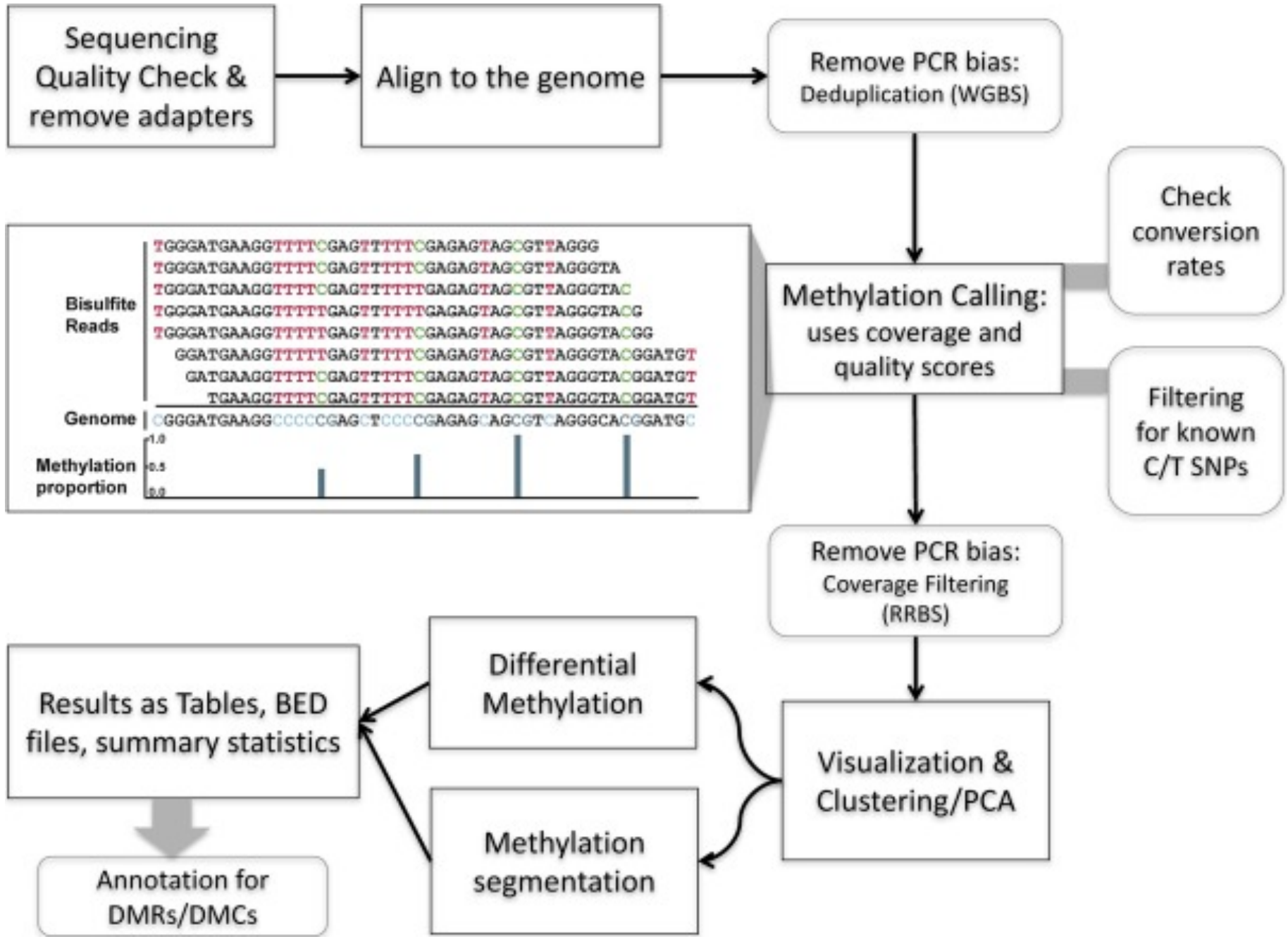
