SciLifeLab

GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

DNA Methylation Methods and Technologies

Jessica Nordlund, PhD
Managing Director
SciLifeLab National Genomics Infrastructure
SNP&SEQ Technology Platform
Uppsala University

NBIS course - Epigenomics Data Analysis Sept 18, 2023

Outline

i. Overview of methods for interrogation of DNA methylation

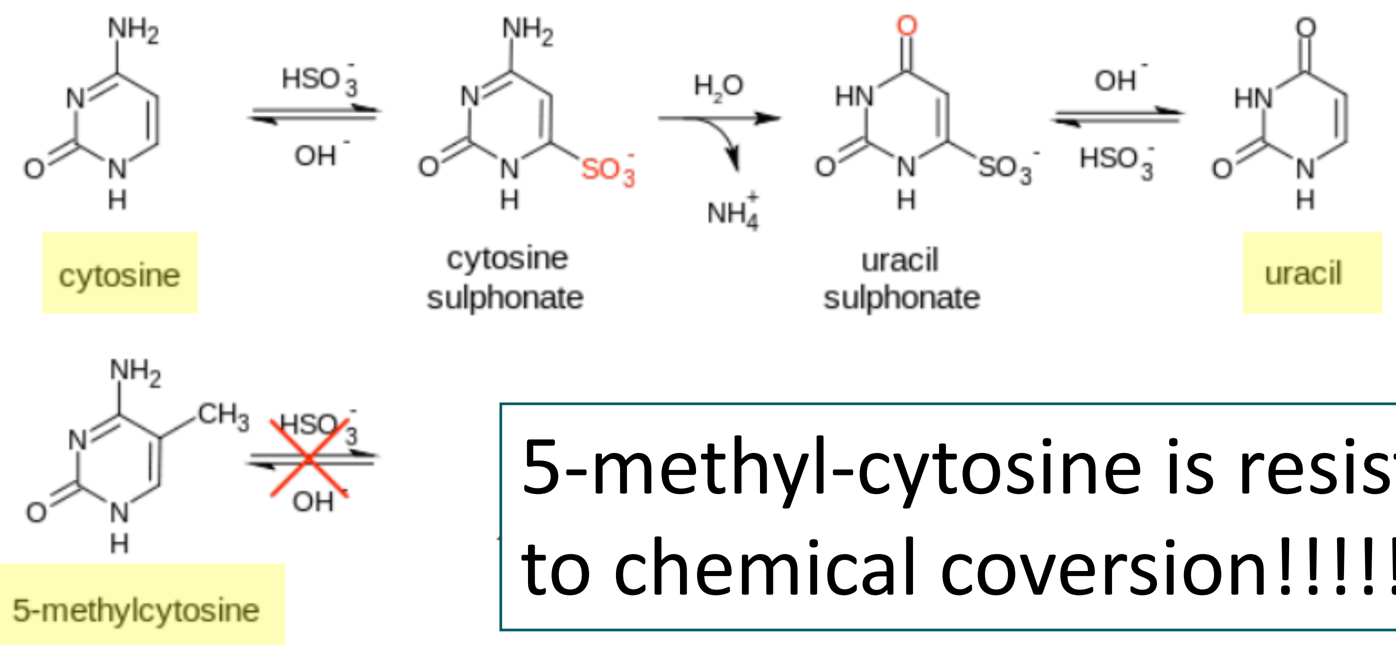
- Overview of important concepts
- Enrichment & targeted-based methods
- Genome-wide methods

ii. How to access epigenomics services for your research project at Sweden’s National Genomics Infrastructure (NGI)

Short intro: Conversion

Bisulfite coversion has been the ”Gold standard” for DNA methylation analysis.
Provides ”single nucleotide resolution”.

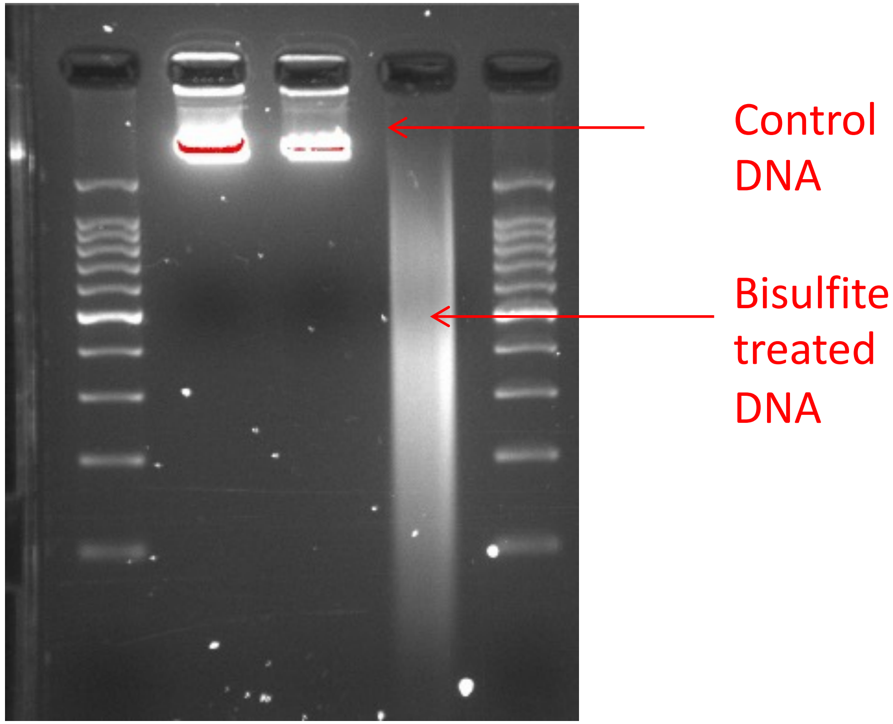
The Chemistry of Bisulfite Conversion of Cytosine to Uracil:



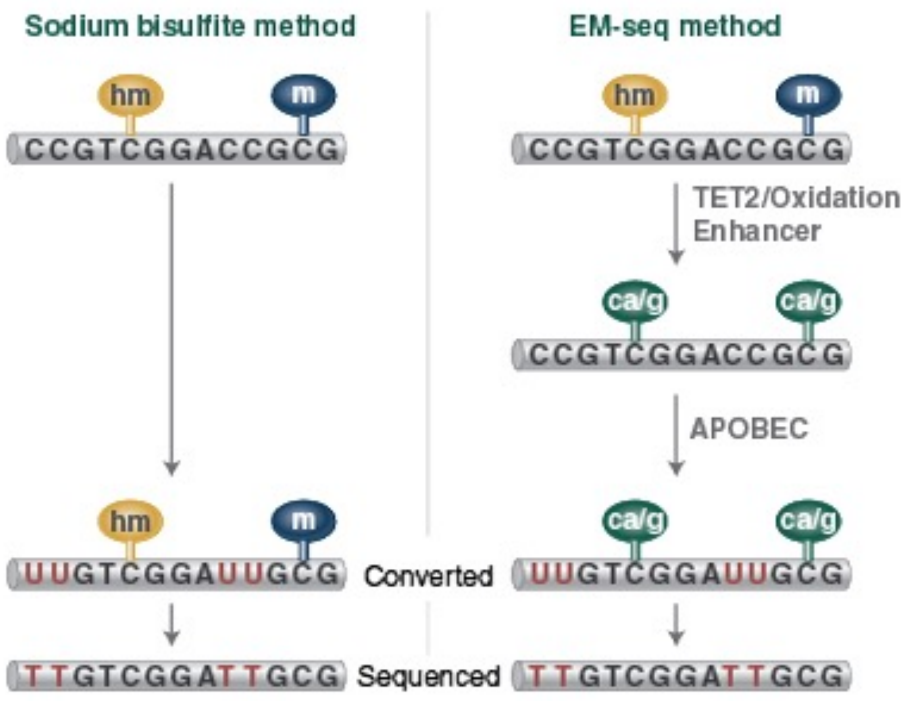
5-methyl-cytosine is resistant to chemical coversion!!!!!

What you need to know about bisulfite conversion

- Very harsh chemical that degrades and fragments DNA



New innovation- Enzymatic conversion!



Gentle with little/no strand breakage!

Figure: www.NEB.com

WGBS is the gold standard for methylome analysis, but the chemical bisulfite reaction:

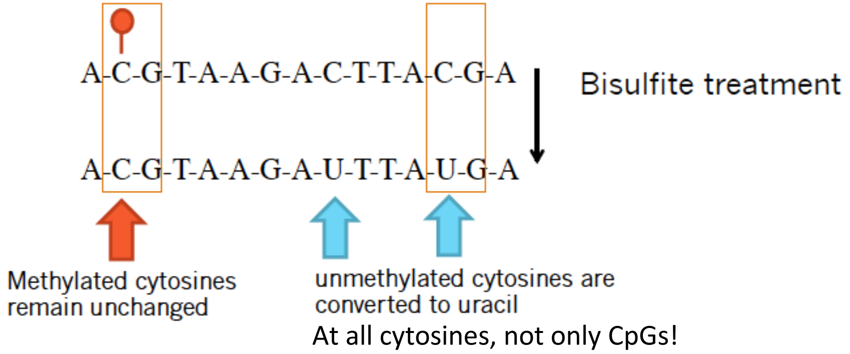
I. Damages / degrades DNA
II. Results in fragmentation / loss
III. Can result in CG bias and uneven genome coverage

Enzymatic methylation sequencing (EM-seq)
TET2 enzymatically oxidizes 5mC and 5hmC through a cascade reaction into 5-carboxycytosine (5caC)

5-methylcytosine (5mC) → 5-hydroxymethylcytosine (5hmC) → 5-formylcytosine (5fC) → 5-carboxycytosine (5caC)

A second enzymatic step uses APOBEC to deaminate cytosine to uracil, but does not affect 5caC.

Base-pair resolution and quantitative measurement of methylation levels



A-C-G-T-A-A-G-A-C-T-T-A-C-G-A

Bisulfite treatment

A-C-G-T-A-A-G-A-U-T-T-A-U-G-A

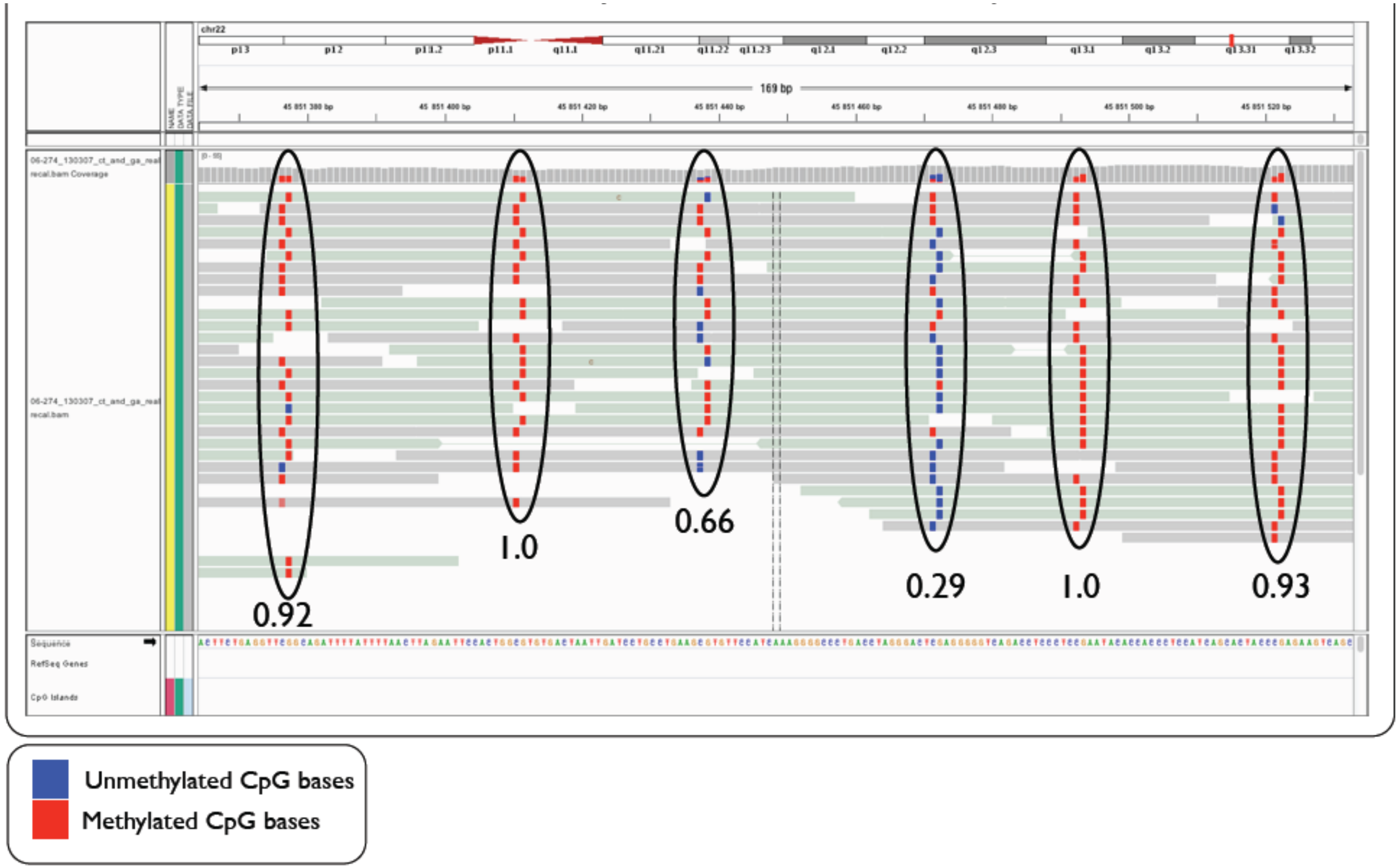
Methylated cytosines remain unchanged

unmethylated cytosines are converted to uracil

At all cytosines, not only CpGs!

