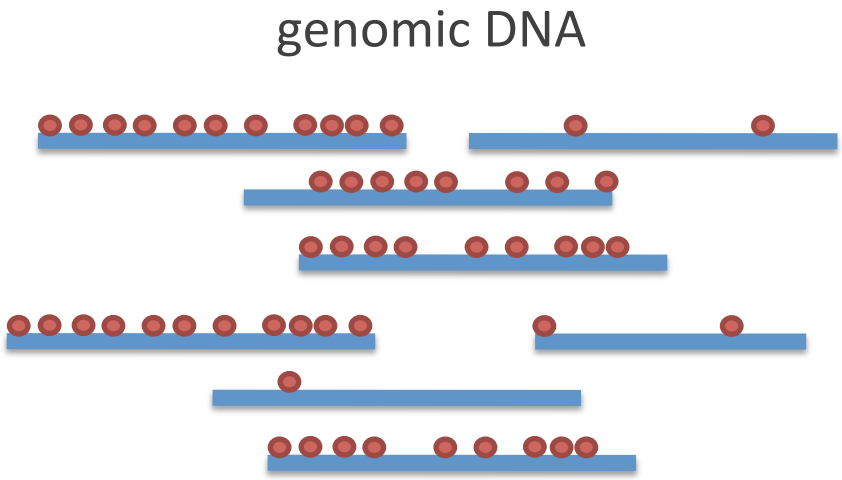
Measuring DNA methylation using high-throughput sequencing

Mackenzie Gavery
11/10/16

DNA methylation – how to we measure it?

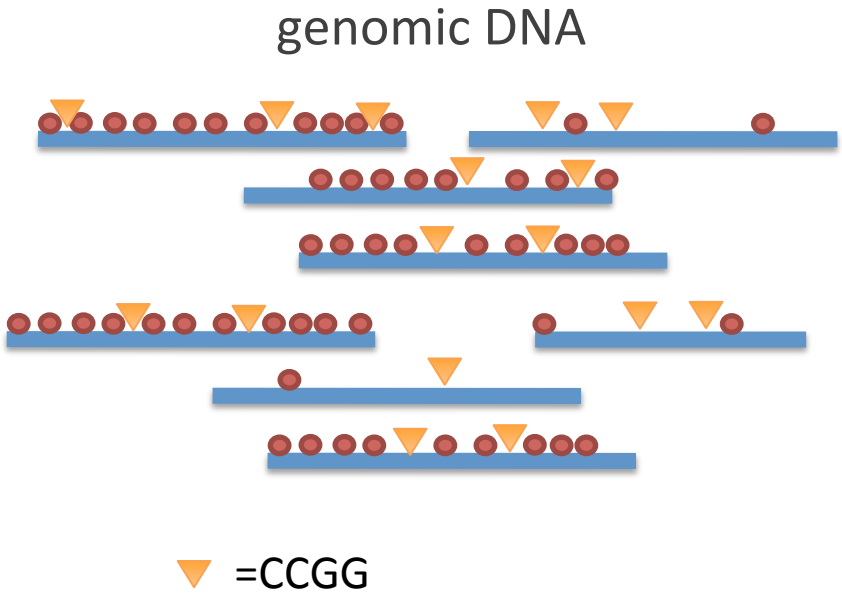
DNA methylation – how to we measure it?

Reduced Representation Bisulfite Sequencing (RRBS)



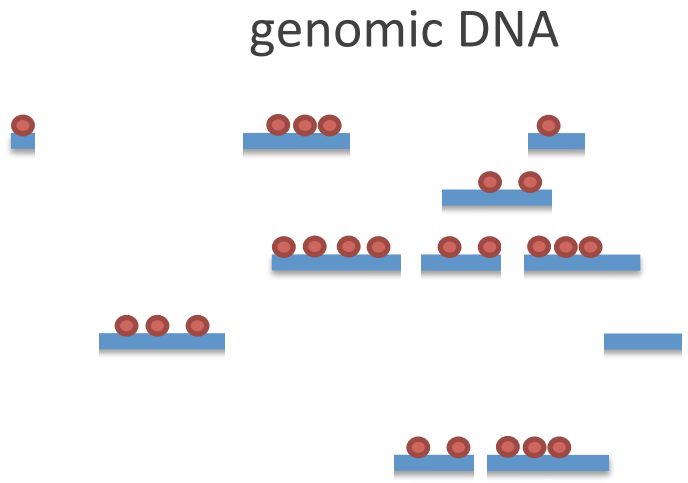
DNA methylation – how to we measure it?

Reduced Representation Bisulfite Sequencing (RRBS)



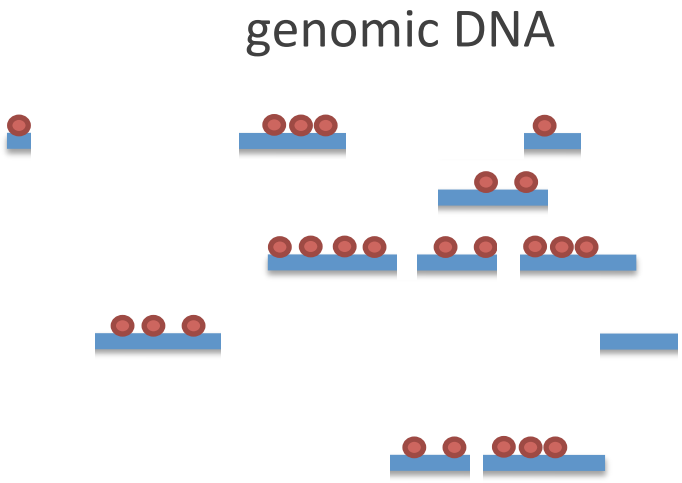
DNA methylation – how to we measure it?

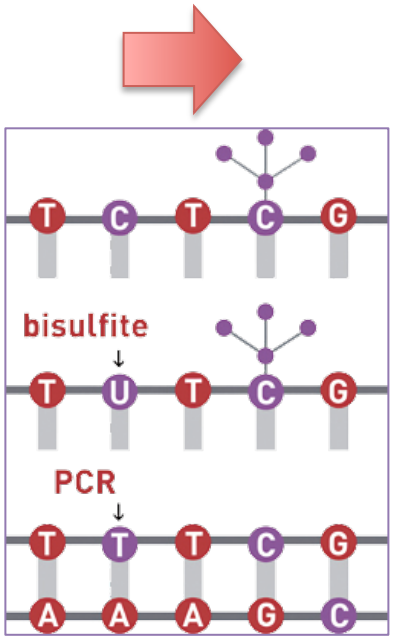
Reduced Representation Bisulfite Sequencing (RRBS)



DNA methylation – how to we measure it?

Reduced Representation Bisulfite Sequencing (RRBS)

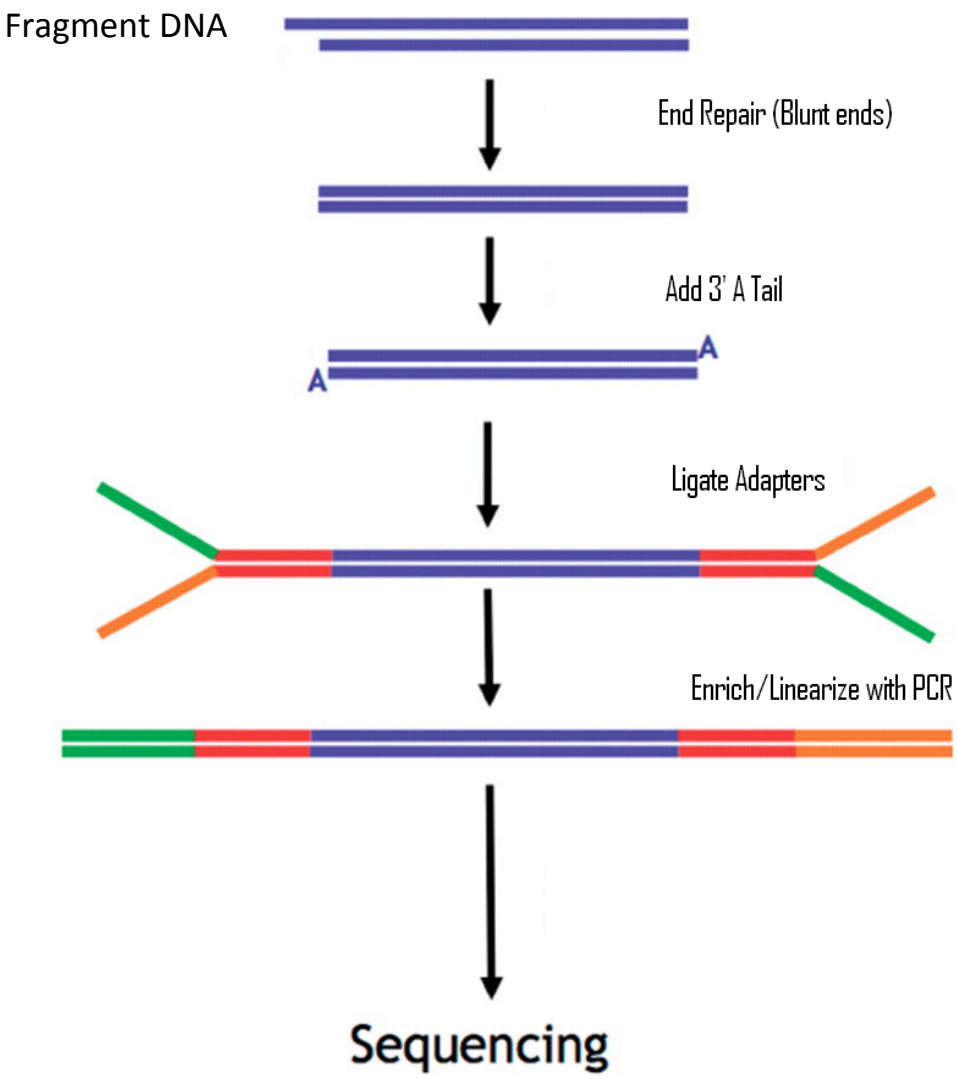




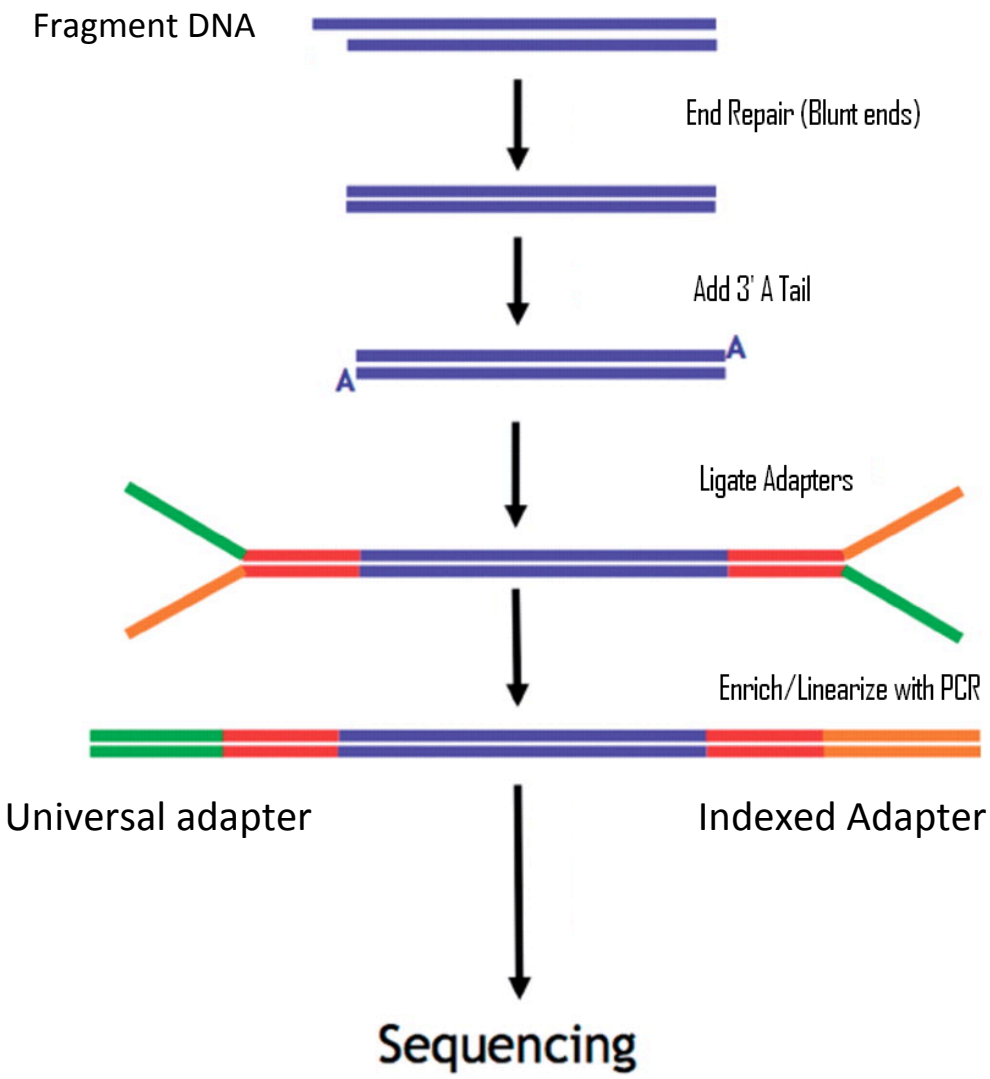
DNA methylation analysis

- Library prep
- Sequencing
- Bioinformatics (DNA methylation)
  - QC/Trimming
  - Mapping
  - Extracting Methylation Data
  - Interpretation (e.g. differential methylation)

