V15: Analysis of DNA methylation data

Epigenetics refers to **alternate phenotypic states**

that are **not based** on **differences in genotype**.

They are potentially reversible,

but are generally stably maintained during cell division.

Examples:

- imprinting (monoallelic expression – one allele silenced by DNA methylation),
- cell differentiation,
- cancer vs. normal cells, ….
- **repetitive** genomic sequences such as human endogenous retroviral sequences (HERVs) are **heavily methylated**, which means transcriptionally silenced.

Laird, Hum Mol Gen 14, R65 (2005)

11.1 What is epigenetics?

Epigenetics is nowadays considered to invovle

multiple mechanisms that interact to collectively establish:

- alternate states of chromatin structure (open – packed/condensed),
- **histone modifications**,
- composition of associated proteins (e.g. histones),
- transcriptional activity,
- activity of microRNAs,
- in mammals, **cytosine-5 DNA methylation** at CpG dinucleotides,
- in bacteria **adenine-6 DNA methylation**.

Laird, Hum Mol Gen 14, R65 (2005)

11.1 Epigenetic marks

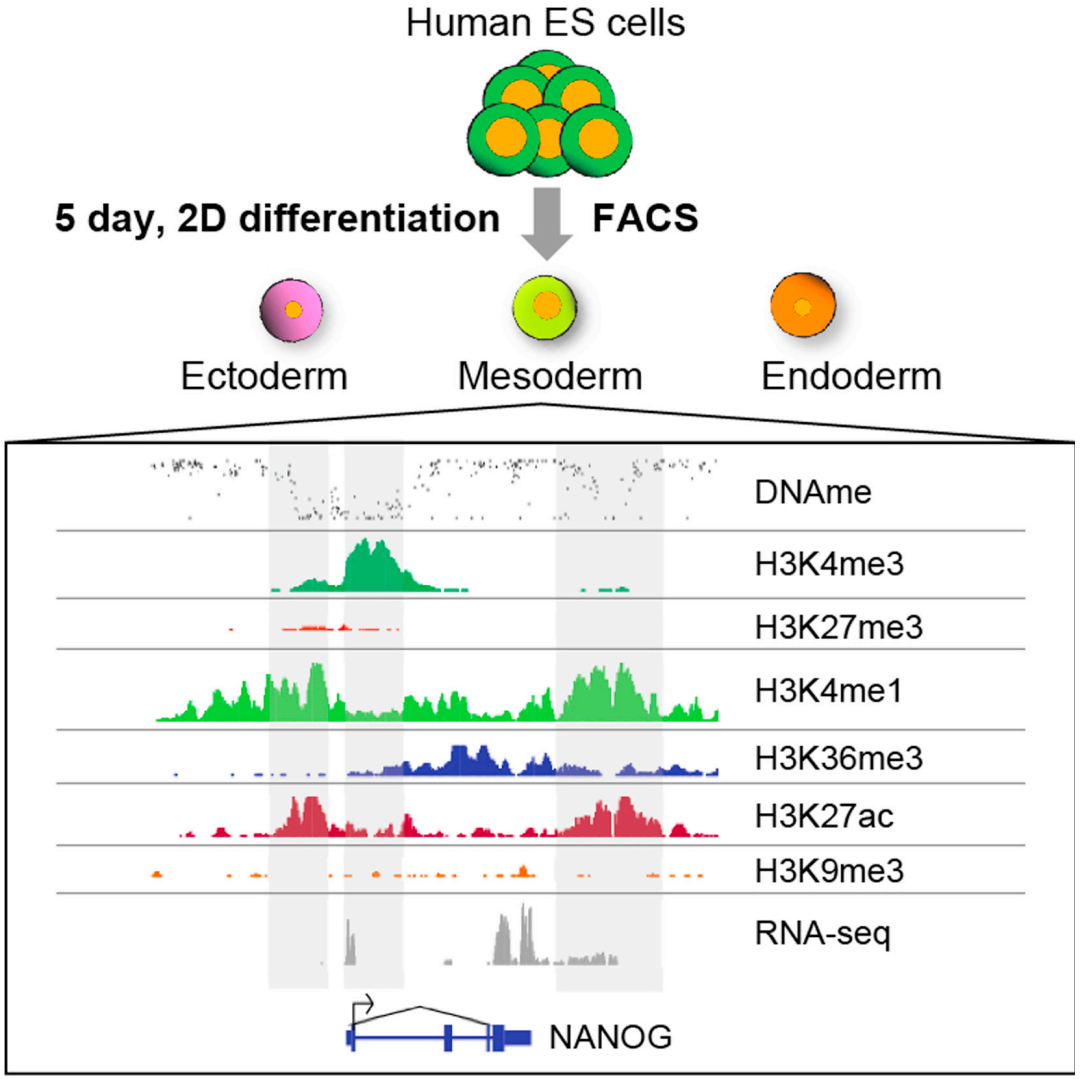
Epigenetic marks around the NANOG gene after 2 days of directed differentiation of human embryonic stem cells into mesoderm tissue.

Top row : DNA methylation level.

Next six rows : presence/absence of specified histone marks.

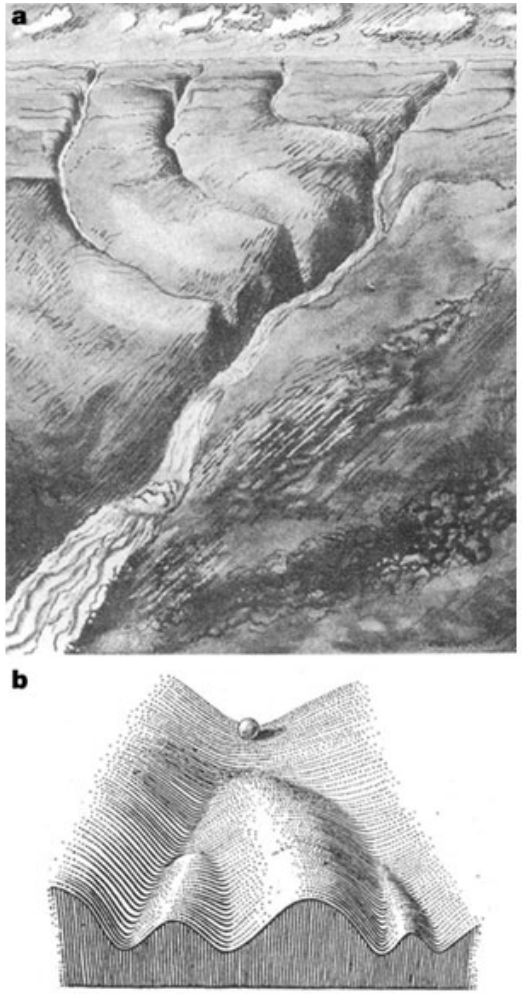
Bottom row : level of gene transcription measured by RNA sequencing.

Shown at the bottom is the exon structure of the gene NANOG that is crucial for development.



Gifford CA et al. (2013) *Cell* **153**, 1149-1163

Waddington epigenetic landscape for embryology



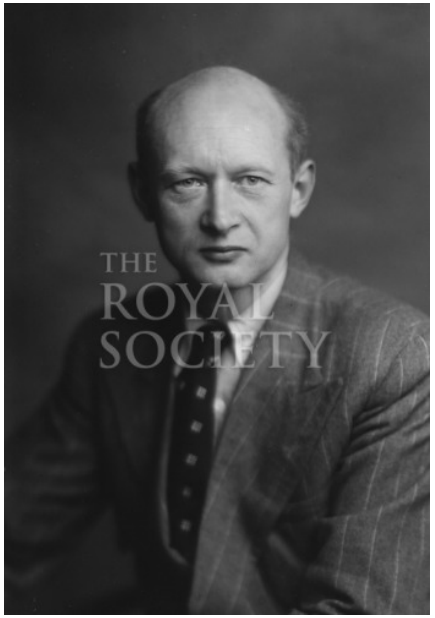
Slack, Nature Rev Genet 3, 889-895 (2002)

Waddington worked in **embryology**

a) is a painting by John Piper that was used as the frontispiece for Waddington's book *Organisers and Genes*.
It represents an epigenetic landscape.

Developmental pathways that could be taken by each cell of the embryo are metaphorically represented by the path taken by water as it flows down the valleys.

b) Later depiction of the epigenetic landscape. The ball represents a cell, and the bifurcating system of valleys represents bundles of trajectories in state space.

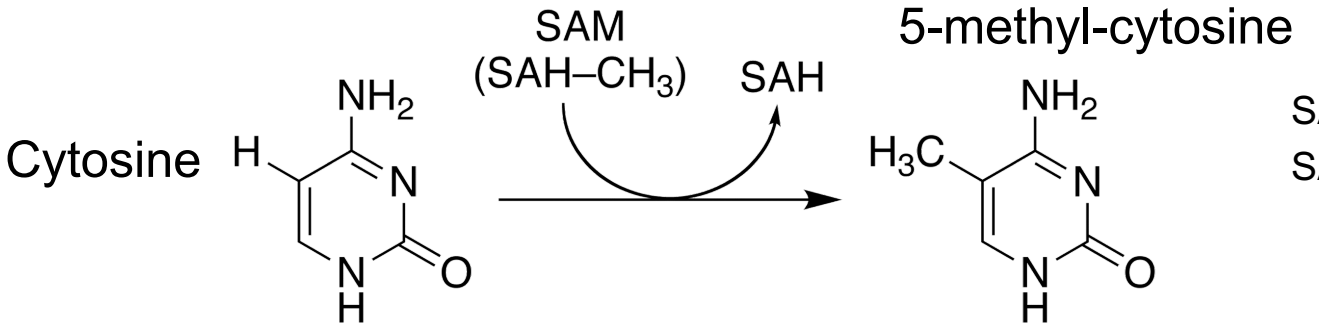


Conrad Hal Waddington (1905 – 1975)
pictures.royalsociety.org

Cytosine methylation

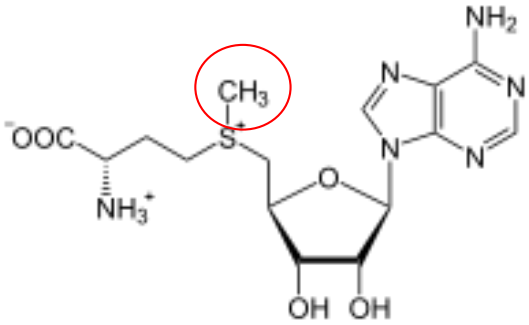
Observation: 3-6 % of all cytosines are methylated in human DNA.

This methylation occurs (almost) exclusively when cytosine is followed by a guanine base -> **CpG dinucleotide**.



SAM: S-adenosyl-methionine

SAH: S-adenosyl-homocysteine



As most CpGs serve as targets of DNA methyltransferases, about 70 - 80% of them are usually methylated.

BUT mammalian genomes contain much fewer (only 20-25 %) of the CpG dinucleotide than is expected by the G+C content (we expect 1/16 ≈ 6% for any random dinucleotide).

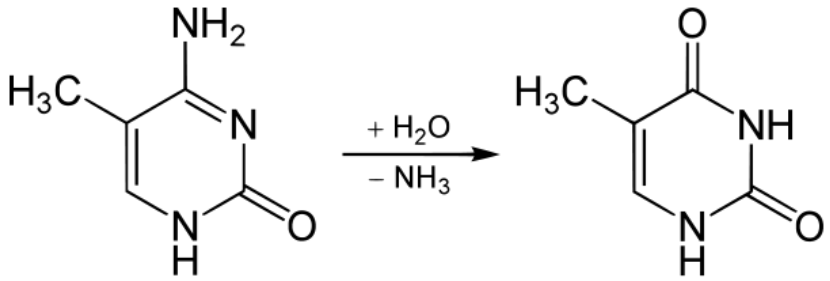
This is typically explained in the following way:

…. (see following page)

Esteller, Nat. Rev. Gen. 8, 286 (2007)
www.wikipedia.org

Cytosine methylation

5-Methylcytosine can easily **deaminate** to **thymine**.

5-methyl-cytosine 

thymine

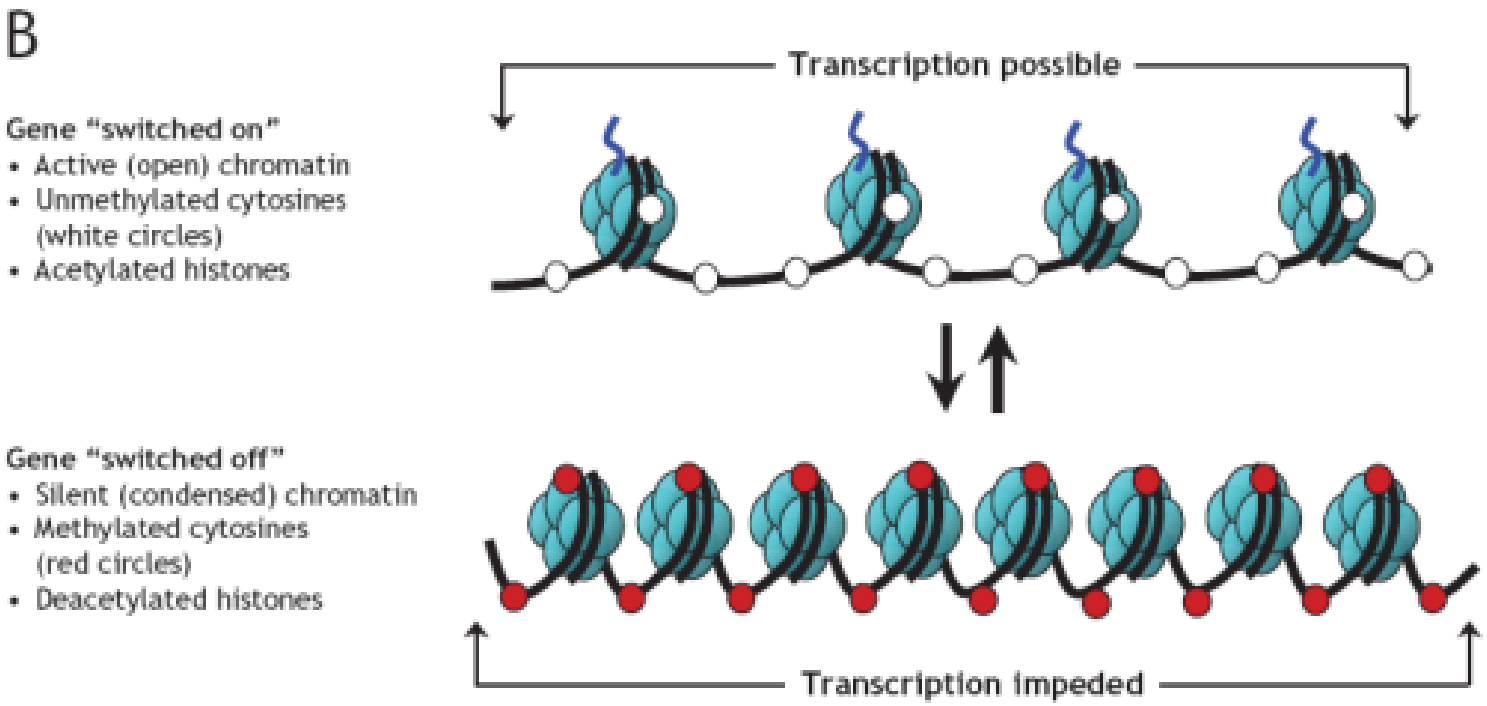
If this mutation is not repaired, the affected CpG is permanently converted to TpG (or CpA if the transition occurs on the reverse DNA strand).