C = methylated

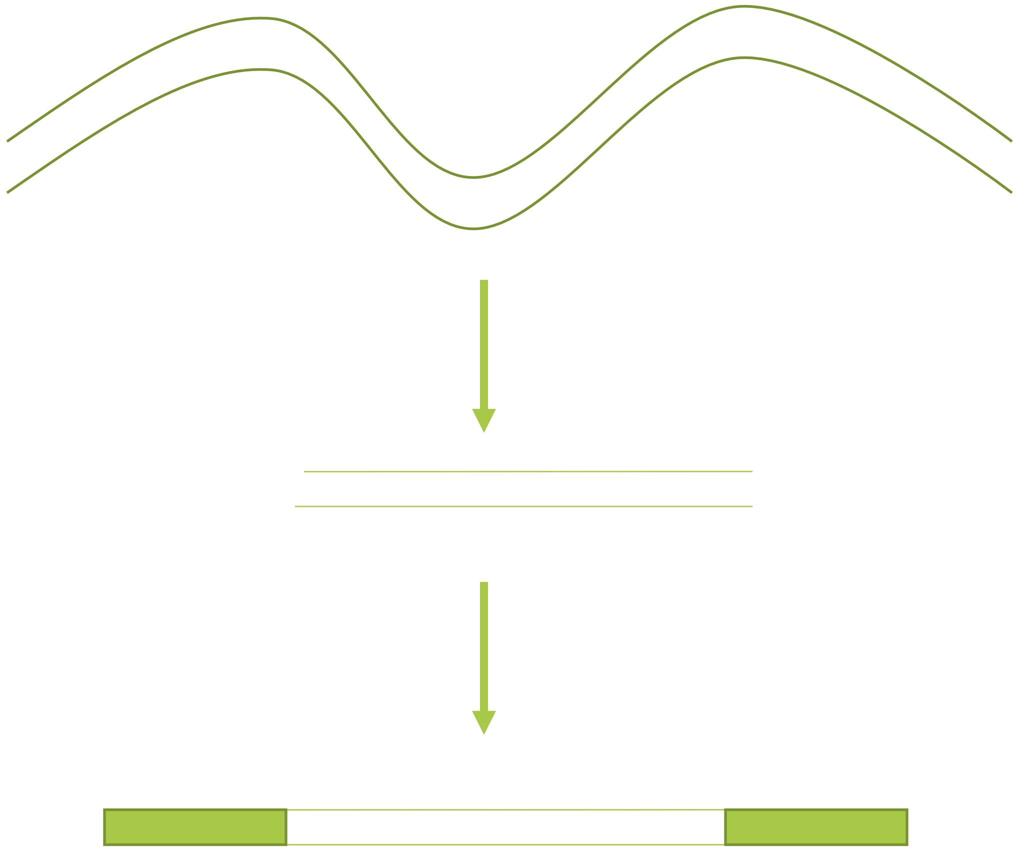
U->T = unmethylated



🟦 Unmethylated CpG bases

🟥 Methylated CpG bases

Short intro: ”NGS” libraries

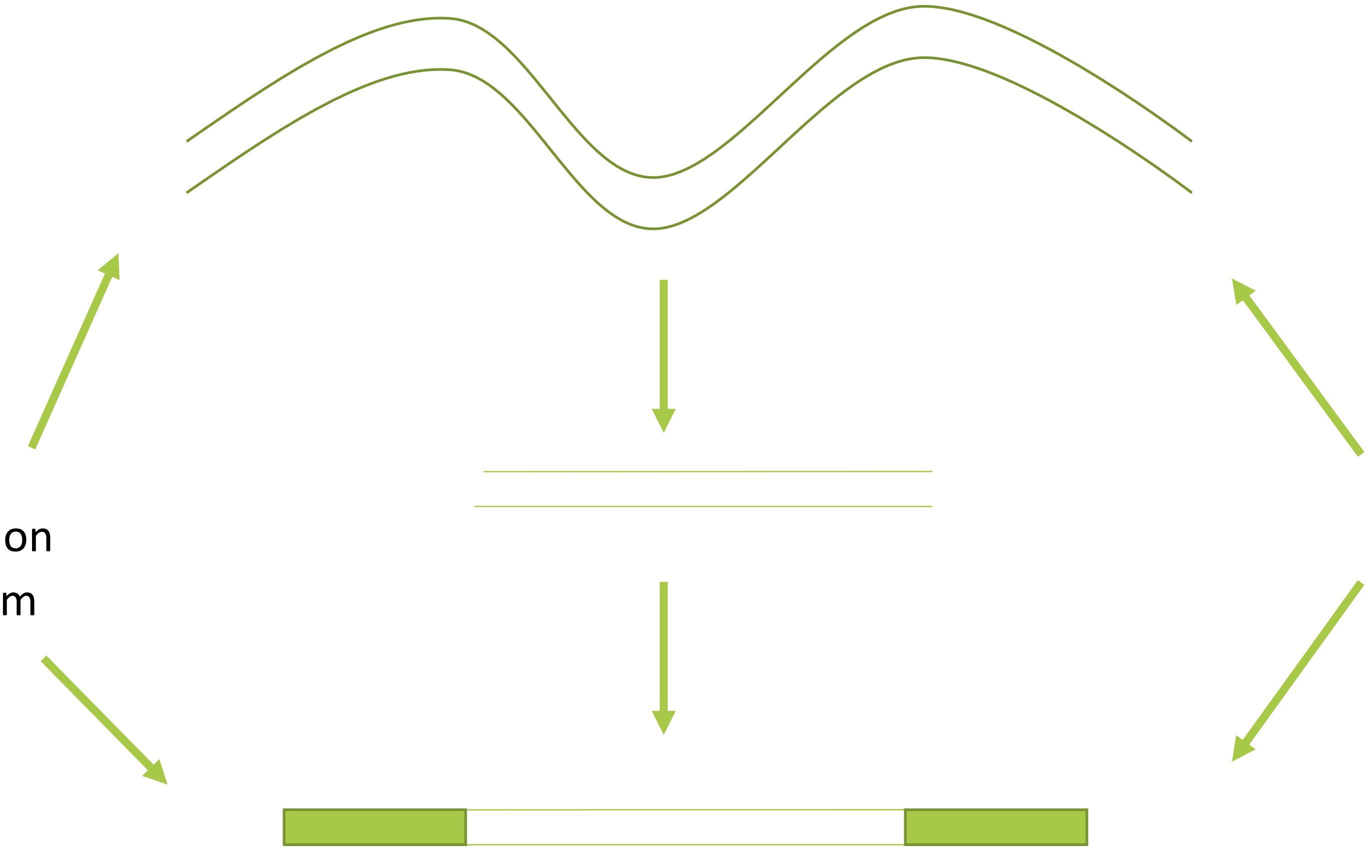


Double stranded genomic DNA

Shearing to make DNA fragments shorter (with bisulfite treatment optional)

Ligate platform-specific sequencing adapters

Short intro: ”NGS” libraries



+ Bisulfite conversion

Do distinguish C from 5mC

- Amplification

PCR and whole genome amplification (WGA) do not copy epigenetic marks like DNA methylation!!!

Enrichment & targeted-based methods

Different approaches to reduce the genome to regions of interest (typically those with many CpG sites)

- Cost saving (less sequencing required)
- Less computationally intensive (less data generated)
- High throughput (some approaches)

Enrichment-based methods

Capture of methylated DNA fragments using methyl-binding protein or a anti-methyl-cytosine antibody

- MeDIP-seq (Methylated DNA immunoprecipitation):
  - ✓ Genome-wide coverage
  - ✓ ~150bp resolution.
  - ✓ Anti-body against 5-Hydroxy-methyl-cytosine
  - ✓ Relatively cost-efficient
- MBD-seq (Methylated DNA binding domain):
  - ✓ Genome-wide coverage
  - ✓ ~150bp resolution.
  - ✓ Only capture CpG methylation not CHH
  - ✓ Relatively cost-efficient

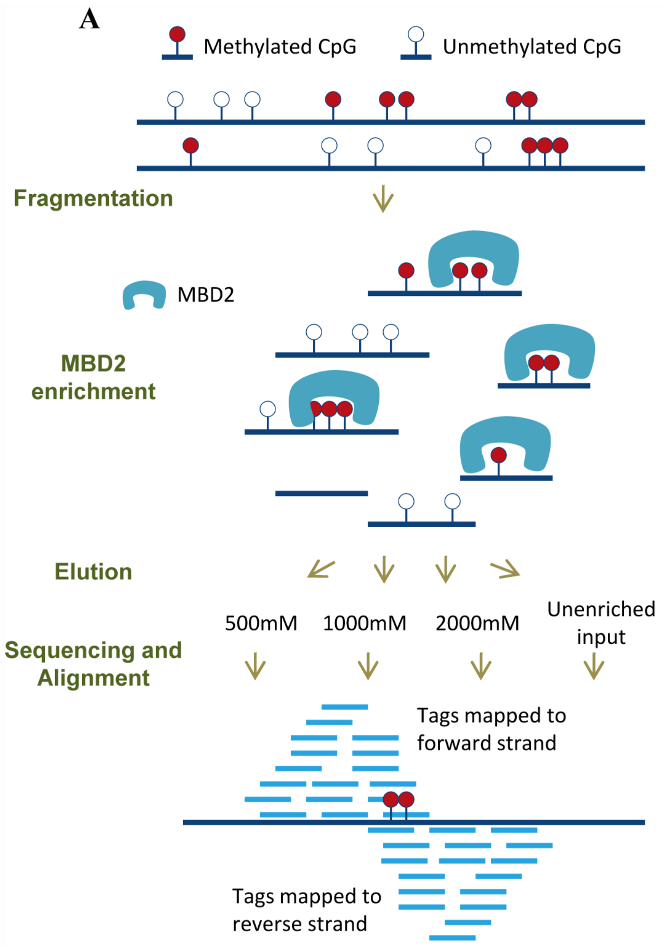
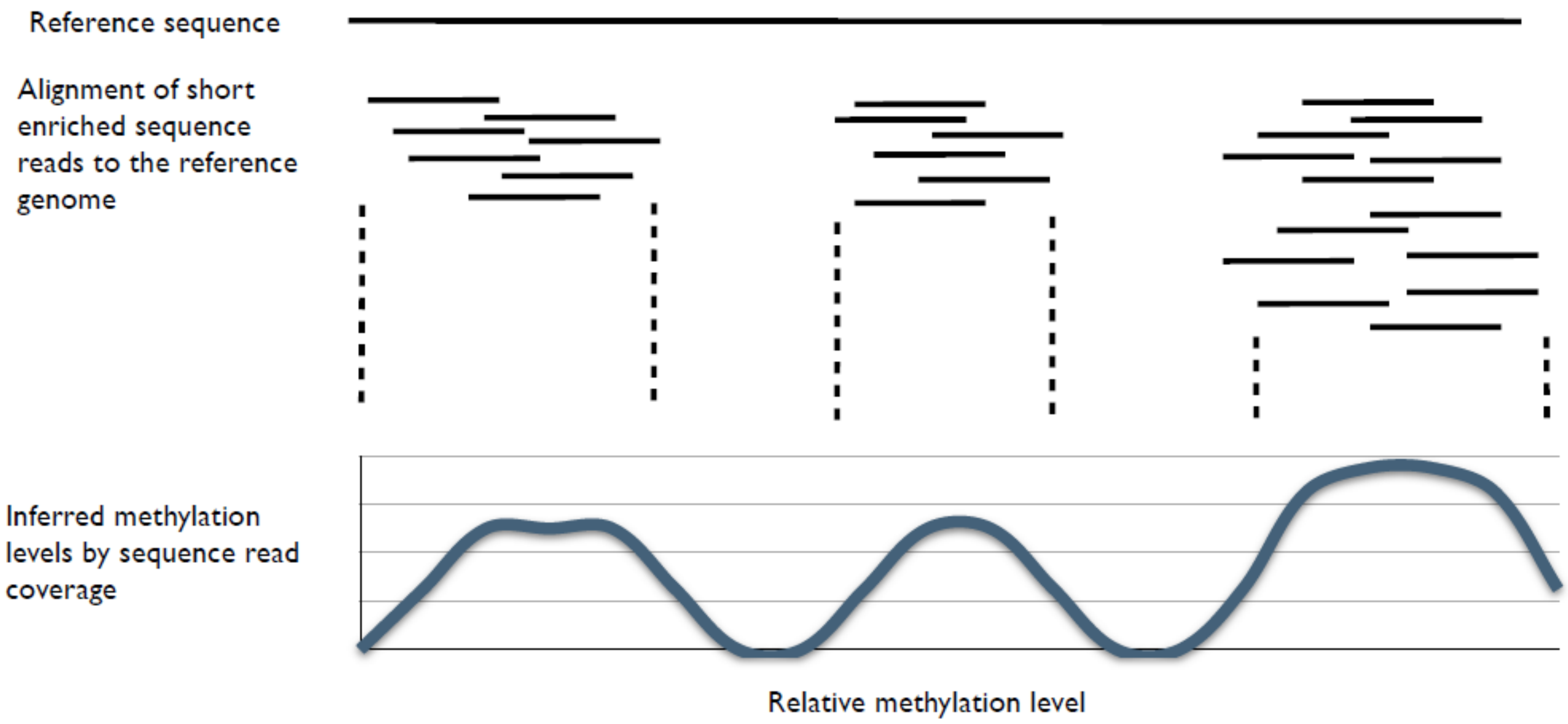


