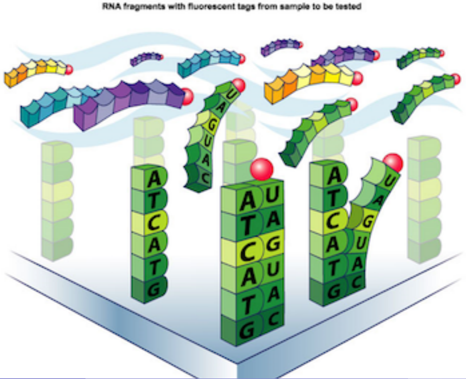
Array-based methylation technologies and analysis

Mikhail Dozmorov

2021-04-12

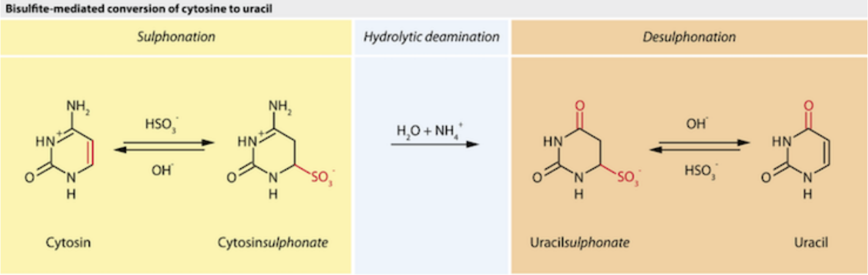
Bisulfite conversion-based Microarray Analysis

- A DNA microarray is a technology that consists of thousands of spots with DNA oligonucleotides (probes) that are used to hybridize a target sequence.
- Probe-target hybridization is usually detected and quantified by detection of fluorophore-, or chemiluminescence-labeled targets.


RNA fragments with fluorescent tags from sample to be tested
ATCATG
UAGUAC
ATCATG
UAGUAC
ATCATG
UAGUAC
UAGUAC

Sodium Bisulfite conversion

- Modifies non-methylated cytosines to uracil (methylation is protective from conversion)
- Differentiation of methylated and non-methylated cytosines at base-pair resolution
- $C \rightarrow U$ - which reads as **T** during sequencing
- $C^M \rightarrow C$ - which reads as **C** during sequencing



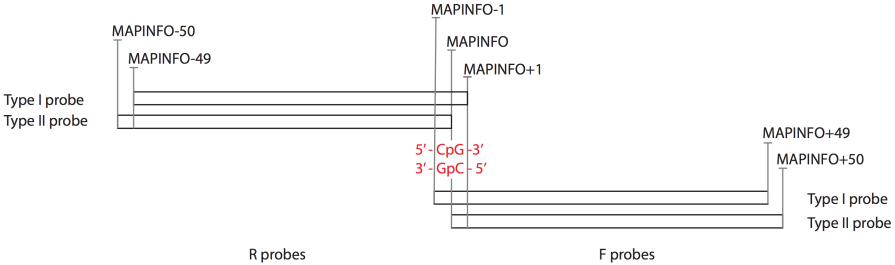
Tollefsbol T (ed.): Handbook of Epigenetics: The New Molecular and Medical Genetics. 1st edition. London, San Diego: Academic Press, 2011.

Illumina Infinium methylation assay

- Unmethylated **cytosines** are chemically deaminated to **uracil** in the presence of bisulfite.
- Methylated cytosines are refractory to the effects of bisulfite and remain cytosine.
- After bisulfite conversion, each sample is whole-genome amplified (WGA) and enzymatically fragmented.
- The bisulfite-converted WGA-DNA samples are purified and applied to the BeadChips.