Hence, methylCpGs represent **mutational hot spots** in the genome.
If such mutations occur in the germ line, they become heritable.

A constant loss of CpGs over thousands of generations
can explain the low frequency of this
special dinucleotide in the genomes of human and mouse.

Esteller, Nat. Rev. Gen. 8, 286 (2007)
www.wikipedia.org

chromatin organization affects gene expression



Schematic of the reversible changes in chromatin organization that influence gene expression:

genes are expressed (switched on) when the chromatin is **open** (active), and they are inactivated (switched off) when the chromatin is **condensed** (silent).

White circles = unmethylated cytosines;

red circles = methylated cytosines.

Rodenhiser, Mann, CMAJ 174, 341 (2006)

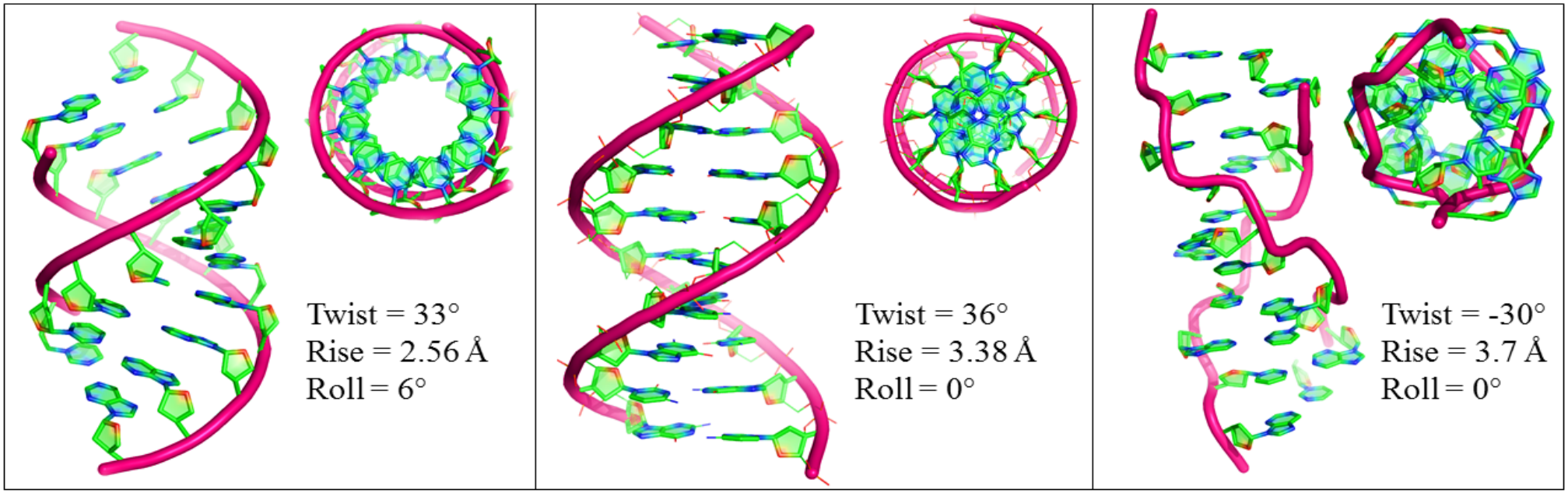
DNA fiber forms

A-DNA

B-DNA

Z-DNA

Requires more methylation, higher concentration of physiological salts

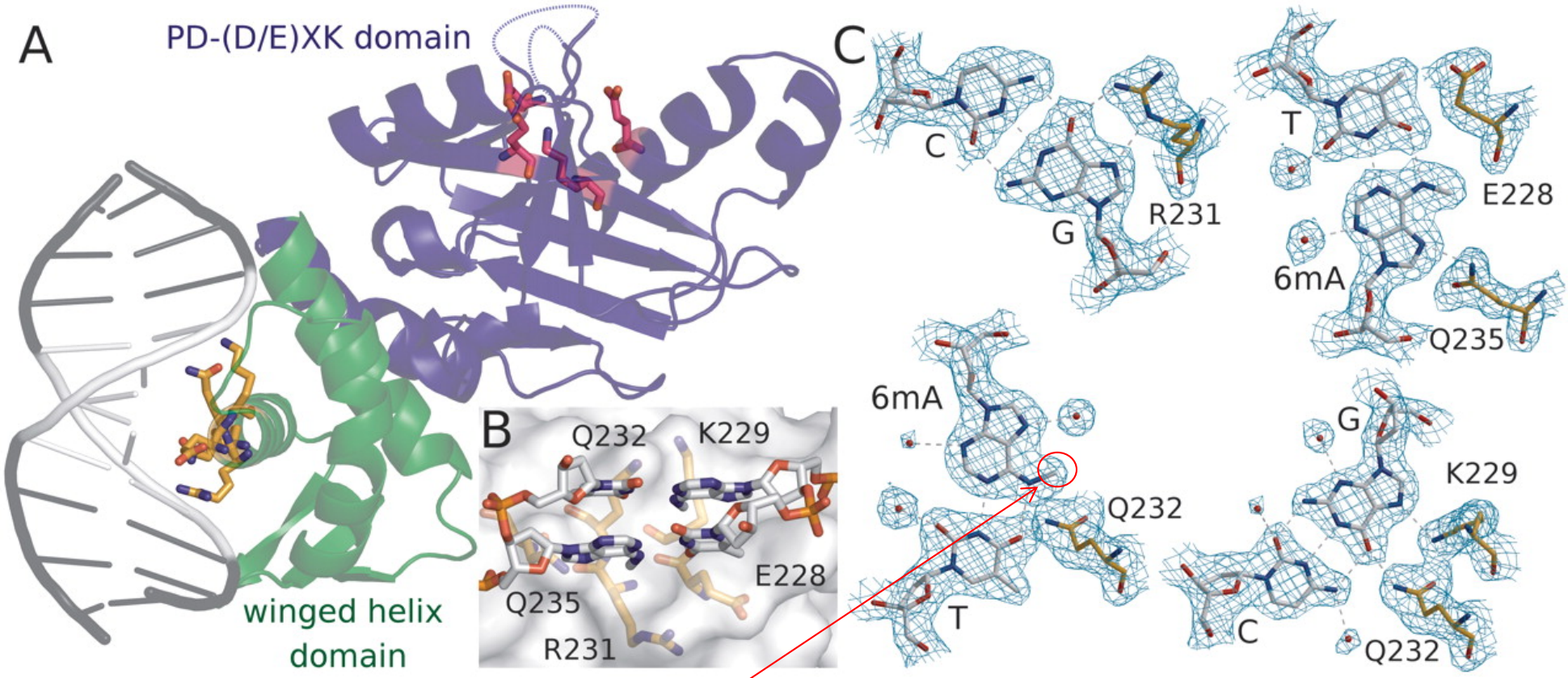


Dry Environment

Most prominent in cellular conditions

Equilibrium shift with specific conditions

Protein-DNA^{Me} interaction (R.DpnI from *E.coli*)

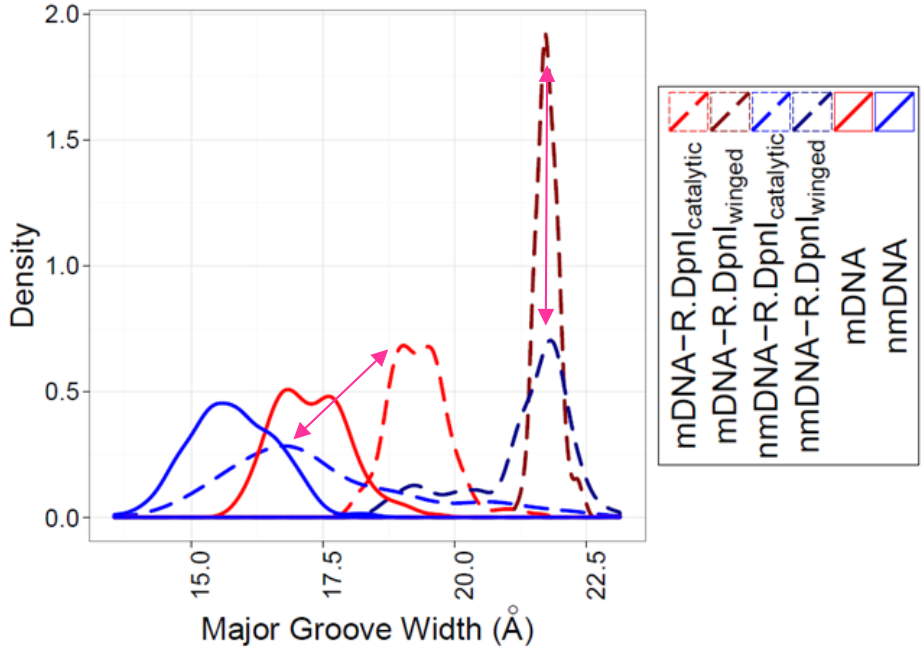


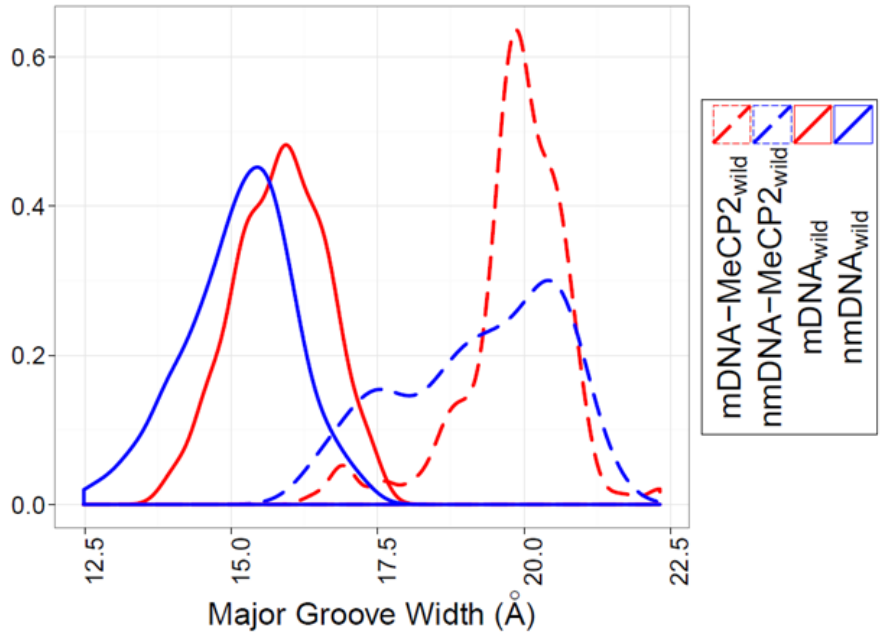
Left: structural transitions of DNA affect accessibility of the base pairs

Right: recognition of 6-methylated adenine (common form of DNA methylation in bacteria)

Siwek et al. Nucl. Acids Res. (2012) 40 (15): 7563-7572.

Protein-DNA^{Me} interaction





Binding of *E.coli* restriction enzyme R.DpnI to adenine-methylated or unmethylated target sequence. R.DpnI has 2 domains that bind DNA, a „catalytic“ domain and a „winged“ domain.

-> methylation linked to increased width of major groove when bound to „catalytic“ domain, not to „winged“ domain.

Solid lines: free DNA

Binding of MeCP2 to cytosine-methylated or unmethylated target BDNF sequence from human

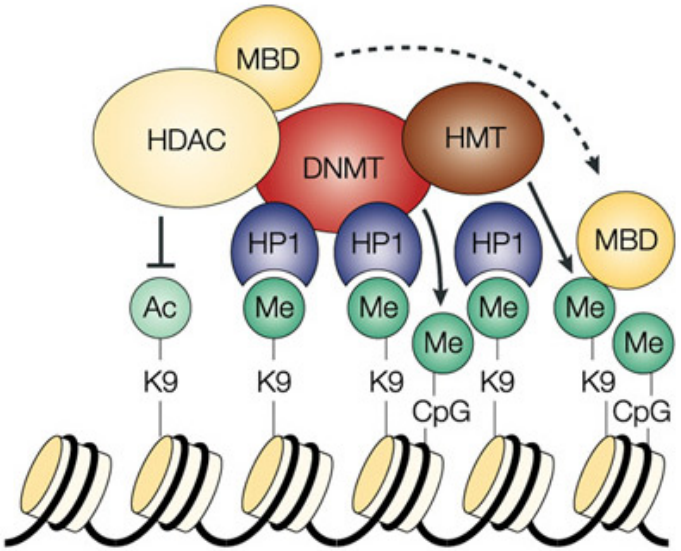
-> methylation has smaller effects on width of major groove

PhD thesis Siba Shanak (2015)

Enzymes that control DNA methylation and histone modfications

The dynamic chromatin states are controlled by reversible epigenetic patterns of **DNA methylation** and **histone modifications**.

Enzymes involved in these processes include
- DNA methyltransferases (DNMTs),
- histone deacetylases (HDACs),
- „writers“ such as histone acetylases and histone methyltransferases and
- „reader“ proteins such as the methyl-binding domain protein MECP2.



Rodenhiser, Mann, CMAJ 174, 341 (2006)
Feinberg AP & Tycko P (2004) Nature Reviews: 143-153

DNA methylation

Typically, unmethylated clusters of CpG pairs are located in
tissue-specific genes and in essential **housekeeping genes.**

(House-keeping genes are involved in routine maintenance roles and are expressed in most tissues.)

These clusters, or **CpG islands**, are targets for proteins
that bind to unmethylated CpGs and initiate gene transcription.

In contrast, **methylated CpGs** are generally associated with silent DNA,
can block methylation-sensitive proteins and can be easily mutated.

The loss of normal DNA methylation patterns is the
best understood epigenetic cause of disease.

In animal experiments, the removal of genes that encode DNMTs is lethal;
in humans, overexpression of these enzymes has been linked
to a variety of cancers.

Rodenhiser, Mann, CMAJ 174, 341 (2006)

CpG islands

CpG islands are characterized by an **elevated density** of **CpG dinucleotides** that can be targeted by DNA methylation (elevated relative to the rest of the genome).

CpG islands are regulatory elements and are often located in the promoter region of genes.