Figure from:
Lan, et al. (2011) High Resolution Detection and Analysis of CpG Dinucleotides Methylation Using MBD-Seq Technology.
https://doi.org/10.1371/journal.pone.0022226

Enrichment-based methods


Reference sequence
Alignment of short enriched sequence reads to the reference genome
Inferred methylation levels by sequence read coverage
Relative methylation level

The depth of sequence reads is taken as an indirect measurement of Methylation levels

Pros:

- Works for different species

Cons:

- Not base-pair resolution
- Indirect measurement of DNA methylation can be more difficult to interpret
- Lab-intensive and not easily automated

Target-Capture

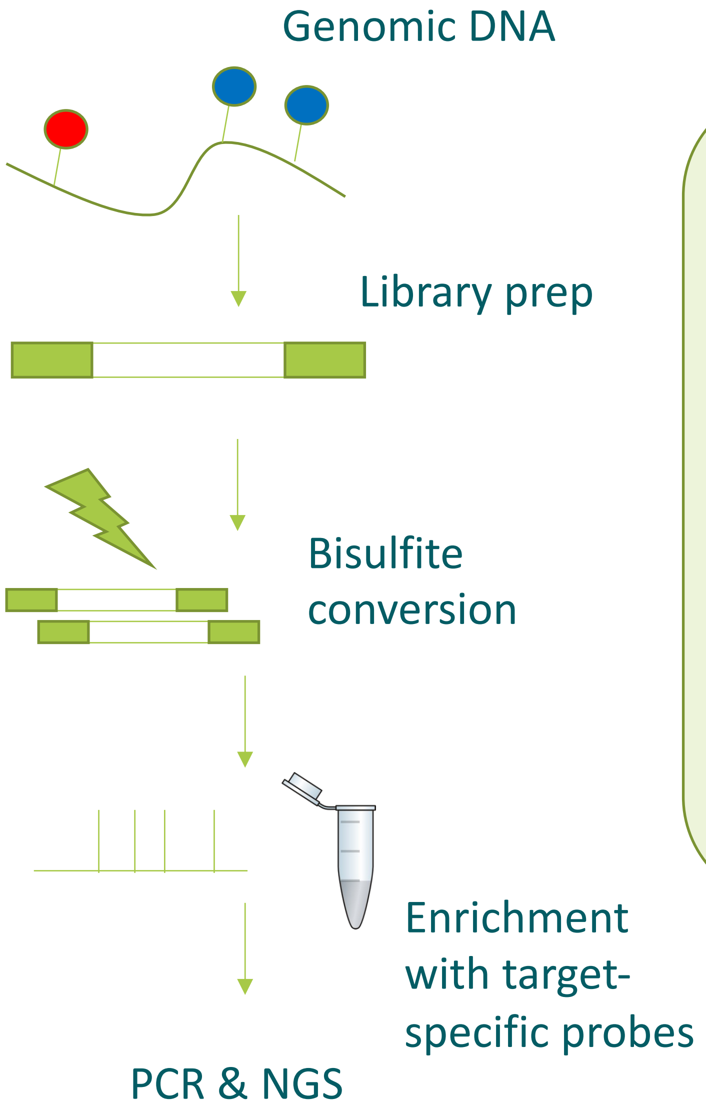
Target-capture of pre-defined genomic regions, NGS library preparation, uses bisulfite conversion.

Pros:

- Focused set of targets regions: can achieve high coverage on target
- "Cost-effective"
- Captures millions of CpG sites (3-5M)

Cons:

- Typically only for Human, other species possible on some platforms
- Standard conversion cannot distinguish between 5mC and 5hmC



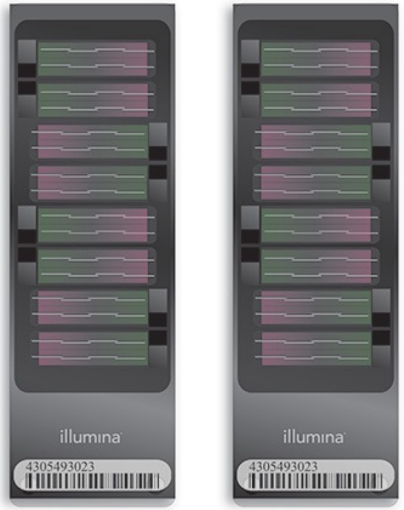
Seq-Cap enrichment (Roche) – 5M CpGs

SureSelect^{XT} Methyl-Seq Target Enrichment Kit (Agilent Technologies) – 3.7M CpGs

Twist Custom Methylation Panel – 3.2M CpGs

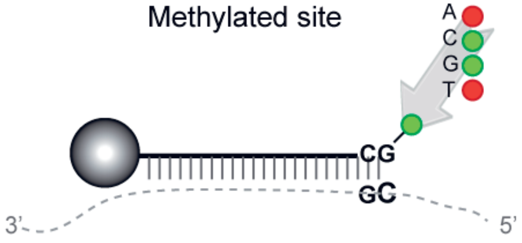
DNA methylation arrays

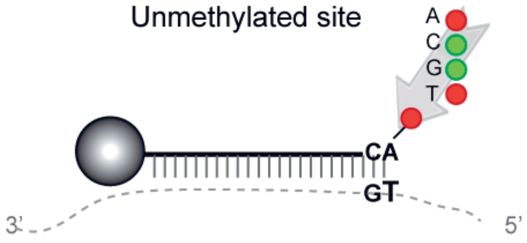
- Bisulfite converted DNA
- >800,000 CpG sites
- 96% CpG islands
- 99% Refseq genes
- CpG sites outside of CpG islands
- Non-CpG methylated sites identified in human stem cells
- Differentially methylated sites found in cancer and several tissue types
- FANTOM 4 promoters
- DNase hypersensitive sites
- miRNA promoters



8 samples per chip

2 chips in parallel





$$\text{Beta value } (\boldsymbol{\beta}) = \frac{M}{M + U + 100}$$

DNA methylation arrays

Pros:

- The most popular method on the market
- Base-pair resolution
- Compatible with FFPE DNA
- Compatible with 5hmC detection
- Many **R packages** available for data analysis and publically available datasets

- Cons
  - Human only* (Mouse Methylation BeadChip available with 285k CpG sites or flexible iSelect Methyl Custom BeadChip, but $$)
  - 850k out of 29M CpG sites

Reduced Representation Bisulfite Sequencing (RRBS)

- ”reduces” the genome to informative regions with high CG content
- Based on restriction digestion with an enzyme that cuts at CCGG (MspI)

Pros:

- Compatible with most species
- Low cost
- Base-pair resolution (bisulfite)
- reads are heavily concentrated to CpG islands
- High throughput

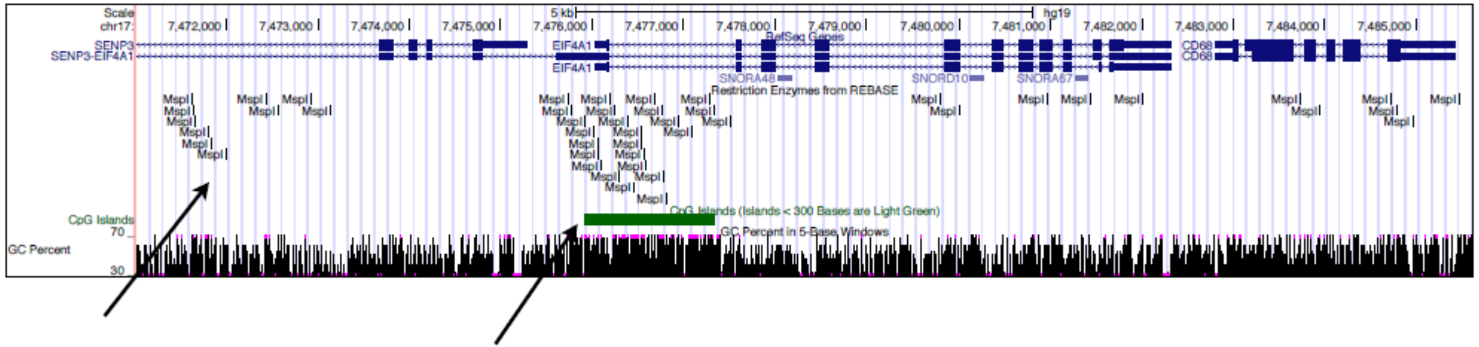
Cons

- Does not capture all promoters or CpG islands
- Results can vary depending on input DNA quality / contaminants in the sample

- MspI (C^CGG)
- TaqI (T^CGA)

