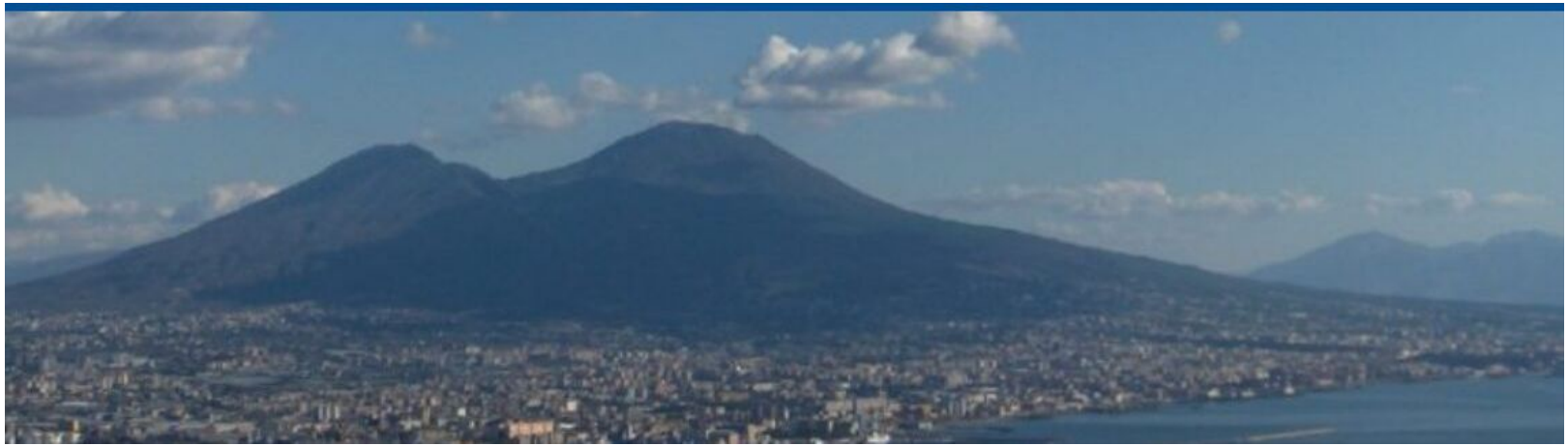
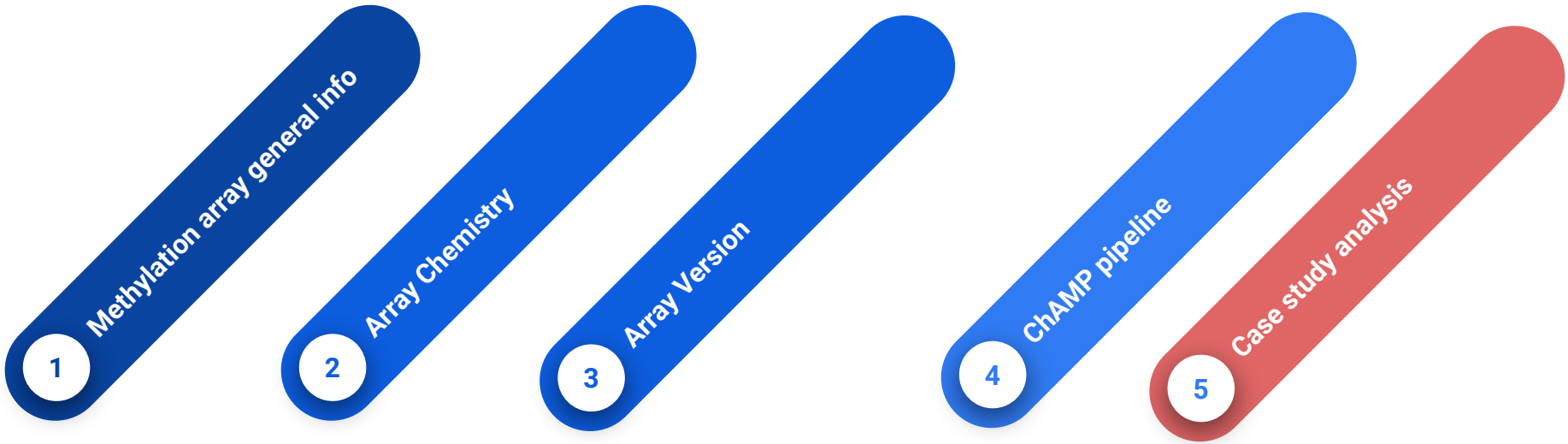
Francesco Cecere



From Raw Matrices to Differential Expression/Methylation Patterns: A Functional Genomics Approach to Detect Molecular Insights

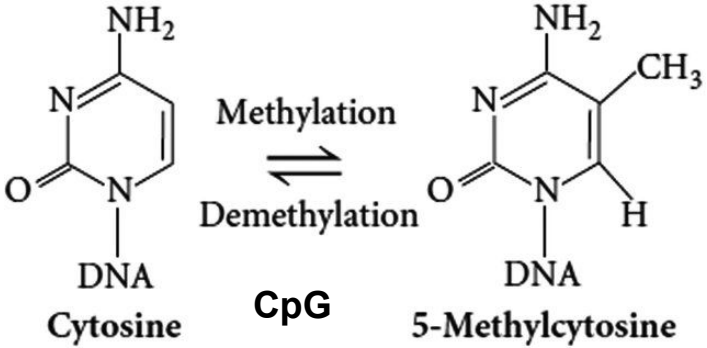
25.11.2024

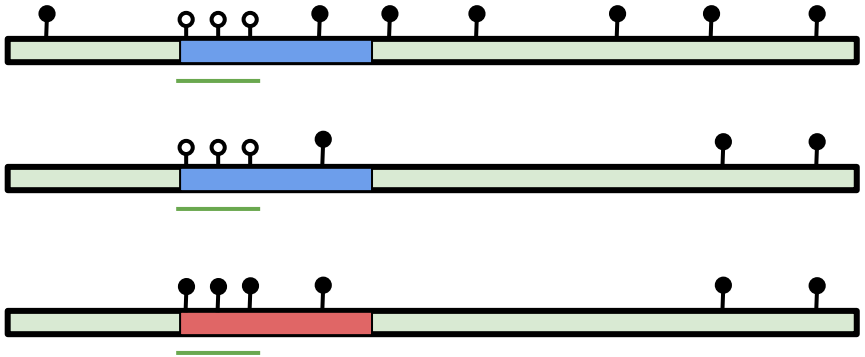
Outline


1
Methylation array general info
2
Array Chemistry
3
Array Version
4
ChAMP pipeline
5
Case study analysis

DNA methylation

DNA methylation is an epigenetic mechanism involving the transfer of a methyl group onto the C5 position of the cytosine to form 5-methylcytosine