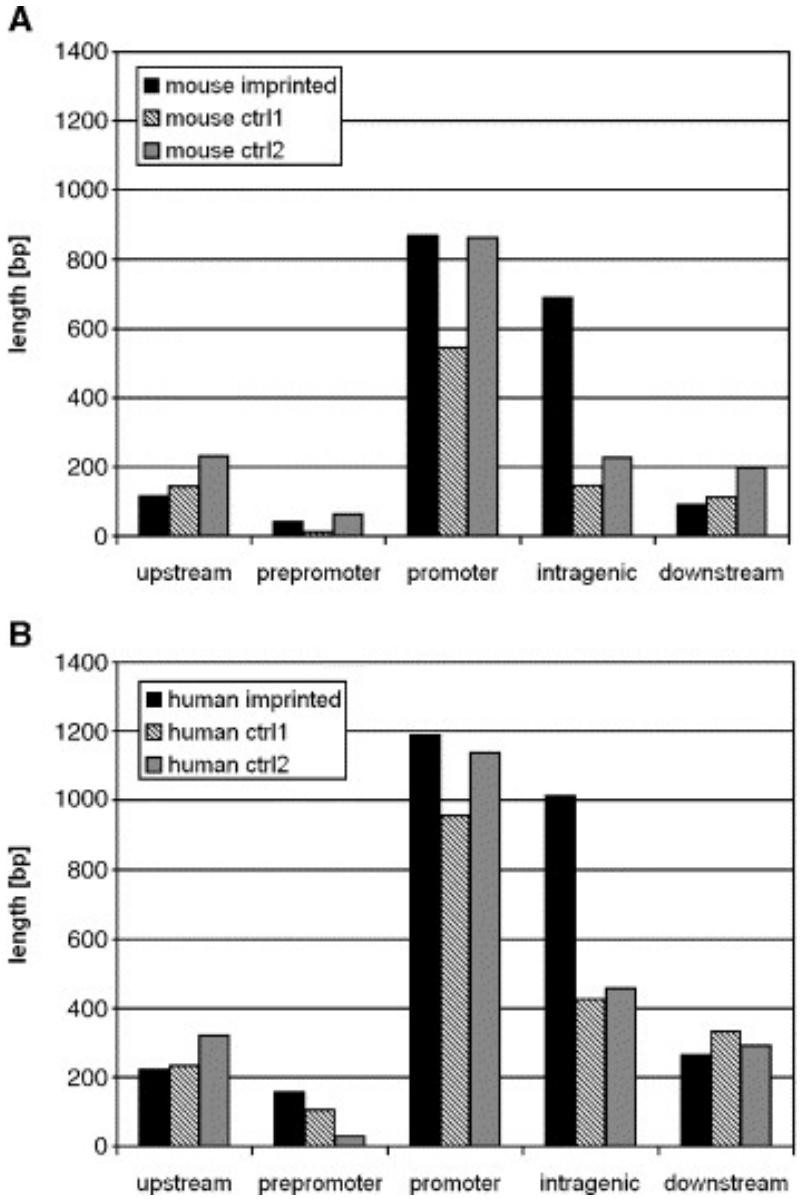
Criteria to define CpG islands:

Gardiner-Garden and Frommer: ≥ 200 bp length,
G + C ≥ 50%
CpG_{obs}/CpG_{exp} ≥ 0.6

Takai and Jones: ≥ 500 bp length
G + C ≥ 55%
CpG_{obs}/CpG_{exp} ≥ 0.65.

Hutter, Helms, Paulsen, Genomics 88, 323 (2006)

CpG islands



Average total length of CpG islands per gene in repeat-masked sequences at five different locations in (A) Mouse, (B) human.

Imprinted genes are monoallelically expressed, the other allele is silenced by DNA methylation. In 2006, about 100 imprinted genes were experimentally confirmed.

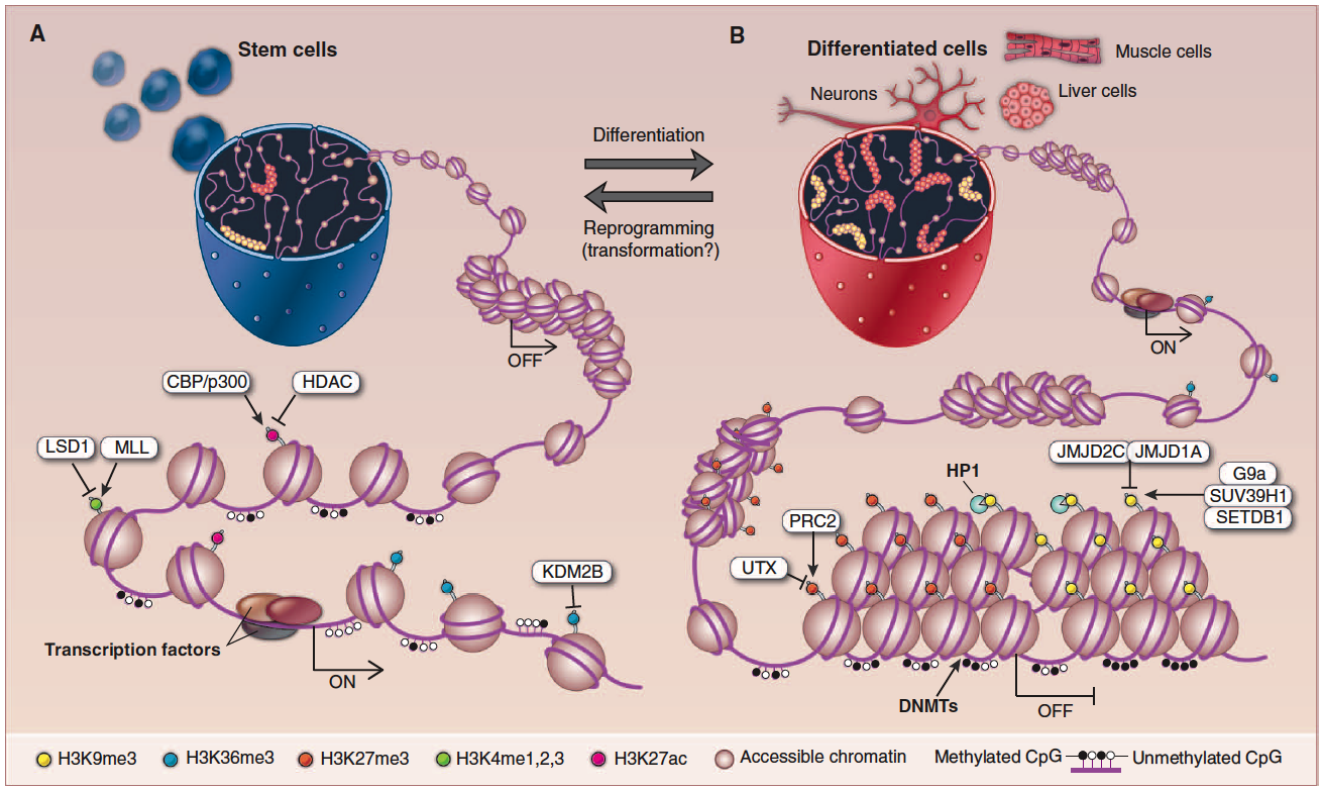
Ctrl1, ctrl2: groups of randomly selected (most likely biallelic) control genes

Takai and Jones parameters

-> CpG islands frequent in promoters and in the gene body of imprinted genes.

Hutter, Helms, Paulsen, Genomics 88, 323 (2006)

Differentiation linked to alterations of chromatin structure



(B) Upon differentiation, inactive genomic regions may be sequestered by repressive chromatin enriched for characteristic histone modifications.

(A) In pluripotent cells, chromatin is hyperdynamic and globally accessible.

ML Suva et al. Science 2013; 339:1567-1570

Altered DNA methylation upon cancerogenesis

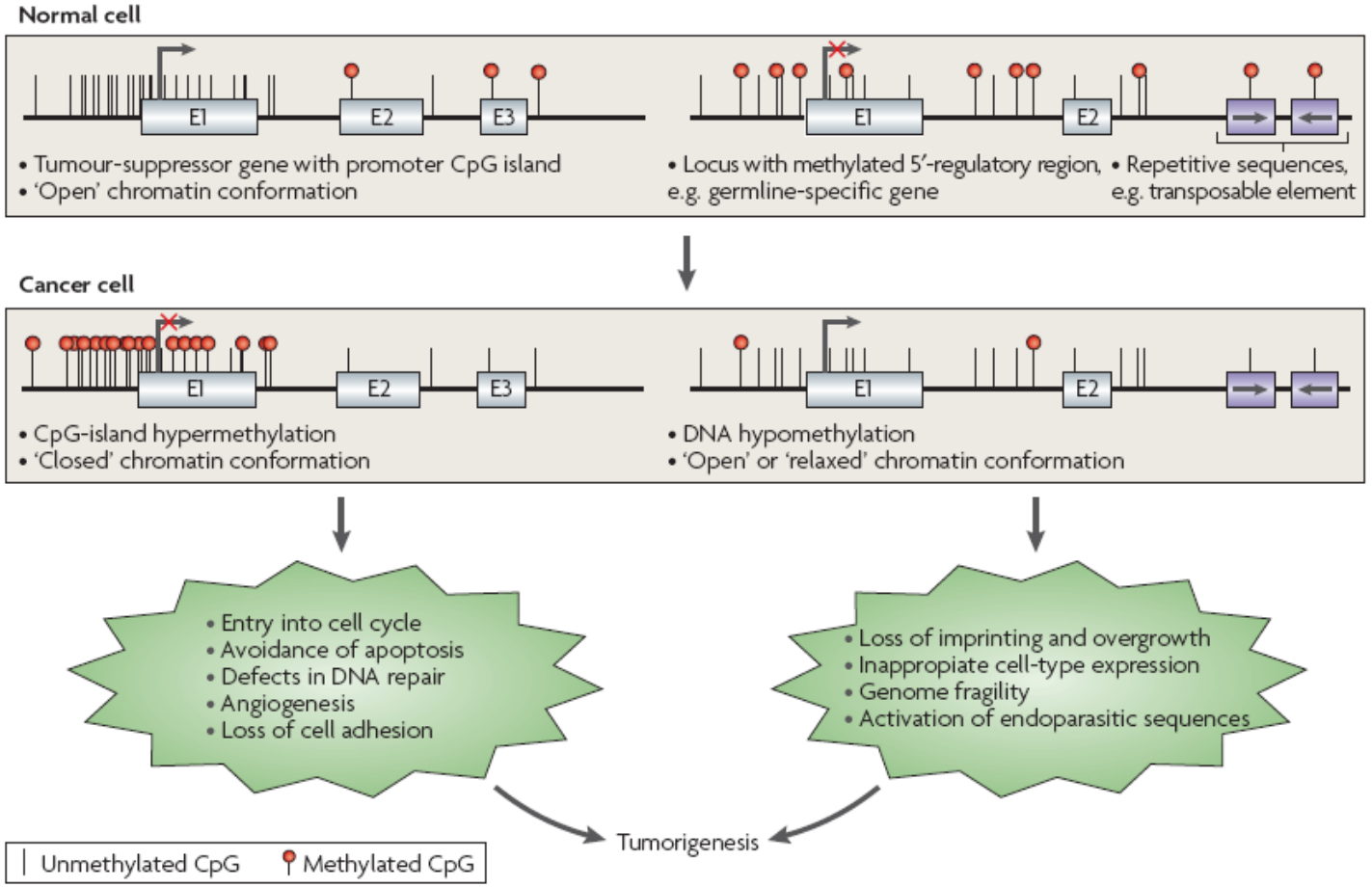
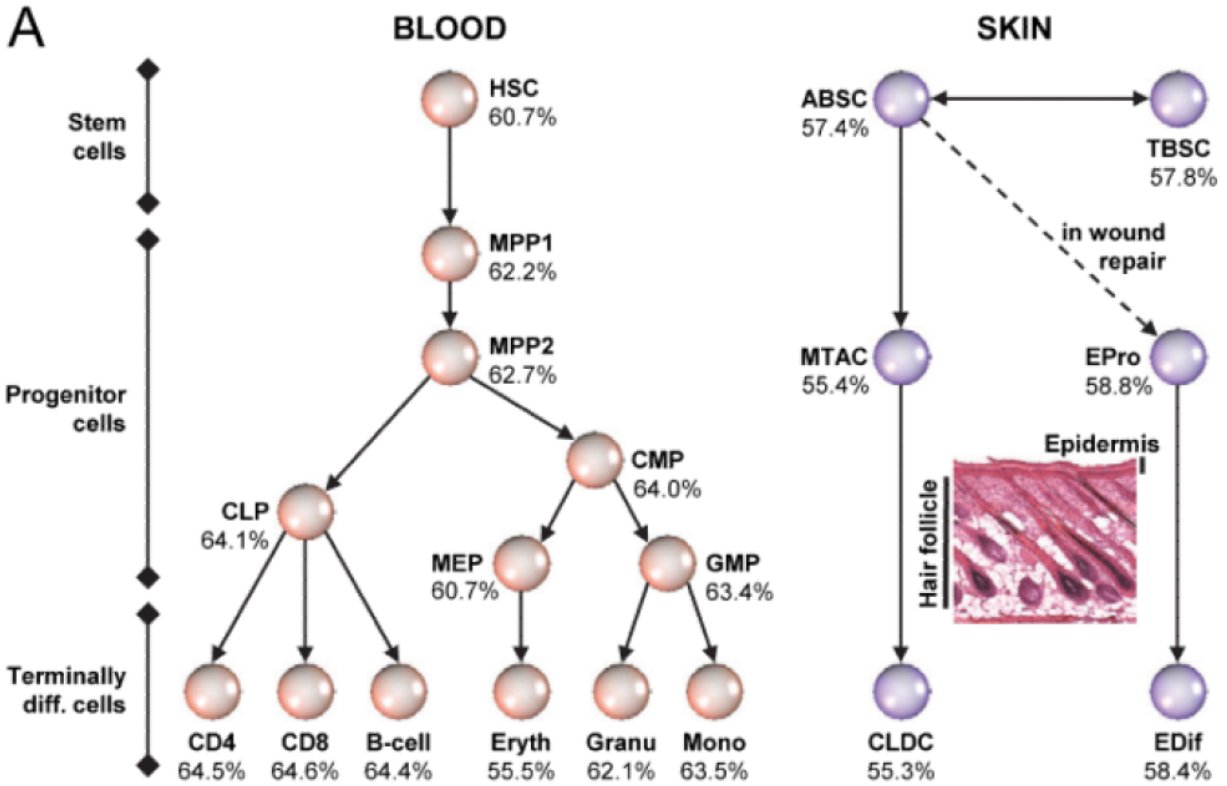
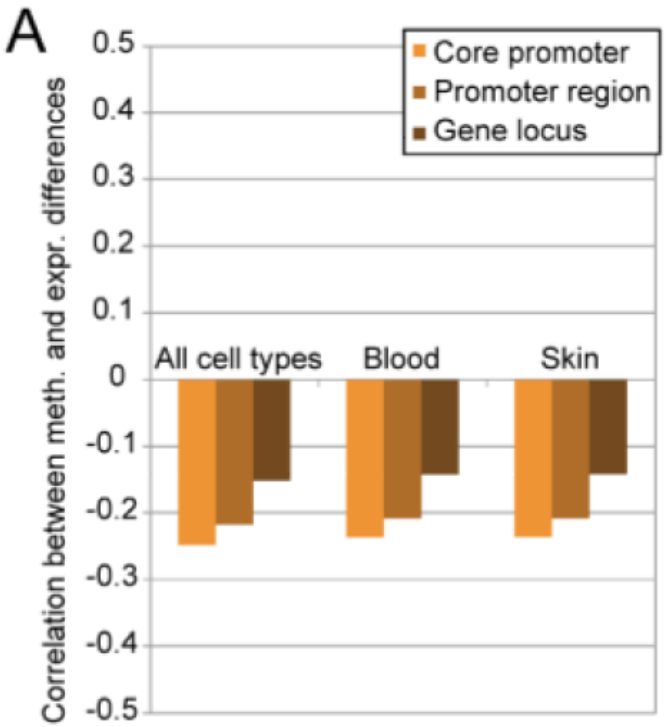


Figure 1 | **Altered DNA-methylation patterns in tumorigenesis.** The hypermethylation of CpG islands of tumour-suppressor genes is a common alteration in cancer cells, and leads to the transcriptional inactivation of these genes and the loss of their normal cellular functions. This contributes to many of the hallmarks of cancer cells. At the same time, the genome of the cancer cell undergoes global hypomethylation at repetitive sequences, and tissue-specific and imprinted genes can also show loss of DNA methylation. In some cases, this hypomethylation is known to contribute to cancer cell phenotypes, causing changes such as loss of imprinting, and might also contribute to the genomic instability that characterizes tumours. E, exon.