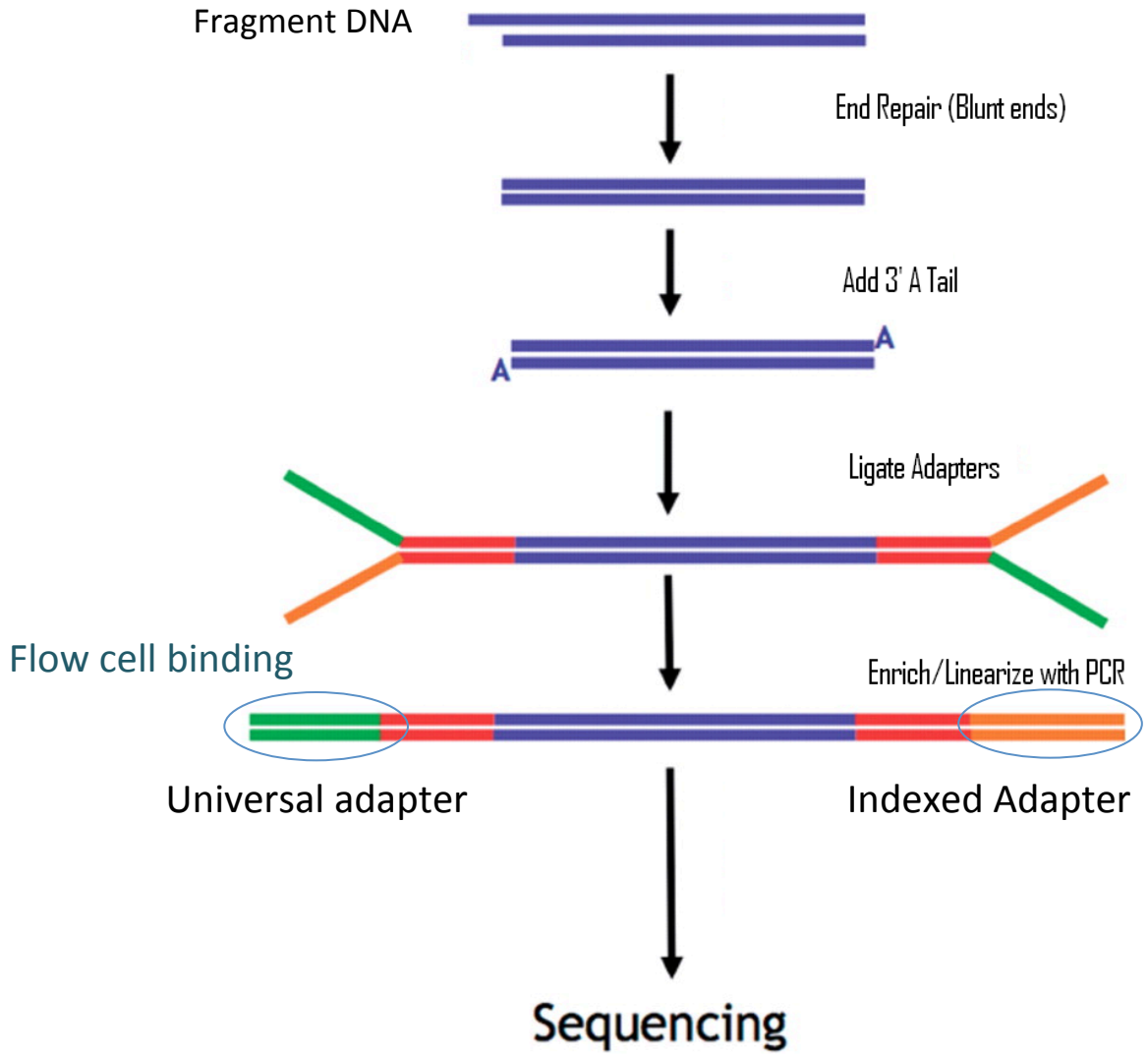
Library Prep



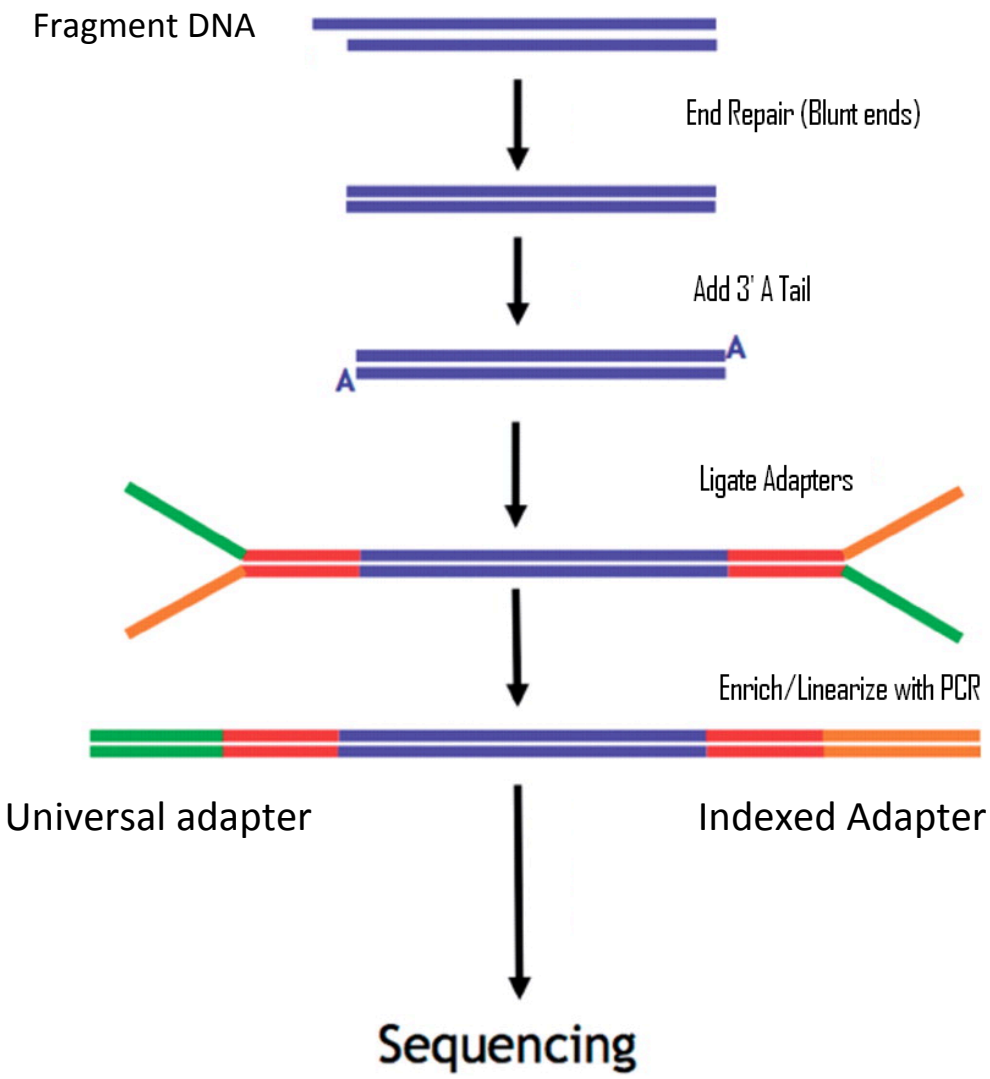
Library Prep



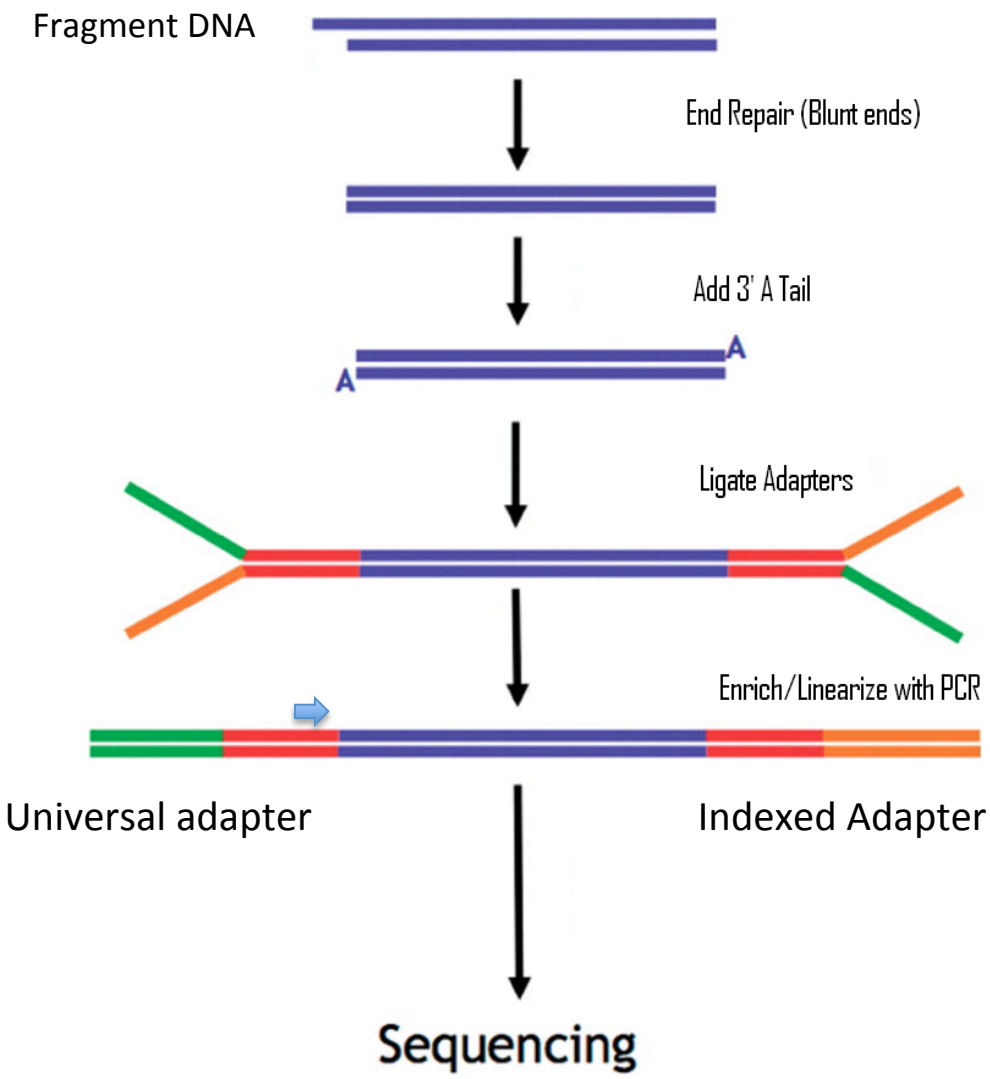
Library Prep



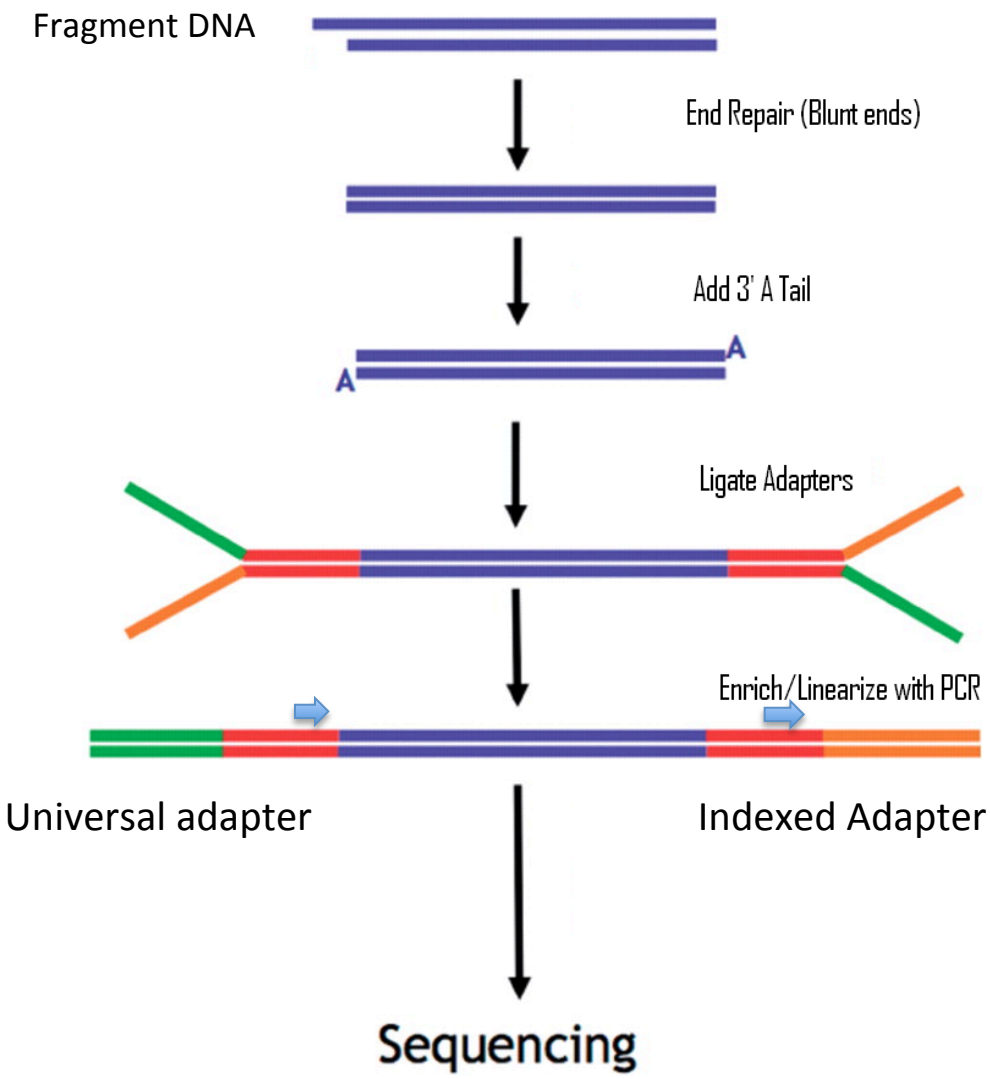
Library Prep



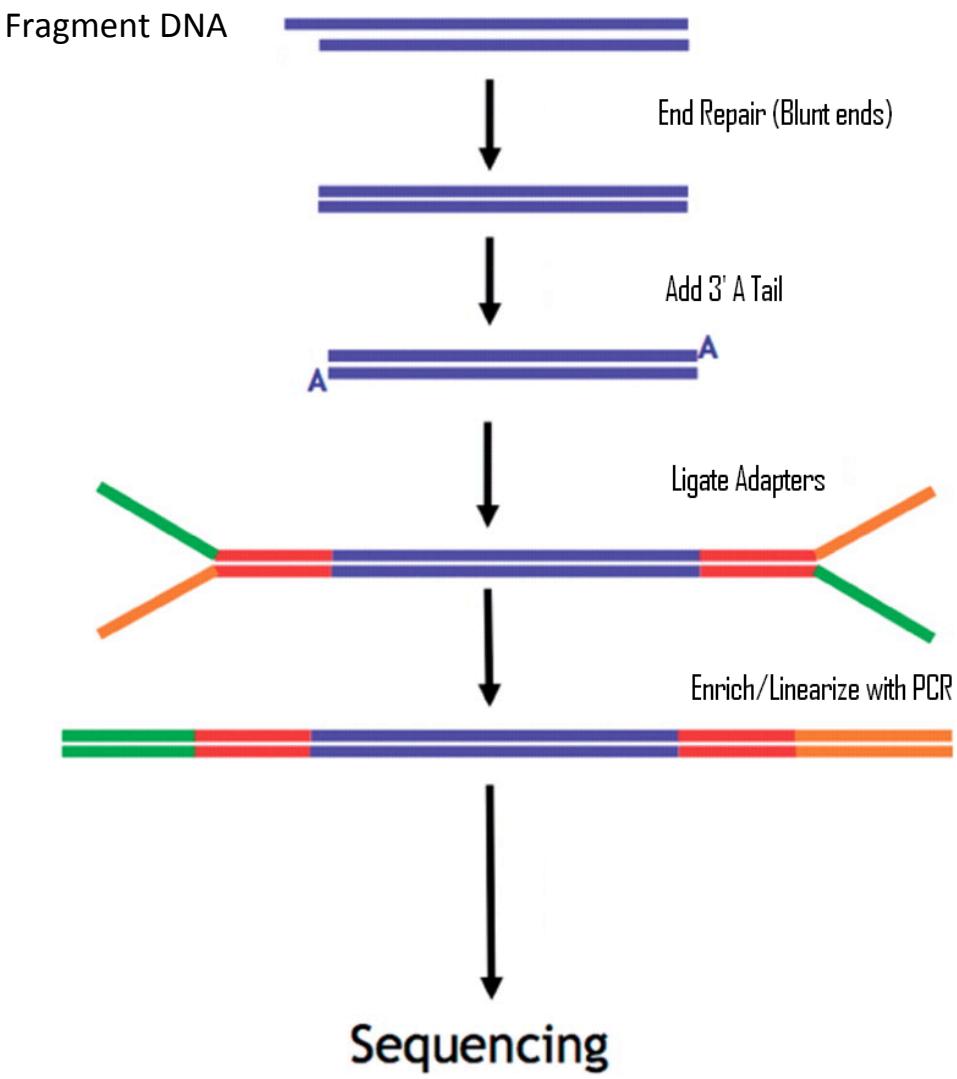
Library Prep



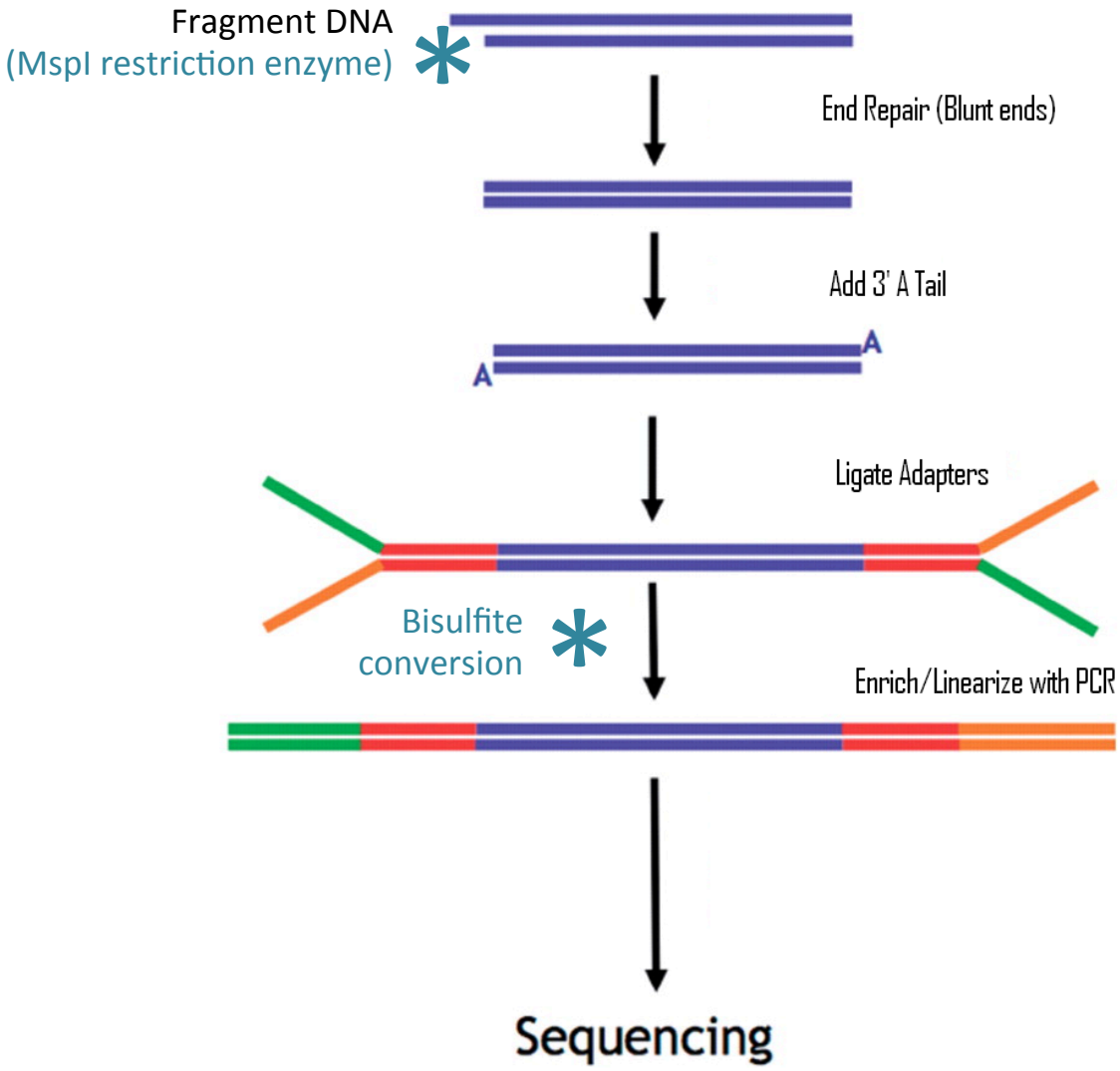
Library Prep



Library Prep



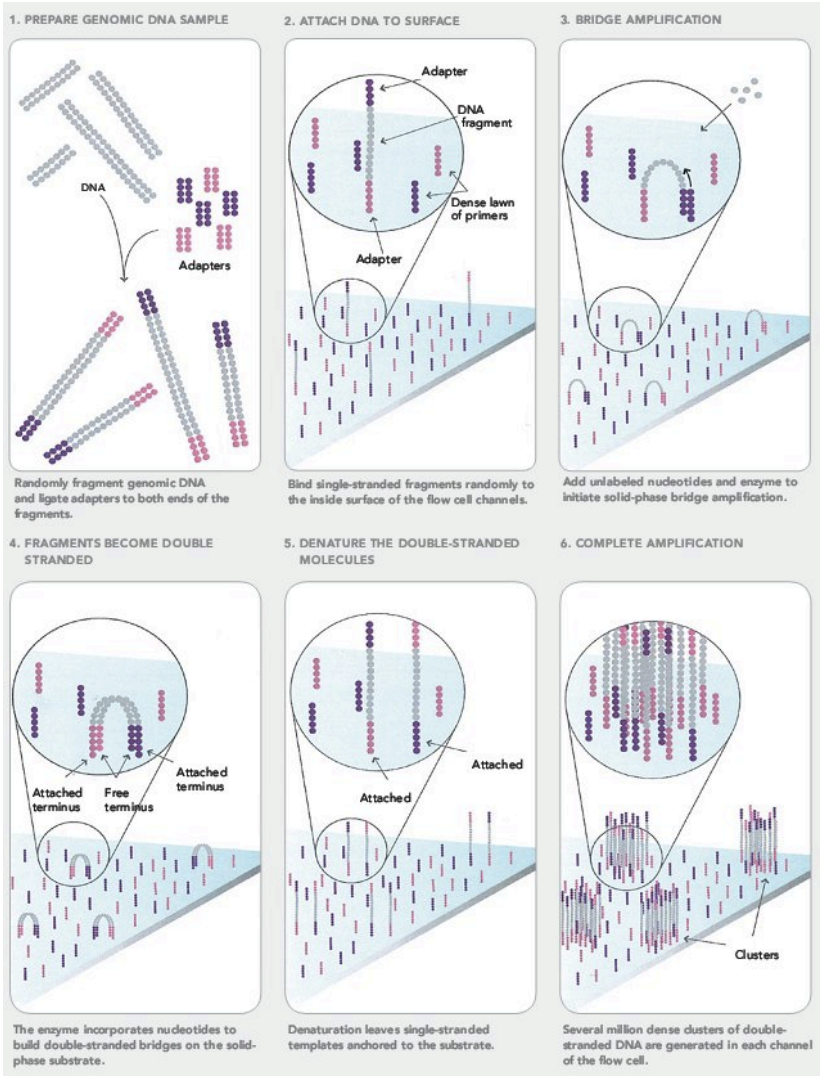
Library Prep



Sequencing

- Indexed libraries are pooled together
- Illumina HiSeq

  (single read 100bp)
- ~200 million reads



Fastq Files

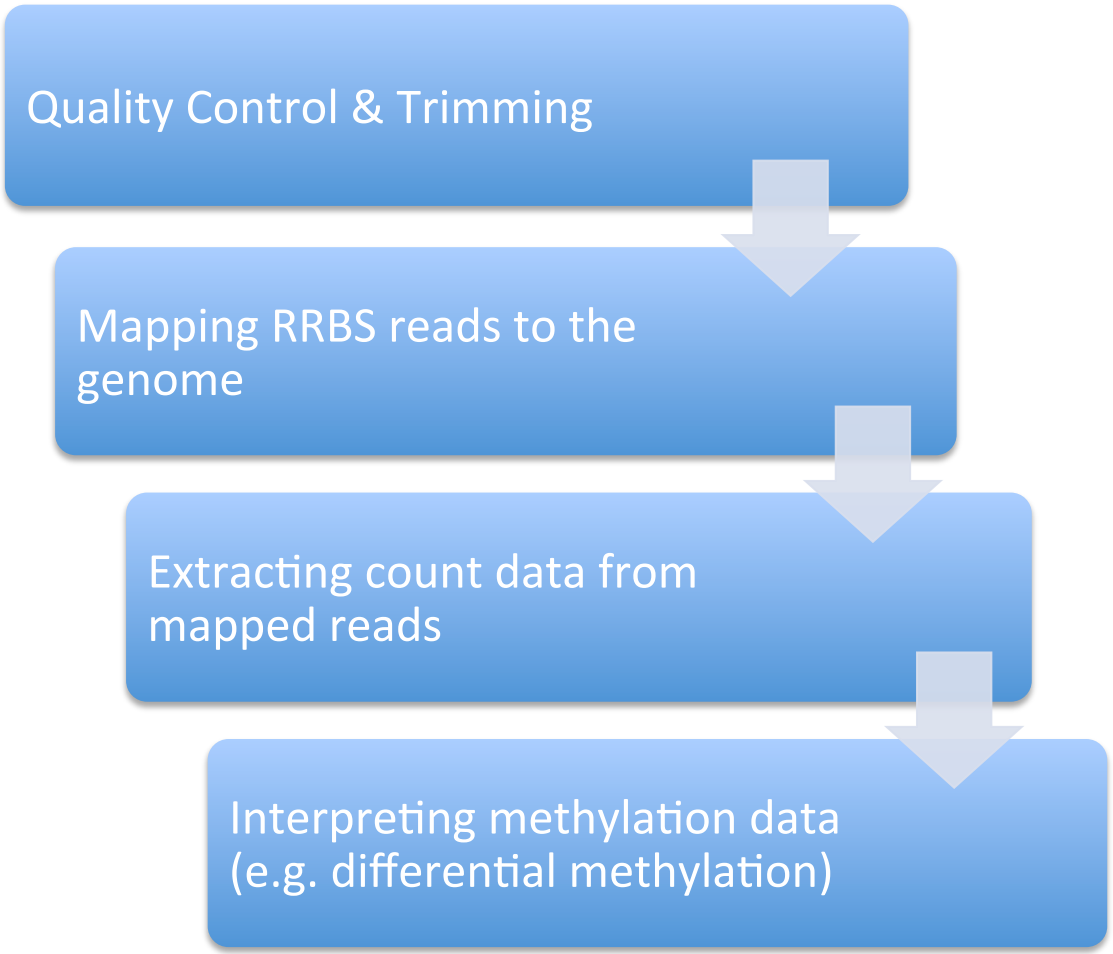
```