CpG Island
Expressed gene
Silenced gene

ENVIRONMENTAL EPIGENETICS

Environmental Epigenetics, 2024, **10(1)**, dvae012

DOI: https://doi.org/10.1093/eep/dvae012
Advance Access Publication 10 August 2024
Review Article

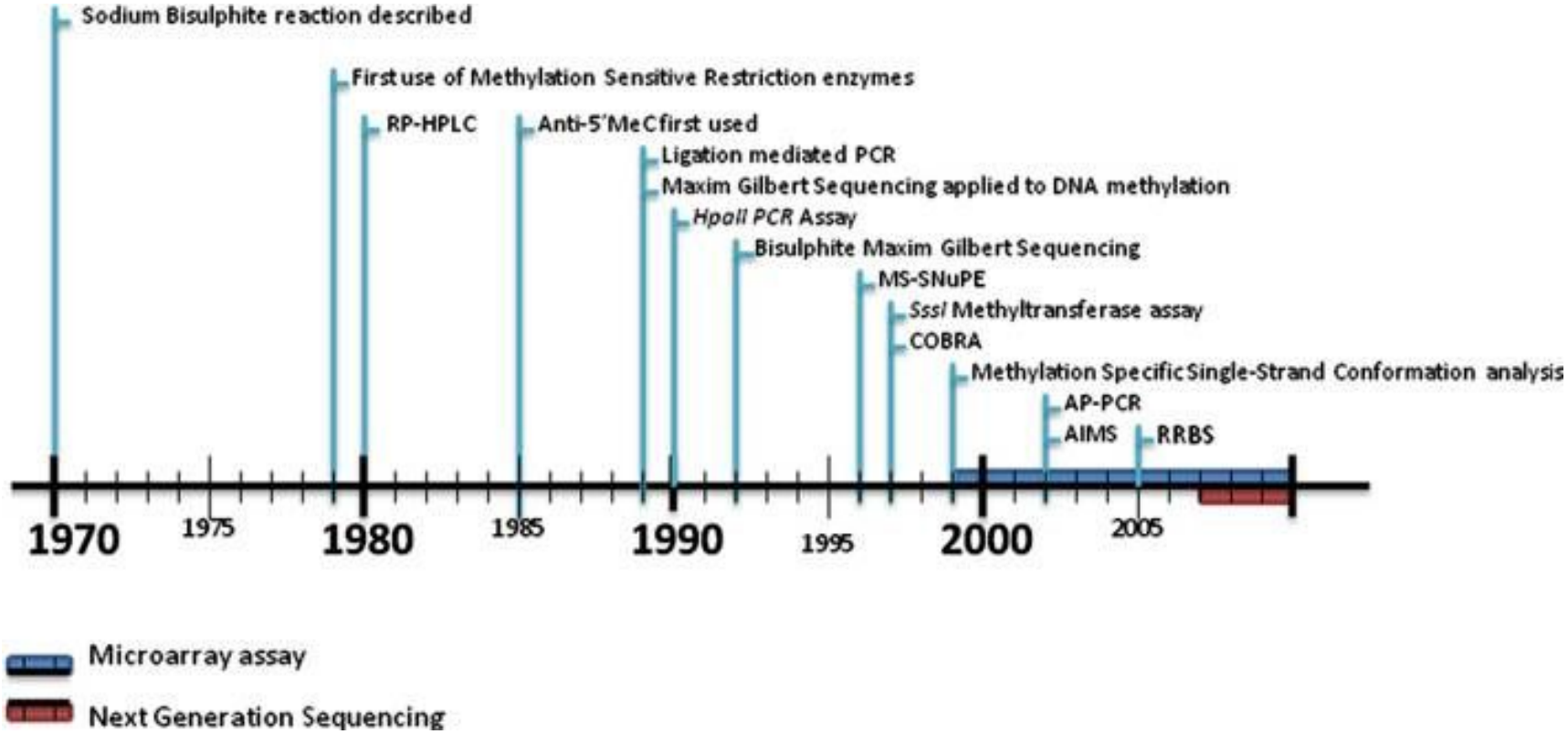
Epigenetics of violence against women: a systematic review of the literature

Paolo Bailo[1], Andrea Piccinini[2,3], Giussy Barbara[3,4,5], Palmina Caruso[2], Valentina Bollati[5,*], Simona Gaudi[6]

[1]Section of Legal Medicine, School of Law, University of Camerino, Camerino 62032, Italy
[2]Department of Biomedical Sciences for Health, Università degli Studi di Milano, Milan 20100, Italy
[3]Service for Sexual and Domestic Violence (SVSeD), Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan 20100, Italy
[4]Gynecology Emergency Unit, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan 20100, Italy
[5]Department of Clinical Sciences and Community Health, Dipartimento di Eccellenza 2023-2027, University of Milan, Milan 20122, Italy
[6]Department of Environment and Health, Italian National Institute of Health, Rome 00161, Italy

*Corresponding author. Department of Clinical Sciences and Community Health, University of Milan, Via S. Barnaba, 8, Milan 20122, Italy. E-mail: valentina.bollati@unimi.it

DNA methylation

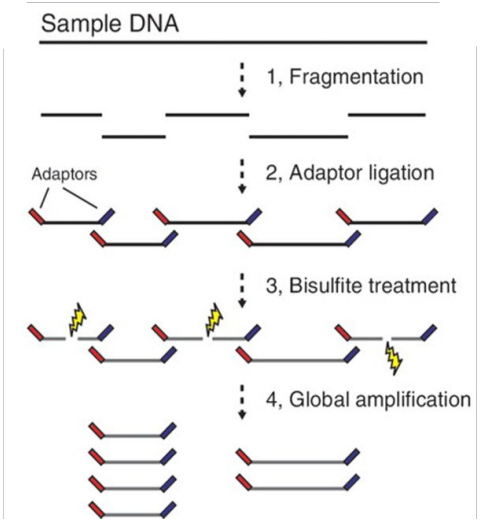


Harrison A, Parle-McDermott A. DNA methylation: a timeline of methods and applications. Front Genet. 2011 Oct 25;2:74. doi: 10.3389/fgene.2011.00074. PMID: 22303369; PMCID: PMC3268627.