Esteller, Nat. Rev. Gen. 8, 286 (2007)

DNA methylation is typically only weakly correlated with gene expression!





Left: different states of hematopoiesis (blood cell differentiation).
HSC: hematopoietic stem cell
MPP1/2: multipotent progenitor cell

Right: skin cell differentiation

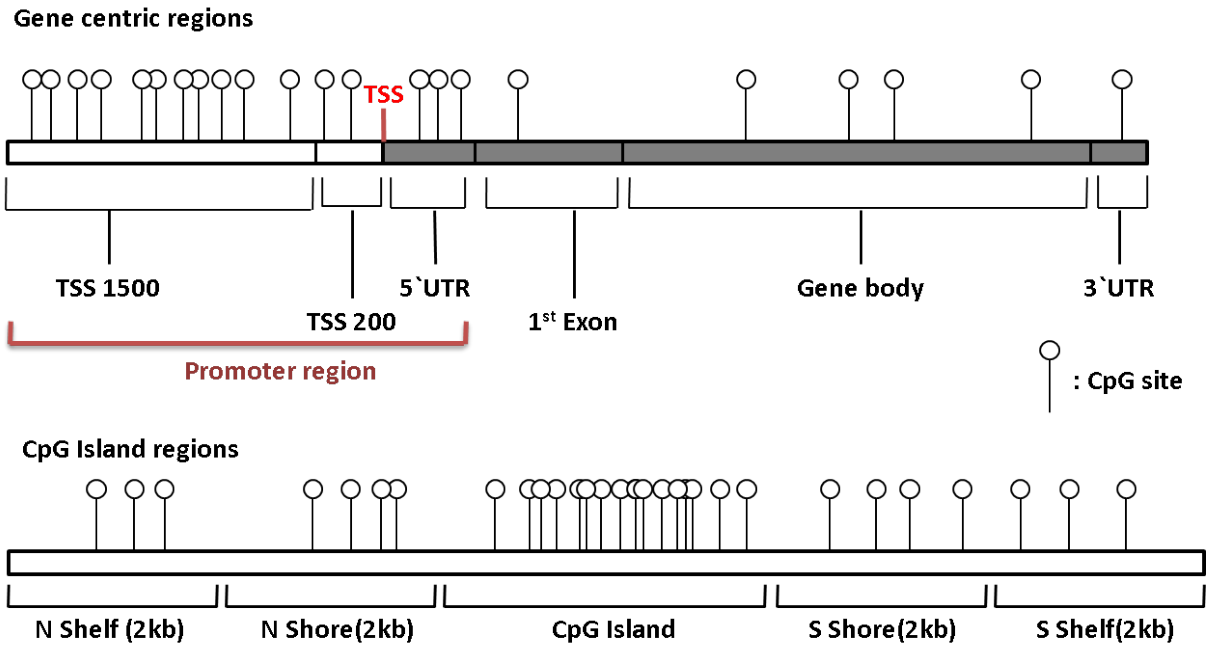
Bock et al. , Mol. Cell. 47, 633 (2012)

Promoter methylation vs. gene-body methylation

The relationship between methylation and gene expression is complex.

High levels of gene expression are often associated with
low **promoter methylation** but elevated **gene body methylation**.

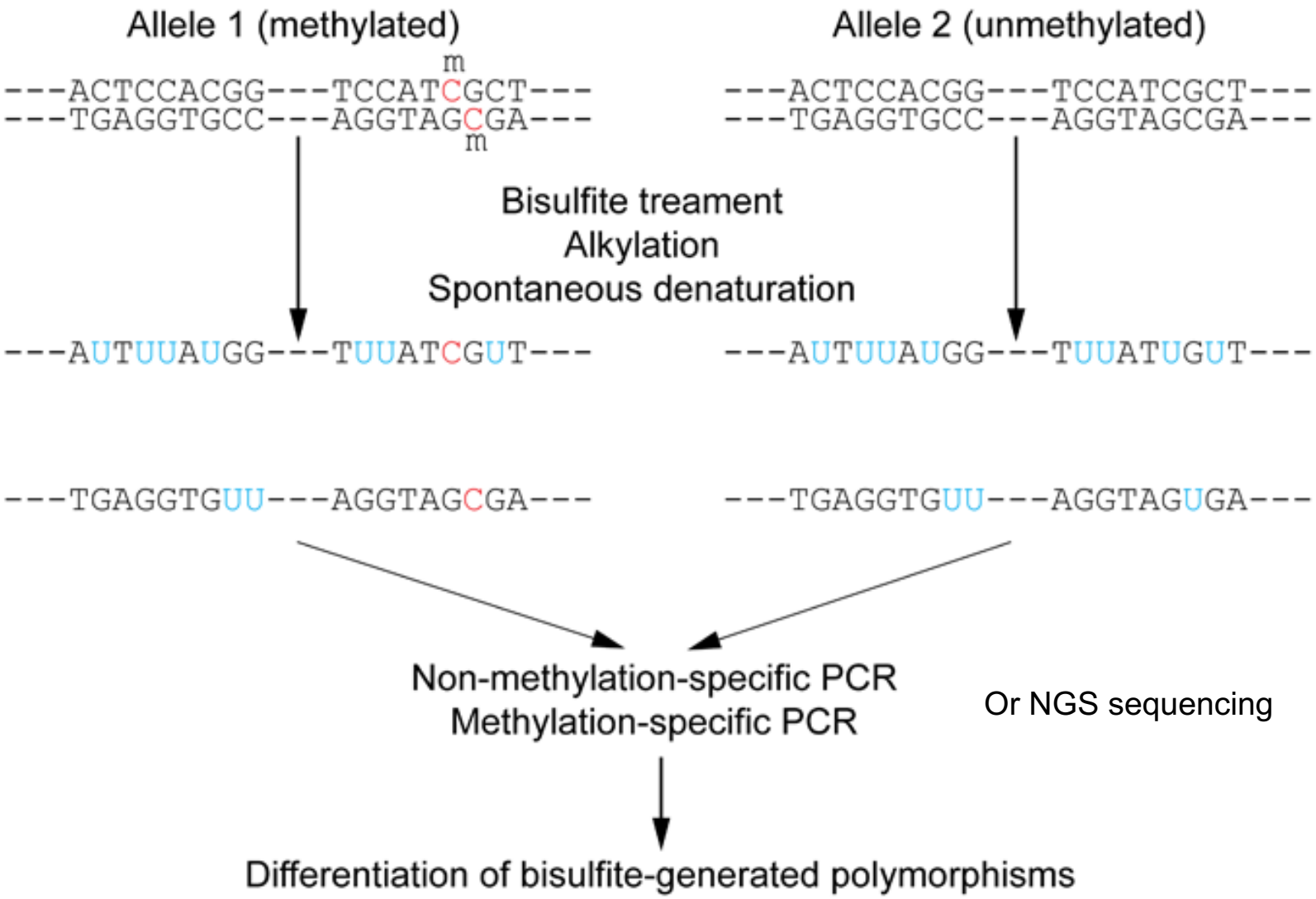
However, the **causality relationships** between expression levels and DNA methylation have not yet been completely determined.



Wagner et al.
Genome Biology (2014) **15**:R37

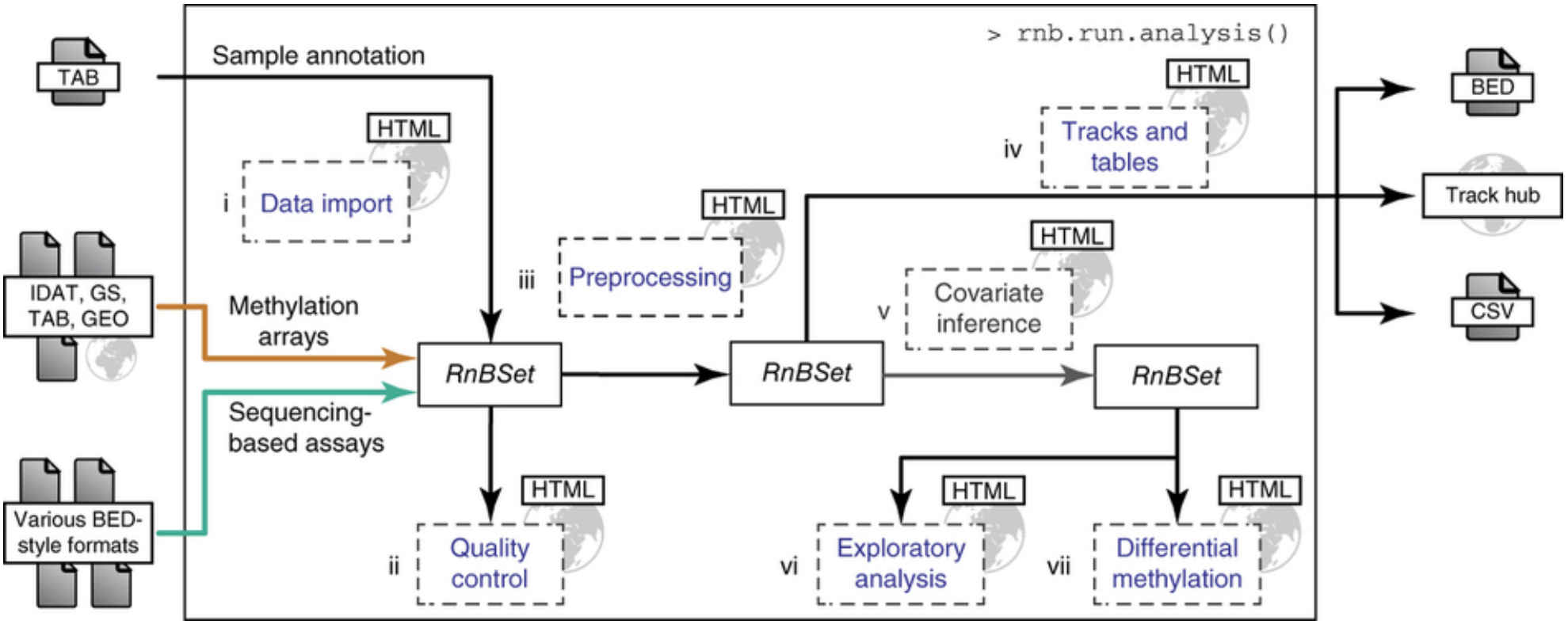
http://methhc.mbc.nctu.edu.tw

Detect DNA methylation by bisulfite conversion



www.wikipedia.org

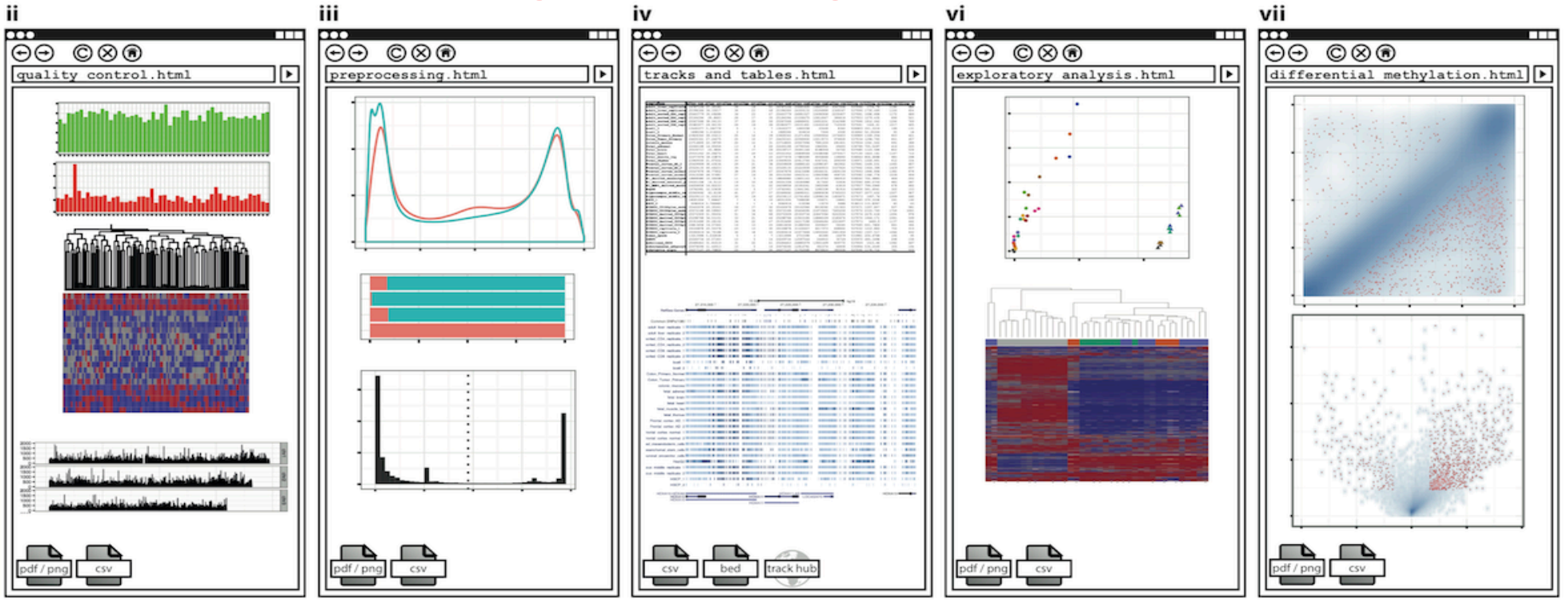
Processing of DNA methylation data with RnBeads



Left stages: processing of raw data (sequencing reads e.g. from bisulfite conversion)

Assenov et al. Nature Methods 11, 1138–1140 (2014)

DNA methylation analysis with RnBeads



Top: read coverage of CpGs

Distribution of beta-values

Bottom: „Volcano“ plot x-axis – difference of methylation site between 2 probes, y-axis – statistical significance of the difference;

Require enough variation and enough significance

Assenov et al. Nature Methods 11, 1138–1140 (2014)

Beta-values measure fractional DNA methylation levels

After analysis of raw sequencing data + filtering of problematic regions etc

the degree of methylation is typically expressed as

fractional **beta value:** %mCG(i) / (%mCG(i) + %CG(i))

A beta value for CpG position *i* takes on values between

0 (position *i* not methylated) and 1 (position *i* fully methylated)

Methylation levels of neighboring sites are correlated

- Observation: methylation levels of **neighboring CpG positions** within 1000 bp are often **correlated**;
- distance between neighboring CpGs is ca. 100 bp (1% frequency)

- Idea: exploit this effect to „smoothen“ experimental data, e.g. when this is obtained at low coverage

Master thesis of Junfang Chen (February 2014):

Journal of Bioinformatics and Computational Biology
Vol. 12, No. 6 (2014) 1442005 (16 pages)

DOI: 10.1142/S0219720014420050

Imperial College Press
www.icpress.co.uk

AKSmooth: Enhancing low-coverage bisulfite sequencing data via kernel-based smoothing

Junfang Chen*,†,‡, Pavlo Lutsik†, Ruslan Akulenko*,
Jörn Walter†,§ and Volkhard Helms*,§

*Center for Bioinformatics, Saarland University
Saarbrücken 66123, Germany

†Department of Genetics, Saarland University
Saarbrücken 66123, Germany
‡s9juchen@stud.uni-saarland.de

Correlated methylation of neighboring CpGs

$$\hat{f}_h(t) = \frac{\sum_i^N K_h(t,i) C_t(i) y_i}{\sum_i^N K_h(t,i) C_t(i)},$$

$$K_h(t,i) = K\left(\frac{|i-t|}{h}\right),$$

$$C_t(i) = \begin{cases} g_t & \text{if } i = t; \\ 1 & \text{if } i \neq t. \end{cases}$$

t : target CpG site

h : „band-width“: size of window
(# of neighboring CpGs around *t*)

y_i : methylation level of *i*-th CpG site within window of given size

C_t(i): weighting factor to consider read coverage of neighboring CpG sites relative to that of target site

K_h(t, i): Kernel function that considers the distance between positions *t* and *i*.