@SN747:551:C99B9ACXX:7:1110:1855:1151 1:N:0:CAGATC
CGGGTGATGTAGATGGTGGTGGGGTGGTTGGATCGATTGTGGGGAGTTGGGAAGGTGGTGTAATTTGTAGTTGGCGTACGATTTGAGATCGGAAGAGCACA
+
@@C?A@?4CFFFFG@EEGG@AGHA?CAD=FHDFB;;FHF=FEBHFD?C;>CAC==CAB8<<ACCDDCEDCDD:A@09<9<@B?B>@<8@:(087&+2<@A<
```

Fastq Files

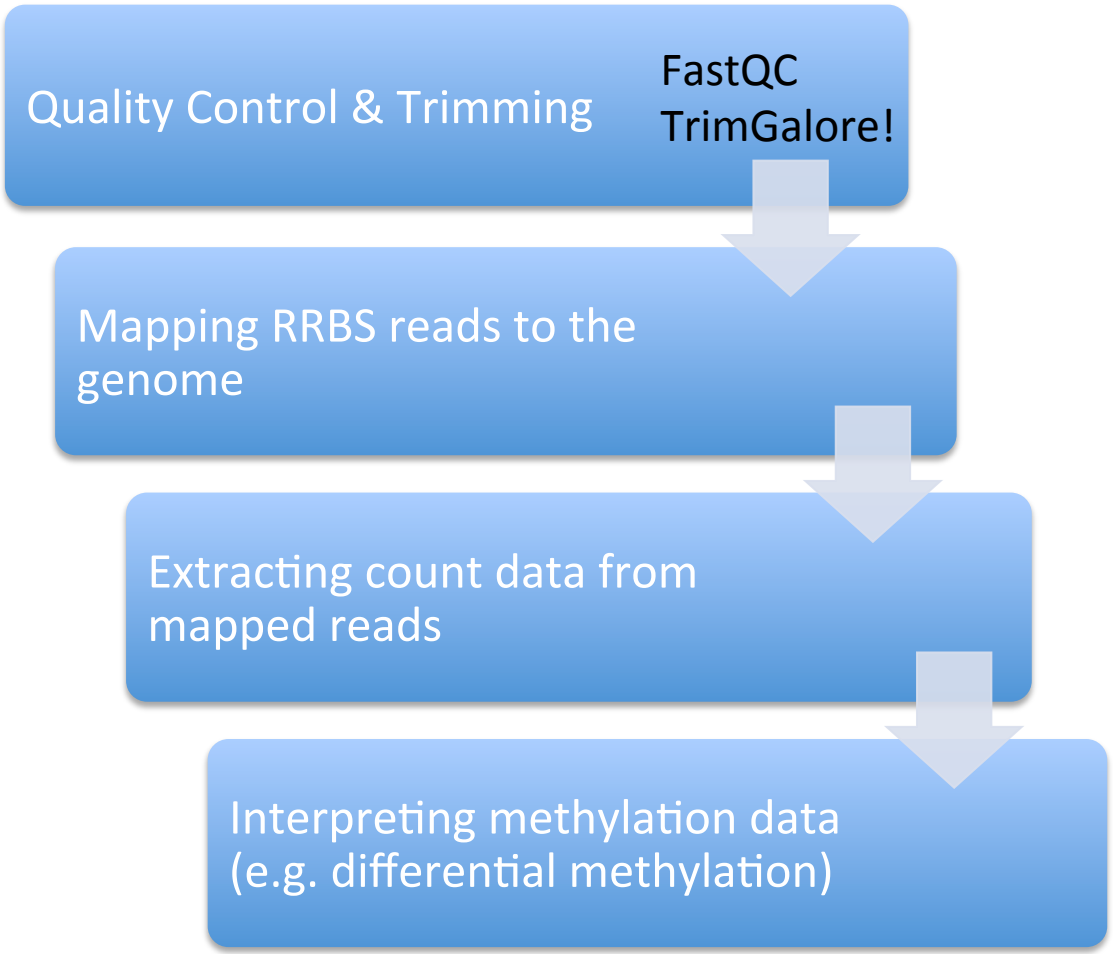
```
@SN747:551:C99B9ACXX:7:1110:1855:1151 1:N:0:CAGATC
CGGGTGATGTAGATGGTGGTGGGGTGGTTGGATCGATTGTGGGGAGTTGGGAAGGTGGTGTAATTTGTAGTTGGCGTACGATTTGAGATCGGAAGAGCACA
+
@@C?A@?4CFFFFG@EEGG@AGHA?CAD=FHDFB;;FHF=FEBHFD?C;>CAC==CAB8<<ACCDDCEDCDD:A@09<9<@B?B>@<8@:(087&+2<@A<
```

How do you analyze 200 million reads!?