1 Methylation array general info

A full understanding of the role of DNA methylation in health and disease requires the development of tools that can simultaneously measure DNA methylation across large portions of the genome.

Whole Genome Bisulphite Sequencing (WGBS)

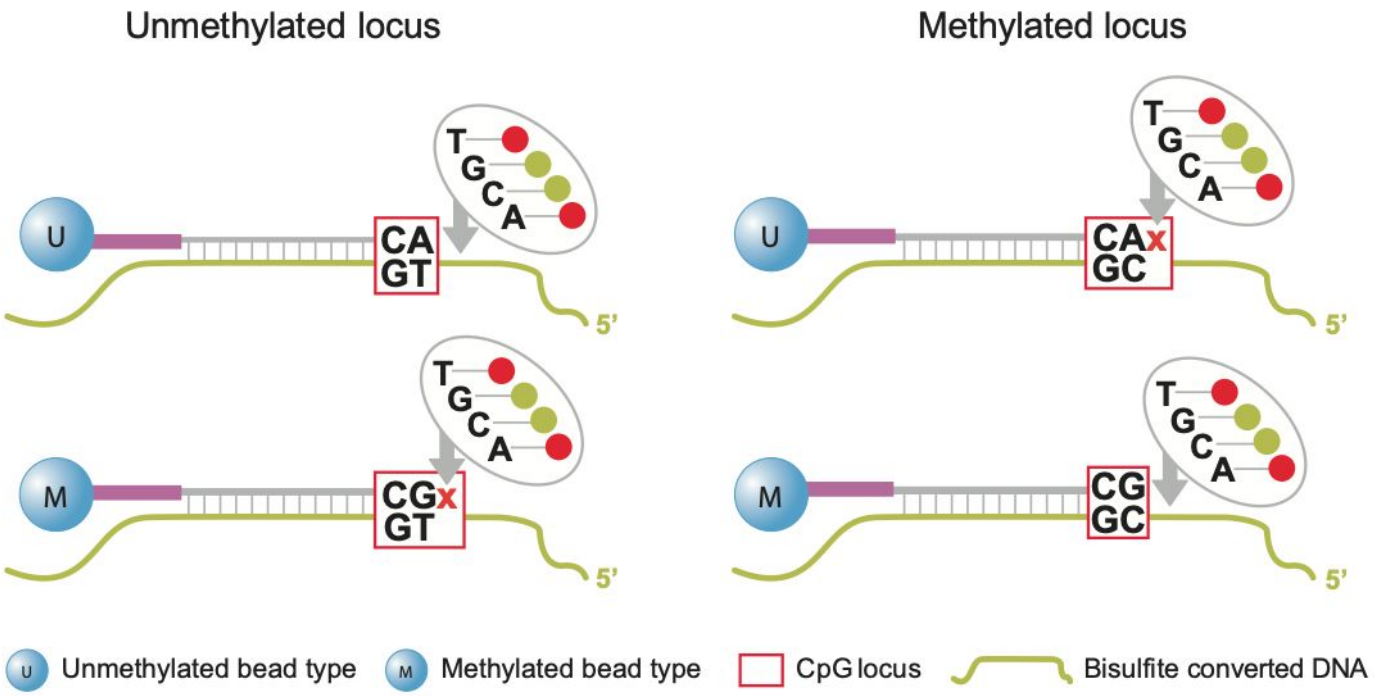


- WGBS has been successfully applied to a range of biological tissues and cell lines.
- Map ~28 million CpG sites
- Due to the high cost and the expertise required, it is not the most feasible method

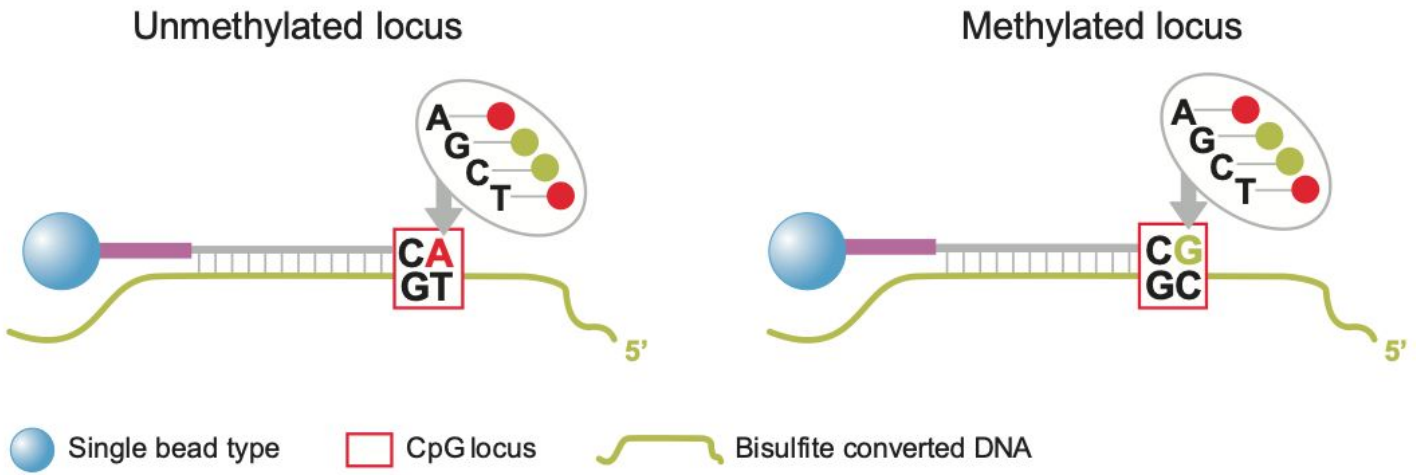
Methylation array

- User-friendly alternative
- Time-efficient
- Cost-effective
- INITIALLY Not properly considerate as genome-wide analysis

Infinium I


Unmethylated locus
Methylated locus
U
M
CA
GT
CAx
GC
CGx
GT
CG
GC
T
G
C
A
5'
Unmethylated bead type
Methylated bead type
CpG locus
Bisulfite converted DNA