Methylation insensitive restriction enzymes

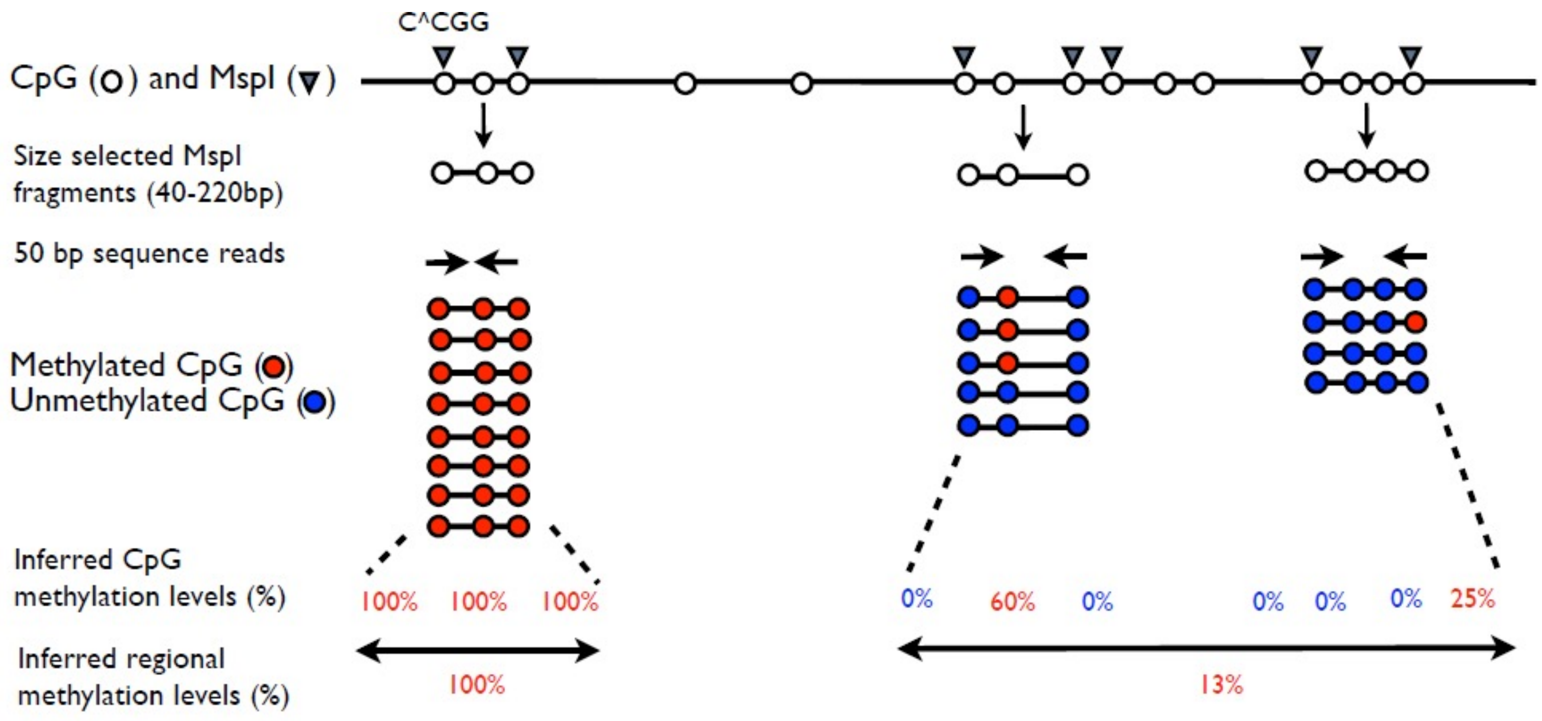
- Base-pair resolution

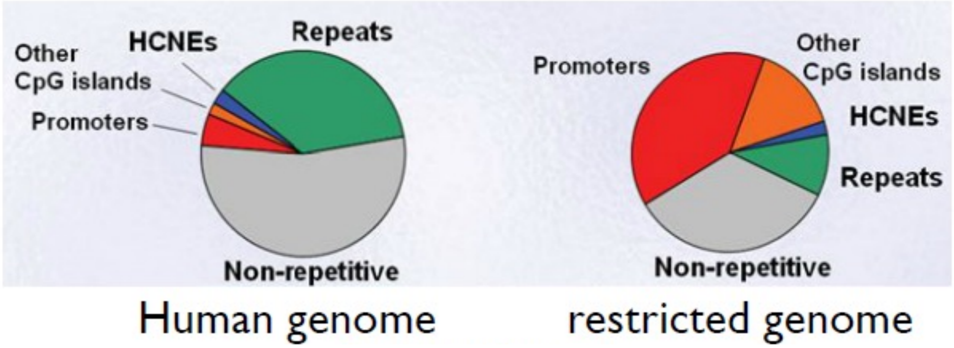


MspI restriction cuts

CpG island

RRBS





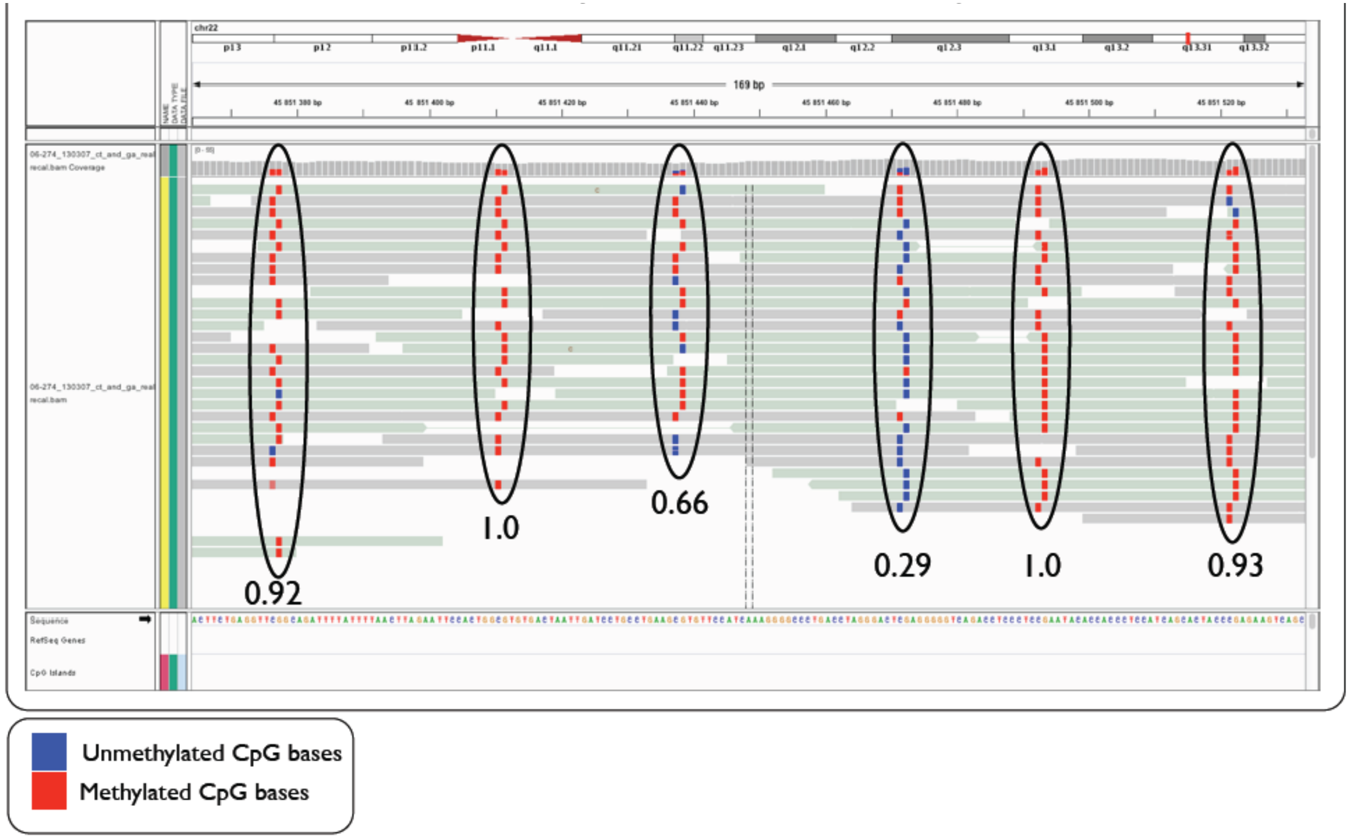
Genomic structures enrichment after digestion and size selection

HCNE:
highly conserved non-coding element

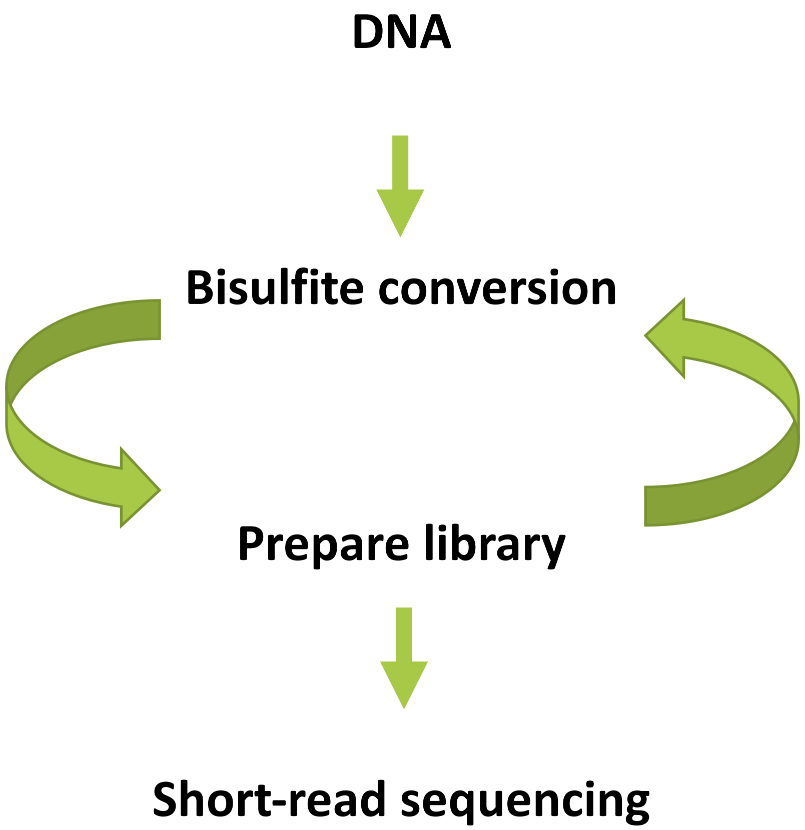
Figure adapted from Meissner et al. *Nature* **454** (2008) https://doi.org/10.1038/nature07107

Whole Genome Methylome Sequencing

- Many acronyms ; WGBS, MethylC-seq, BS-seq
- ”Unbiased” – no selection or enrichment
- Genome-wide coverage of all cytosines
- Base-pair resolution
- Uses bisulfite conversion or enzymatic conversion to distinguish methylated from unmethylated cytosines



Many different approaches …

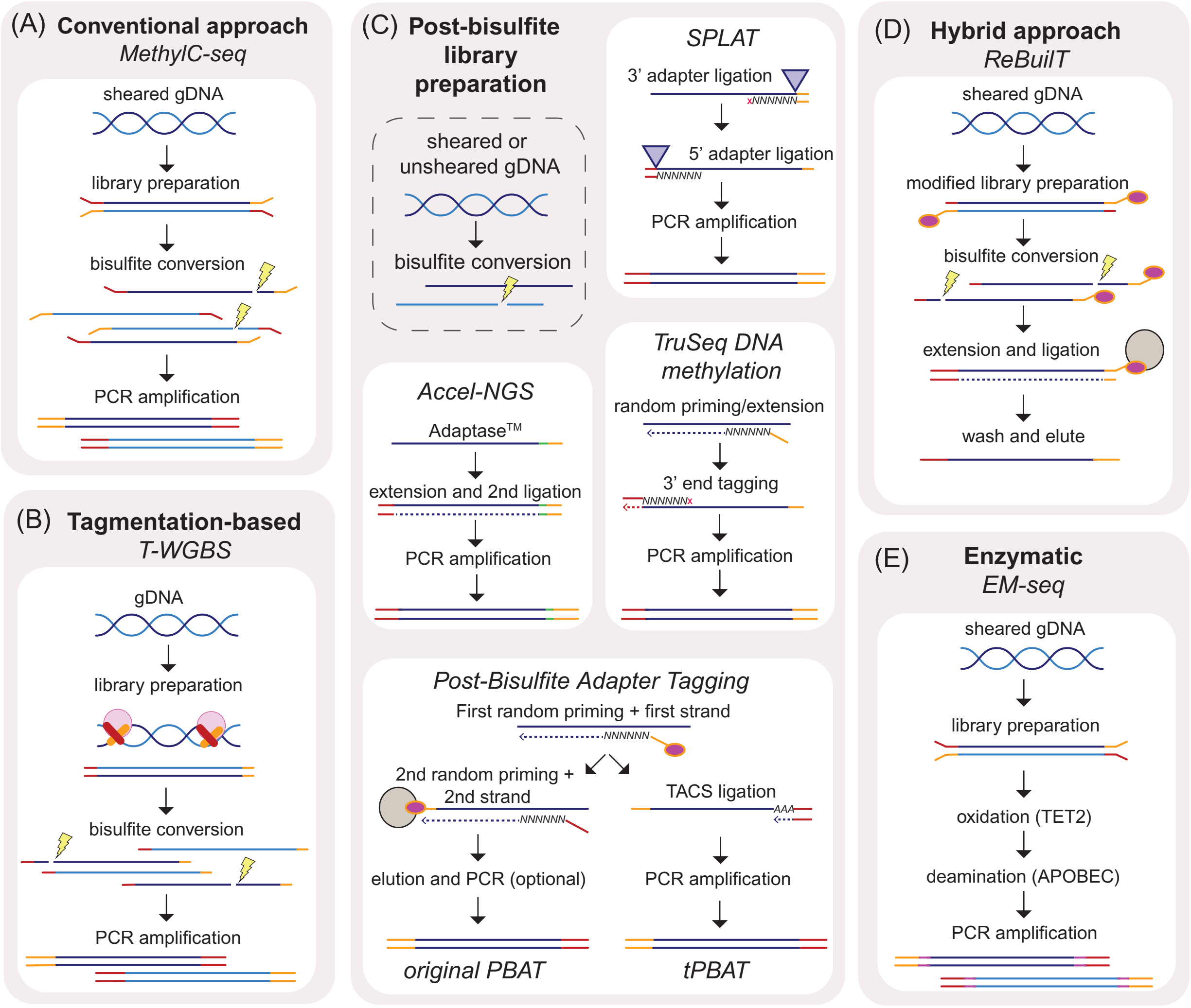


DNA

Bisulfite conversion

Prepare library

Short-read sequencing



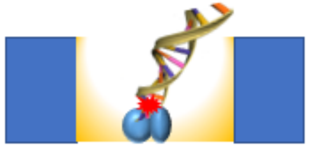
J Nordlund, Chapter Eleven - Advances in whole genome methylomic sequencing, Epigenetics Methods, Academic Press (2020), https://doi.org/10.1016/B978-0-12-819414-0.00011-2.