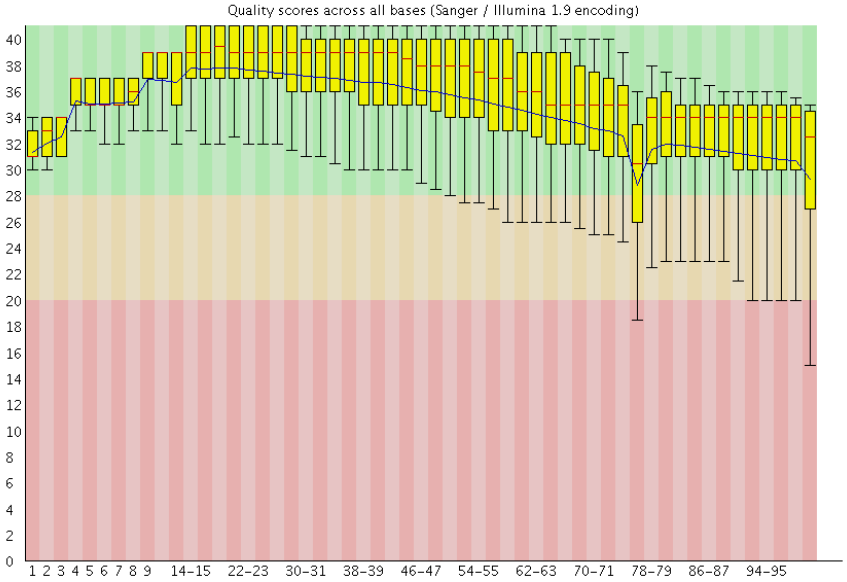
Bioinformatics (DNA methylation)



Bioinformatics (DNA methylation)

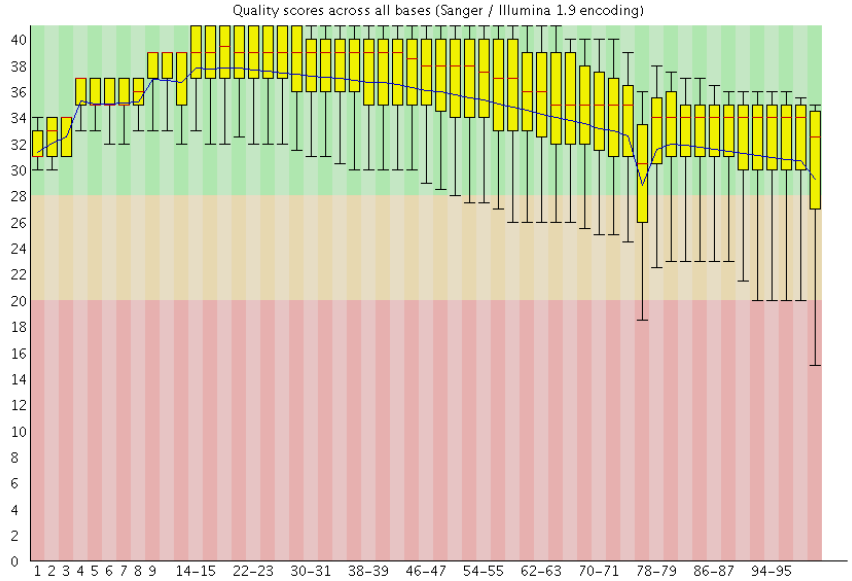


Per base sequence quality



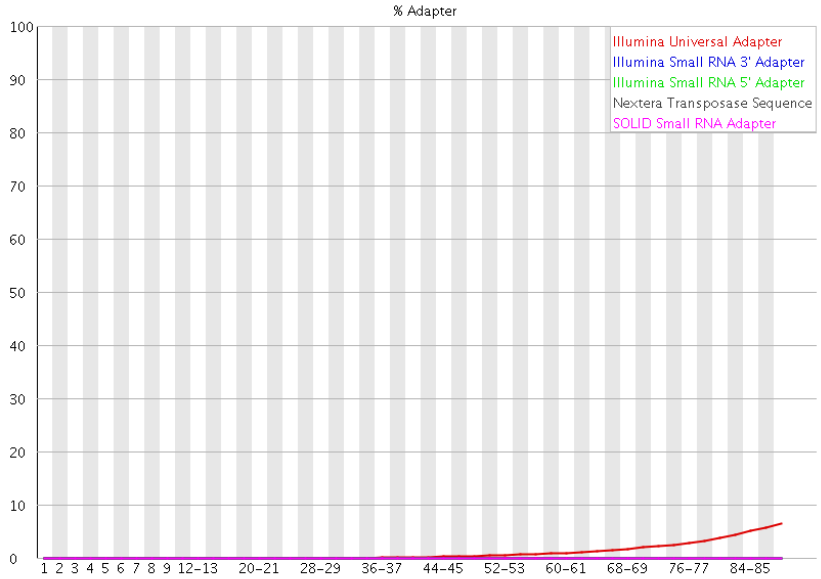
Position in read (bp)

Per base sequence quality



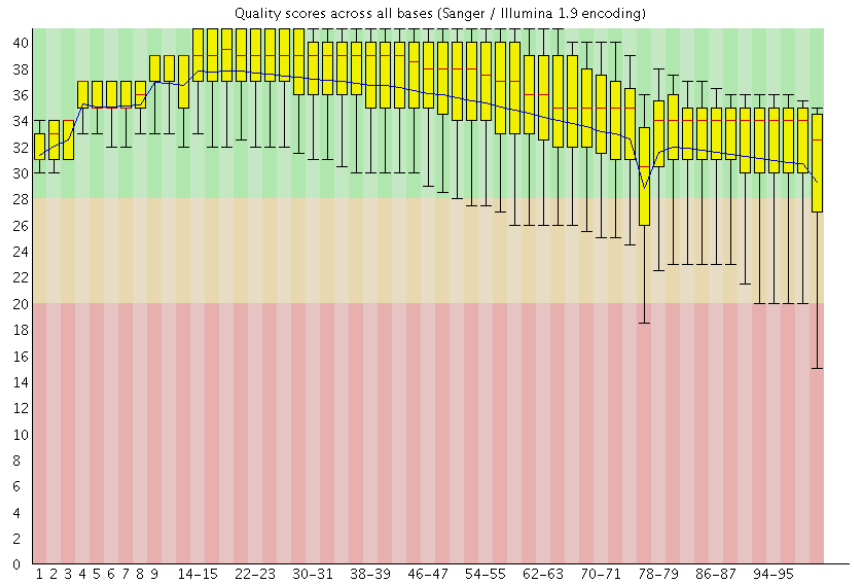
Position in read (bp)

Adapter Content



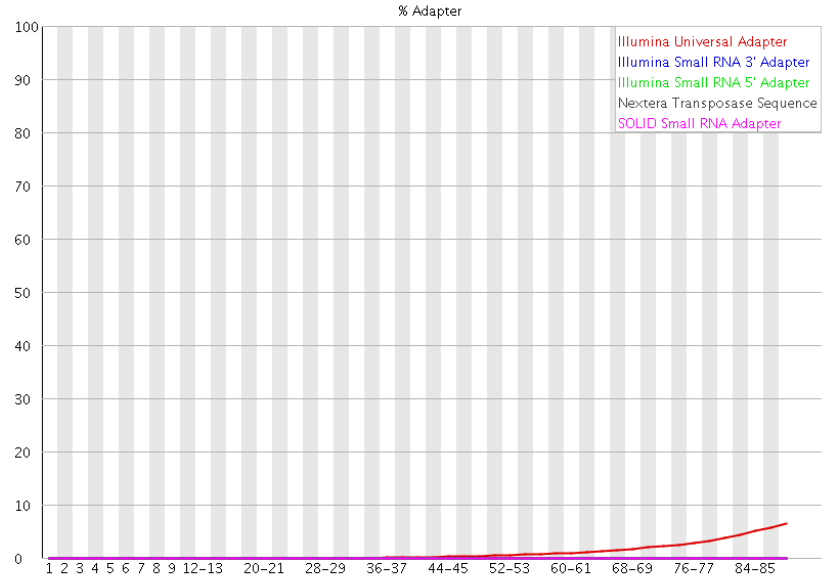
Position in read (bp)

Per base sequence quality



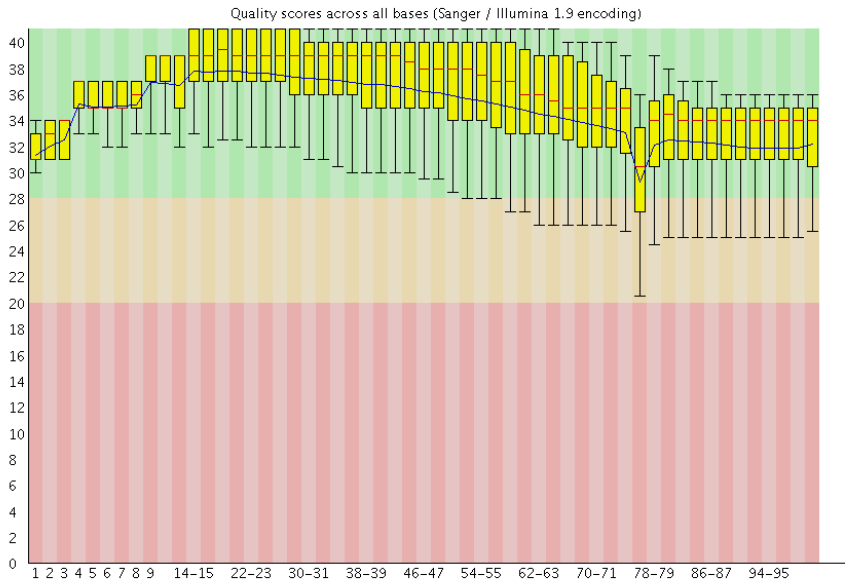
Position in read (bp)

Adapter Content

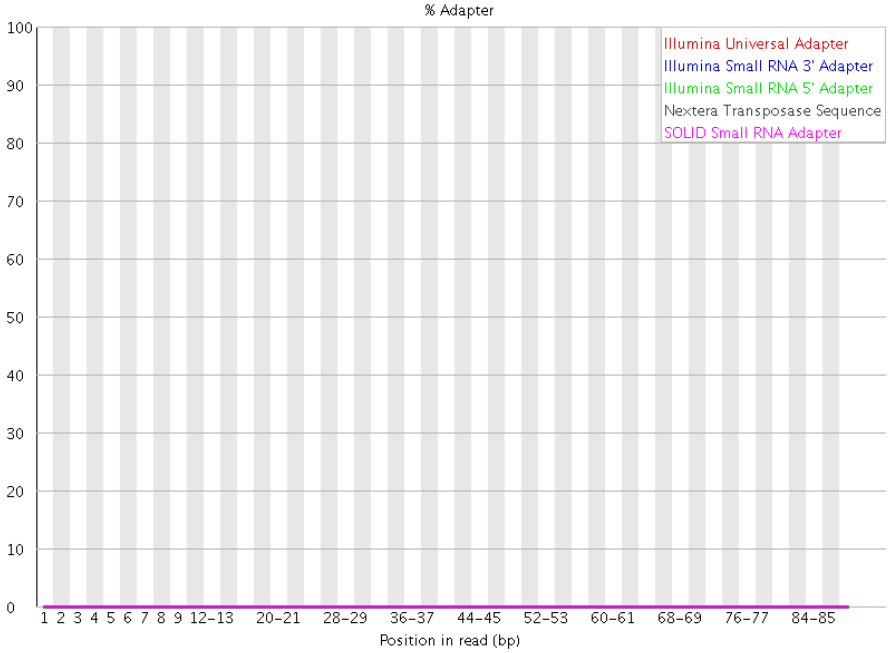


Position in read (bp)