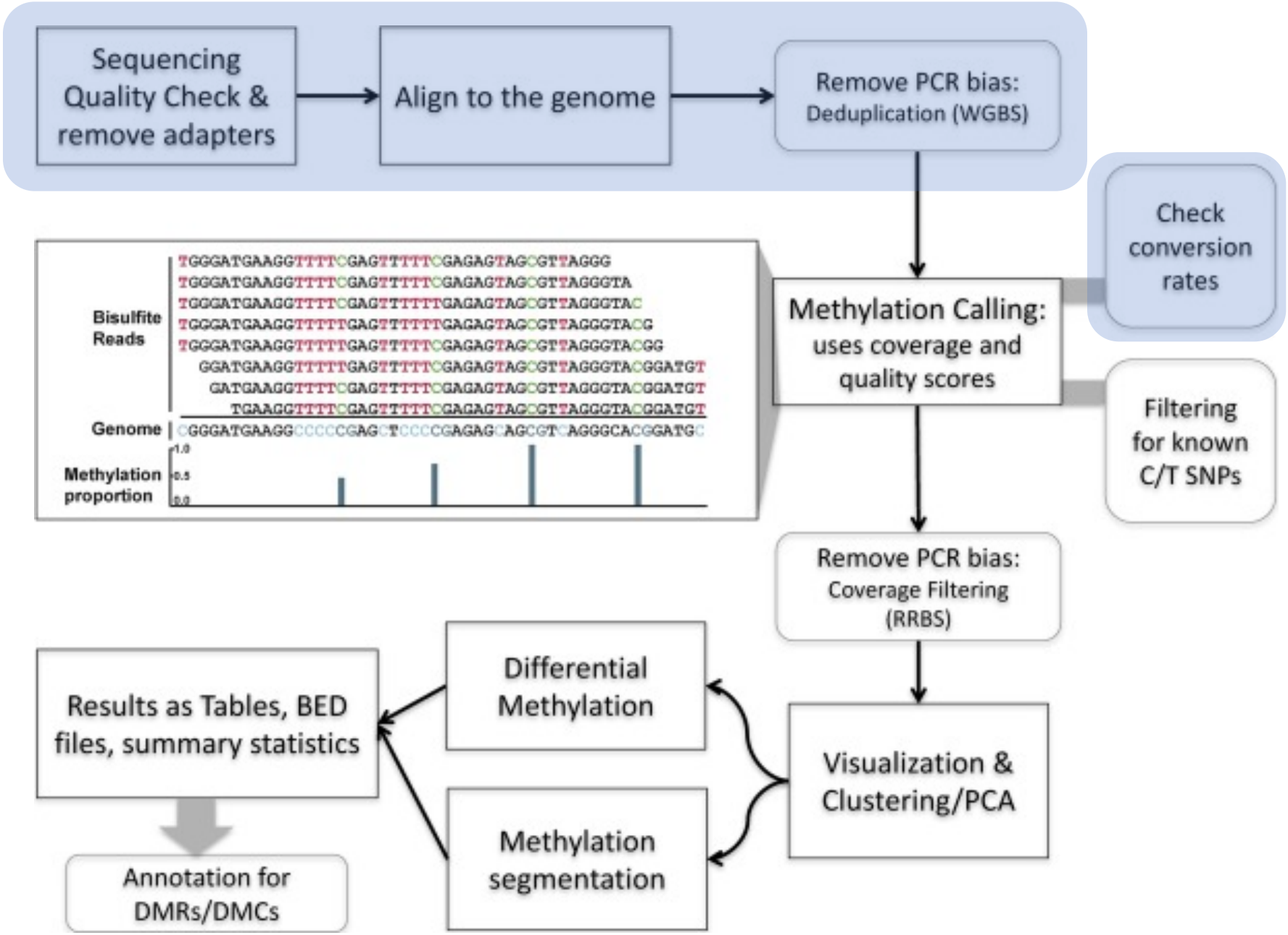
Methylation Sequencing Workflow