Infinium II


Unmethylated locus
A
G
C
T
CA
GT
5'
Methylated locus
A
G
C
T
CG
GC
5'
Single bead type
CpG locus
Bisulfite converted DNA

3 **Array Type**
HumanMethylation27kBeadChip

Infinium technology I
Cover ~27000 CpGs

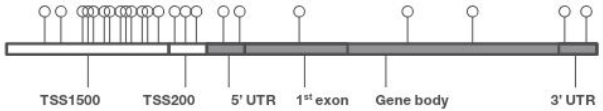
Table 1: Markers of the HumanMethylation27 BeadChip.[3]		
Type of target	**CpG sites present**	**Avg # of CpG sites per target**
RefSeq Genes	14,475	1.9 sites
Well-annotated genes described in the NCBI CCDS database (Genome build 36)	12,833	1.9 sites
Methylation hotspots in cancer genes	144	7.6 sites
Cancer-related targets	982	1.9 sites
miRNA promoters	110	2.3 sites

Important for investigation of the role of DNA methylation in carcinogenesis and identification of cancer biomarker. EWAS study for the association of ageing and smoking with DNA methylation.

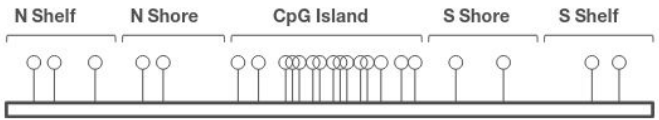
not truly considered “genome-wide” !

3 Array Type HumanMethylation450kBeadChip

Infinium technology I and II
Cover ~450000 CpGs



Feature Type	Genes Mapped	Percent Genes Covered	Number of Loci on Array
NM_TSS200	14895	0.79	2.56
NM_TS1500	17820	0.94	3.41
NM_5'UTR	13865	0.78	3.34
NM_1stExon	15127	0.80	1.62
NM_3'UTR	13042	0.72	1.02
NM_GeneBody	17071	0.97	8.97
NR_TSS200	1967	0.65	1.84
NR_TSS1500	2672	0.88	2.92
NR_GeneBody	2345	0.77	5.34



Feature Type	Islands Mapped	Percent Islands Covered	Average Number of Loci on Array
Island	26153	0.94	5.08
N_Shore	25770	0.93	2.74
S_Shore	25614	0.92	2.66
N_Shelf	23896	0.86	1.97
S_Shelf	23968	0.86	1.94

The HumanMethylation450 BeadChip offers broad coverage across gene regions, as well as CpG islands/CPG island regions, shelves, and shores for the most comprehensive view of methylation state.

- Differentially methylated sites identified in tumor versus normal (multiple forms of cancer) and across several tissue types
- FANTOM 4 promoters
- DNase hypersensitive sites
- miRNA promoter regions
- ~ 90% of content contained on the Illumina HumanMethylation27 BeadChip

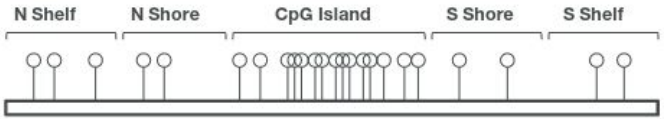
3 Array Type EPICv1 BeadChip

Infinium technology I and II Cover ~850000 CpGs

Feature Type	# Features Mapped	% Features Covered	Avg # Loci/Feature
RefSeq			
NM_TSS200[a]	> 20,000	> 88%	3
NM_TSS1500	> 23,000	> 97%	5
NM_5'UTR	> 20,000	> 85%	7
NM_1stExon	> 20,000	> 85%	2
NM_3'UTR	> 14,000	> 70%	1
NM_ExonBoundaries	> 8000	> 35%	0.5
NR_TSS200	> 4000	> 65%	1
NR_TSS1500	> 5000	> 80%	3
NR_ExonBoundaries	> 500	> 15%	0.2
GenCode Basic v12			
TSS200	> 65,000	> 86%	2
TSS1500	> 80,000	> 95%	5
5'UTR	> 50,000	> 75%	7
First Exon	> 45,000	> 60%	2
3'UTR	> 35,000	> 65%	3
Exon Boundaries	> 8000	> 30%	0.5
Enhancers			
ENCODE Open Chromatin[b] Evidence ≥ 4	> 150,000	> 65%	2
ENCODE TFBS in Open Chromatin[c] Evidence ≥ 3	> 220,000	> 50%	1
ENCODE TFBS in Open Chromatin Evidence ≥ 4	> 150,000	> 75%	3
FANTOM5 Enhancers[d]	> 23,000	> 80%	1

Infinium MethylationEPIC BeadChip Highlights

- **Unique Combination of Coding Region and Enhancer-Wide Coverage, High-Throughput, and Low Cost**
  Over 850,000 methylation sites per sample at single-nucleotide resolution
- **High Assay Reproducibility**
  > 98% reproducibility for technical replicates
- **Simple Workflow**
  PCR-free protocol with the powerful Infinium HD Assay
- **Compatible with FFPE Samples**
  Protocol available for methylation studies on FFPE samples



Feature Type	# Covered	% Covered	Avg # Loci/Feature
Island	26,000	> 95%	6
North Shore	25,000	> 90%	3.5
South Shore	25,000	> 90%	3.5
North Shelf	22,000	> 80%	2
South Shelf	22,000	> 80%	2