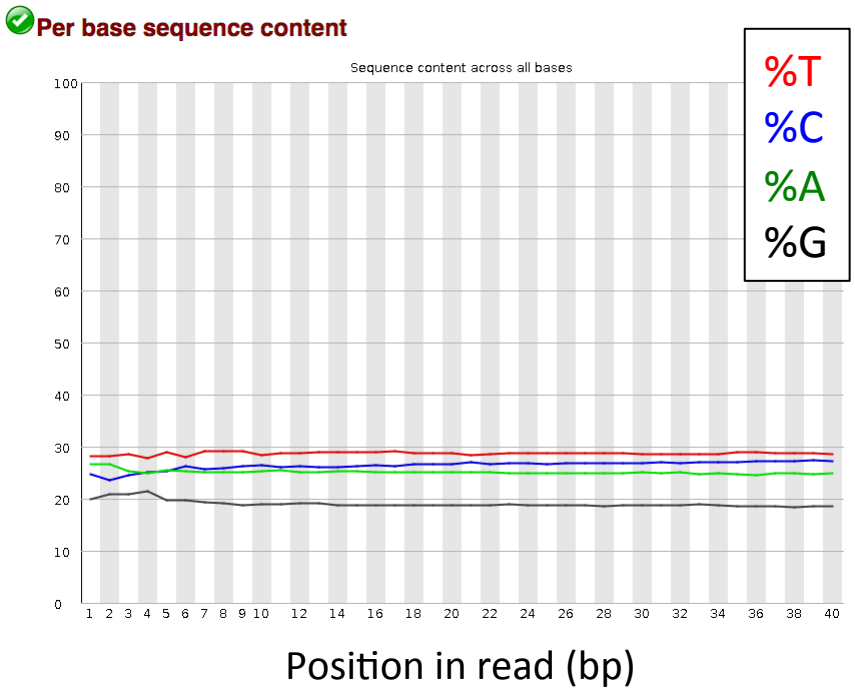
Per base sequence quality



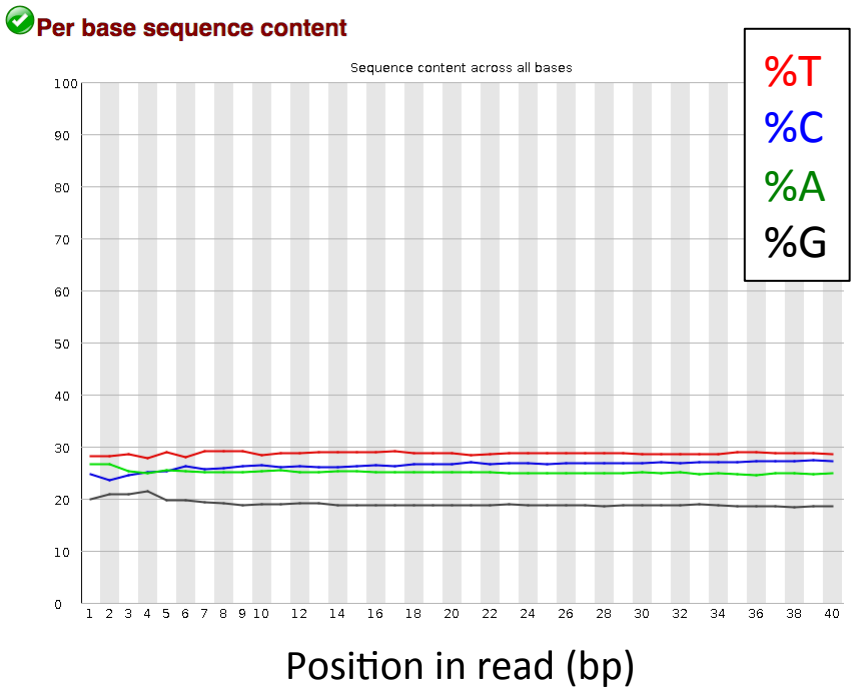
Adapter Content

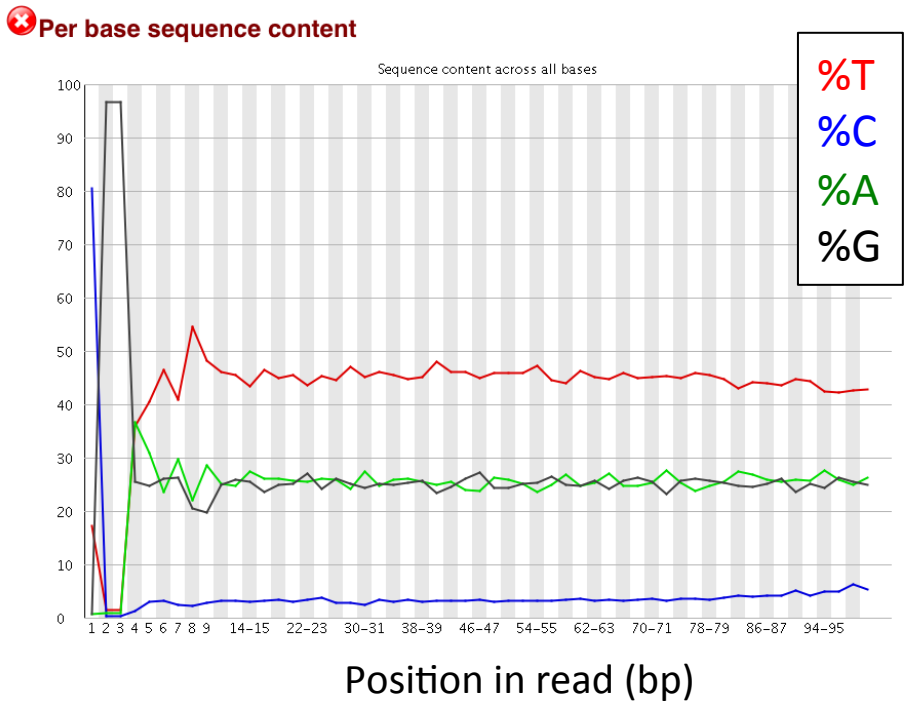


QC

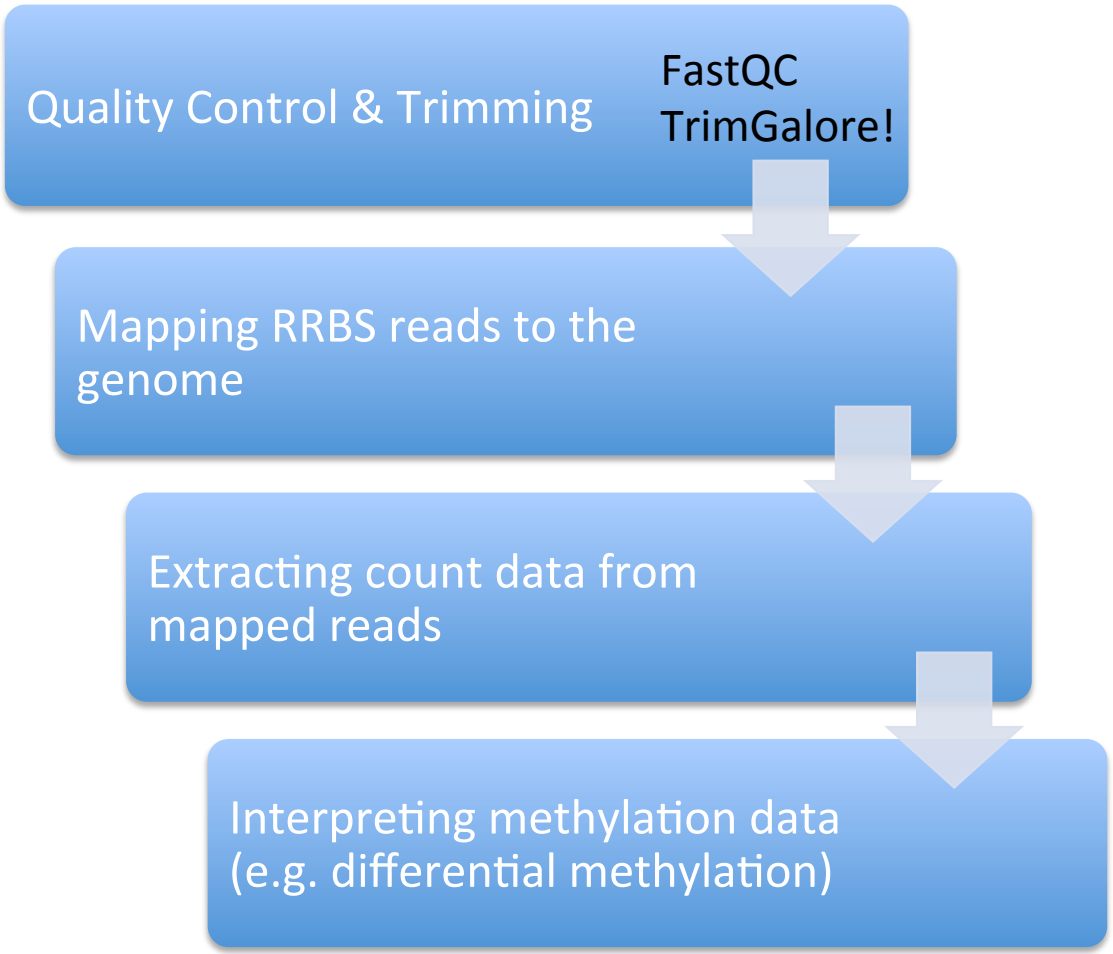

Per base sequence content
Sequence content across all bases
%T
%C
%A
%G
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)

QC

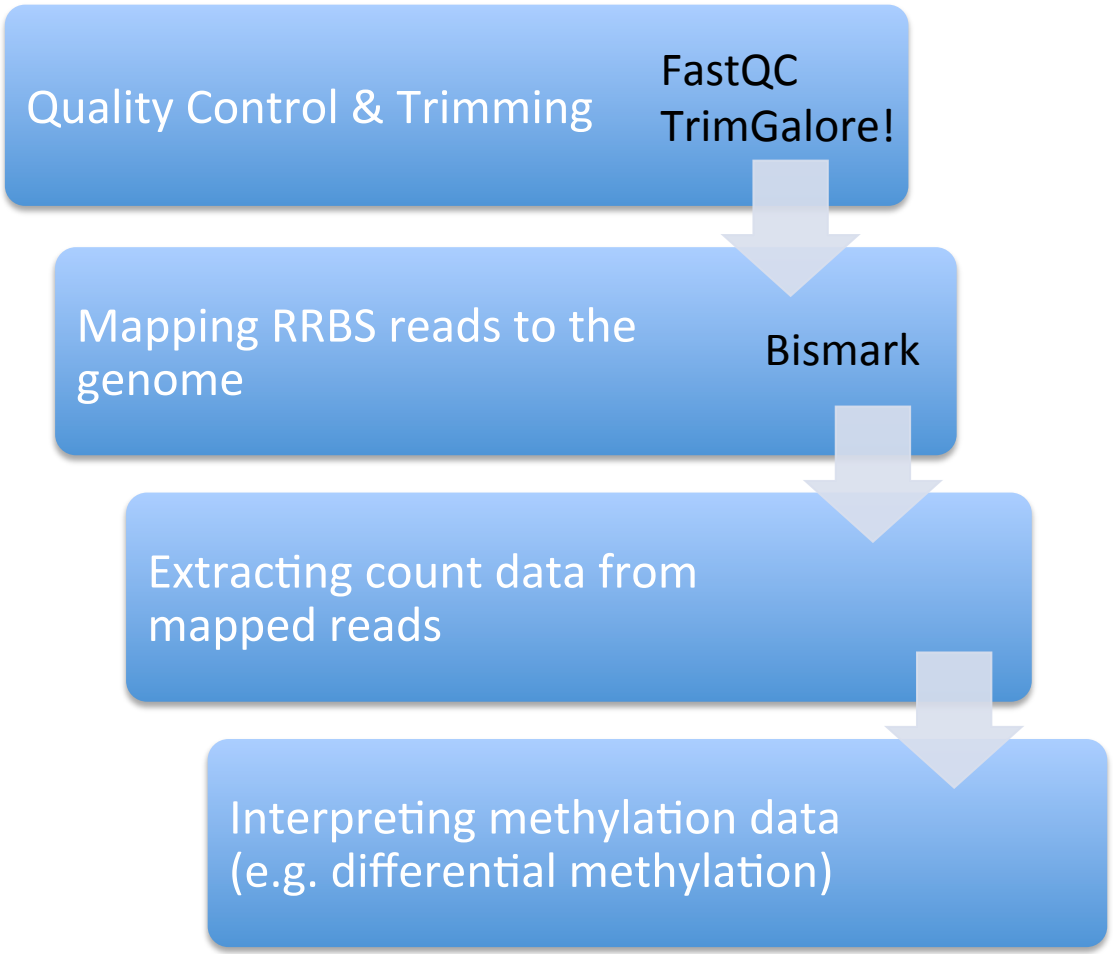

Per base sequence content
Sequence content across all bases
%T
%C
%A
%G
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)


Per base sequence content
Sequence content across all bases
%T
%C
%A
%G
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 14-15 22-23 30-31 38-39 46-47 54-55 62-63 70-71 78-79 86-87 94-95
Position in read (bp)

Bioinformatics (DNA methylation)



Bioinformatics (DNA methylation)



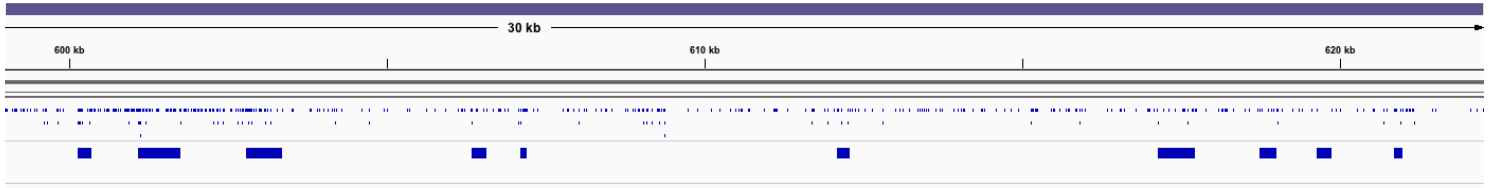
O. mykiss scaffold 13

SN747:551:C99B9ACXX:8:2315:19986:99910_1:N:0:ACAGTG 16 scaffold_570 338476 4 73M * 0 0
ATCACAAAACGCGCTAACCAAAATTACCAAATACACGATATTTCCTCCAACACATTAACGCGACTAACTTCCG
EDDDDDDDDFGHGIEHCJJJJIGHFFJJJIIFDHGIJJJJIGGJJJJIIGHGJJIJIJIJHDFHHFDFFFCCC NM:i:18 MD:Z:
0G0G4G1G12G0G2G3G0G1G4G1G8G7G0G4G2G0G6 XM:Z:x.....x.h.Z.Z.....hh..h...xh.h...Zx.h.....z.....hh.Z.Zx..xh.....Z XR:Z:CT
XG:Z:GA
SN747:551:C99B9ACXX:8:2315:19801:99933_1:N:0:ACAGTG 0 scaffold_2861 36547 42 97M * 0 0
CGGAAGGTGTTATAGAGGGTAGTGCGTACGGTTTAGTATATTATTGGGGTAAATTTTTTTGATATTTAGGGGAAGGTTTAAAAAATTGTGAAAGATT
@@@DDDDBB:CFA22AGE6+<:FFH8FFE@<FFFEA0BFFCECBD4BFB;;@CF>FCEFDDCCEDB@@@B?B==?<>@BB>@>@@BBAA@>>>:3A@
NM:i:17 MD:Z:9C2C18C0C0C4C2C1C10C2C0C3C3C9C0C0C16C1
XM:Z:Z.....h..x.....Z...Z..hhx....h..h.x.....h..hx...h...x.....hhh.....h. XR:Z:CT XG:Z:CT
SN747:551:C99B9ACXX:8:2315:20190:99881_1:N:0:ACAGTG 16 scaffold_614 418590 0 101M * 0 0
CAAATACAAAAAATCAAAACAAATAAATACCCACCCAAAACCTCAAACCATAAAACAAACAAAAAACATCATCTTATAAATAATCAAATAACCAACCACCA
>EEDA?5BBBDA>1595BB@EB@;@5;(,89,,''FFFE@3D=52FBFFIIFFFFF??9G:FFGD?F>@FFFCE9HCAEA?CCIIIFF<C2)DB@B;A@?: NM:i:24
MD:Z:
2G0G1G3G7G15A6A5G2G8G1G0G0G4G9G3G0G2G0G0G3G2G2G2G0XM:Z:..xh.h...h.....h.....h..z.....h.zxh....z.....h...hh..zxh
.....x..z..z XR:Z:CT XG:Z:GA
SN747:551:C99B9ACXX:8:2315:20074:99951_1:N:0:ACAGTG 16 scaffold_1684 42005 3 98M * 0 0
CATTAAACGACGAATTAATCTTCGTCACCGAACTCGTAAAAAAAAAACTCACGCTCTCCGTCGCCCCCCTCTCCGCATCACTACCATCTTCCTCTACCG
CDDBDB@BBCDDEEDDDDDDB?<C??8DDDD@DBDDDDDDDBCC@>D@<7DDDDDDDDDBDBBIGIGIGJIJJIJGHHFHGCJHFADDBGFFDBDBB?
NM:i:15 MD:Z:4G1G5G0G2G14G5G0G3G0G1G3G2T7T33G3
XM:Z:....h.h.Z..Zxh..h.....Z.....Z.h...Z.hh...hh.h...z.Z.....Z..H.....Z.....x..Z XR:Z:CT XG:Z:GA