Whole Genome Methylome Sequencing

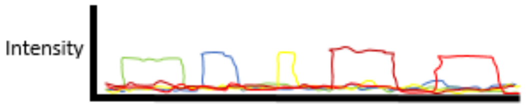
Direct read out of DNA modifications by single molecule, long read technologies (PacBio, Oxford Nanopore)

PacBio SMRT seq

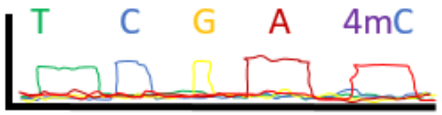
DNA passes thru polymerase in an illuminated volume



Raw output is fluorescent signal of the nucleotide incorporation, specific to each nucleotide

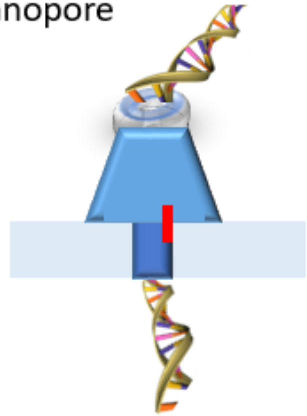


A,C,T,G have known pulse durations, which are used to infer methylated nucleotides

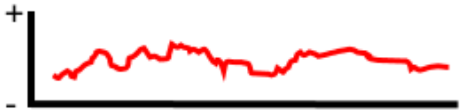


Oxford Nanopore

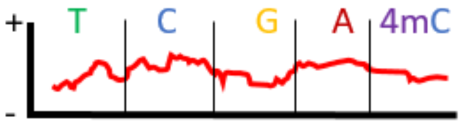
DNA passes thru nanopore



Raw output is electrical signal caused by nucleotide blocking ion flow in nanopore



Each nucleotide has a specific electric “signature”



In theory can detect all sorts of DNA modification-Challenge is to train models to correctly detect specific modifications

Cons; need a lot of native DNA for sequencing + compute

Pros: Phased information! Allele specific methylation. Imprinting

Short vs long-read sequencing, what's the difference?

Short-read

Illumina

Pros:

- Low cost
- High throughput
- Detect 5mC & 5hmC *depending on library prep applied
- Species agnostic

Cons:

- Requires conversion of (un)modificed bases DNA with chemicals or enzymes
- 5mC cannot be distingushed from 5hmC (and other types of marks) without specific conversion approaches

Long-read

PacBio/ONT

Pros:

- Base modification can be read directly from sequencing
- Maintain phasing information
- Detect 4mC, 5mC, 5hmC, 5fC, 5caC, 6mA, etc
- Species agnostic

Cons:

- Cost (high coverage needed) – limiting for large genomes
- Difficult to detect signals
- Low throughput

Reproducibility & quality

Foox J, Nordlund J, et al. The SEQC2 epigenomics quality control (EpiQC) study. Genome Biol 2021: https://doi.org/10.1186/s13059-021-02529-2

EPIC arrays

- duplicate/triplicate at 3 labs

WGBS

- TruSeq DNA methylation (Illumina)
- Accel-NGS methyseq (Swift)
- SPLAT (Raine et al, NAR 2017)

OXBS

- TrueMethyl oxBS-seq (NuGEN)

Enzymatic deamination

- EM-seq (NEB)

ONT: direct methylation calling

7 cell lines

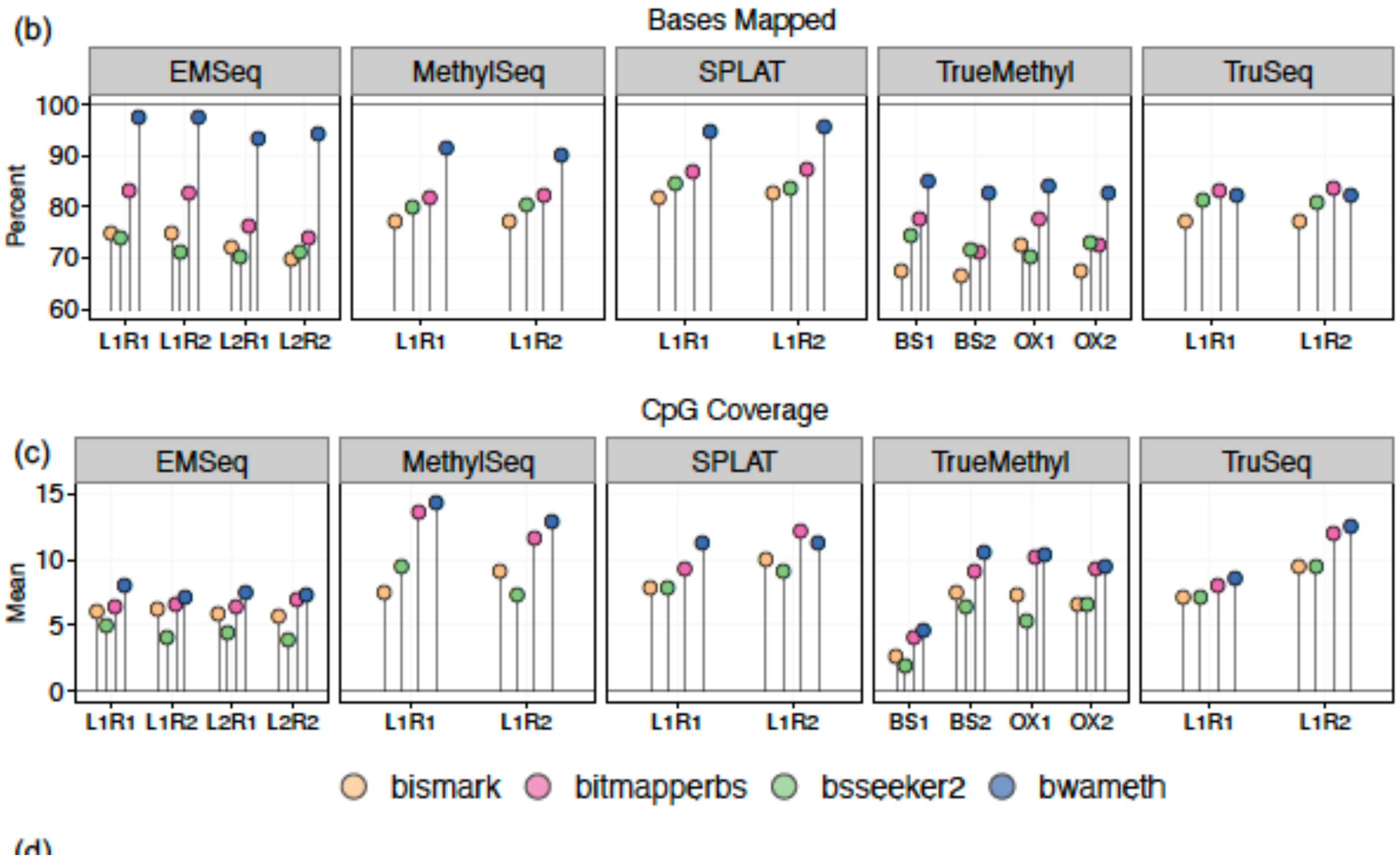
Alignment and methylation calling:

- BISMARK
- BitMapperBS
- BSSeeker2
- Bwa-meth
- Gem-bs

Microarray normalization

- 26 between-array and within-array normalization methods

Reproducibility & quality

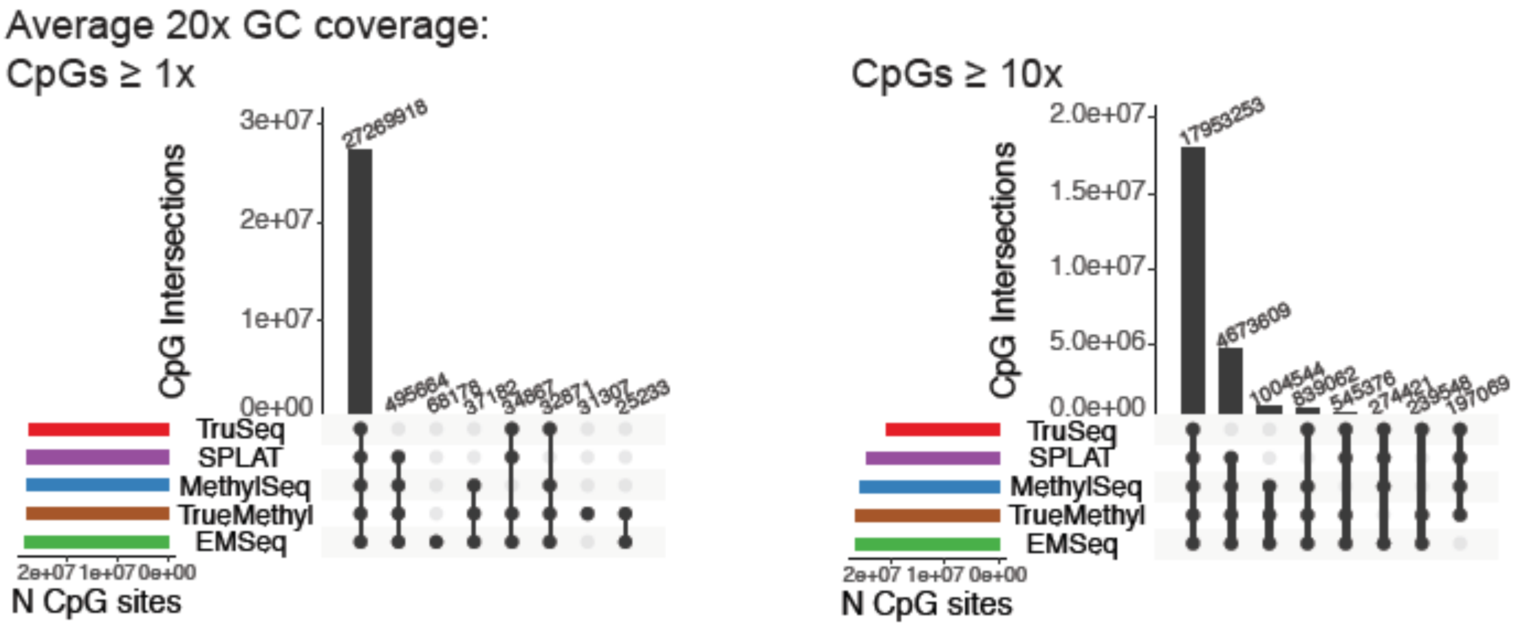


Overall, no major quantitative difference between pipelines but bwa-meth was easiest to implement and retained most data.

Noticeable inter- and intra-library differences

Foox J, Nordlund J, et al. The SEQC2 epigenomics quality control (EpiQC) study. Genome Biol 2021: https://doi.org/10.1186/s13059-021-02529-2

Reproducibility & quality



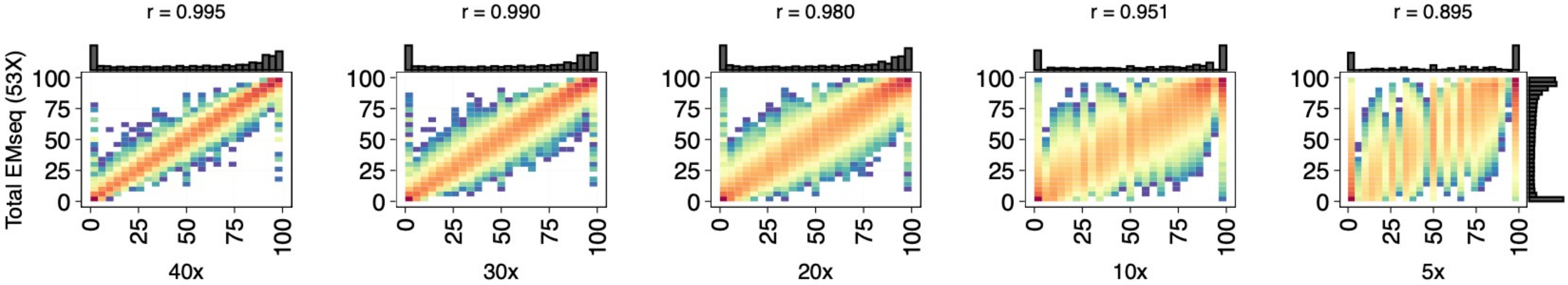
Overall, no major quantitative difference between methylation (beta-values) called after libraries were normalized for nr reads mapped (see next slide).

But they did differ in number of CpG sites detected!