3 **Array Type**
EPICv2 BeadChip

Infinium technology I and II
Cover ~935000 CpGs

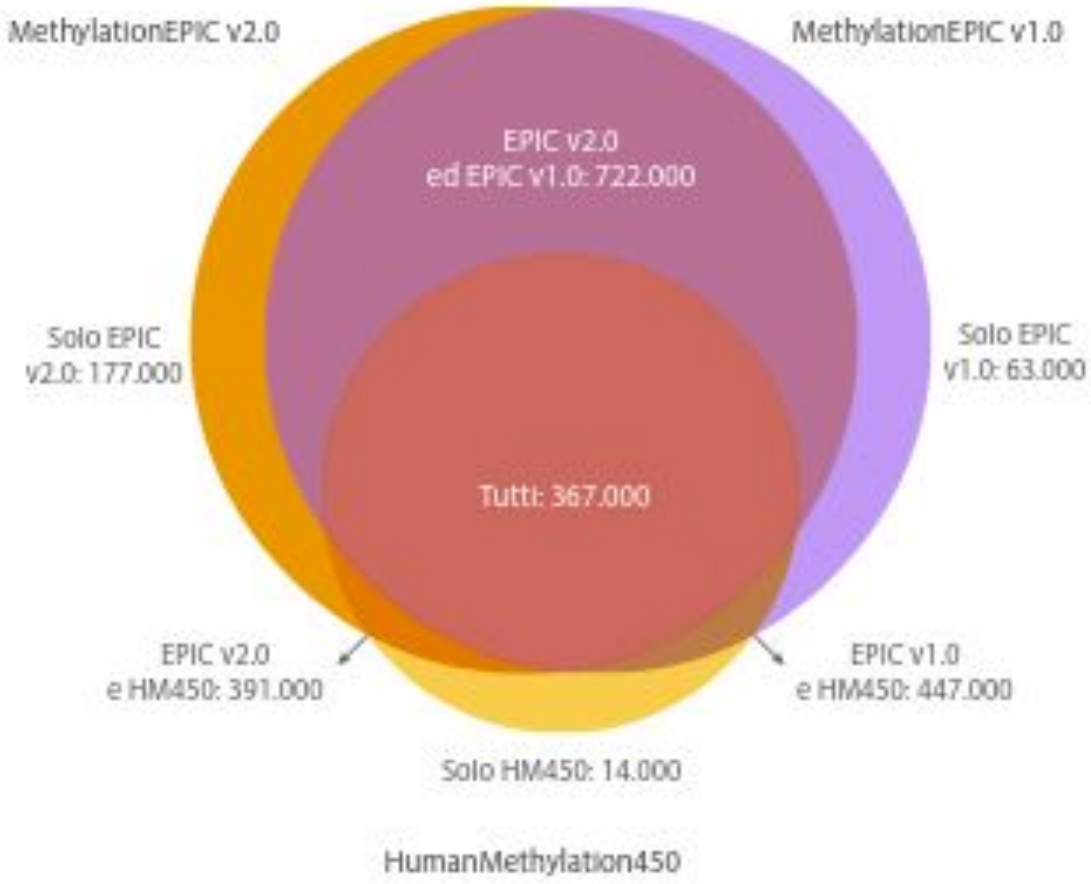
Tabella 2: Copertura densa delle isole CpG

Caratteristica	N. coperto	% coperta	N. medio di loci/ caratteristica
Isola	25.381	91%	5,4
Sponda nord	25.115	90%	3,5
Sponda sud	24.870	89%	3,6
Sito nord	21.719	78%	2,1
Sito sud	21.677	78%	2,1

Tabella 1: Informazioni sul prodotto

Caratteristica	Descrizione
Specie	Umana
N. totale di marcatori[a]	> 935.000
N. di campioni per BeadChip	8
Requisito di input di DNA	250 ng
Tipi di campione specializzati	Tessuto FFPE
Chimica del saggio	Infinium HD
Supporto strumento	iScan System, NextSeq 550 System

a. Siti di metilazione interrogati.


MethylationEPIC v2.0
MethylationEPIC v1.0
EPIC v2.0
ed EPIC v1.0: 722.000
Solo EPIC
v2.0: 177.000
Solo EPIC
v1.0: 63.000
Tutti: 367.000
EPIC v2.0
e HM450: 391.000
EPIC v1.0
e HM450: 447.000
Solo HM450: 14.000
HumanMethylation450

4 ChAMP pipeline

ChAMP

Chip Analysis Methylation Pipeline for Illumina HumanMethylation450 and EPIC

Bioconductor version: Release (3.16)

The package includes quality control metrics, a selection of normalization methods and novel methods to identify differentially methylated regions and to highlight copy number alterations.

Author: Yuan Tian [cre,aut], Tiffany Morris [ctb], Lee Stirling [ctb], Andrew Feber [ctb], Andrew Teschendorff [ctb], Ankur Chakravarthy [ctb]

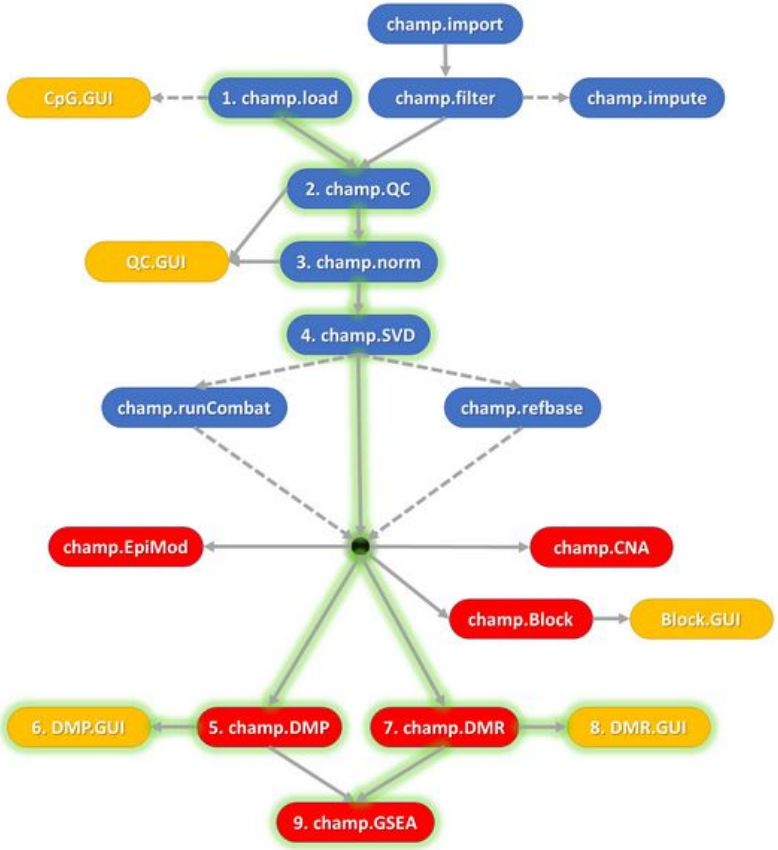
Maintainer: Yuan Tian <champ450k at gmail.com>

Citation (from within R, enter `citation("ChAMP")`):

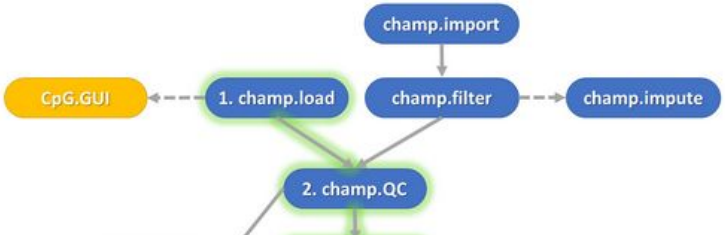
Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Andrew F, Teschendorff AE (2017). "ChAMP: updated methylation analysis pipeline for Illumina BeadChips." *Bioinformatics*, btx513. doi: 10.1093/bioinformatics/btx513.