“Looking” at data using IGV

“Looking” at data using IGV

O. mykiss scaffold 13



CG motif

RRBS fragments

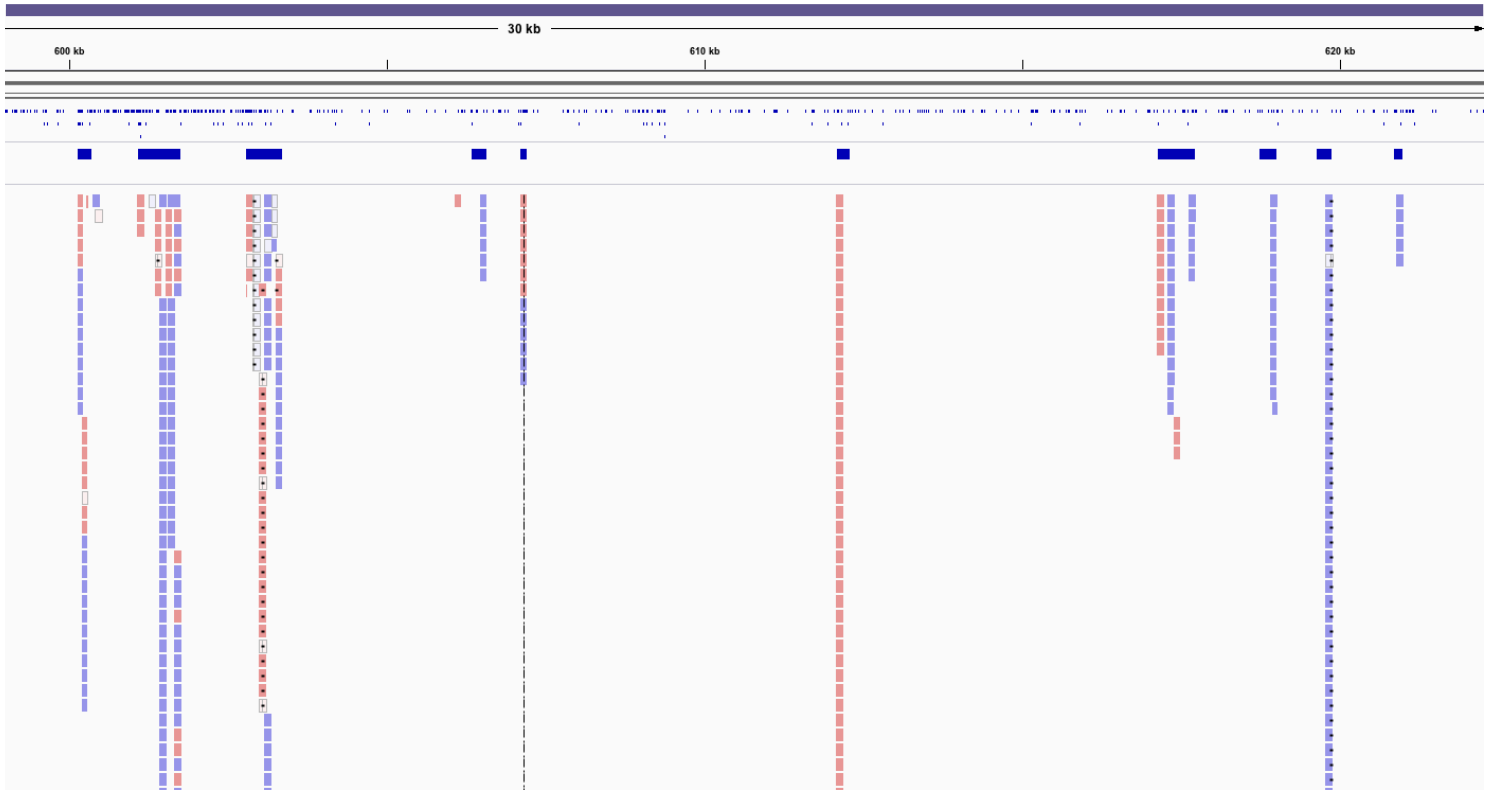
“Looking” at data using IGV

O. mykiss scaffold 13

CG motif

RRBS fragments

mapped reads

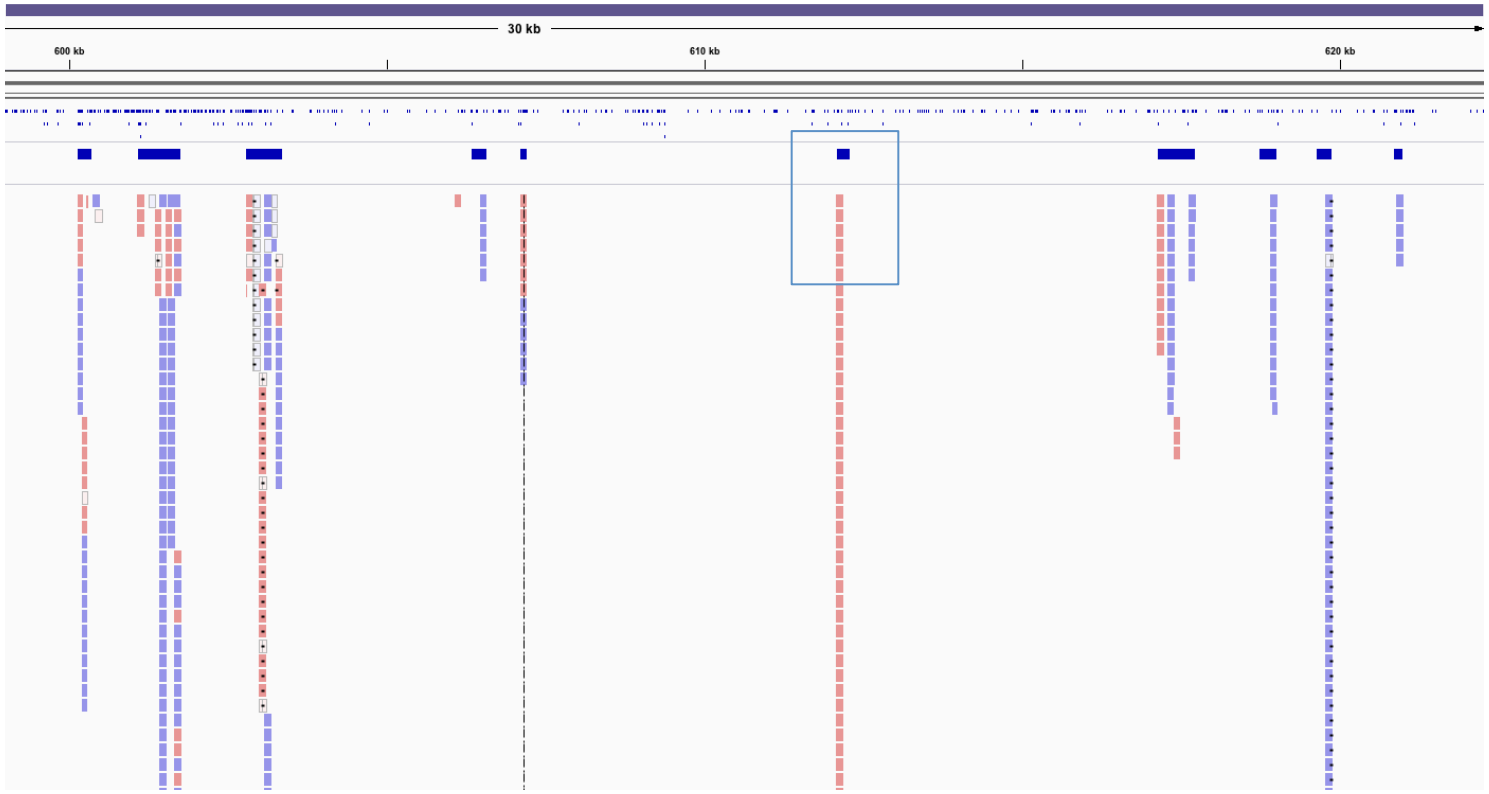


“Looking” at data using IGV

O. mykiss scaffold 13

CG motif
RRBS fragments

mapped reads

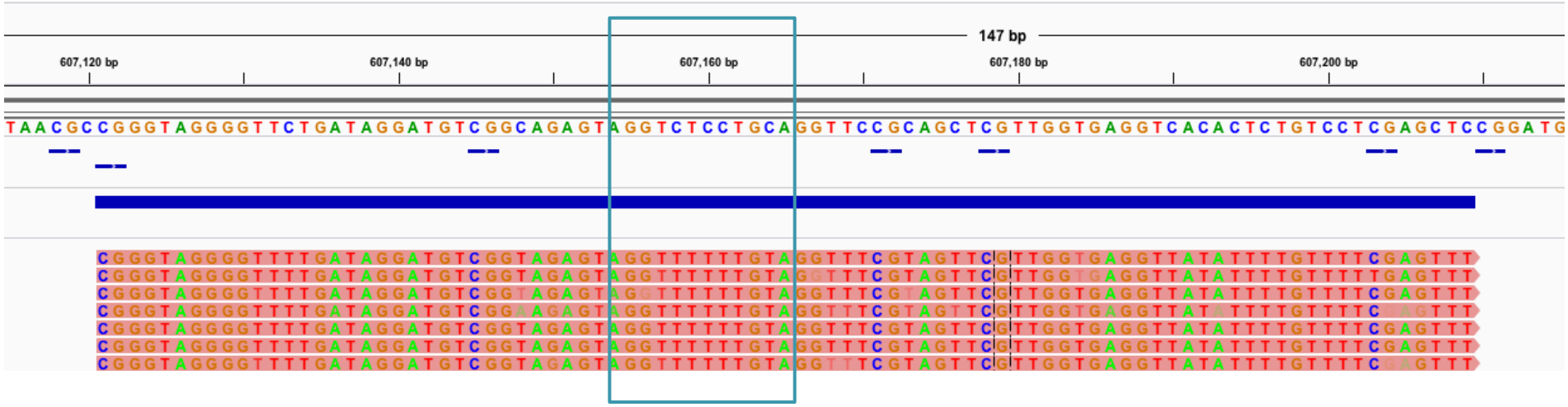


“Looking” at data using IGV

O. mykiss scaffold 13

“Looking” at data using IGV

O. mykiss scaffold 13



“Looking” at data using IGV

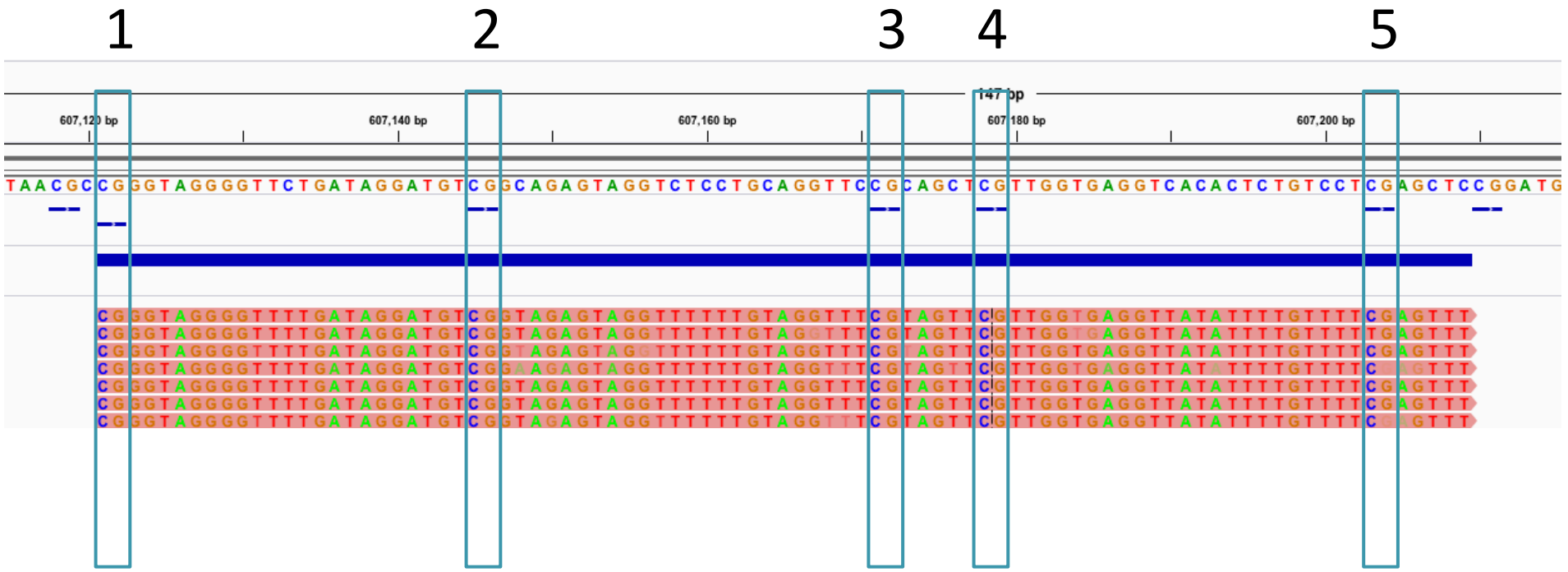
O. mykiss scaffold 13

“Looking” at data using IGV

