Ilots CpG et méthylation

CpG

dinucléotides C-G sur le même brin : CpG

la distribution de CpG varie dans le génome humaine (et mammifères en général)

juste avant le début d'un gène (ATG), les CpG sont souvent dense

dans le génome, très rares⋆

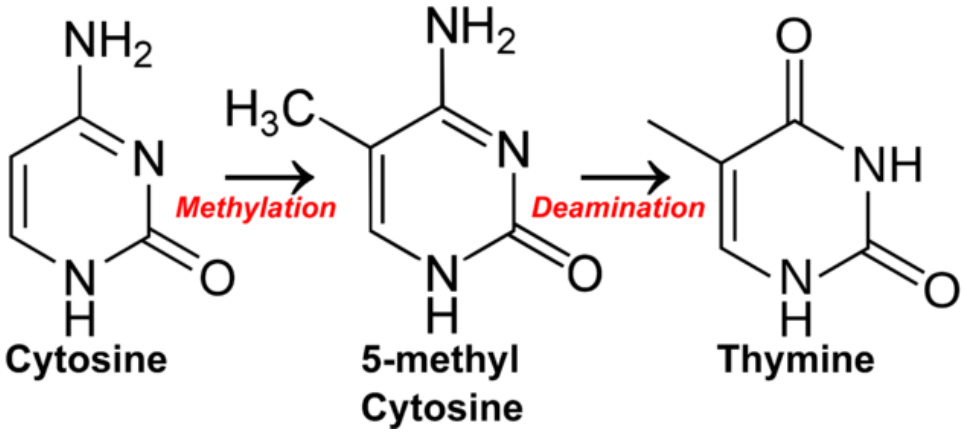
⋆ dans le modèle iid, la fréquence de GpC $= \pi_C \pi_G$

	nucléotides	composition	GpC	CpG
chrX	151.10M	C = 29.81M, G = 29.87M (19.7%)	5.94M (3.9%)	1.25M (0.8%)
chr12	130.5M	C = 26.63M, G = 26.61M (20.4%)	5.54M (4.2%)	1.28M (1.0%)
chr5	177.7M	C = 35.09M, G = 35.13M (19.8%)	7.11M (4.0%)	1.51M (0.8%)
chr1	225.3M	C = 47.02M, G = 47.02M (20.9%)	9.95M (4.4%)	2.28M (1.0%)

Wikipedia

Méthylation

Typiquement, la cytosine de CpG est méthylée, et peut se transformer en thymine facilement



Wikipedia

Ilots CpG

enrichissement dans une région de longueur ℓ : compter les occurrences n.

$$\mathsf{CpG(o/e)} = \ell \frac{n_{\mathtt{C}p\mathtt{G}}}{n_{\mathtt{C}} \cdot n_{\mathtt{G}}}$$

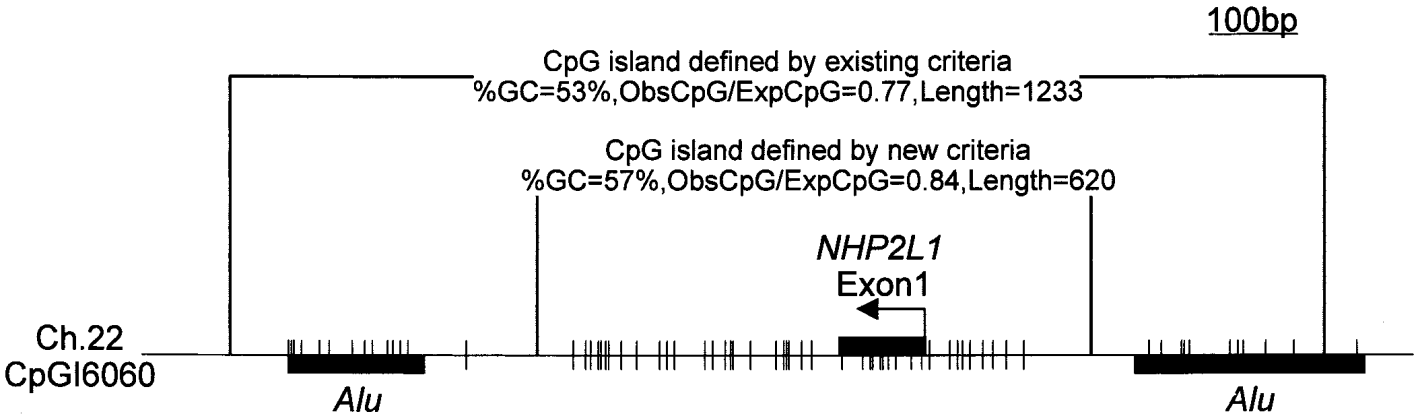
ilot CpG (*CpG island*) : définitions par $(\mathtt{G}+\mathtt{C})\%$ élevée, enrichissement de CpG, longueur minimale