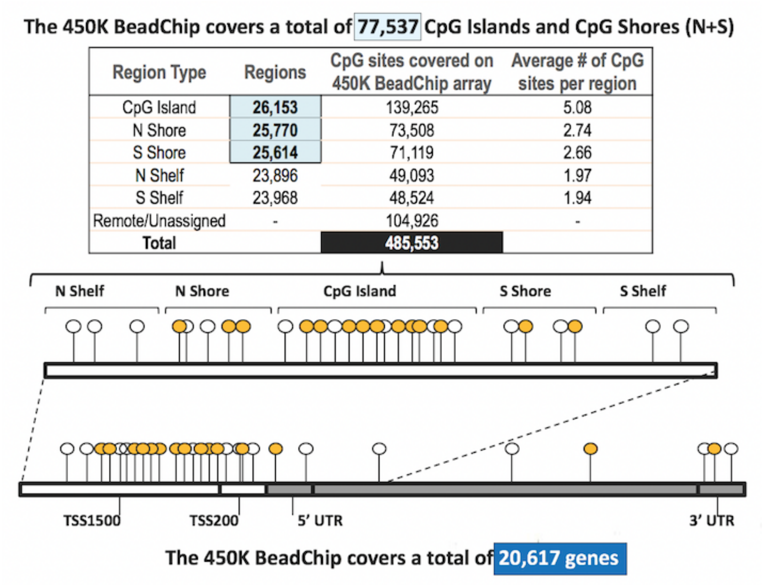
Illumina Infinium methylation assay

- Bead technology
- Each bead has oligos containing 23-base address + 100-base probe complementary to bisulfite converted DNA with the CpG site in the center



Illumina Infinium evolution

- 2008: **HumanMethylation27K**. 25,578 probes targeting CpG sites within the proximal promoter regions.
- 2011: **HumanMethylation450K**. 485,577 probes targeting additional CpG islands, shores and shelves, the 5' and 3' UTRs, gene bodies, some enhancer regions. Covers 99% of RefSeq genes.
- 2015: **MethylationEPIC**. $>$850,000 probes. Additional cooverage of regulatory elements. 58% of FANTOM5 enhancers, 7% distal and 27% proximal ENCODE regulatory elements.



The 450K BeadChip covers a total of 77,537 CpG Islands and CpG Shores (N+S)

Region Type	Regions	CpG sites covered on 450K BeadChip array	Average # of CpG sites per region
CpG Island	**26,153**	139,265	5.08
N Shore	**25,770**	73,508	2.74
S Shore	**25,614**	71,119	2.66
N Shelf	23,896	49,093	1.97
S Shelf	23,968	48,524	1.94
Remote/Unassigned	-	104,926	-
Total		485,553	

The 450K BeadChip covers a total of 20,617 genes

Measurement of methylation level

Illumina 450K and 850K use two types of probes:

- **Type I probes** have two separate probe sequences per CpG site (one each for methylated and unmethylated CpGs). ~28% of probes. Suggested to be more stable and reproducible than the Type II probes
- **Type II probes** have just one probe sequence per CpG site. Use half of the physical space. ~ 72% of probes. Have a decreased quantitative dynamic range compared to Type I probes.

Measurement of methylation level

Beta-value - bimodal distribution within [0,1] range

$$\beta = \frac{M}{U + M}$$

- M - signal from methylated probes
- U - signal from unmethylated probes

$\beta = 0/1$ - all probes are non-methylated/fully methylated, respectively

Measurement of methylation level

Beta-value - bimodal distribution within [0,1] range

$$\beta = \frac{M}{U + M}$$

- M - signal from methylated probes
- U - signal from unmethylated probes

M-value - centered around 0, $[-\infty, +\infty]$ range

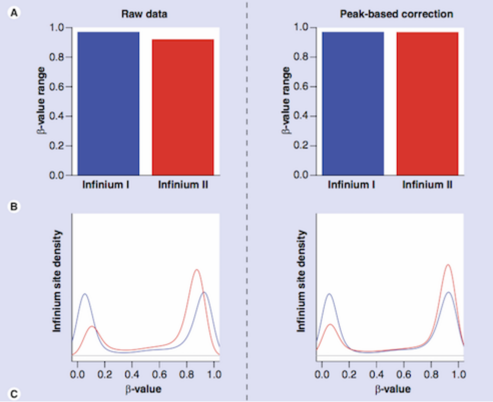
$$Mvalue = log\left(\frac{M}{U}\right) = log\left(\frac{\beta}{1 - \beta}\right)$$

$M = -\infty$ - all probes are non-methylated

$M = +\infty$ - all probes are methylated

Measurement of methylation level

- β values obtained from Infinium II probes are slightly less accurate and reproducible than those obtained from Infinium I probes (Dedeurwaerder et al. 2011)
- Peak correction methods (normalization) are available



Filter questionable probes

- Remove probes that have failed to hybridize (detection p-value)
  - Detection p-value represents the probability the target signal was distinguishable against background noise
- Drop probes that failed in n^{th} percent of samples
  - Common thresholds are 20%, 10%, 5% of probes at $>$0.05, $>$0.01
- Drop samples that failed in n^{th} percent of probes
  - Common thresholds are 50%, 20% at $>$0.05, $>$0.01