O. mykiss scaffold 13

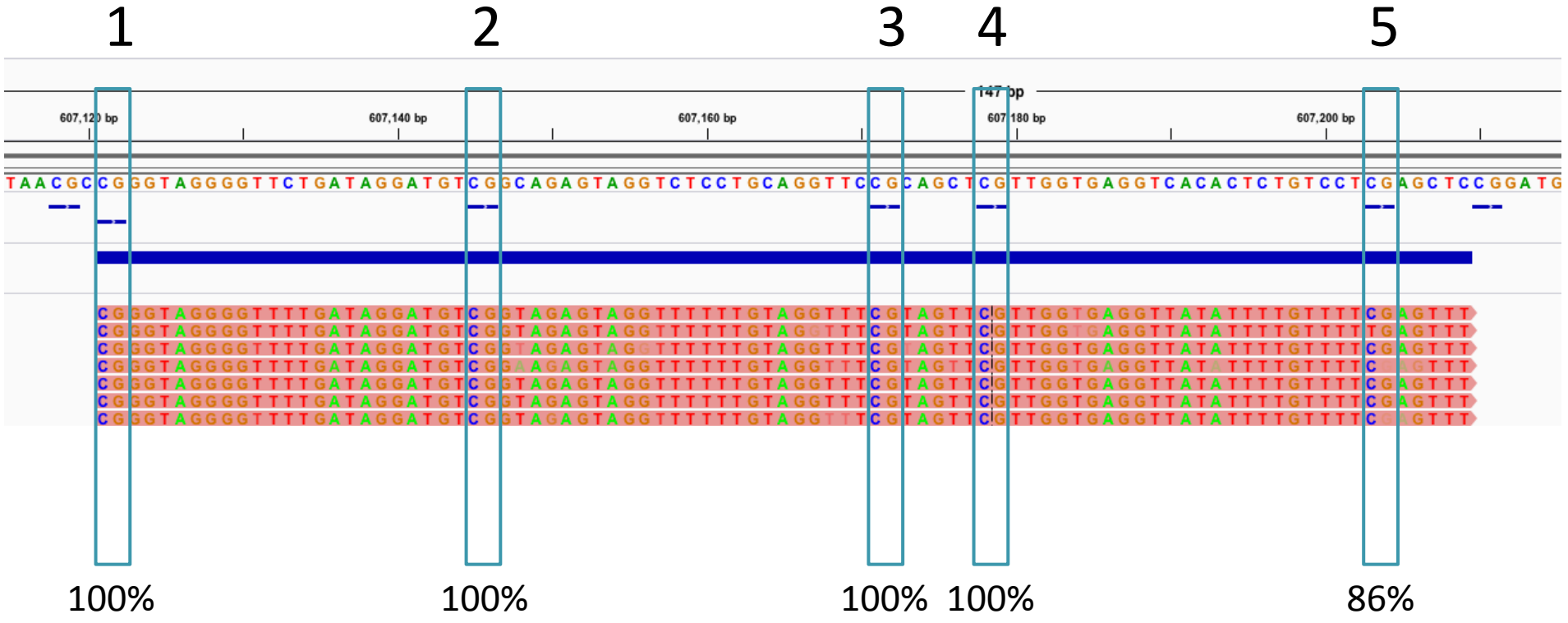
“Looking” at data using IGV

O. mykiss scaffold 13

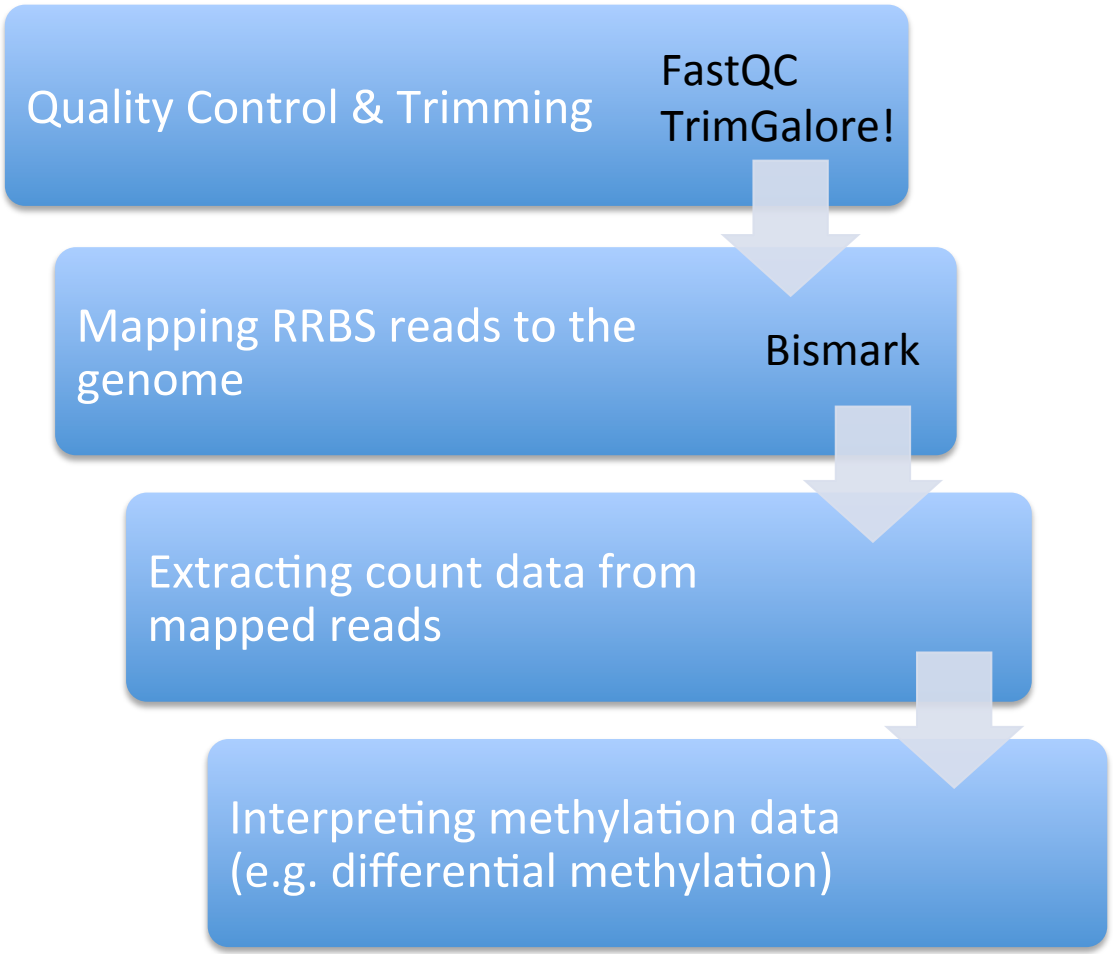


“Looking” at data using IGV

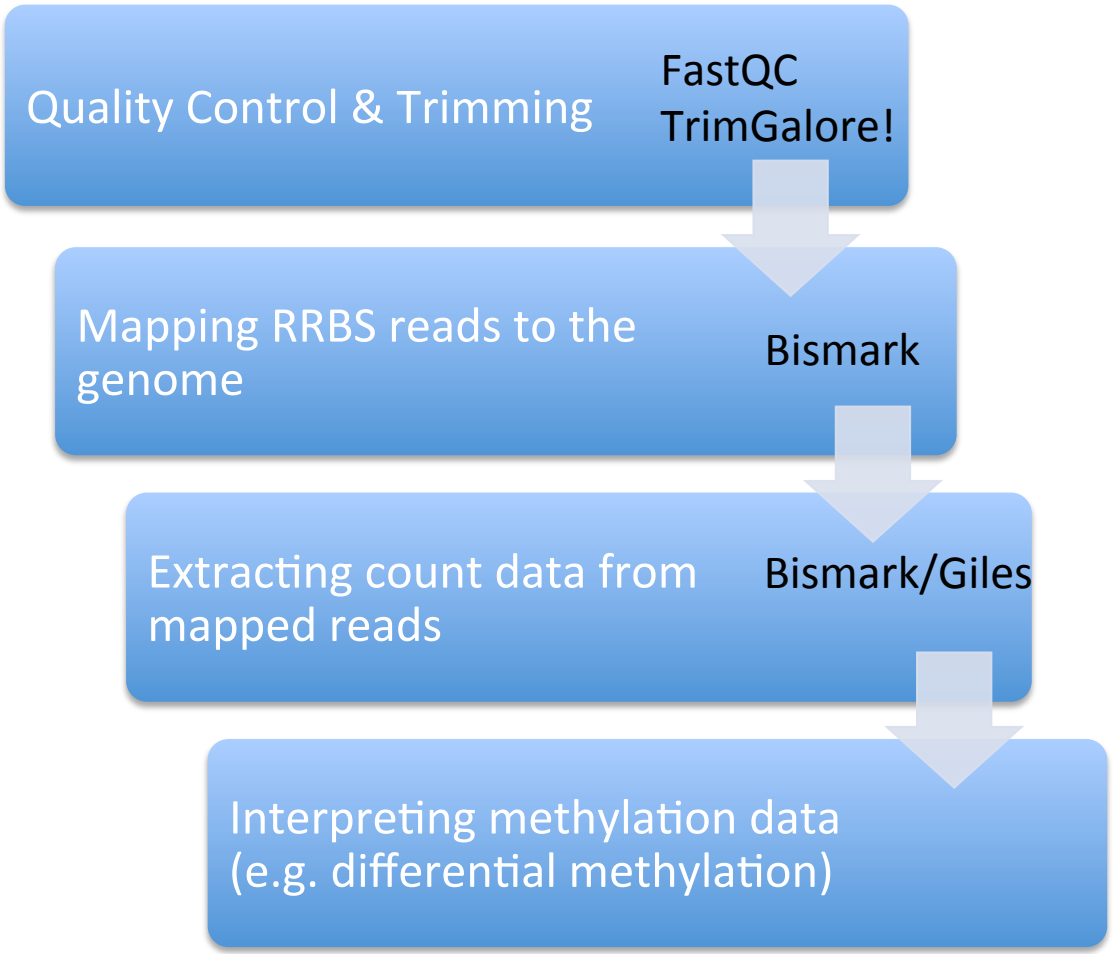
O. mykiss scaffold 13



Bioinformatics (DNA methylation)



Bioinformatics (DNA methylation)



Bioinformatics (DNA methylation)

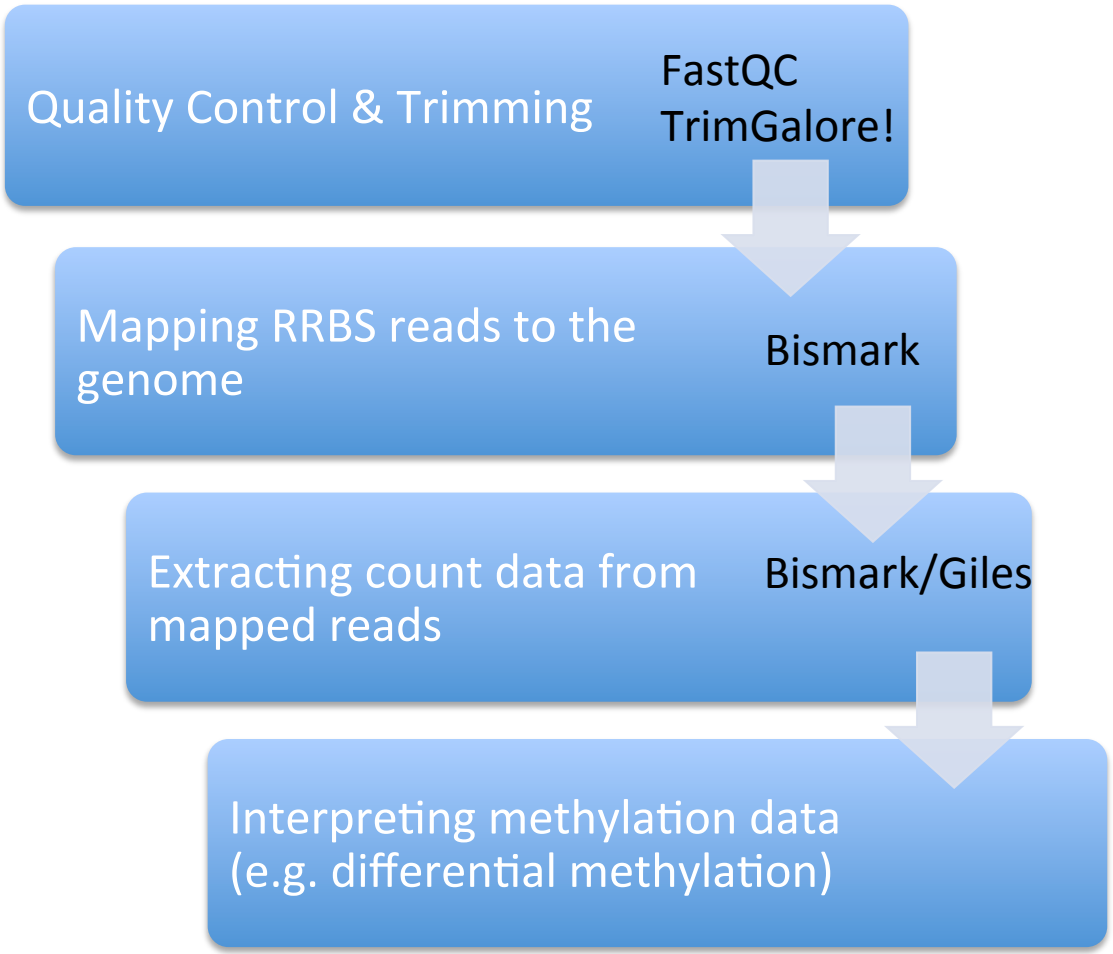
- Quality Control & Trimming — FastQC, TrimGalore!
- Mapping RRBS reads to the genome — Bismark
- Extracting count data from mapped reads — Bismark/Giles
- Interpreting methylation data (e.g. differential methylation)

Bioinformatics (DNA methylation)