Morris TJ, Butcher LM, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S (2014). "ChAMP: 450k Chip Analysis Methylation Pipeline." *Bioinformatics*, **30**(3), 428-430. doi: 10.1093/bioinformatics/btt684.

Butcher LM, Beck S (2015). "Probe Lasso: A novel method to rope in differentially methylated regions with 450K DNA methylation data." *Methods*, **72**, 21-28. doi: 10.1016%2Fj.ymeth.2014.10.036.



4 ChAMP pipeline - Load Data



Loading data is always the first step. ChAMP provides a loading function to get data from .idat files (with pd file (Sample_Sheet.csv) in it.

Illumina idat files: The Illumina Intensity Data (idat) file format is used to summarize the data from the Illumina BeadArray platforms including the 450K and EPIC methylation arrays. For each sample, a "Red" and a "Green" idat file are available representing the intensities.

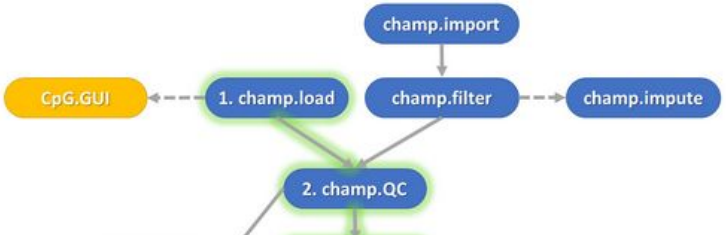
$\beta = M / (M + U + \alpha)$

M = methylated intensity
U = unmethylated intensity
α = *constant offset default as 100*

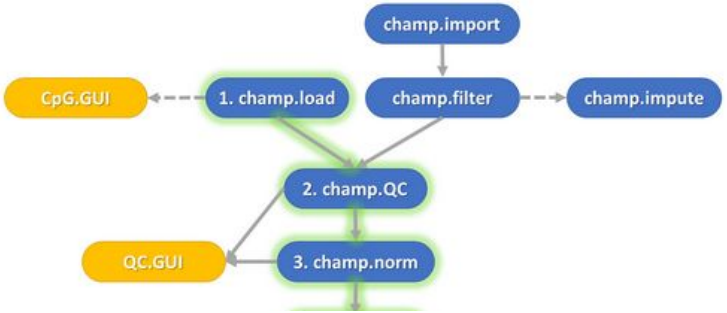
	S1	S2
cg07881041	0.946525423	0.9082143763
cg03513874	0.954758496	0.8649431033
cg05451842	0.035386102	0.0214317673
cg14797042	0.961342732	0.7921114094
cg09838562	0.005279178	0.0095512204
cg25458538	0.625504222	0.5601842907
cg09261072	0.667497712	0.5738260821
cg02404579	0.936233239	0.8968762818
cg04118974	0.904251639	0.8500860028
cg01236347	0.724598193	0.5695492801
cg22585117	0.868421838	0.4644636067
cg25552317	0.942922436	0.9302927783
cg23875663	0.681850383	0.3108122340
cg07659892	0.185790859	0.0990251867
cg15995909	0.977874451	0.9722461508
cg23728960	0.355524478	0.3445295273
cg11993619	0.765112776	0.6192689007
cg01925883	0.864812931	0.4767928379
cg03452160	0.653043420	0.3292471557
cg09430819	0.442579962	0.4100515444
cg13871826	0.787152839	0.8245848827

4 ChAMP pipeline - Filter Data

850K probes ----------> ~700K probes



- **First filter is for probes with detection p-value (default > 0.01)**
- **Second, ChAMP will filter out probes with <3 beads in at least 5% of samples per probe.** This default can be changed with the filterBeads parameter or the frequency can be adjusted with the beadCutoff parameter.
- **Third, ChAMP will by default filter out all non-CpG probes contained in your dataset.**
- **Fourth, by default ChAMP will filter all SNP-related probes.** The SNP list comes from Zhou's Nucleic Acids Research paper in 2016. Note that if you know which population your data is from, you can choose certain populations to do the filtering. Otherwise ChAMP would use General Recommended Probes provided by Zhou to do filtering. You just need to assign "population" parameter to achieve this.
- **Fifth, by default setting, ChAMP will filter all multi-hit probes.** The multi-hit probe list comes from Nordlund's Genome Biology Paper in 2013[22].
- **Sixth, ChAMP will filter out all probes located in chromosome X and Y.** This is also a default setting, but user can change it with filterXY parameter.



BMIQ - beta-mixture quantile normalization

SWAN - Subset-quantile Within Array Normalization

PBC - peak-based correction

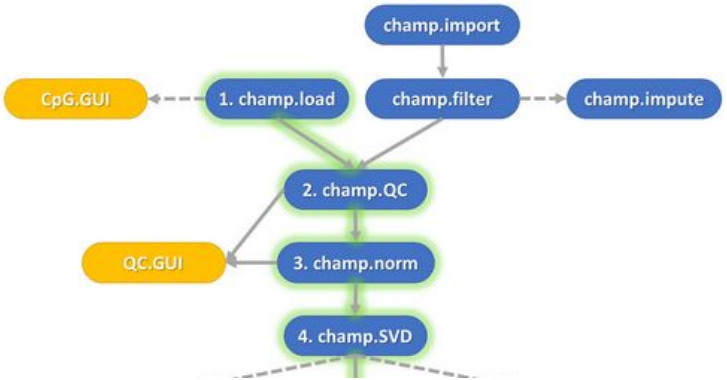
FunctionalNormalization

Type I probes have a broader range and better-defined distributions.
Type II probes have a compressed range and look different statistically.