P.e. : $(\mathtt{G}+\mathtt{C})\% \geq 50\%$, $\ell \geq 200$, $\mathsf{CpG(o/e)} \geq 0.6$

Ilot CpG : paramètres

→ paramètres différents . . .

→ CpG fréquent dans éléments mobiles (p.e., Alu)



Overview of CpG island prediction algorithms.

Database/prediction	Length	G + C	CpG[*o*/*e*]	RM[a]	Comments	Reference
ENSEMBL	⩾400	⩾50%	⩾0.6	N	Stringent length constraint	[88]
NCBI relaxed	⩾200	⩾50%	⩾0.6	N	Total CGIs = 307 193	
NCBI strict	⩾500	⩾50%	⩾0.6	N	Total CGIs = 24 163	
USCS[b]	>200	⩾50%	>0.6	Y	Total CGIs = 28 226	[89]
EMBOSS	UD[c]	UD	UD	NA	Variable parameters	[90]
CpGProD	>500	>50%	>0.6	Y	Total CGIs = 76 793	[23]
CpGcluster	NA	NA	NA	N	Clustering Total = 197 727	[25]

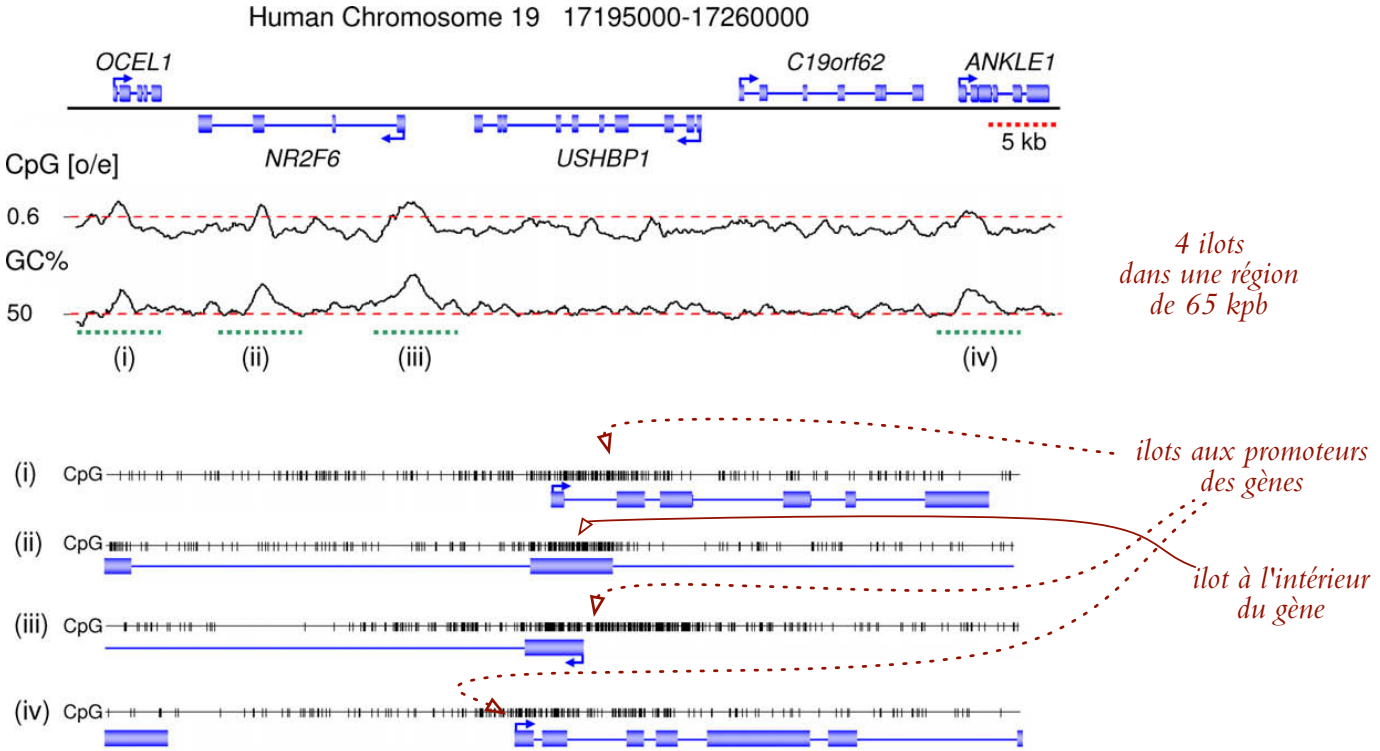
[a] RM, repeat masked; Y, yes; N, no; NA, non applicable.
[b] Parameters used for CGI identification for the ENCODE project although totals vary due to repeat masking differences between hg17 and hg18 builds [87].
[c] UD, user defined.