-> more distant positions get smaller weight.

Choice of kernel function

The kernel K

$$K_h(t,i) = D\left(\frac{|i-t|}{h}\right),$$

is either a standard Gaussian function

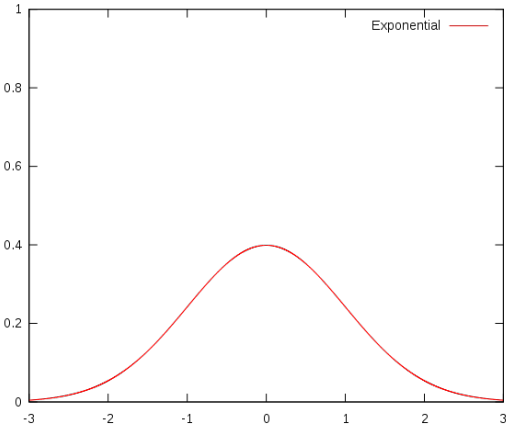
$$D(\mu) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2}$$

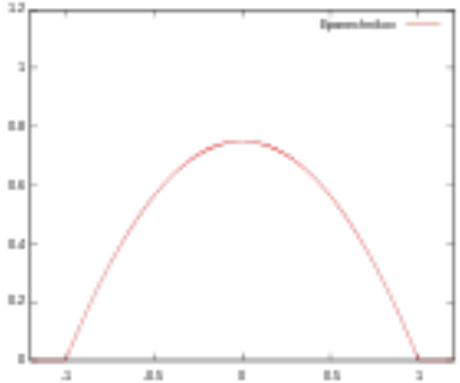
or the Epanechnikov kernel

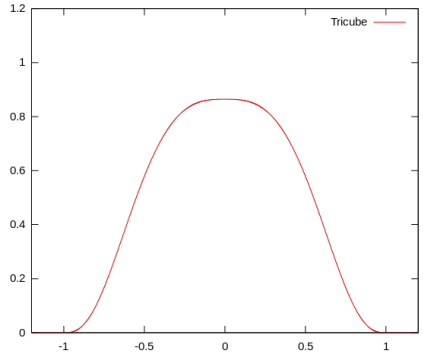
$$D(\mu) = \begin{cases} \frac{3}{4}(1-\mu^2) & \text{if } |\mu| \leq 1; \\ 0 & \text{otherwise} \end{cases}$$

or the tricubic kernel

$$D(\mu) = \begin{cases} \frac{70}{81}(1-|\mu|^3)^3 & \text{if } |\mu| \leq 1; \\ 0 & \text{otherwise.} \end{cases}$$

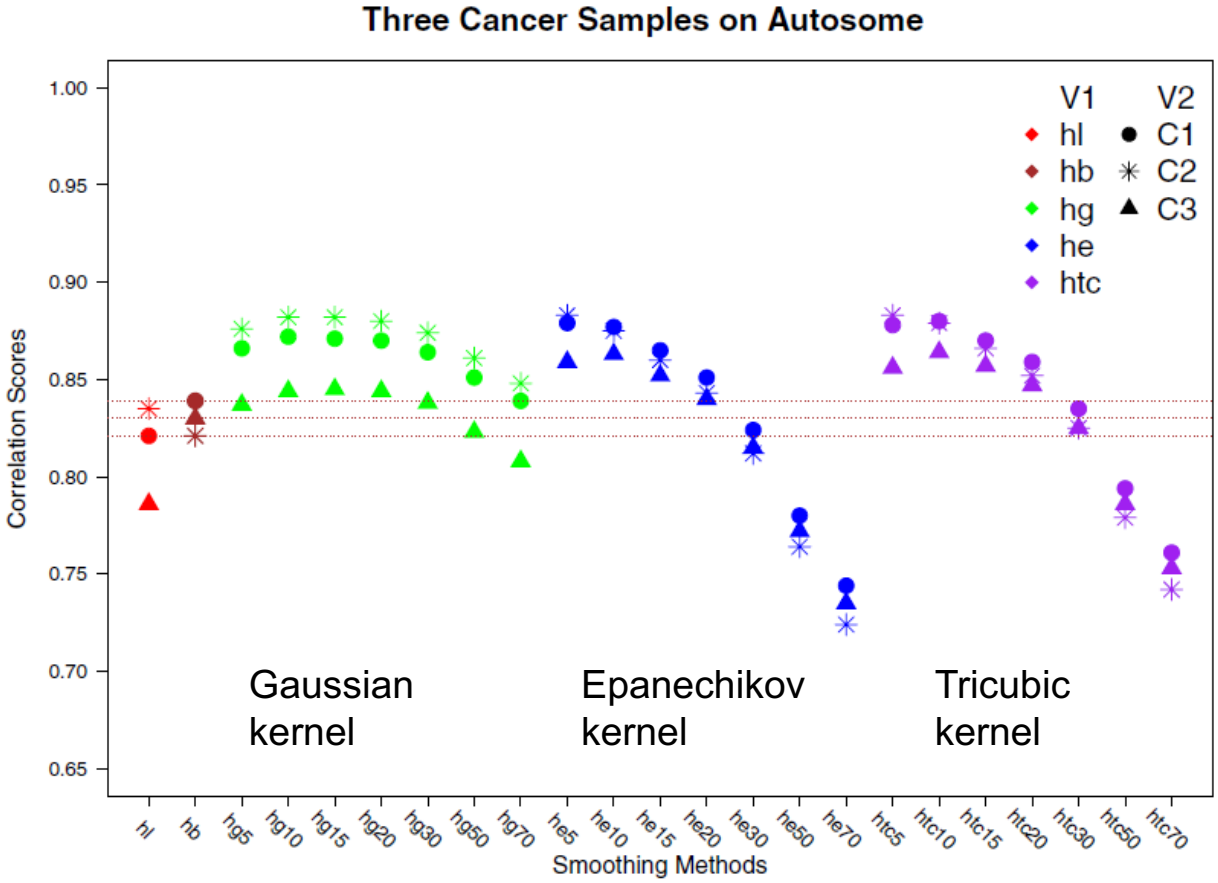






www.wikipedia.org

Correlation of low-coverage and high-coverage data



C1, C2, C3 are three different samples.

Best results for window considering nearby 10-20 CpGs.

Gaussian kernel („hg“) more robust with distance (exponential weighting).

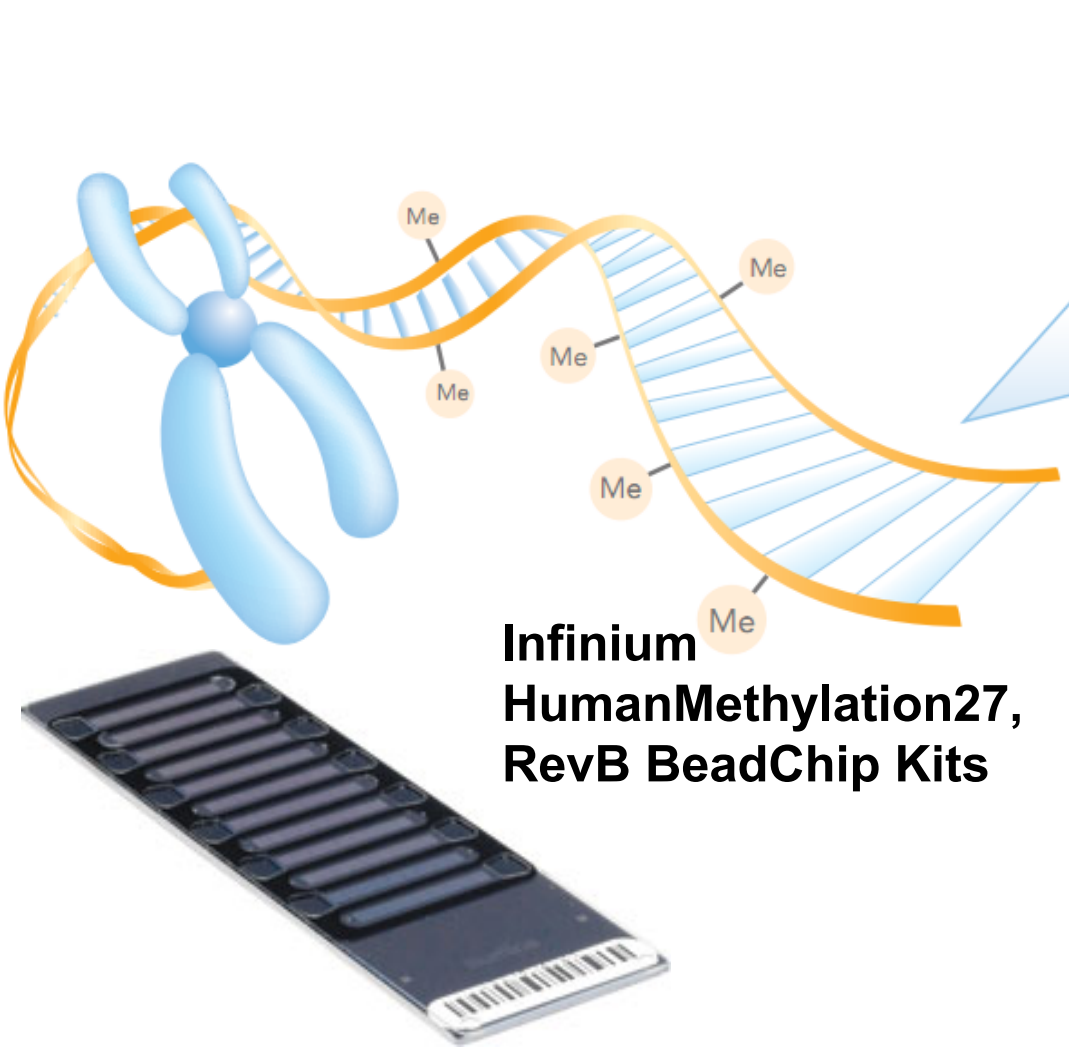
Tricubic and Epanechikov kernels show stronge decrease for large windows.

Every method was tested for including neighboring 5, 10, 15, … 70 CpGs.

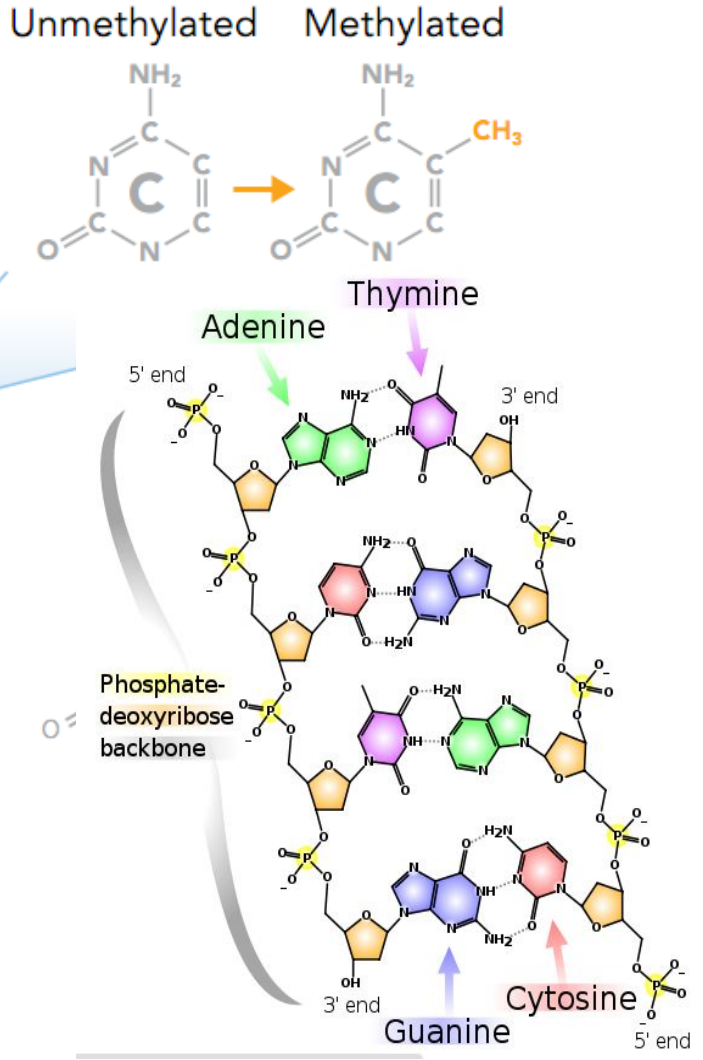
Red symbols „hl“ : low-coverage data (unsmoothened)