Foox J, Nordlund J, et al. The SEQC2 epigenomics quality control (EpiQC) study. Genome Biol 2021: https://doi.org/10.1186/s13059-021-02529-2

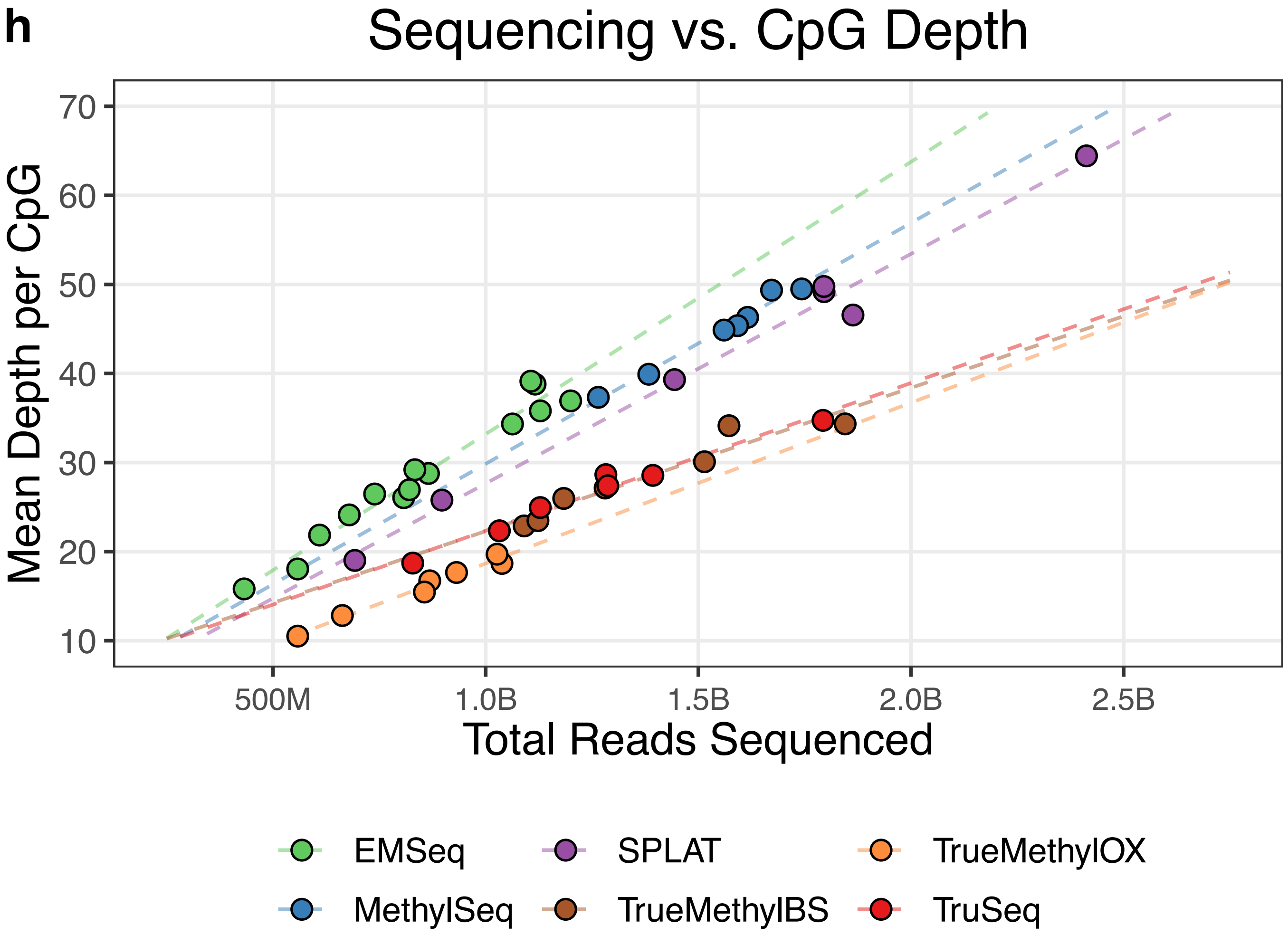
Reproducibility & quality

Correlation in DNA methylation estimation decreases as coverage decreases



Foox J, Nordlund J, et al. The SEQC2 epigenomics quality control (EpiQC) study. Genome Biol 2021: https://doi.org/10.1186/s13059-021-02529-2

Reproducibility & quality

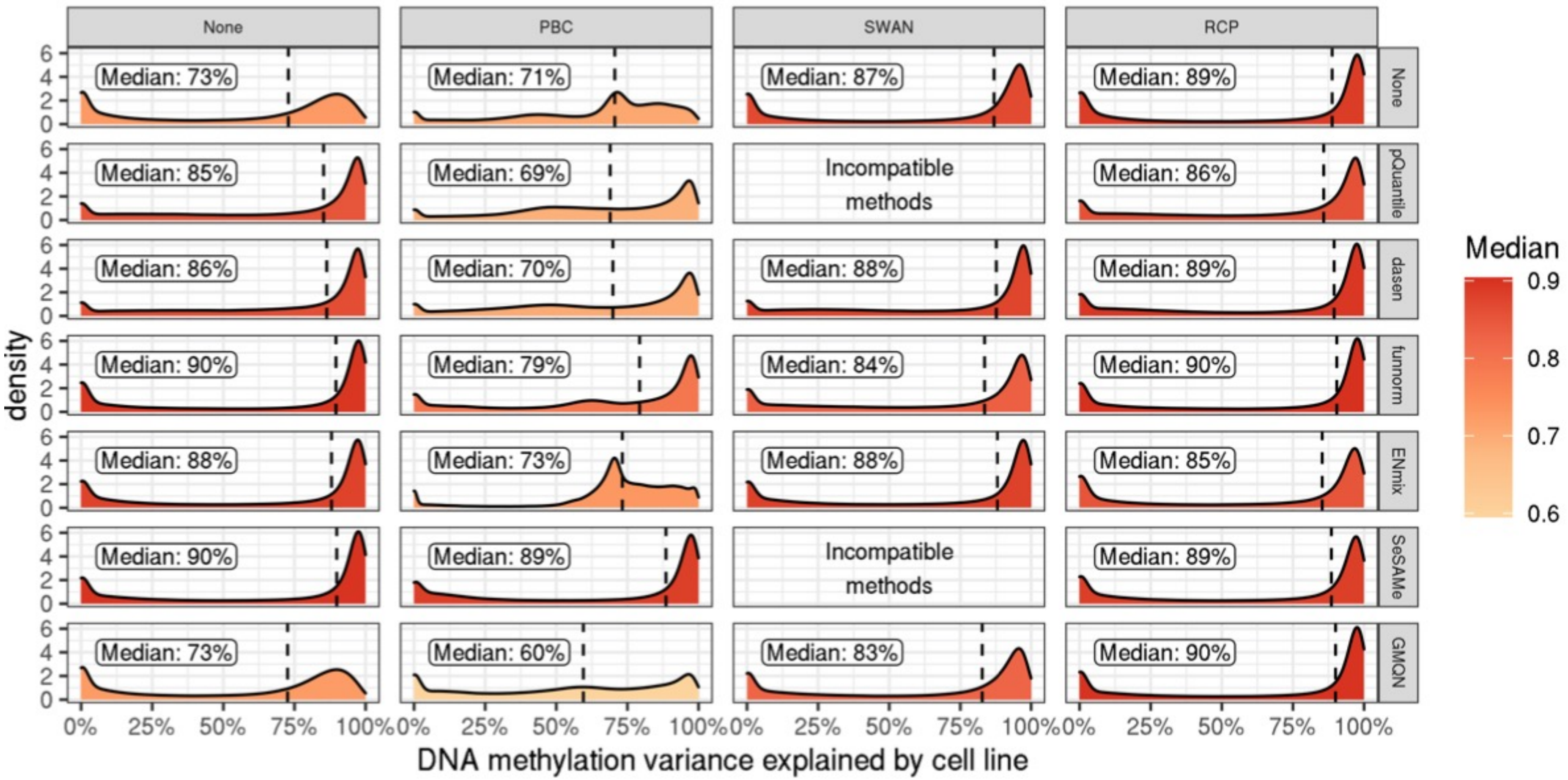


Foox J, Nordlund J, et al. The SEQC2 epigenomics quality control (EpiQC) study. Genome Biol 2021: https://doi.org/10.1186/s13059-021-02529-2

Reproducibility & quality

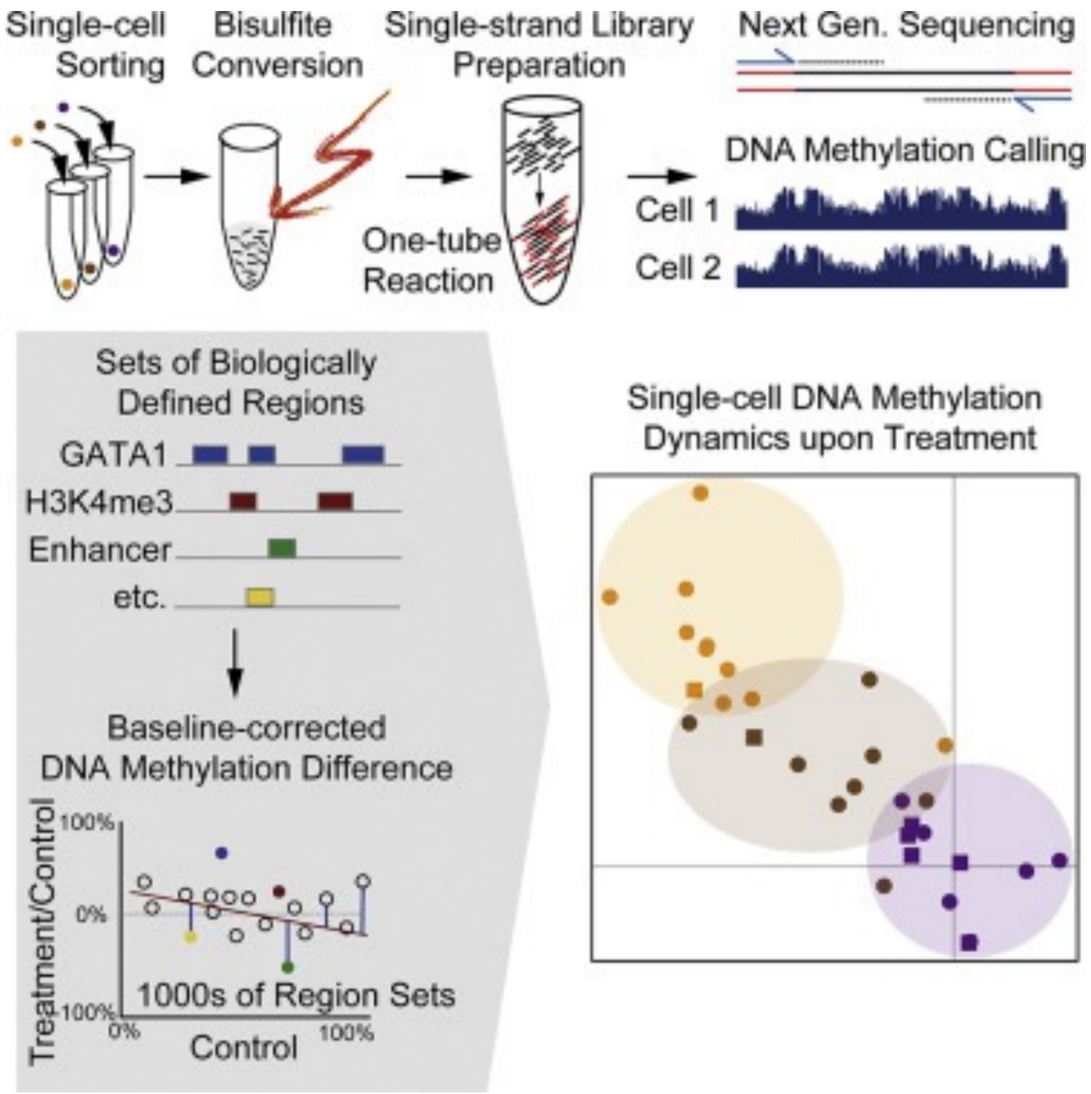
funnorm + RCP worked best on these samples

(a) Concordance between microarray replicates across the epigenome, by normalization pipeline



Foox J, Nordlund J, et al. The SEQC2 epigenomics quality control (EpiQC) study. Genome Biol 2021:
https://doi.org/10.1186/s13059-021-02529-2

Single-cell WGBS



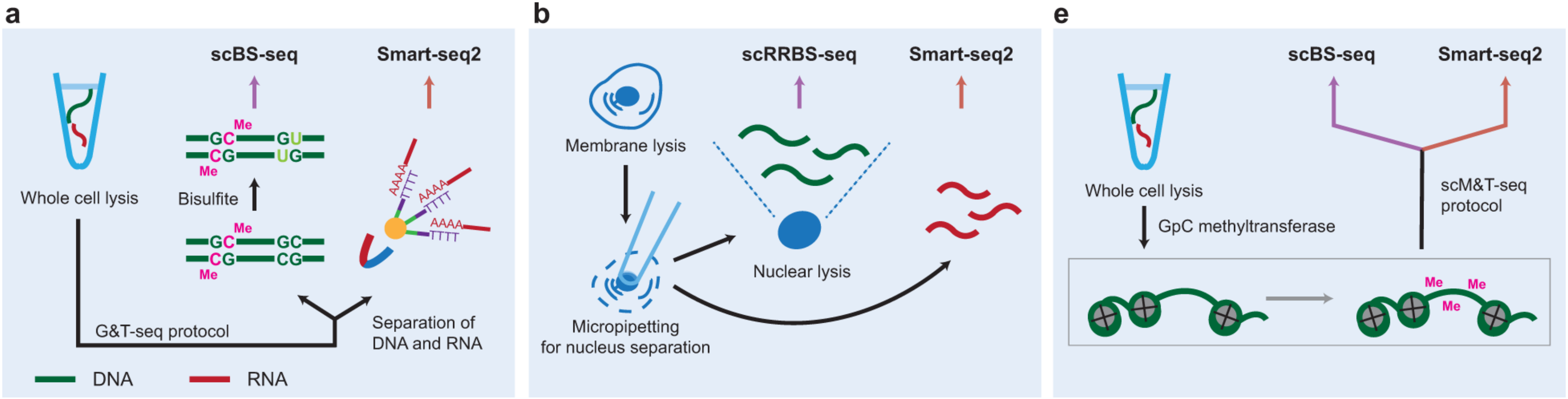
Single cell WGBS

- ✓ Single stranded library prep
- ✓ FACS sorting required (384 plates)
- ✓ Plate- based low throughput (although autmation enable troughput of >1000 cells/exp)
- ✓ Expensive
- ✓ Sparse information-At most 50% CpG sites coverage, usually a lot less