Filter questionable probes

- Probes on X and Y chromosomes
- Probes with the lowest variation
- Probes with extreme methylation level (e.g. median $= 0\%$ or 100%)
- Keep only those in regions of interest (e.g., CpG islands, shores)

Filter questionable probes

- Data from Chen YA et al. “Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray.” Epigenetics.
  - List of non-specific probes - 29,233 non-specific ‘cg’ probes, 1,736 non-specific ‘ch’ probes;
  - List of polymorphic CpGs - 70,899 records (66,877 unique probes) about CpGs containing SNPs at or near single base extension (SBE) position, 316,034 records (220,582 unique probes) having SNPs in probe sequences.
- More for MethylationEPIC at https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1066-1

A typical pipeline

1. Filtering non-specific, polymorphic, SNP, chromosome Y probes
2. Pre-processing and QC
   - `dasen` (background correction and quantile normalization)
   - `BIMQ` (Beta-mixture quantile normalization, correcting batch effect of Infinium I and II chemistries)
   - Principal Components Analysis to detect batch effects
   - `ComBat`, `ISVA` (removing batch effect)
3. Association analysis, or differential methylation
   - `betareg` regression model
   - Pearson correlation coefficient
   - `limma`, `minfi` for differentially methylated regions
   - Benjamini-Hochberg adjusted p-values < 0.05
4. Functional enrichment analyses of genes associated with differentially methylated probes

Interpretation

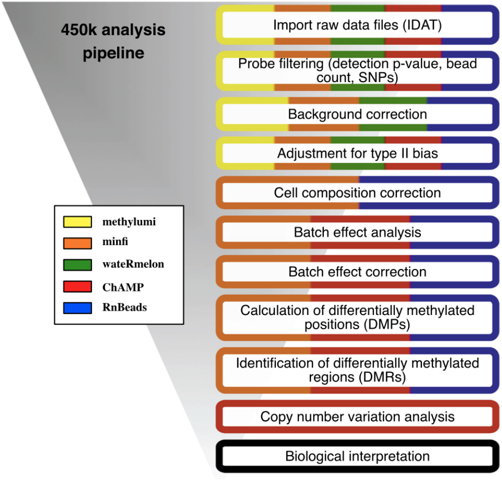
- Map CpG sites of interest to the nearby genes, analyze genes for functional enrichment
- Analyze genomic location of CpG sites, using genomic coordinates
  - **GREAT** predicts functions of cis-regulatory regions, http://bejerano.stanford.edu/great/public/html/
  - **Enrichr**, gene- and genomic regions enrichment analysis tool, http://amp.pharm.mssm.edu/Enrichr/#
  - **GenomeRunner**, Functional interpretation of SNPs (any genomic regions) within regulatory/epigenomic context, http://integrativegenomics.org/

R packages for Illumina Infinium array analysis

- **lumi** - normalization, vusualization, gene annotation https://www.bioconductor.org/packages/release/bioc/html/lumi.html
- **methylumi** - normalization and general data handling http://www.bioconductor.org/packages/release/bioc/html/methylumi.html
- **minfi** - normalization, analysis and visualization http://www.bioconductor.org/packages/release/bioc/html/minfi.html, or **ChAMP** - eight functions to run *minfi* pipelines, https://bioconductor.org/packages/release/bioc/html/ChAMP.html
- **RnBeads** - works for 450K arrays, BS-Seq, MeDIP or MBD-Seq data https://bioconductor.org/packages/release/bioc/html/RnBeads.html
- **wateRmelon** - 15 normalization methods, other QC metrics https://bioconductor.org/packages/release/bioc/html/wateRmelon.html

Morris TJ, Beck S "**Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data**" Methods. 2015 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4304832/

R packages for Illumina Infinium array analysis



450k analysis pipeline

- Import raw data files (IDAT)
- Probe filtering (detection p-value, bead count, SNPs)
- Background correction
- Adjustment for type II bias
- Cell composition correction
- Batch effect analysis
- Batch effect correction
- Calculation of differentially methylated positions (DMPs)
- Identification of differentially methylated regions (DMRs)
- Copy number variation analysis
- Biological interpretation

🟨 methylumi
🟧 minfi
🟩 wateRmelon
🟥 ChAMP
🟦 RnBeads

Morris TJ, Beck S "**Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data**" Methods. 2015 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4304832/