Brown symbols „hb“: low-coverage data processed with (another) Bsmooth-program

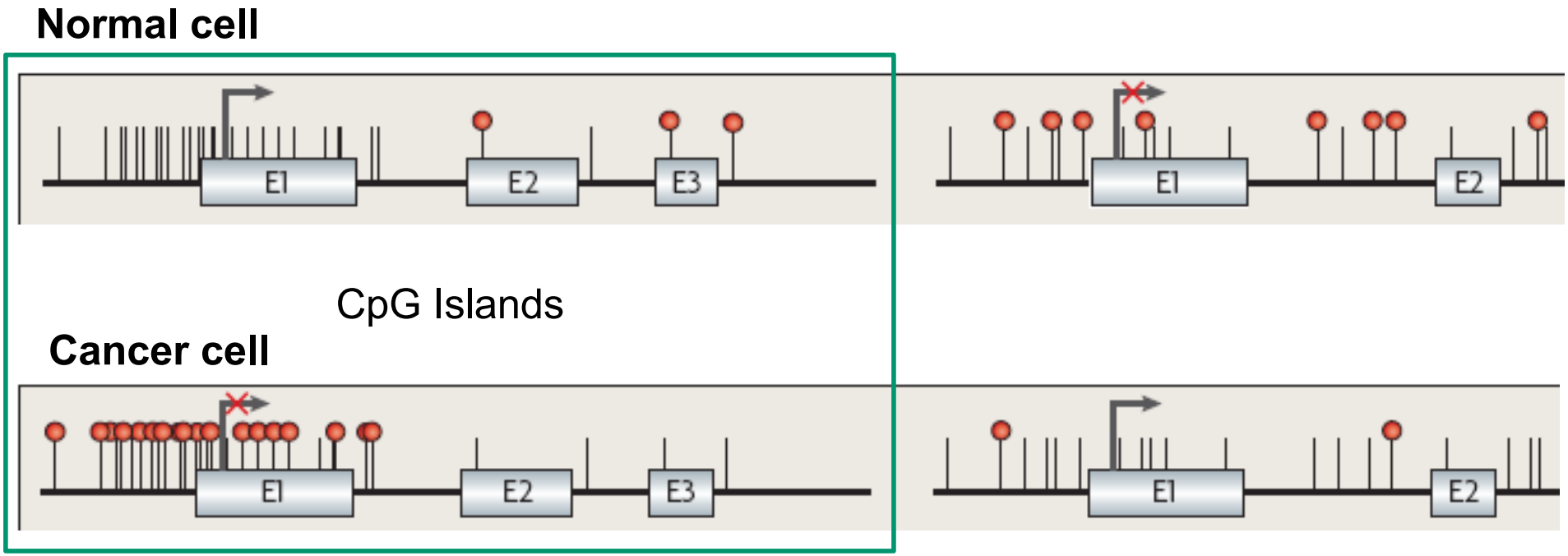
DNA methylation in breast cancer



Infinium HumanMethylation27, RevB BeadChip Kits



DNA methylation in cancer



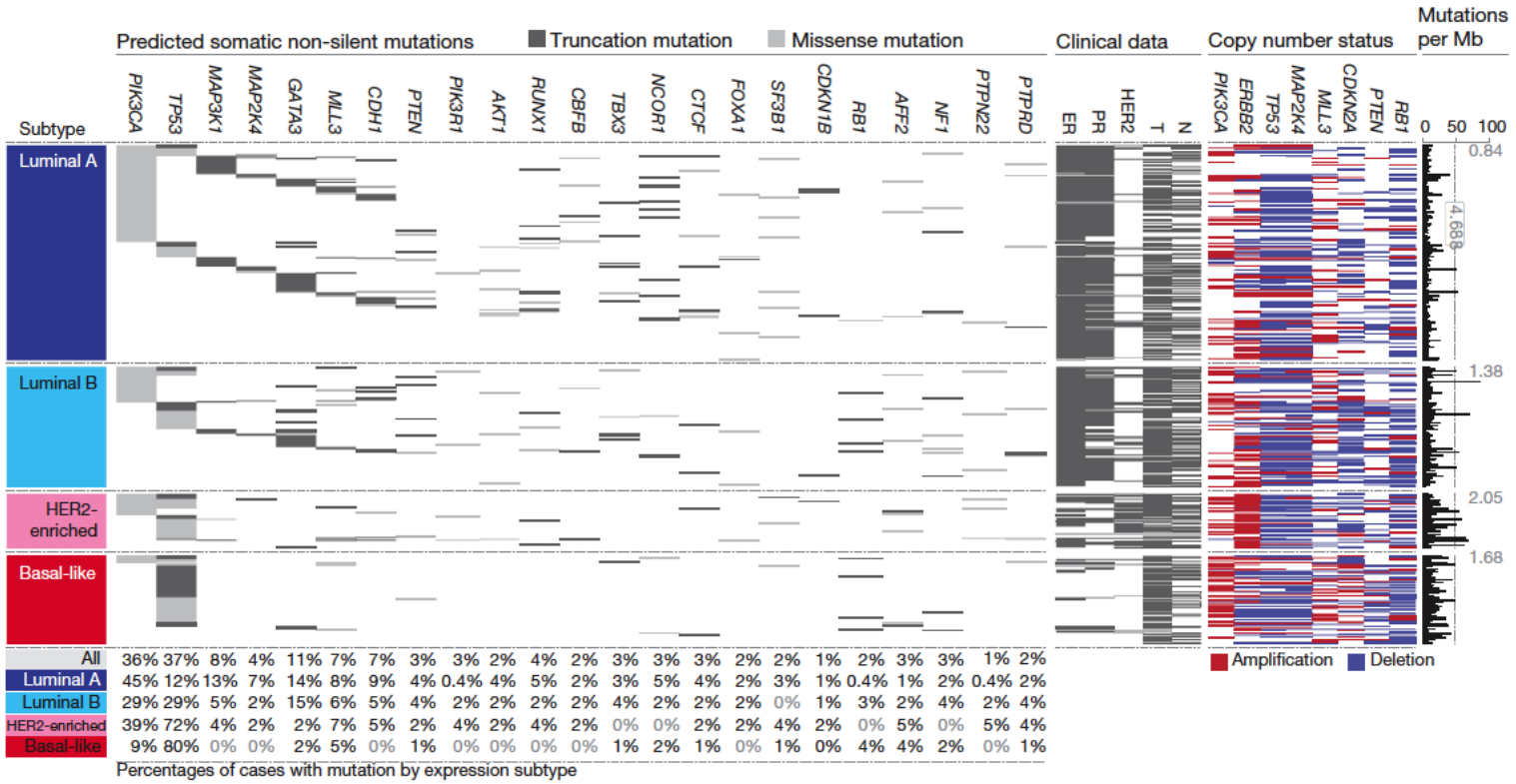
The Cancer Genome Atlas

ARTICLE

doi:10.1038/nature11412

Comprehensive molecular portraits of human breast tumours

The Cancer Genome Atlas Network*

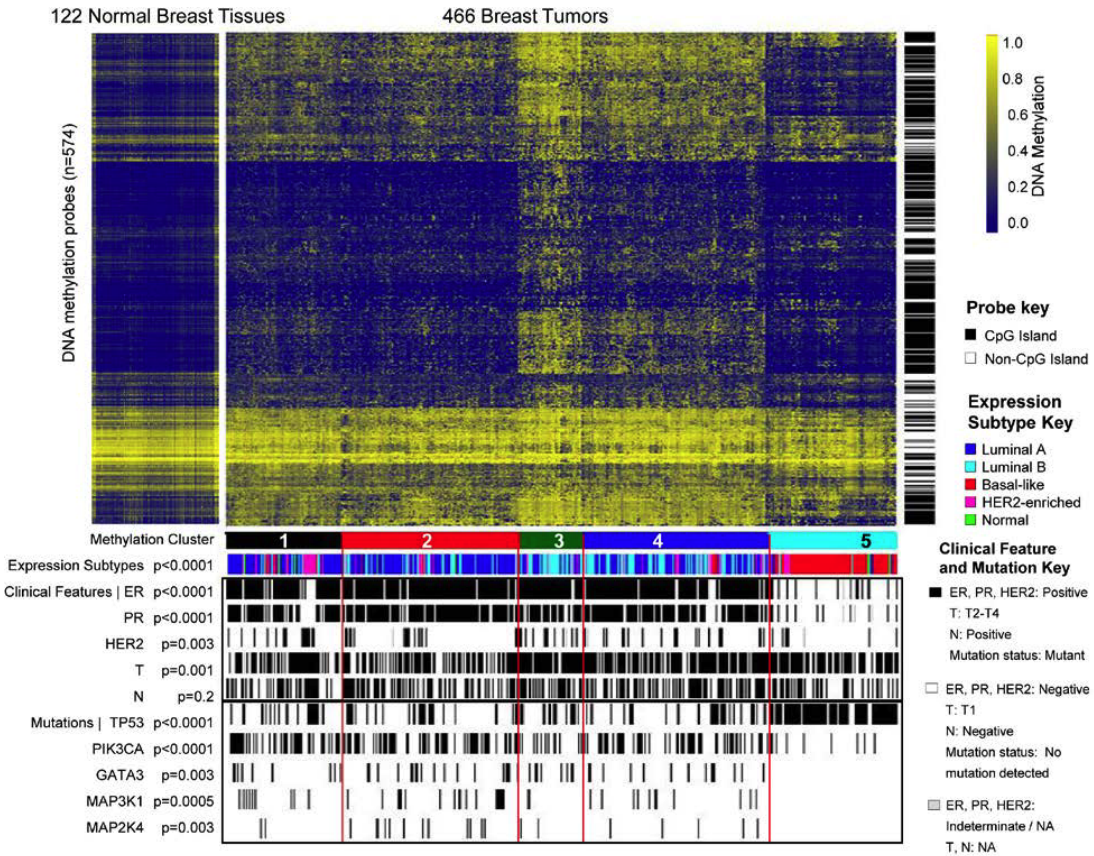


The Cancer Genome Atlas

DNA methylation

Illumina Infinium DNA methylation arrays were used to assay 802 breast tumours. Data from HumanMethylation27 (HM27) and HumanMethylation450 (HM450) arrays were combined and filtered to yield a common set of 574 probes used in an unsupervised clustering analysis, which identified five distinct DNA methylation groups (Supplementary Fig. 8). Group 3 showed a hypermethylated phenotype and was significantly enriched for luminal B mRNA subtype and under-represented for *PIK3CA*, *MAP3K1* and *MAP2K4* mutations. Group 5 showed the lowest levels of DNA methylation, overlapped with the basal-like mRNA subtype, and showed a high frequency of *TP53* mutations. HER2-positive ($HER2^+$) clinical status, or the HER2E mRNA subtype, had only a modest association with the methylation subtypes.

A supervised analysis of the DNA methylation and mRNA expression data was performed to compare DNA methylation group 3 ($N = 49$) versus all tumours in groups 1, 2 and 4 (excluding group 5, which consisted predominantly of basal-like tumours). This analysis identified 4,283 genes differentially methylated (3,735 higher in group 3 tumours) and 1,899 genes differentially expressed (1,232 downregulated); 490 genes were both methylated and showed lower expression in group 3 tumours (Supplementary Table 4). A DAVID (database for annotation, visualization and integrated discovery) functional annotation analysis identified 'extracellular region part' and 'Wnt signalling pathway' to be associated with this 490-gene set; the group 3 hypermethylated samples showed fewer *PIK3CA* and *MAP3K1* mutations, and lower expression of Wnt-pathway genes.



Supplemental Figure 8. DNA methylation subtypes and comparison to normal breast tissues. DNA methylation cluster membership was determined by a Recursively Partitioned Mixture Model (RPMM) for 466 breast tumors using 574 selected probes and compared to 122 breast normal samples in the same probe order. DNA methylation levels (beta value) are shown with a color spectrum; blue, no methylation to yellow, full methylation. Cluster memberships are indicated by the horizontal color bar: black Cluster 1 (n=80); red Cluster 2 (n=123); green Cluster 3 (n=44) blue Cluster 4 (n=128); cyan Cluster 5 (n=91). Molecular and clinical features as indicated in the color key. P-values for association with molecular and clinical features were calculated using a Chi-square test or Fisher's exact test, wherever applicable.

11.2 Differential methylation analysis

After quantification of methylation levels, one typically detects **differentially methylated regions (DMRs)** that show consistent differences between sample groups (e.g. cases versus controls).

Length of DMRs ranges from a single cytosine base to an entire gene locus.

In some cases a single methylated CpG may be involved in regulating gene expression and may thus affect disease risk.

The vast majority of known DMRs have a size between a few hundred and a few thousand bases. This range matches that of gene-regulatory regions.

It is assumed that DMRs can regulate transcriptional repression of an associated gene in a cell-type-specific manner.

11.2 Differential methylation analysis

Given sufficient data for 2 groups of samples, DMRs can be detected by t-tests or Wilcoxon rank-sum tests (see differential expression analysis, V10).

Importantly, when differences in DNA methylation are detected by a statistical test at a large number of genomic loci, the results need to be corrected for **multiple hypothesis testing** so that a false-discovery rate is inferred for each DMR.

As there exists a large number of CpGs in the genome, often only the most pronounced single-CpG differences are kept as significant after such an adjustment.

11.2 Differential methylation analysis

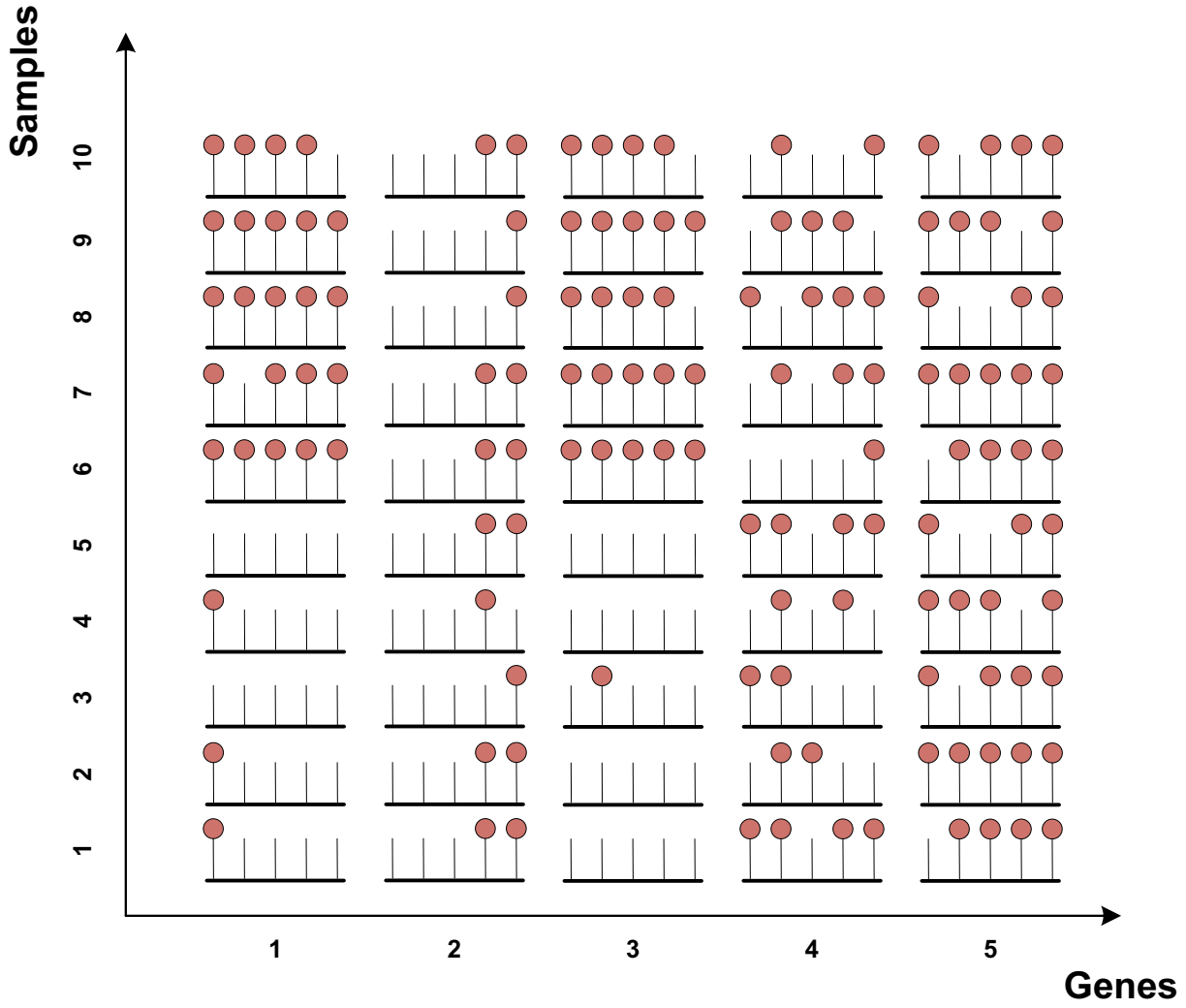
One can apply 2 complementary strategies to enhance the statistical power while detecting weak differences in DNA methylation.

