.

Si tiene réplicas, la función utilizará automáticamente la regresión logística.

Metilación diferencial

6. Creación de ventanas de tamaño fijo

```
tiles = tileMethylCounts(filtered.obj, win.size=200, step.size=100, cov.bases = 5)
```

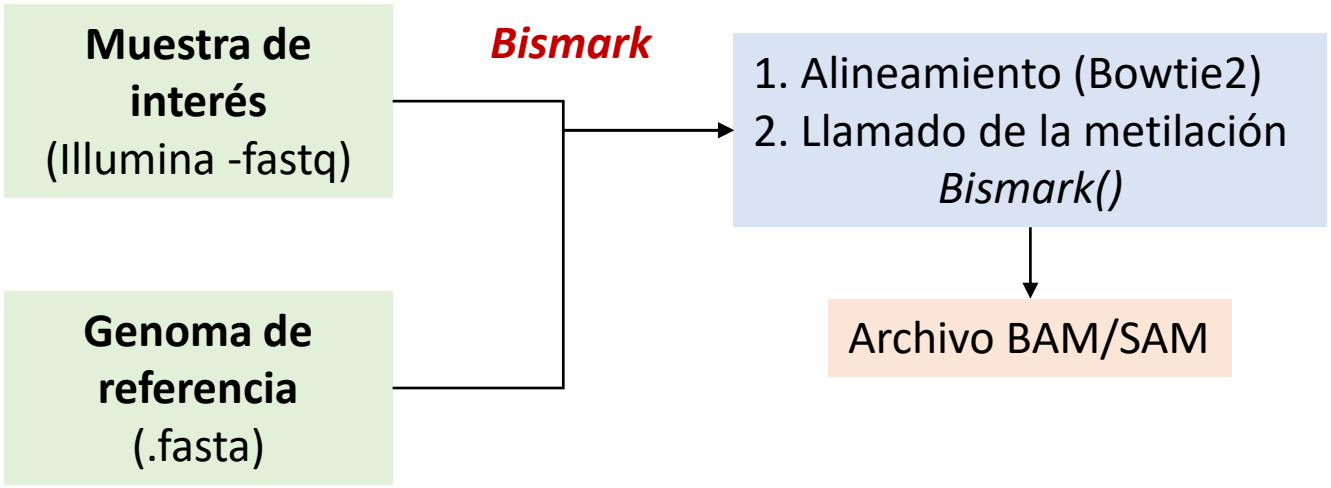
7. Prueba estadística para evaluar regiones diferencialmente metiladas

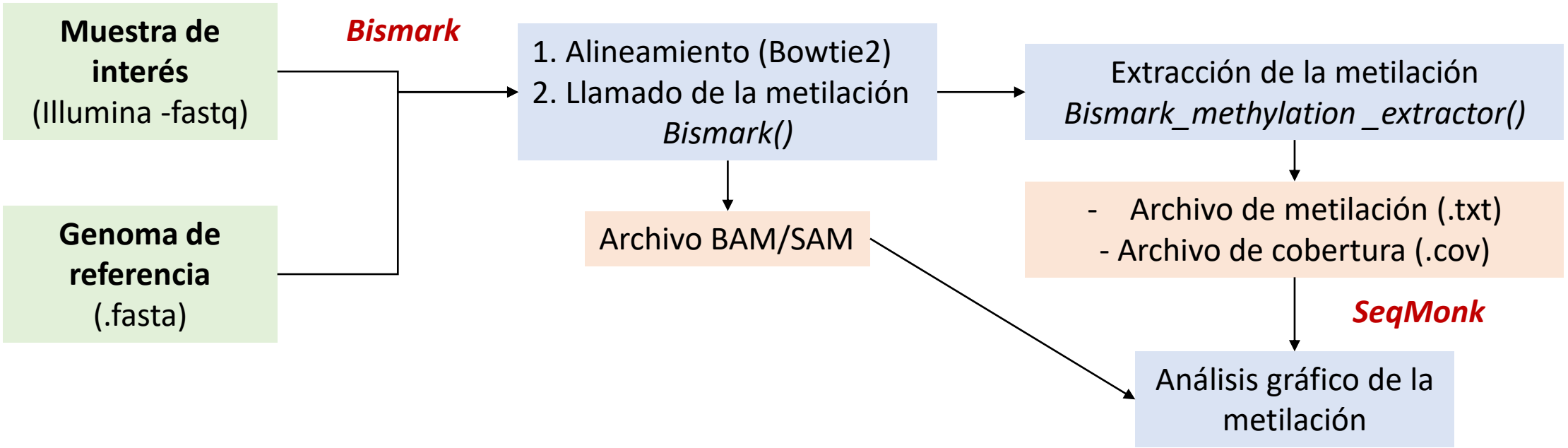
```
myDiff=calculateDiffMeth(meth)

myDiff_25.hyper = getMethylDiff(myDiff,difference=25, qvalue=0.01, type="hyper")
myDiff_25.hypo = getMethylDiff(myDiff,difference=25, qvalue=0.01, type="hypo")
myDiff_25.all = getMethylDiff(myDiff,difference=25, qvalue=0.01)
```