Sequencing Quality Check & remove adapters
Align to the genome
Remove PCR bias: Deduplication (WGBS)
Check conversion rates
Bisulfite Reads
Genome
Methylation proportion
1.0
0.5
0.0
Methylation Calling: uses coverage and quality scores
Filtering for known C/T SNPs
Remove PCR bias: Coverage Filtering (RRBS)
Differential Methylation
Results as Tables, BED files, summary statistics
Visualization & Clustering/PCA
Methylation segmentation
Annotation for DMRs/DMCs

Quality Control is important!

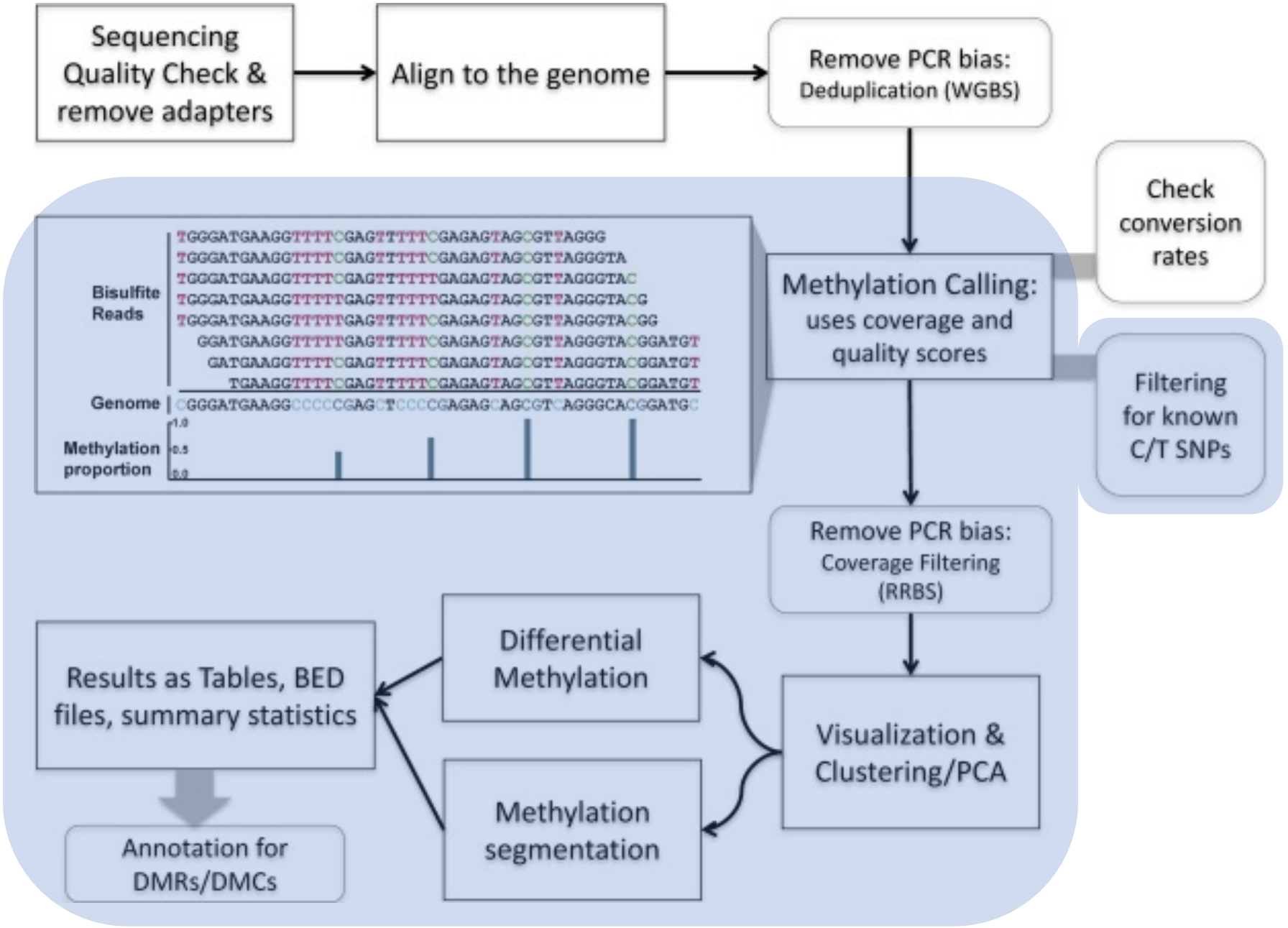
- Relies on accurate C > T detection
- Pre-alignment:
  - Base quality
  - Trim adapters
- Right aligner for the job: Bismark, bwa, bsmap, BS-seeker, …
- Post-alignment:
  - Incomplete conversion? Check non-CpG methylation, should be near 100%
  - Degradation? Check alignment rates and read length
  - PCR bias? Deduplicate (maybe not for RRBS)



nf-core

A community effort to collect a curated set of analysis pipelines built using Nextflow

- Nf-core pipeline: methylseq (see Thursday)
- Preprocessing + QC
  - 2 aligners: Bismark or bwa-meth/MethylDackel
  - QC: qualimap, preseq and multiqc
- Output of this can be used as starting point downstream analysis