(1) one can apply the statistical tests to **longer genomic regions** rather than to individual CpG sites. (Reason: there are much fewer of them. Not so much statistical power is lost due to multiple testing correction.)

If neighbouring CpGs show similar differences of DNA methylation levels, this reduced „resolution“ leads to more significant results.

(2) small standard deviations frequently arise by chance and may yield spurious results.
When the standard deviation of a given CpG or genomic region is estimated by taking the average of observed and expected values, more robust p-values can be obtained for DNA methylation comparisons with many measurements and few samples per sample group.

Idea: identify co-methylation of genes in TCGA samples



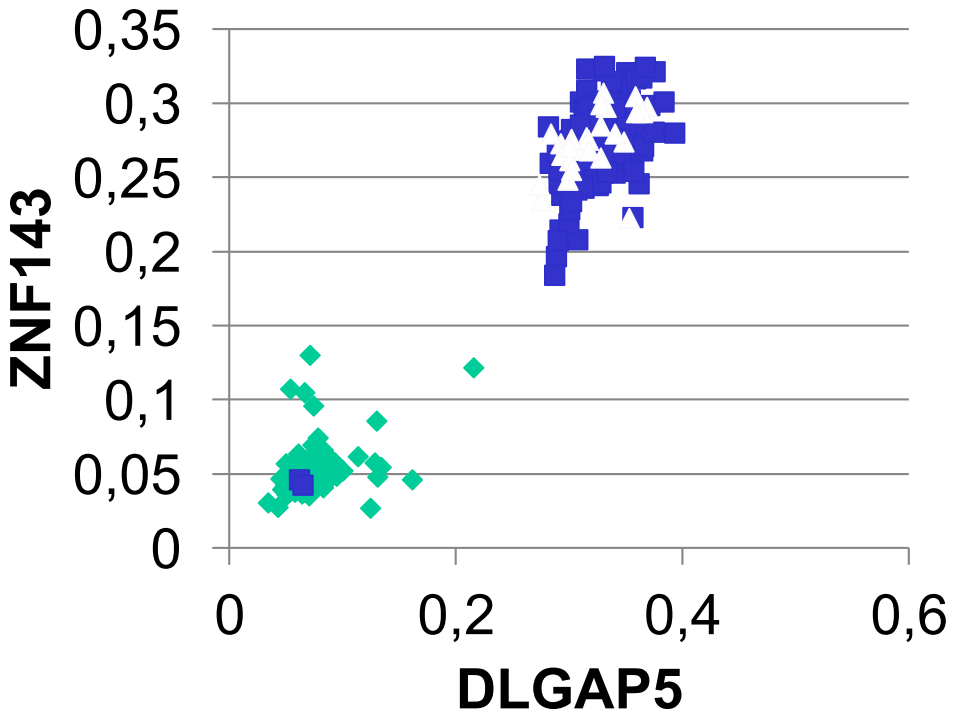
Co-methylation of genes 1 and 3 across samples

Tumor data

Data Type (Base-Specific)	Level 1 (Raw Data)	Level 2 (Normalized/ Processed)	Level 3 (Segmented/ Interpreted)	Level 4 (Summary Finding/ROI)
DNA Methylation	Raw signals per probe	Normalized signals per probe or probe set and allele calls	Methylated sites/genes per sample	Statistically significant methylated sites/genes across samples

- 183 tumor samples deposited in Sept 2011 (tumor group 1);
- 134 tumor samples deposited in Oct 2011 (tumor group 2) and
- 27 matched normal samples from Oct 2011.

Difficulties: batch effect

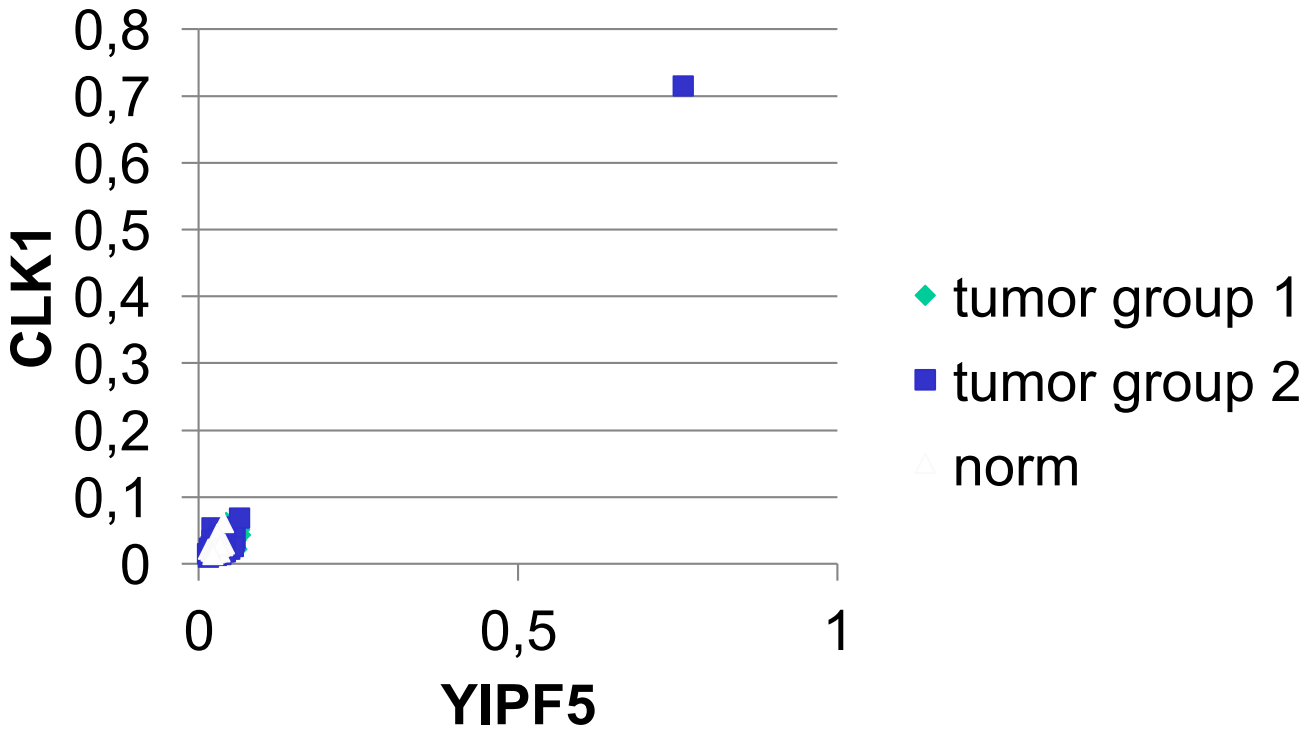


$$beta = \frac{M}{M + U}$$

- tumor group 1 — Sept. 2011
- tumor group 2 — Oct. 2011
- norm

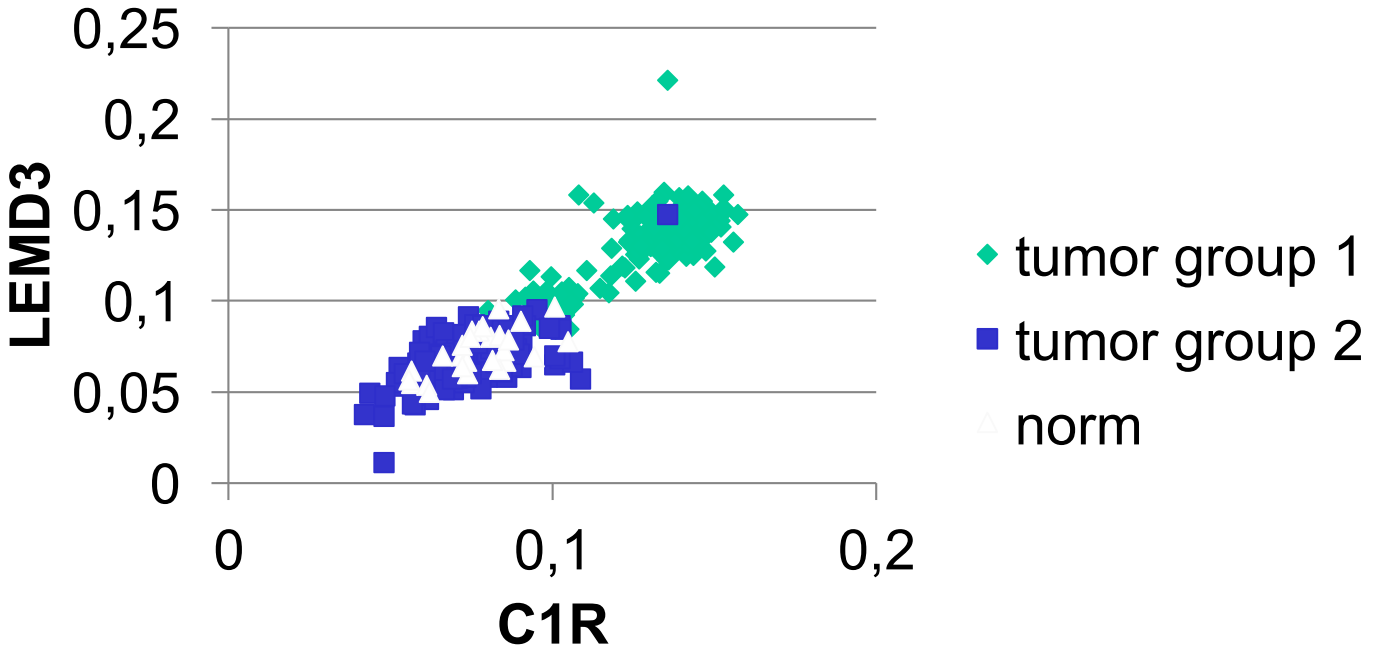
Filter 1: delete genes affected by batch effect