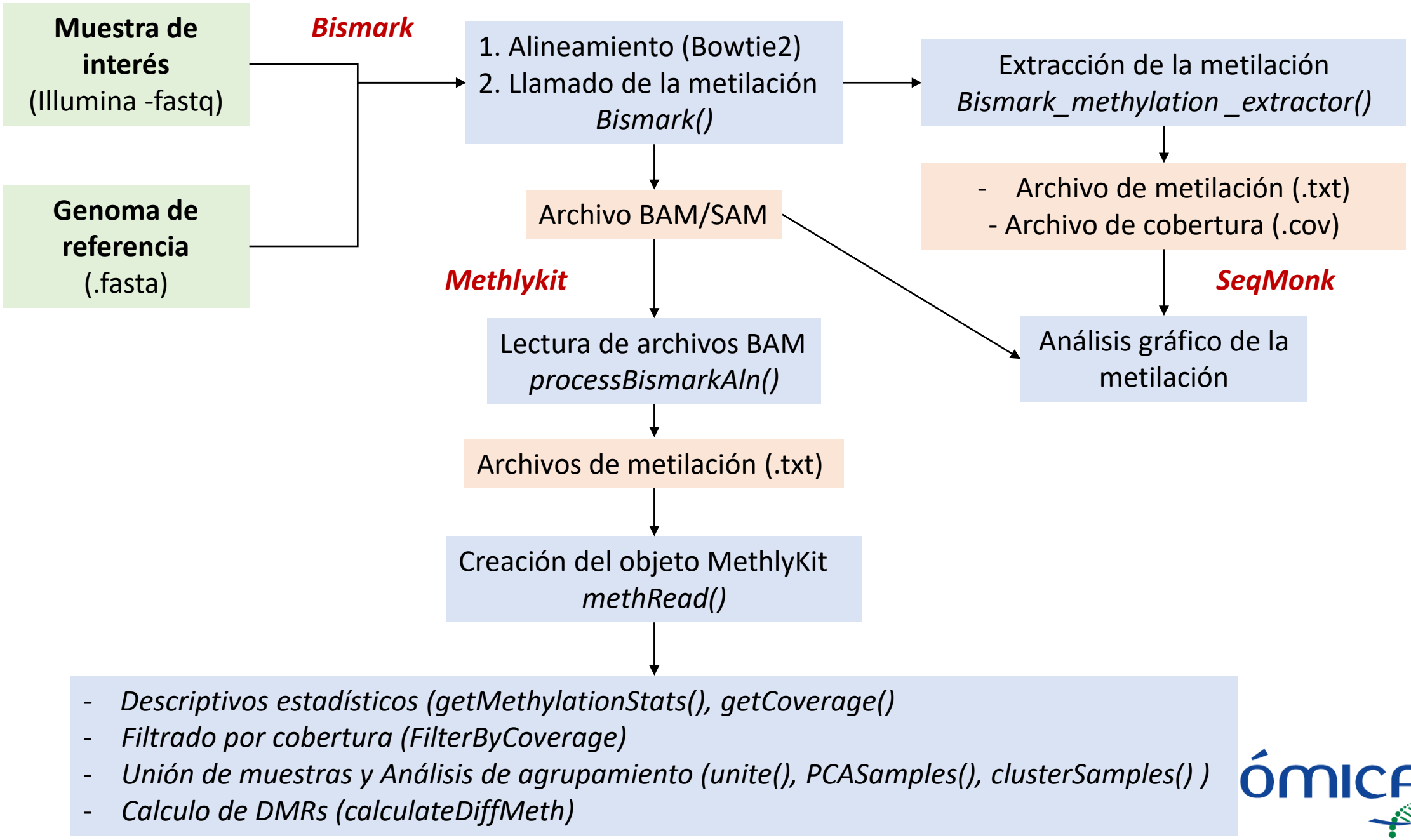
Archivo de DMRs generado en Methylkit

chr	start	end	strand	pvalue	qvalue	meth.diff
Chr1	187001	187200	*	1.937647531153 02e-87	1.99415148344453e-85	82.258064516129
Chr1	462901	463100	*	6.571564477090 56e-113	1.05162024158083e-110	77.6978417266187
Chr1	1568901	1569100	*	2.531527139153 54e-90	2.73362509307498e-88	76.926630499159
Chr1	1569001	1569200	*	2.531527139153 54e-90	2.73362509307498e-88	76.926630499159
Chr1	1796401	1796600	*	7.979595069481 84e-80	7.01657687799872e-78	78.5371800081268
Chr1	2727401	2727600	*	3.296118939331 31e-157	1.14938655527887e-154	76.0563891193848
Chr1	3252101	3252300	*	2.882364216021 44e-158	1.01788331029976e-155	89.9892536396899
Chr1	3252201	3252400	*	5.897903626510 76e-232	6.02377375603014e-229	88.6363636363636
Chr1	3252301	3252500	*	3.586794659109 54e-170	1.54725014245526e-167	86.4347826086956
Chr1	3252401	3252600	*	1.516396826668 67e-200	1.00620012407043e-197	93.2238193018481
Chr1	3252501	3252700	*	2.266348727118 75e-258	3.44673041781404e-255	90.1129943502825
Chr1	3308401	3308600	*	2.576304535337 04e-129	5.48016841641078e-127	80.058524173028
Chr1	3498601	3498800	*	2.684197848382 44e-81	2.43369218350417e-79	77.382618159204
Chr1	3498701	3498900	*	2.342187742790 22e-107	3.41619463935642e-105	83.1172839506173
Chr1	3499301	3499500	*	1.055133053657 92e-131	2.33624697716072e-129	79.1142857142857
Chr1	3499401	3499600	*	2.777448995726 57e-235	2.96764965250914e-232	80.7422166160197
Chr1	3499501	3499700	*	2.388130362034 15e-228	2.33056542111386e-225	75.7549570637308
Chr1	3994901	3995100	*	5.915817439915 58e-155	1.97213160789924e-152	75.6989247311828

Un resumen general del flujo de trabajo…


Muestra de interés
(Illumina -fastq)
Genoma de referencia
(.fasta)
Bismark
1. Alineamiento (Bowtie2)
2. Llamado de la metilación
Bismark()
Archivo BAM/SAM


Muestra de interés
(Illumina -fastq)
Genoma de referencia
(.fasta)
Bismark
1. Alineamiento (Bowtie2)
2. Llamado de la metilación
Bismark()
Archivo BAM/SAM
Extracción de la metilación
Bismark_methylation _extractor()
- Archivo de metilación (.txt)
- Archivo de cobertura (.cov)
SeqMonk
Análisis gráfico de la metilación