Slide courtesy of Amanda Raine

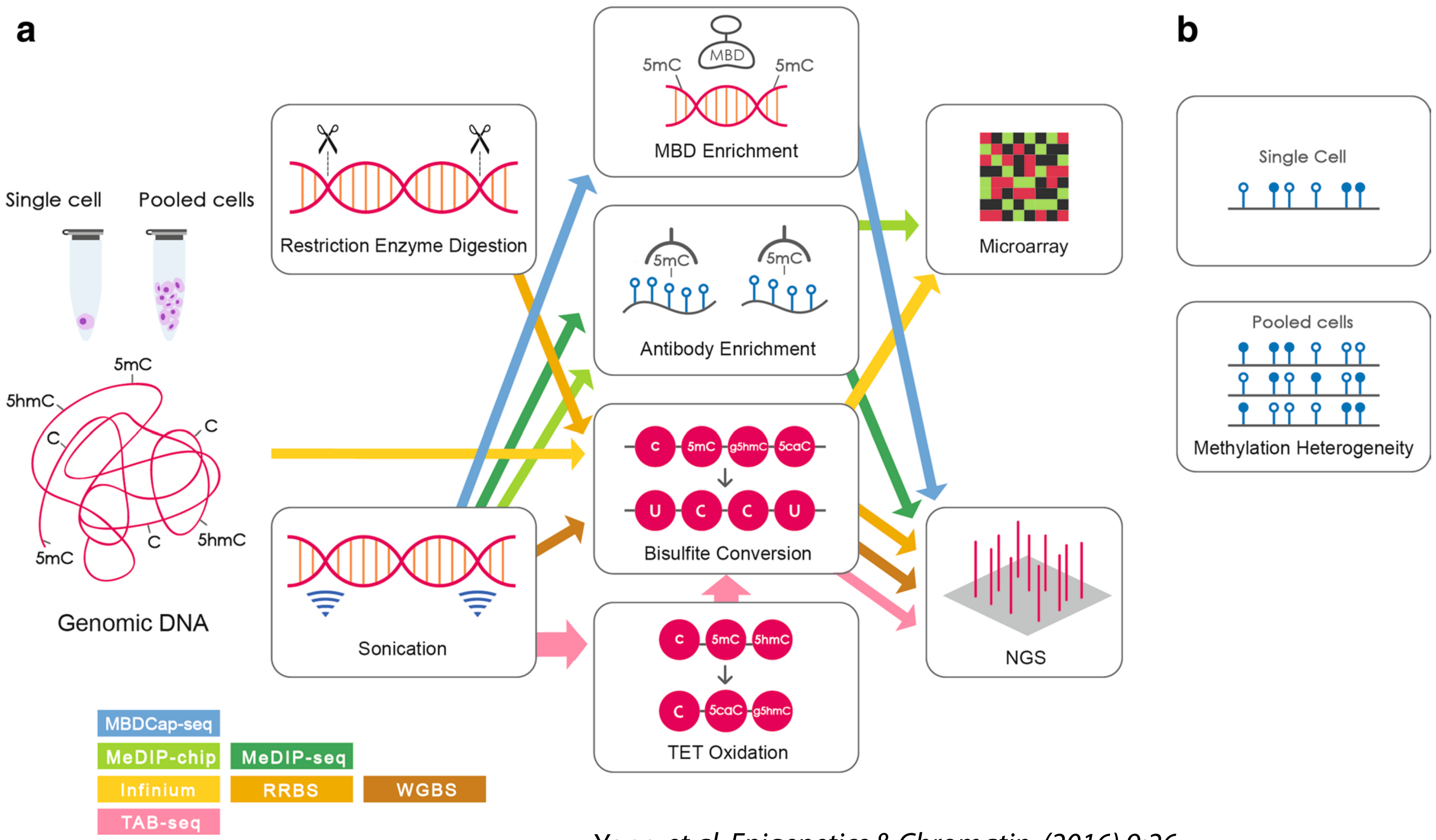
From ”bulk” to single cells

Numerious protocols exist for scWGSB, RRBS, etc – and even integrate transcriptomics in and DNA methylation measurements from the same cell!

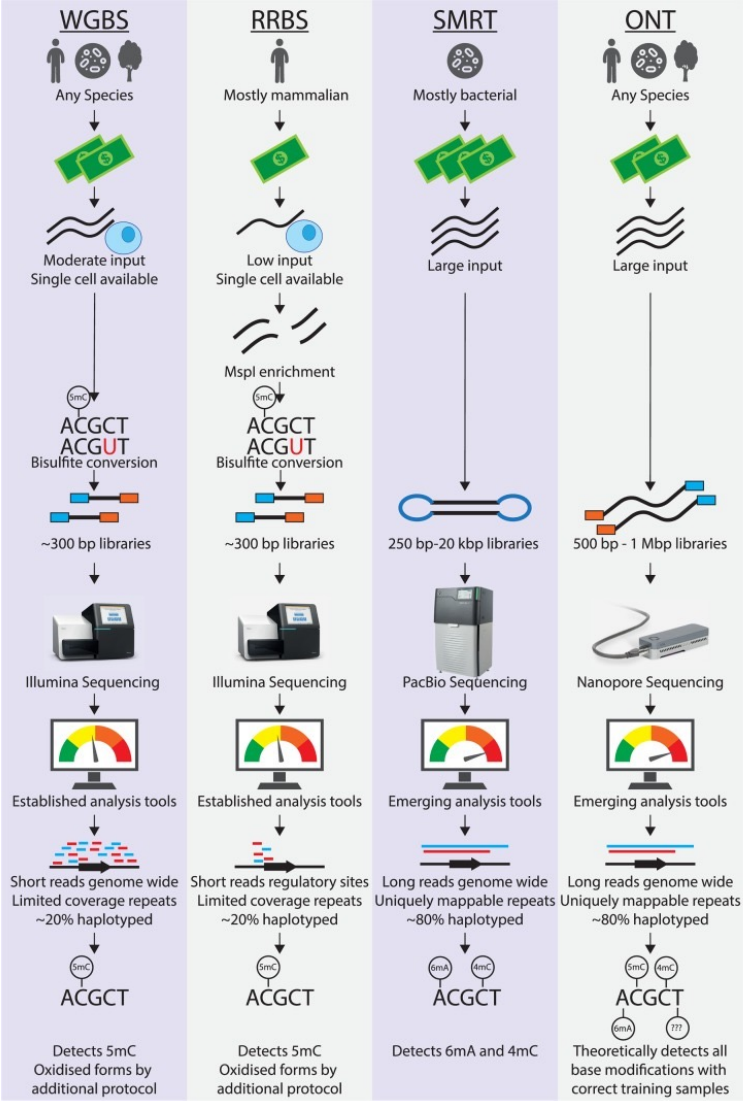


Lee, J. et al. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med* **52,** 1428–1442 (2020). https://doi.org/10.1038/s12276-020-0420-2

In summary, there are many approaches for studying DNA methylation



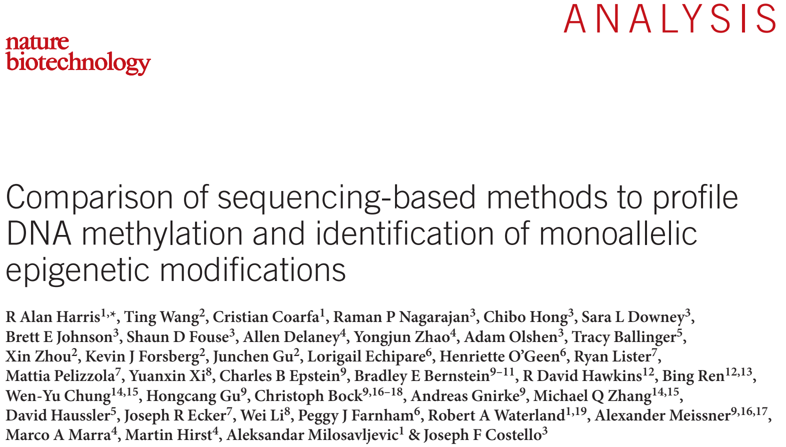
Yong *et al. Epigenetics & Chromatin (2016) 9:26*
DOI 10.1186/s13072-016-0075-3



WGBS	RRBS	SMRT	ONT
Any Species	Mostly mammalian	Mostly bacterial	Any Species
Moderate input Single cell available	Low input Single cell available	Large input	Large input
	Mspl enrichment		
5mC ACGCT ACGUT Bisulfite conversion	5mC ACGCT ACGUT Bisulfite conversion		
~300 bp libraries	~300 bp libraries	250 bp-20 kbp libraries	500 bp - 1 Mbp libraries
Illumina Sequencing	Illumina Sequencing	PacBio Sequencing	Nanopore Sequencing
Established analysis tools	Established analysis tools	Emerging analysis tools	Emerging analysis tools
Short reads genome wide Limited coverage repeats ~20% haplotyped	Short reads regulatory sites Limited coverage repeats ~20% haplotyped	Long reads genome wide Uniquely mappable repeats ~80% haplotyped	Long reads genome wide Uniquely mappable repeats ~80% haplotyped
5mC ACGCT	5mC ACGCT	6mA 4mC ACGCT	5mC 4mC ACGCT 6mA ???
Detects 5mC Oxidised forms by additional protocol	Detects 5mC Oxidised forms by additional protocol	Detects 6mA and 4mC	Theoretically detects all base modifications with correct training samples

Gouil and Keniry, Essays in Biochemistry, 2019. *https://doi.org/10.1042/EBC20190027*

So which method should I choose?



ANALYSIS

nature biotechnology

Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications

R Alan Harris[1,*], Ting Wang[2], Cristian Coarfa[1], Raman P Nagarajan[3], Chibo Hong[3], Sara L Downey[3], Brett E Johnson[3], Shaun D Fouse[3], Allen Delaney[4], Yongjun Zhao[4], Adam Olshen[3], Tracy Ballinger[5], Xin Zhou[2], Kevin J Forsberg[2], Junchen Gu[2], Lorigail Echipare[6], Henriette O'Geen[6], Ryan Lister[7], Mattia Pelizzola[7], Yuanxin Xi[8], Charles B Epstein[9], Bradley E Bernstein[9–11], R David Hawkins[12], Bing Ren[12,13], Wen-Yu Chung[14,15], Hongcang Gu[9], Christoph Bock[9,16–18], Andreas Gnirke[9], Michael Q Zhang[14,15], David Haussler[5], Joseph R Ecker[7], Wei Li[8], Peggy J Farnham[6], Robert A Waterland[1,19], Alexander Meissner[9,16,17], Marco A Marra[4], Martin Hirst[4], Aleksandar Milosavljevic[1] & Joseph F Costello[3]

Foox *et al. Genome Biology* (2021) 22:332
https://doi.org/10.1186/s13059-021-02529-2