Difficulties: outliers



Filter 2: require zero outliers

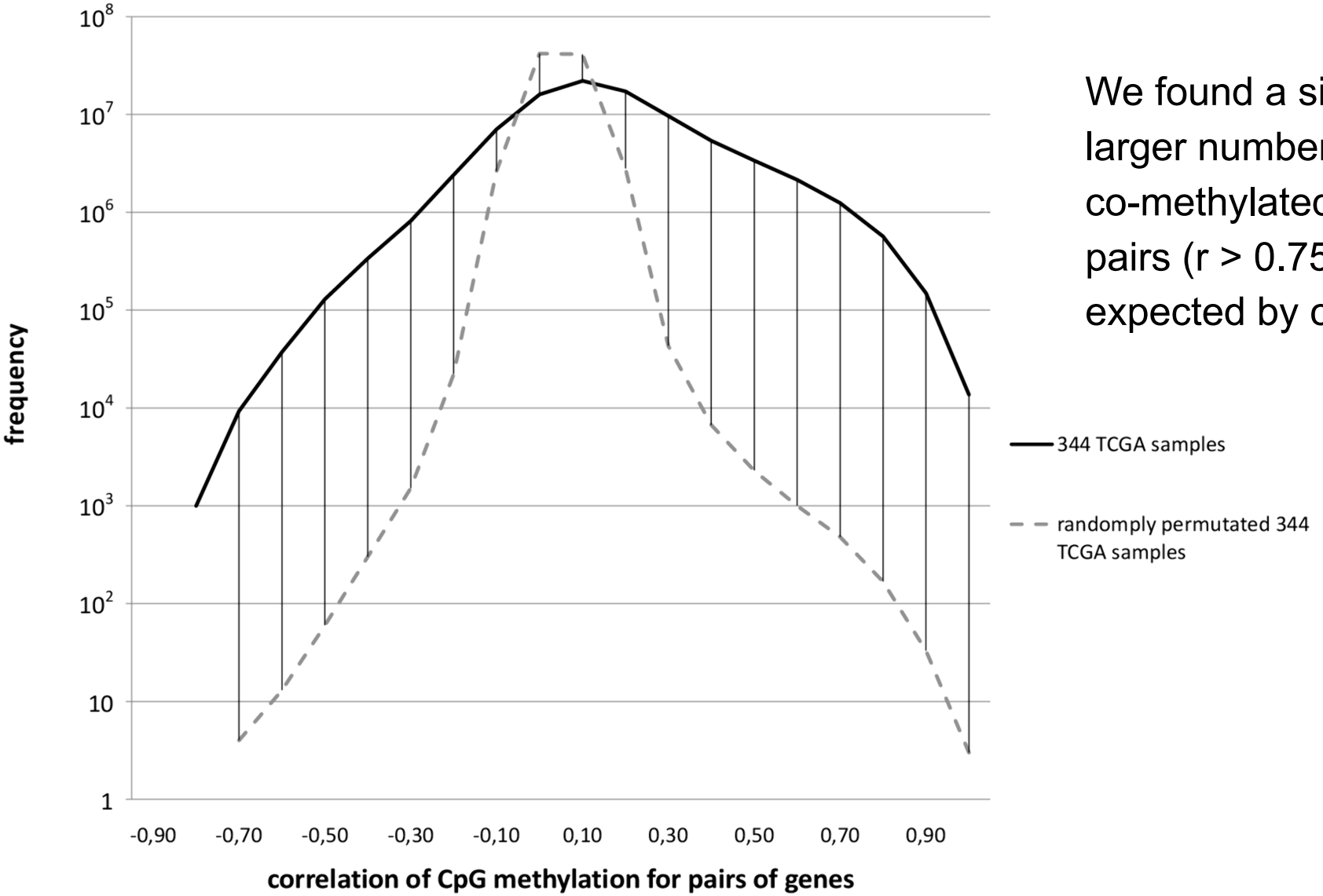
Difficulties: low variance



Filter 3: delete genes with low variance

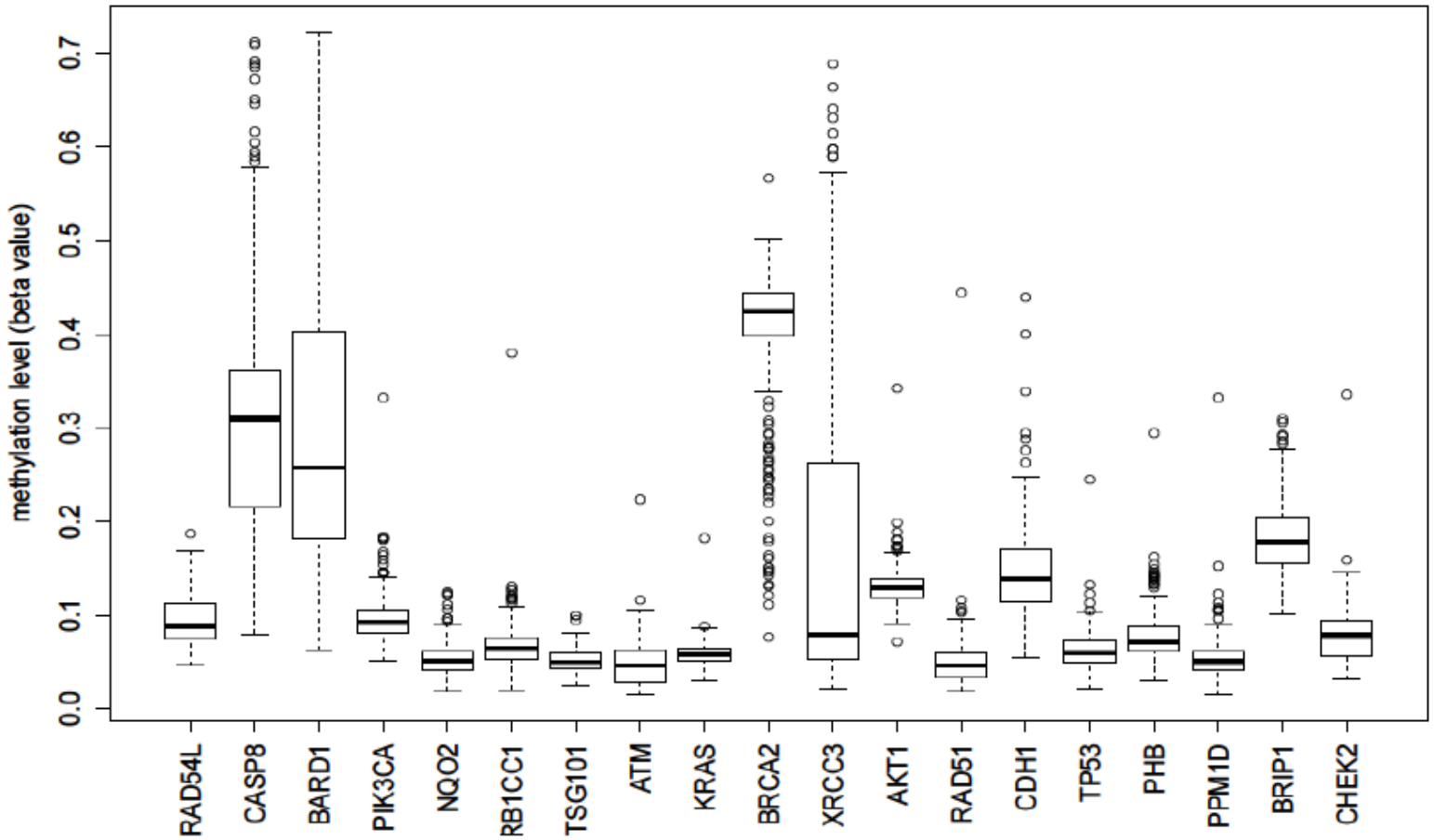
$$\underset{i \in T}{quartile3}(beta_i) - \underset{i \in T}{quartile1}(beta_i) \geq 0{,}1$$

Comparison against randomized data



We found a significantly larger number of co-methylated gene pairs ($r > 0.75$) than expected by chance.

Known breast cancer genes in OMIM: mostly unmethylated



These 19 genes are associated with breast cancer in the Online version of the Mendelian Inheritance in Man (OMIM) database.

They are not involved in co-methylation because most of them show little changes of their (low) methylation levels

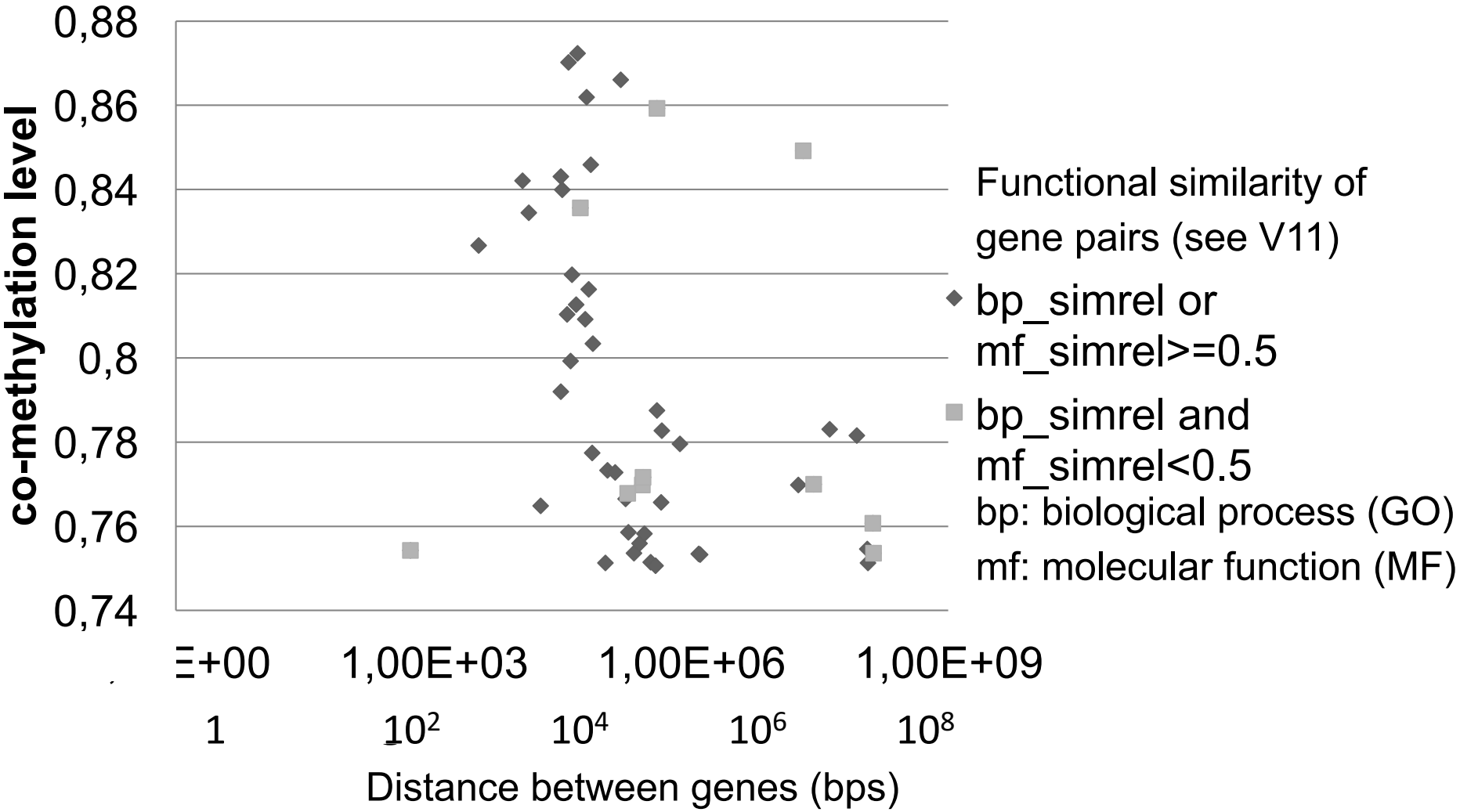
top 10 co-methylated gene pairs

First gene	Second gene	Pearson correlation	Related genes?
SPRR1B	SPRR1A	0,872	Yes
FCN2	FCN1	0,870	Yes
CD244	CD48	0,866	Yes
SPRR1B	SPRR4	0,862	Yes
TAS2R13	PRB4	0,859	**No**
F7	TFF1	0,856	**No**
SH3TC2	SPARCL1	0,853	**No**
ABCE1	SC4MOL	0,849	**No**
REG1B	REG1P	0,846	Yes
SPRR3	SPRR4	0,843	Yes

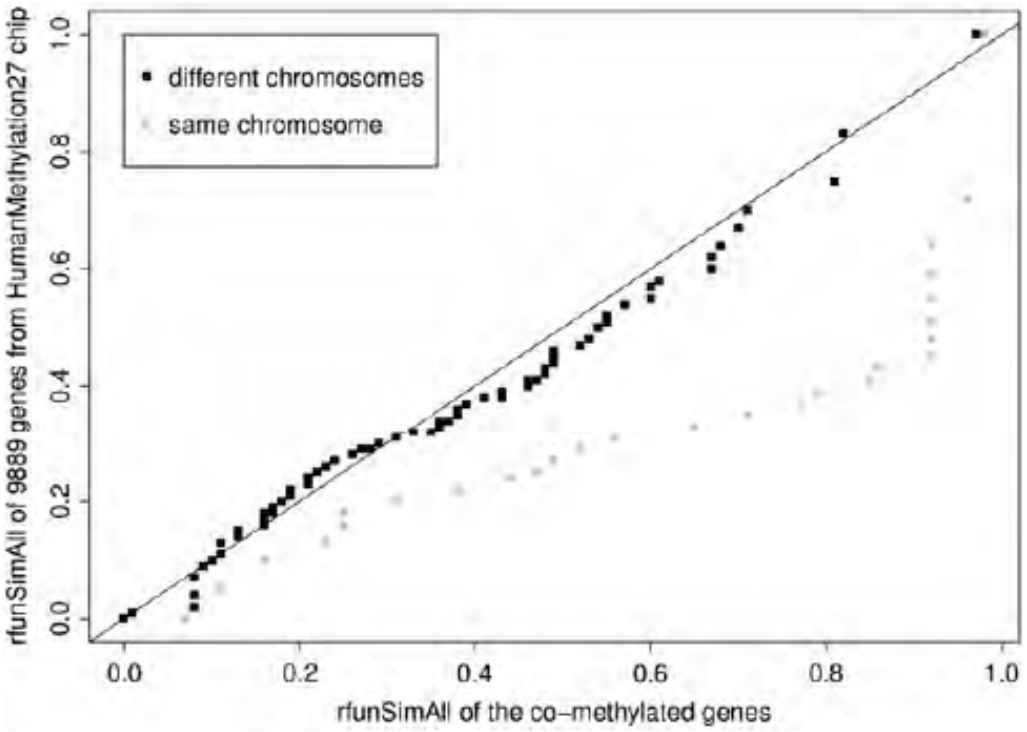
Some genes have related names -> co-methylation may be expected

Are all co-methylated genes neighbors?

Less than half of all co-methylated gene pairs lie on the same chromosome



Functional similarity of co-methylated genes



Co-methylated gene pairs on the same chromosome have higher functional similarity (determined by FunSimMat) than between random pairs of genes

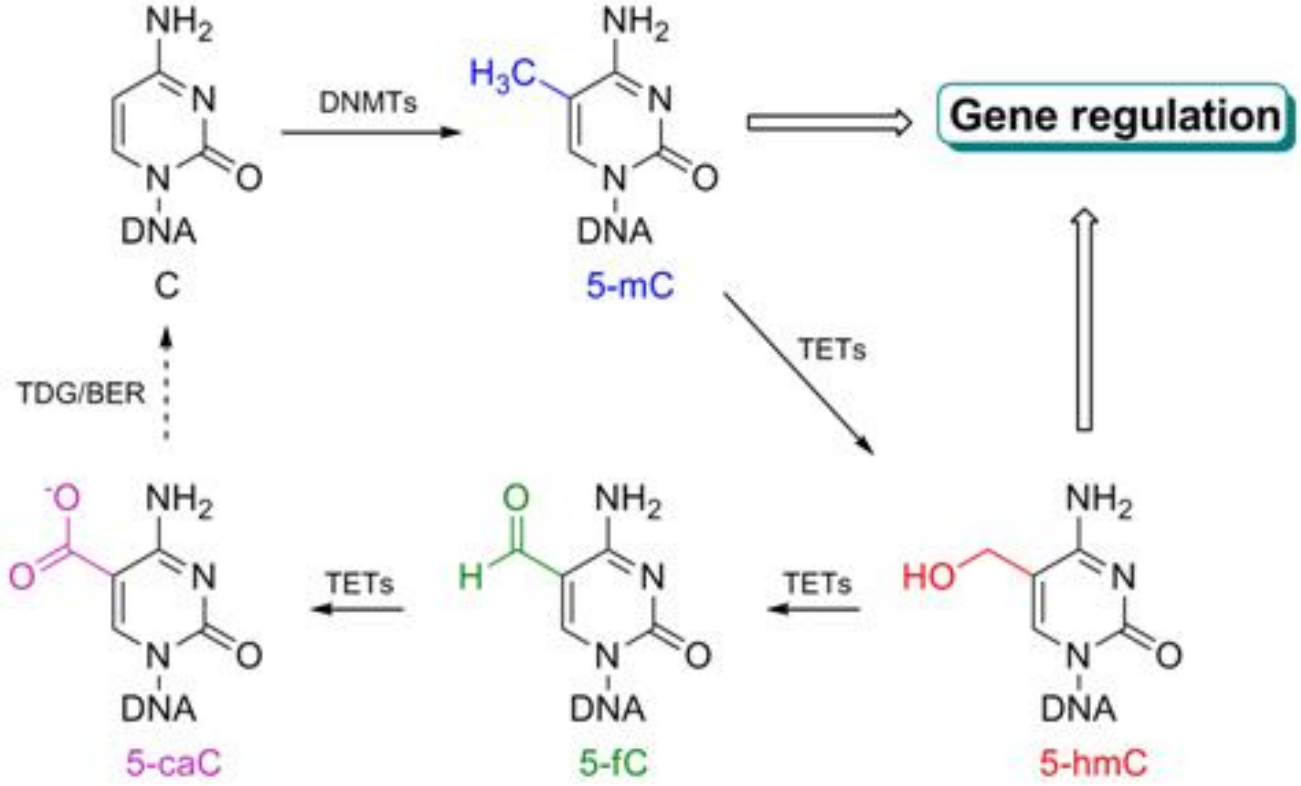
Not the case for co-methylated gene pairs on different chromosomes

Enriched pathways in co-methylated gene clusters

Cluster ID	KEGG pathways	p-value	Genes involved in pathways	FDR
8	hsa04950:Maturity onset diabetes of the young	0.003	HNF1B, FOXA2, NEUROD1	2.622
9	hsa04640:Hematopoietic cell lineage	0.009	CD1A, CD1E, CD1D	6.229
15	hsa04730:Long-term depression	0.004	GRM5, C7ORF16, PRKG2	2.952
22	hsa04060:Cytokine-cytokine receptor interaction	0.047	EGF, TNFSF18, IL20	31.263
27	hsa04512:ECM-receptor interaction	0.005	COL5A2, COL11A1, SPP1	3.500
27	hsa04510:Focal adhesion	0.029	COL5A2, COL11A1, SPP1	17.498

Table S2. The results of pathway enrichment analysis of 29 gene clusters obtained using DAVID. These clusters were formed by applying Affinity Propagation clustering to 779 genes, which were left after three-stage filtered of all 13,313 genes from methylation data samples.

Further modifications of cytosine bases



Further modifications were discovered in the last few years. They are present in cells in much smaller fractions than 5-mC.

Tet enzymes catalyze the conversions.
The biological roles of these modifications are mostly unclear.

http://he-group.uchicago.edu

Summary

DNA methylation and histone marks are epigenetic modifications of genomic DNA and nucleosomes that appear to have regulatory roles in a broad range of biological processes and diseases.

Detection of **DMRs** allows to distinguish and classify different developmental stages of cell differentiation or to distinguish tumor tissue from normal tissue.

DNA methylation levels are generally higher in condensed chromatin regions and in differentiated cells than in open chromatin regions and in stem cells.

Our understanding of the relationship between epigenetic modifications and their effects on gene expression levels is still limited.

DNA methylation levels of **promoter regions** only show **weak anticorrelation** of around 0.15 with the expression levels of the respective genes.