MethylKit

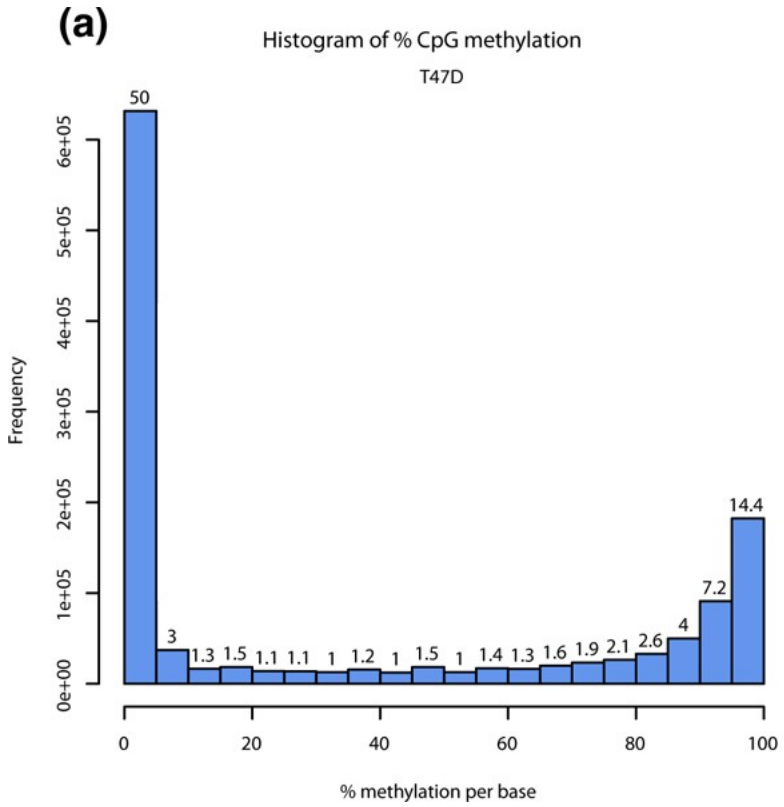
R package for downstream analysis of bisulfite data

Can read in Bismark coverage files as input…

Descriptive statistics

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

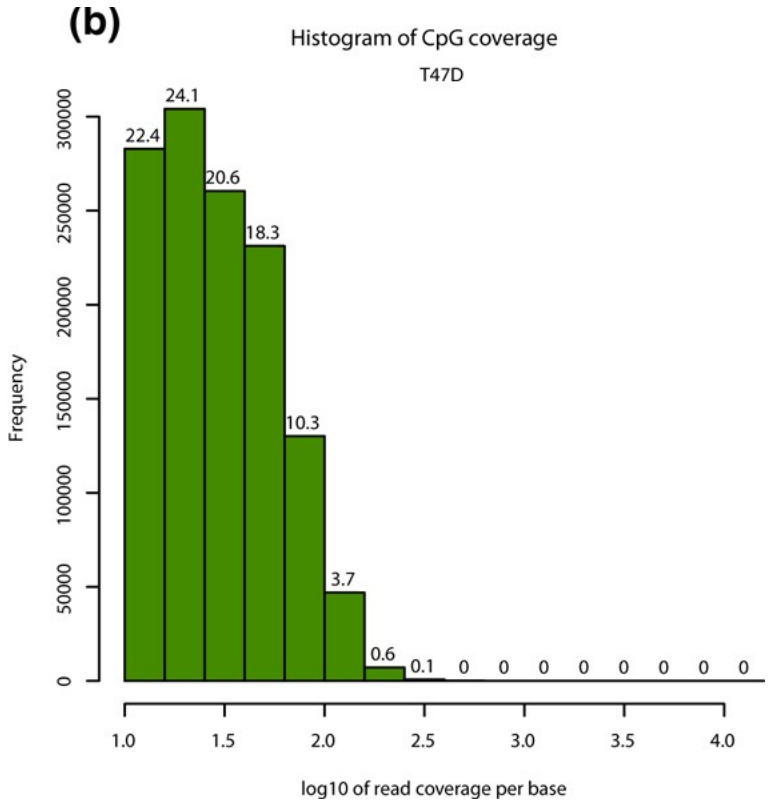
- After reading in the data…
- Distribution percentage methylation