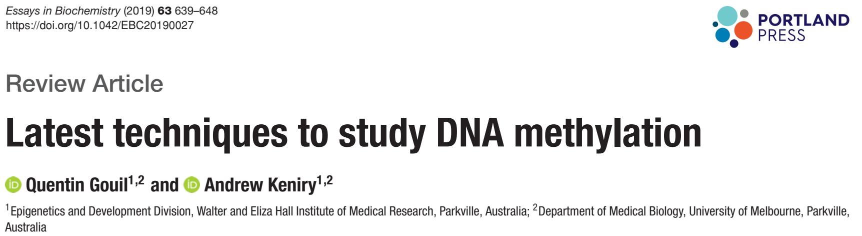
Genome Biology

RESEARCH Open Access

The SEQC2 epigenomics quality control (EpiQC) study

Check for updates

Jonathan Foox[1,2†], Jessica Nordlund[3,4†], Claudia Lalancette[5†], Ting Gong[6†], Michelle Lacey[7†], Samantha Lent[8†], Bradley W. Langhorst[9], V. K. Chaithanya Ponnaluri[9], Louise Williams[9], Karthik Ramaswamy Padmanabhan[5], Raymond Cavalcante[5], Anders Lundmark[3,4], Daniel Butler[1], Christopher Mozsary[1], Justin Gurvitch[1], John M. Greally[10], Masako Suzuki[10], Mark Menor[6], Masaki Nasu[6], Alicia Alonso[1,1], Caroline Sheridan[1,11], Andreas Scherer[4,12], Stephen Bruinsma[13], Gosia Golda[14], Agata Muszynska[15], Paweł P. Łabaj[15], Matthew A. Campbell[9], Frank Wos[16], Amanda Raine[3,4], Ulrika Liljedahl[3,4], Tomas Axelsson[3,4], Charles Wang[17], Zhong Chen[17], Zhaowei Yang[17,18], Jing Li[17,18], Xiaopeng Yang[19], Hongwei Wang[20], Ari Melnick[1], Shang Guo[21], Alexander Blume[22], Vedran Franke[22], Inmaculada Ibanez de Caceres[4,23], Carlos Rodriguez-Antolin[4,23], Rocio Rosas[4,23], Justin Wade Davis[8], Jennifer Ishii[16], Dalila B. Megherbi[24], Wenming Xiao[25], Will Liao[16], Joshua Xu[26], Huixiao Hong[26], Baitang Ning[26], Weida Tong[26], Altuna Akalin[22], Yunliang Wang[21*], Youping Deng[6*] and Christopher E. Mason[1,2,27,28*]



Essays in Biochemistry (2019) **63** 639–648
https://doi.org/10.1042/EBC20190027

PORTLAND PRESS

Review Article

Latest techniques to study DNA methylation

Quentin Gouil[1,2] and Andrew Keniry[1,2]

[1]Epigenetics and Development Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Australia; [2]Department of Medical Biology, University of Melbourne, Parkville, Australia

- Species
- Sample availability
- DNA quality
- Scientific question(s)
- Budget

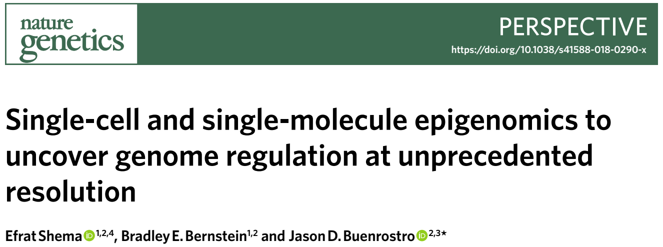
Lee et al. *Experimental & Molecular Medicine* (2020) 52:1428–1442
https://doi.org/10.1038/s12276-020-0420-2

Experimental & Molecular Medicine

REVIEW ARTICLE Open Access

Single-cell multiomics: technologies and data analysis methods

Jeongwoo Lee[1], Do Young Hyeon[1] and Daehee Hwang[1]



nature genetics

PERSPECTIVE
https://doi.org/10.1038/s41588-018-0290-x

Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution

Efrat Shema[1,2,4], Bradley E. Bernstein[1,2] and Jason D. Buenrostro[2,3*]

Yong *et al. Epigenetics & Chromatin* (2016) 9:26
DOI 10.1186/s13072-016-0075-3

Epigenetics & Chromatin

REVIEW Open Access

CrossMark

Profiling genome-wide DNA methylation

Wai-Shin Yong[1†], Fei-Man Hsu[2†] and Pao-Yang Chen[1*]

Epigenomics services offered by the National Genomics Infrastructure (NGI)

GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

NGI is a facility within the **SciLifeLab Genomics Platform** located at two nodes:

NGI-Uppsala

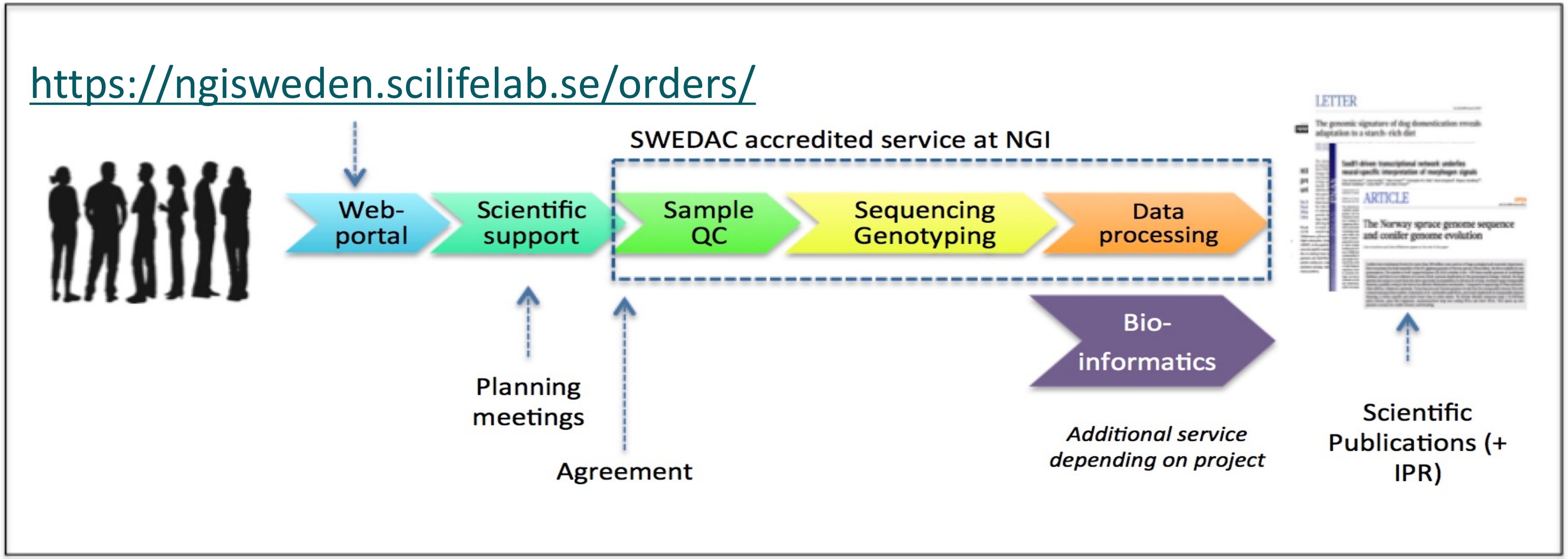
- SNP&SEQ Technology Platform (UU)
- Uppsala Genome Centre (UU)

NGI-Stockholm

- SciLifeLab Solna (KTH, KI, SU)

NGI’s project portal

- All projects submitted through a **common order system**
- Projects are dynamically allocated between Stockholm/Uppsala depending on type of application, queue situation, or request by researcher



https://ngisweden.scilifelab.se/orders/

SWEDAC accredited service at NGI

Web-portal

Scientific support

Sample QC

Sequencing Genotyping

Data processing

Bio-informatics

Planning meetings

Agreement

Additional service depending on project

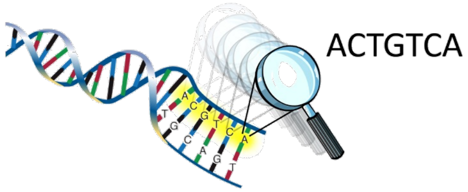
Scientific Publications (+ IPR)

LETTER
The genomic signature of dog domestication reveals adaptation to a starch-rich diet
ARTICLE
The Norway spruce genome sequence and conifer genome evolution

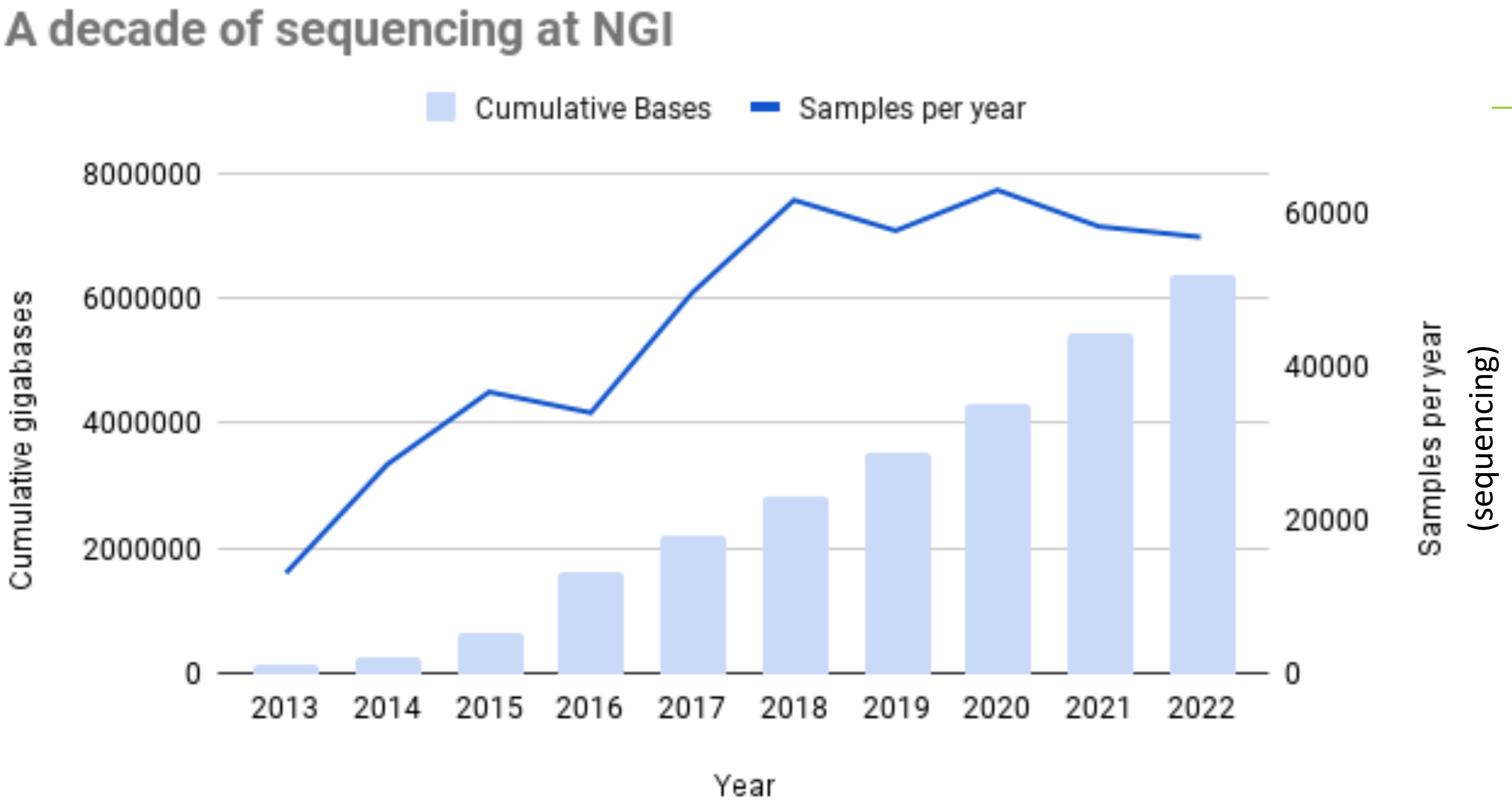
Genotyping and sequencing on all scales

Genotyping

Short-reads



Long-reads



Statistics for 2022:

- 1000 projects / 90,000 samples
- **912 Terabases** (10^{12}) of sequence data

As of Jan 1, 2022 NGI has delivered a total of 6.3 Petabases (10^{15}) of sequencing data

Support

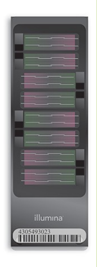
Pre support

- **Project design** via discussions with expert project coordinators
- **Advise** in sample collection and/or preparation
- **DNA extraction services available** for specific applications
- **Sample quality (QC)** for all incoming samples and user-made libraries

Post support

- Control over produced data: making sure data meet our **high standards** in terms of quality and yield.
- Open source Bioinformatic pipelines for a wide range of applications: *NF-core lecture*
- Data delivered via **UPPMAX**

Epigenetic methods available at NGI



EPIC Arrays:

500 ng DNA

Minimum sample size 15 samples: lower cost per sample for large projects

Short-read
Whole genome methylome sequencing with SPLAT (WGBS) or EM-Seq

Twist targeted methylation

~500 ng DNA

Long-read
whole genome sequencing (+base modifications)

PacBio Sequell II / Oxford Nanopore PromethIOn

Cost depends on genome size and epigenetic marks analyzed

Single-cell:

scATAC-seq (10x Genomics)

scWGBS with SPLAT

RRBS:

500 ng DNA
~2000 SEK/sample

**limited availablility*

ATAC-seq

>50.000 cells
~2000 SEK/sample

**limited availablility*

HiC

method for mapping genome-wide DNA contacts

**limited availablility*

Contact information:

Additional information about sequencing applications that NGI supports:

https://ngisweden.scilifelab.se

Don’t hesitate to reach out to NGI’s project coordinators:
support@ngisweden.se

-or me-

jessica.nordlund@medsci.uu.se / seq@medsci.uu.se