Muestra de interés (Illumina -fastq)
Genoma de referencia (.fasta)
Bismark
1. Alineamiento (Bowtie2)
2. Llamado de la metilación
Bismark()
Extracción de la metilación
Bismark_methylation _extractor()
- Archivo de metilación (.txt)
- Archivo de cobertura (.cov)
Archivo BAM/SAM
Methlykit
SeqMonk
Lectura de archivos BAM
processBismarkAln()
Análisis gráfico de la metilación
Archivos de metilación (.txt)
Creación del objeto MethlyKit
methRead()
- Descriptivos estadísticos (getMethylationStats(), getCoverage()
- Filtrado por cobertura (FilterByCoverage)
- Unión de muestras y Análisis de agrupamiento (unite(), PCASamples(), clusterSamples())
- Calculo de DMRs (calculateDiffMeth)

GRACIAS…

PREGUNTAS??

ómicas

Aliados

CIAT
International Center for Tropical Agriculture
Since 1967 Science to cultivate change

cenicaña

FEDEARROZ
FEDERACIÓN NACIONAL DE ARROCEROS

NIAB

intelecto
Soluciones y Tecnología

HITECH
AUTOMATIZACION S.A.

Caltech

GHENT UNIVERSITY

東京大学
THE UNIVERSITY OF TOKYO

ILLINOIS

Pontificia Universidad JAVERIANA Cali

IES Ancla

UNIVERSIDAD DEL QUINDÍO

Universidad de Ibagué

UNILLANOS
UNIVERSIDAD DE LOS LLANOS

Universidad de los Andes
Colombia

UNIVERSIDAD ICESI

Pontificia Universidad JAVERIANA Bogotá

[VIGILADA MINEDUCACIÓN Res. 12220 de 2016.]

Apoyan

El futuro es de todos
Gobierno de Colombia

COLOMBIA CIENTÍFICA
Conocimiento Global para el Desarrollo

Referencias…

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., & Mason, C. E. (2012). MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, *13*(10), R87. https://doi.org/10.1186/gb-2012-13-10-R87

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, *27*(11), 1571–1572. https://doi.org/10.1093/bioinformatics/btr167

Shafi, A., Mitrea, C., Nguyen, T. and Draghici, S. 2018. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. Briefings in Bioinformatics. 737–753

Wreczycka, K., Gosdschan, A., Yusuf, D., Grüning, B., Assenov, Y. & Akalin, A. (2017) Strategies for analyzing bisulfite sequencing data. Journal of biotechnology. 261: 105-115.

Zhang, H., Lang, Z. & Zhu J.K. (2018). Dynamics and function of DNA methylation in plants Huiming. Molecular cell biology. Pag: 489-506.