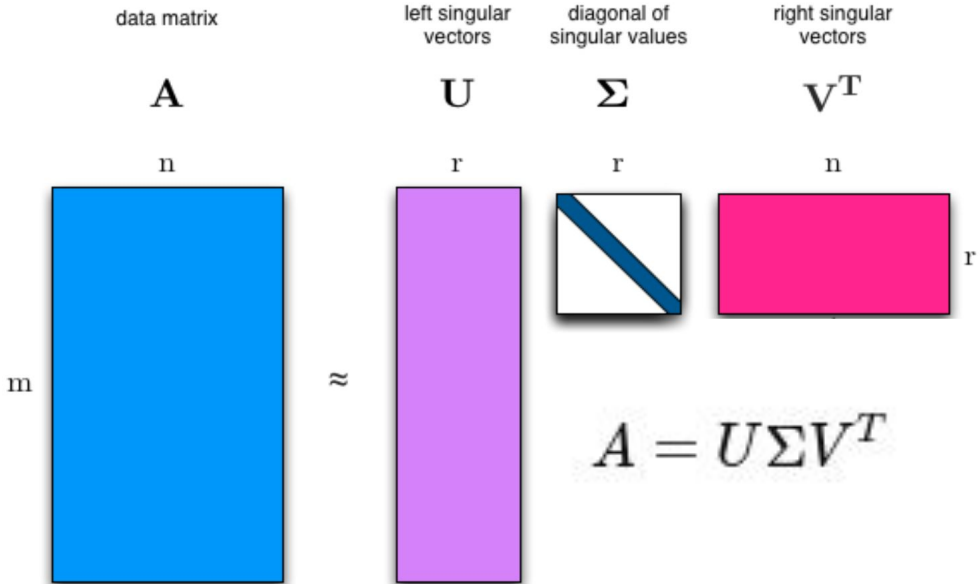
1) Three-State Beta Mixture Model
2) Assign Type II Probes to Methylation States:
3) Quantile Normalization for U and M States:
4) Dilation Transformation for H-State:

$$p(\beta^*) = \pi_U \cdot B(\beta|\alpha_U, \beta_U) + \pi_H \cdot B(\beta|\alpha_H, \beta_H) + \pi_M \cdot B(\beta|\alpha_M, \beta_M)$$