- Equivalent to Beta value in array data
- Expect a peak at low and high ends of the distribution

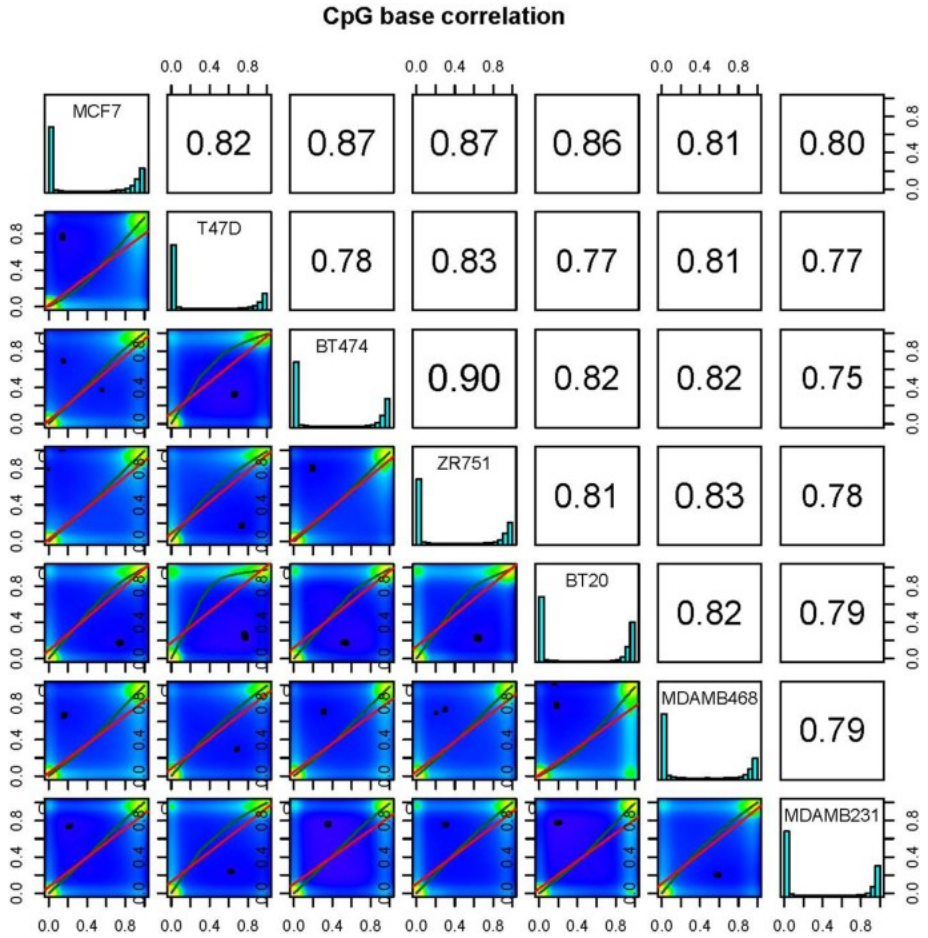
Descriptive statistics

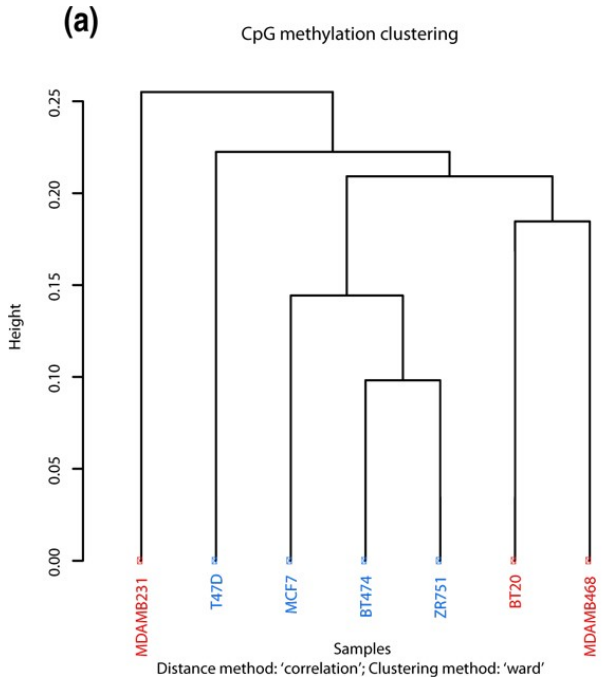
- After reading in the data…
- Coverage distribution