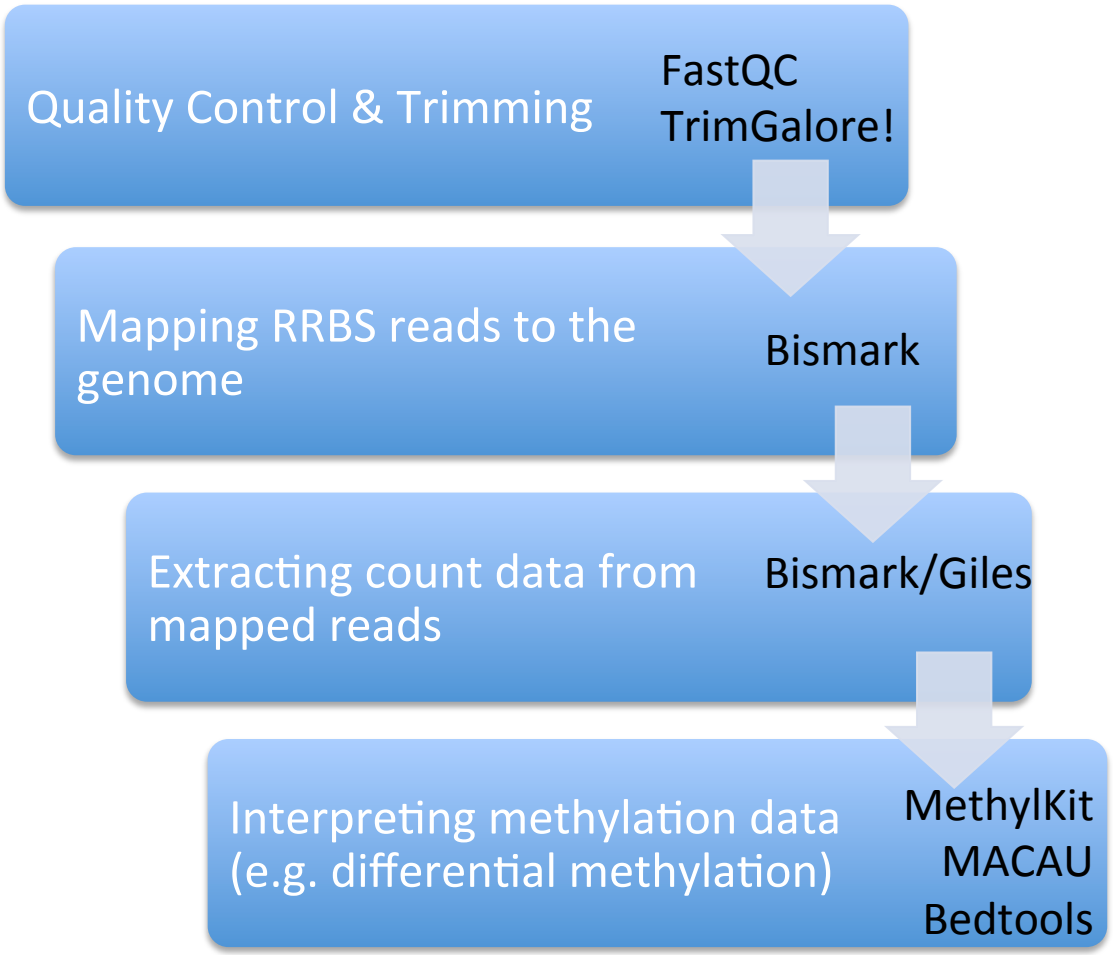
- Quality Control & Trimming — FastQC, TrimGalore!
- Mapping RRBS reads to the genome — Bismark
- Extracting count data from mapped reads — Bismark/Giles
- Interpreting methylation data (e.g. differential methylation)

chr	base	coverage	freqC	freqT
scaffold_1	5941	20	1.00	0.00
scaffold_1	5973	20	0.95	0.05
scaffold_1	5982	20	0.95	0.05
scaffold_1	5994	20	0.95	0.05
scaffold_1	5998	8	0.88	0.12
scaffold_1	6012	20	1.00	0.00
scaffold_1	6101	2	1.00	0.00
scaffold_1	6103	2	1.00	0.00
scaffold_1	6278	9	0.89	0.11
scaffold_1	6285	42	0.98	0.02

Bioinformatics (DNA methylation)



Bioinformatics (DNA methylation)

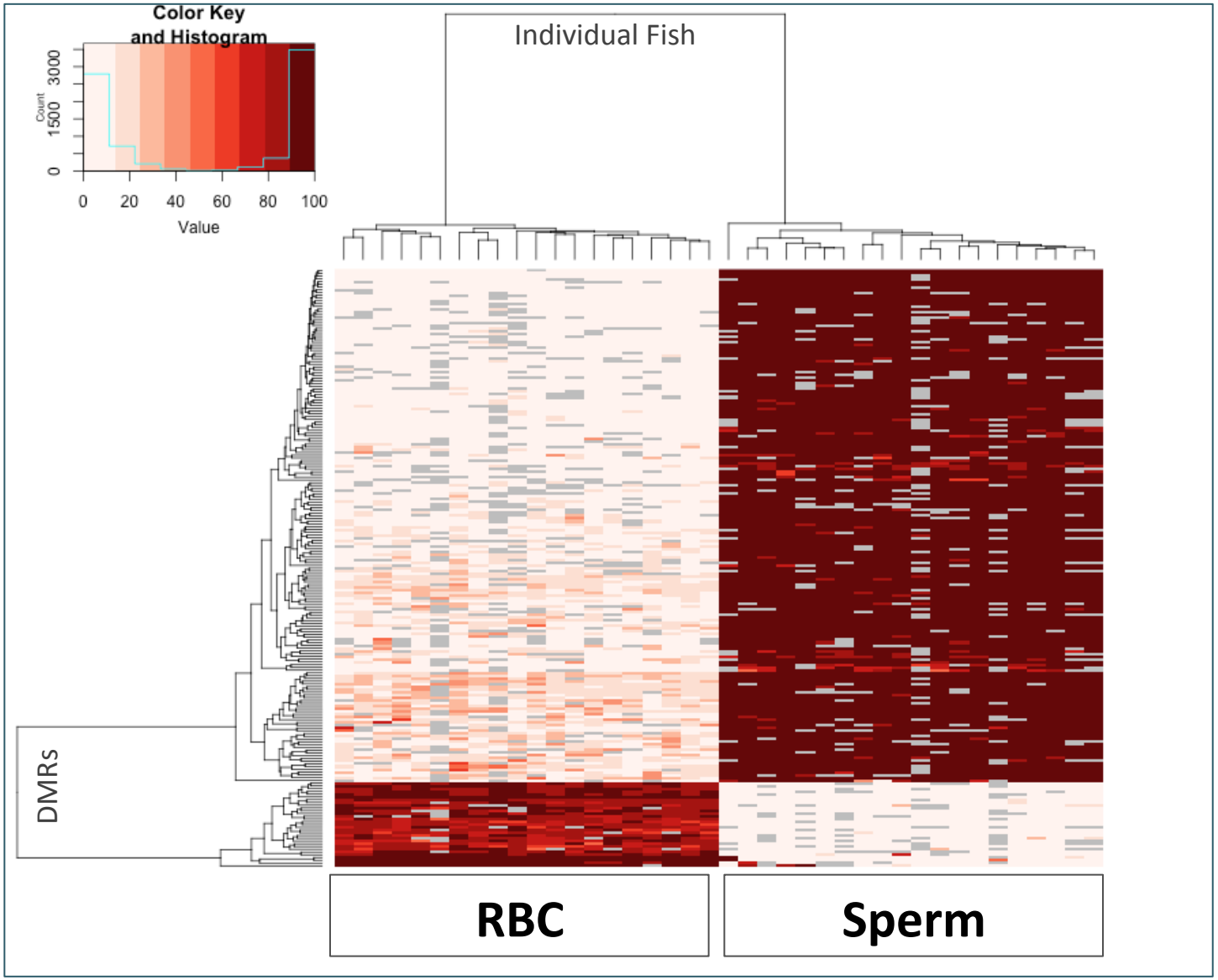


Differential Methylation Analysis

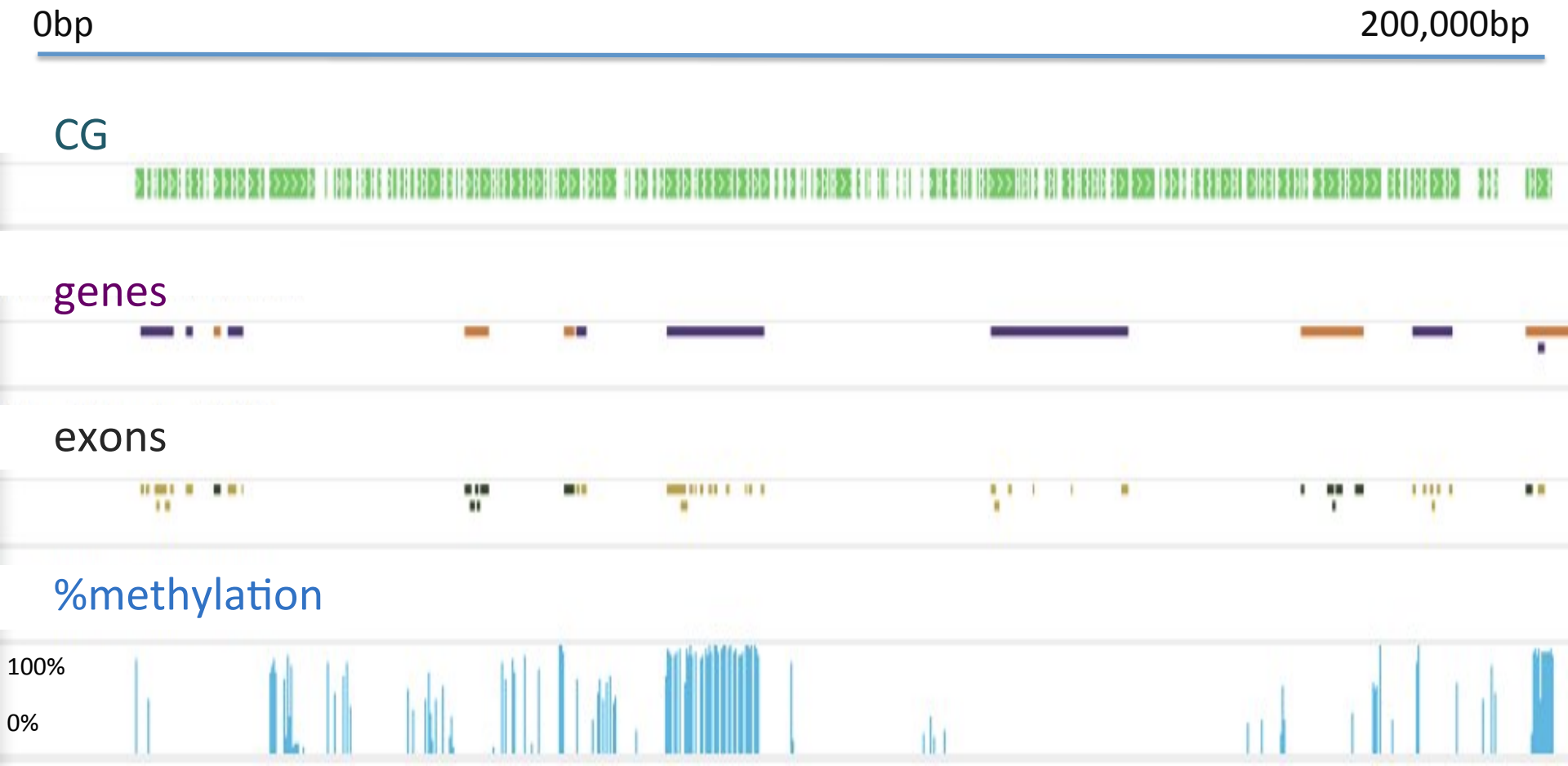
Differential Methylation Analysis

- MethylKit
  - R package
  - Logistic regression
  - Single variable (e.g. treatment v. control)
- MACAU
  - Beta-binomial mixed model
  - Multiple variables including relatedness

Differential Methylation Analysis

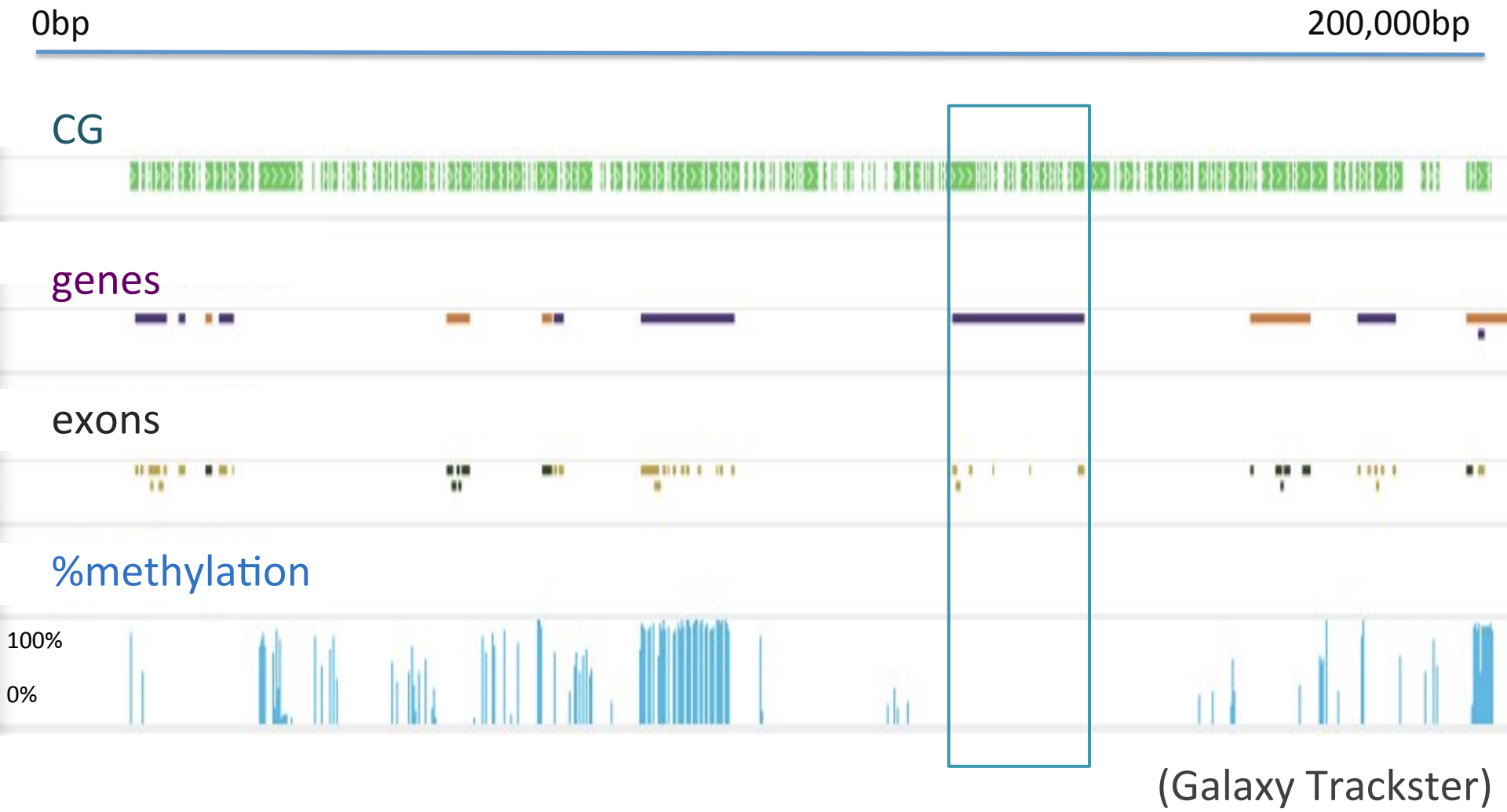


Descriptive Methylome Data



(Galaxy Trackster)

Descriptive Methylome Data



(Galaxy Trackster)

Summary

- It can be hard to work with data that you can't **see**
  - Moving data takes a lot of time and space
  - Formatting data is a constant challenge
  - Software packages are great, but it's always important to understand what they are doing
  - Bisulfite sequencing has unique challenges