4 ChAMP pipeline - SVD



$$A_{ij} = \beta_{ij} - \frac{1}{n}\sum_{k=1}^{n}\beta_{ik}$$

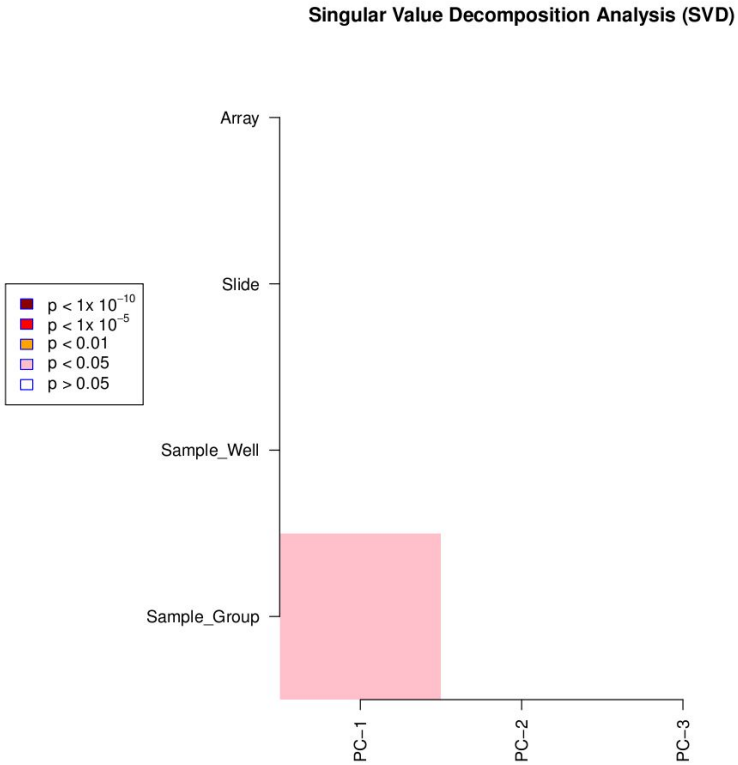


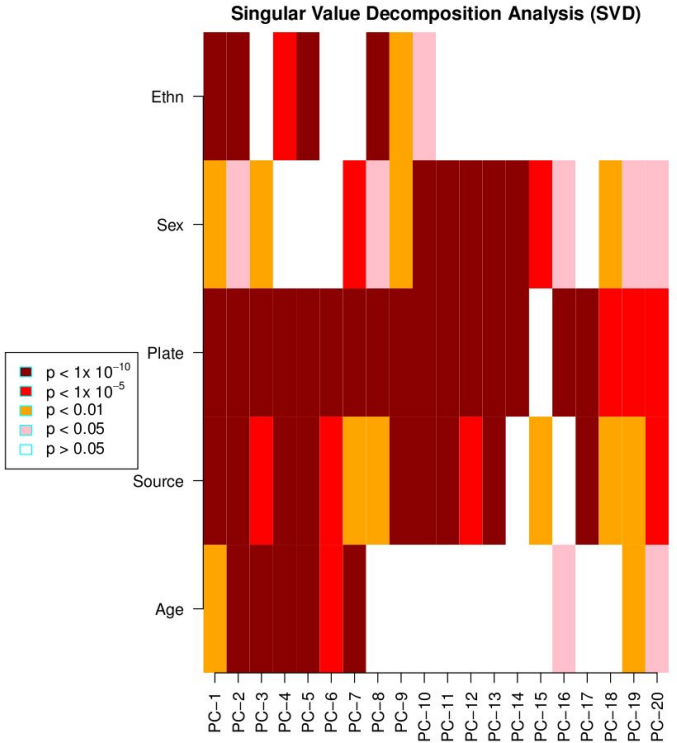
$$A = U\Sigma V^T$$

singular value decomposition

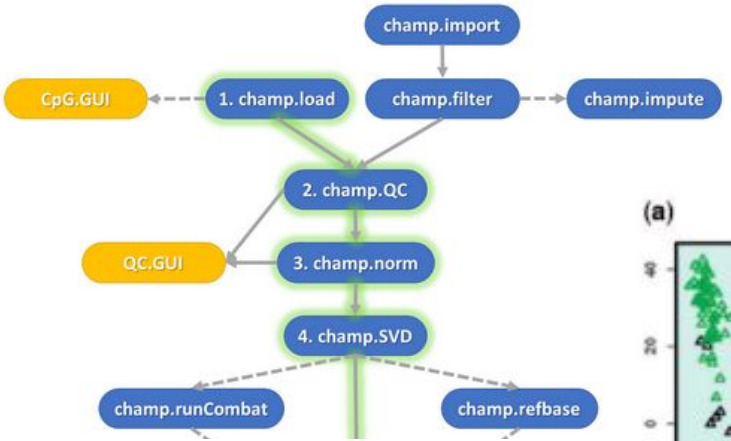
A = Beta-value matrix
U = Left singular vector (cpg variation)
VT = right singular vector (sample contribution)
Σ = singular values - PCs components

4 ChAMP pipeline - SVD

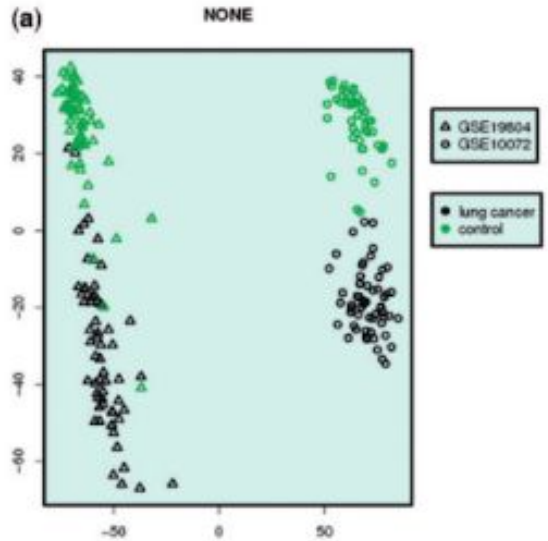


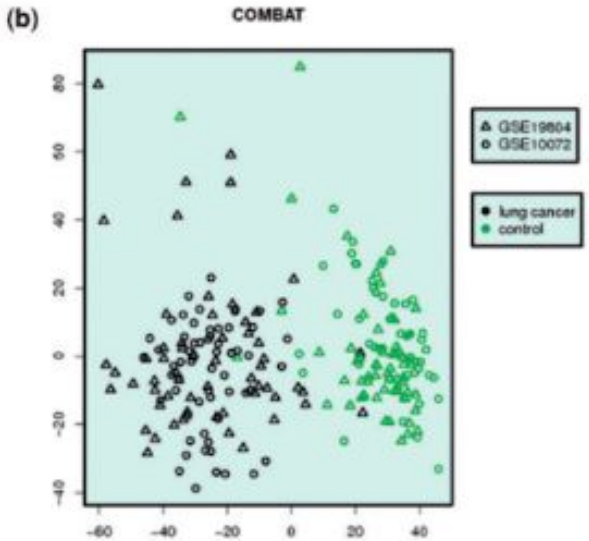


4 ChAMP pipeline - Combat



The Combat function ensures that you're comparing **methylation levels** fairly by removing the influence of batch effects

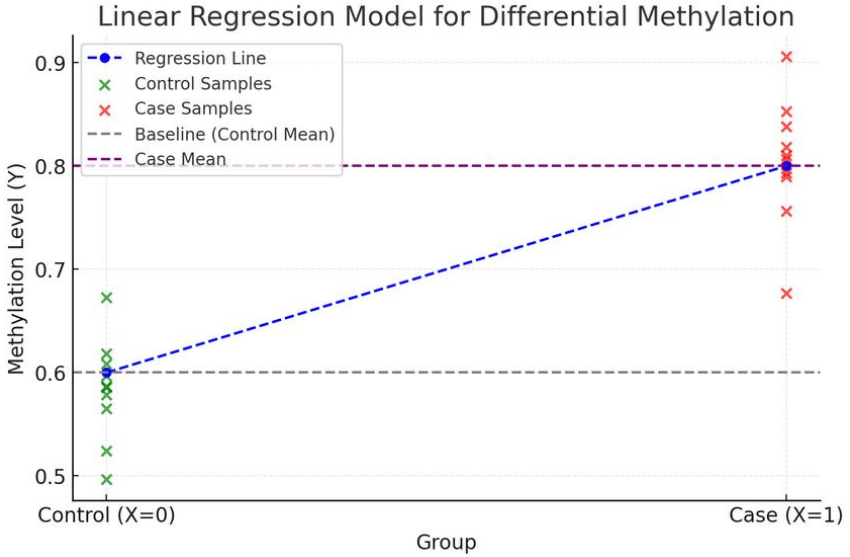




Lazar et.al., "Batch effect removal methods for microarray gene expression data integration: a survey" Brief Bioinform 2013

4 ChAMP pipeline - DMPs

$$\text{Beta Value} = \frac{M}{M + U + \alpha} \qquad M = \log_2\left(\frac{\beta}{1-\beta}\right)$$



$$Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$$

4 ChAMP pipeline - DMRs

Applying Bumphunter, DMRcate or ProbeLasso Algorithms to Estimate regions for which a genomic profile deviates from its baseline value. Originally implemented to detect differentially methylated genomic regions between two populations. By default, we recommend user do champ.DMR on normalized beta value on two populations, like case to control. The function will return detected DMR and estimated ρ value. The three algorithms specified in this function is different, while Bumphunter and DMRcate calcuated averaged candidate bumps methylation value between case and control.

Case study

Sample_Name	ID	Slide	Array	Sample_Group	Treatment
C1	GSM4332034	200863770015	R01C01	Control	Untreated
C2	GSM4332035	200863770015	R02C01	Control	Untreated
C3	GSM4332054	200661940025	R05C01	Control	Untreated
C4	GSM4332055	200661940025	R06C01	Control	Untreated
P1	GSM4332038	200863770015	R05C01	Case	Untreated
P2	GSM4332039	200863770015	R06C01	Case	Untreated
P3	GSM4332050	200661940025	R01C01	Case	Untreated
P4	GSM4332051	200661940025	R02C01	Case	Untreated