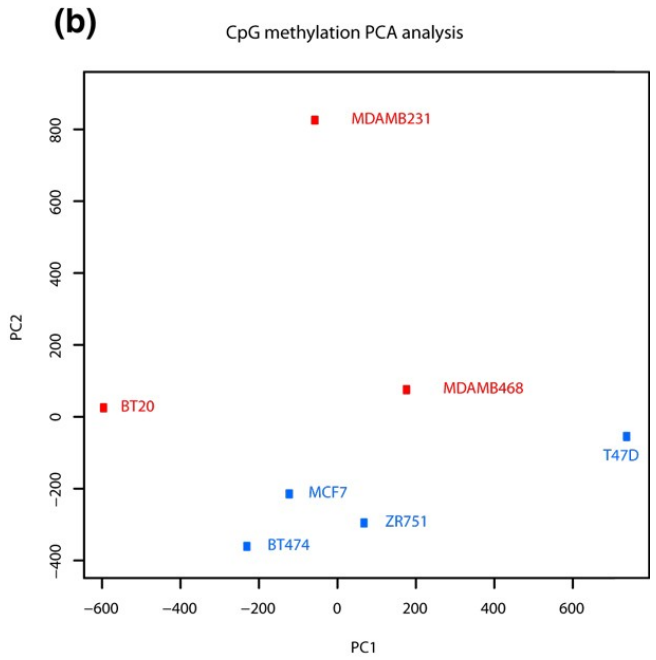


- Experiments that suffer from PCR duplication will have a secondary peak towards the right hand site of the histogram
- Can be used to determine filter cutoff; both very low and very high coverage

Sample Structure



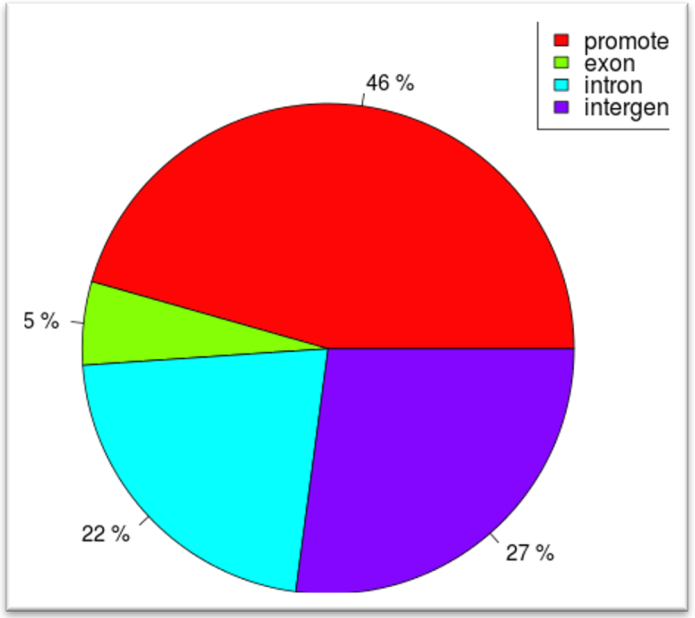




Differential analysis

- Again, many choices; usually calculated by comparing the proportion of methylated Cs in a test sample relative to a control
- No replicates: Fisher's Exact Test
- Replicates:
  - Linear regression (limma as for arrays)
  - Logistic regression (works with [0-1] data)
  - Beta-binomial (deals better with count data)
- Regression models can add covariates (batch, age, sex, …)
- Can also aggregate in regions (see tutorial)

Annotation



- How to interpret the DMR/DMPs?
- Integrate with genome annotation datasets
  - Where in relation to gene/regulatory region
  - Promoter or intron or exon?
- Genomation R package: toolkit for annotation and in bulk visualization of genomic intervals
- Tutorial: basic annotation transcripts and CpG islands
- Requires some knowledge of R (especially the GenomicRanges package)

Remarks…

- Normalization somewhat less important for bisulfite sequencing (Fisher’s Exact is sensitive to sequencing depth though)
- Gene enrichment is as difficult as for arrays; not many implemented methods (rGREAT, Goseq).

Dataset

- Small dataset of mouse mammary gland cells
- 4 samples: 2 luminal, 2 basal
- Bismark coverage files

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