Takai & Jones *PNAS* 99 :3740 (2002) ; Illingworth & Bird *FEBS Lett* 583 :1713 (2009)

Ilots CpG : distribution

→ longueur ∼ 1kpb

→ parfois se trouvent à l'intérieur d'un gène ou dans une région intergénique



Illingworth & Bird *FEBS Lett* 583 :1713 (2009)

Méthylation et transcription

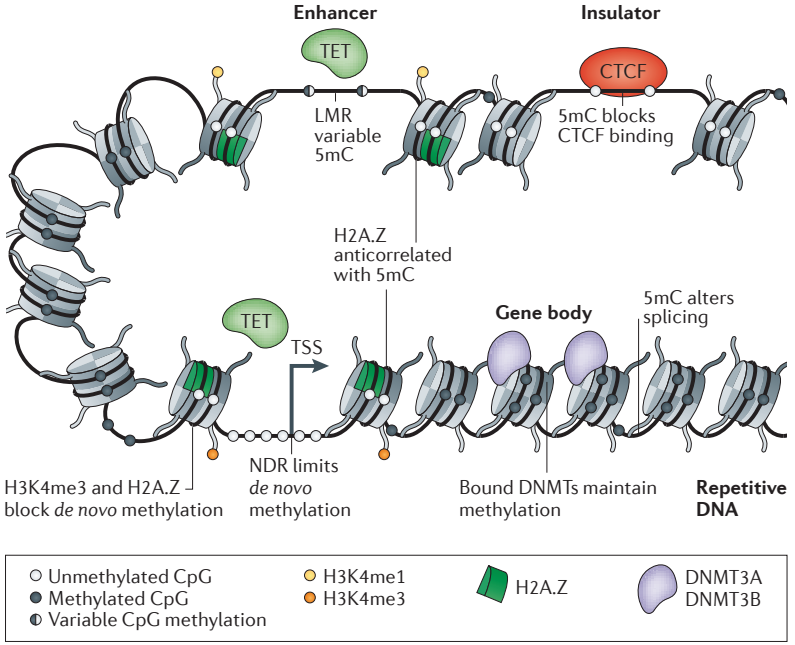


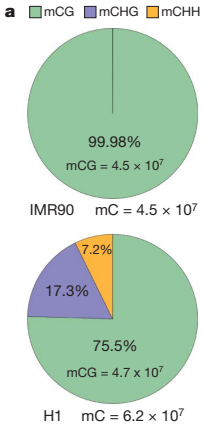
Figure 1 | **Molecular anatomy of CpG sites in chromatin and their roles in gene expression.** About 60% of human genes have CpG islands (CGIs) at their promoters and frequently have nucleosome-depleted regions (NDRs) at the transcriptional start site (TSS). The nucleosomes flanking the TSS are marked by trimethylation of histone H3 at lysine 4 (H3K4me3), which is associated with active transcription, and the histone variant H2A.Z, which is antagonistic to DNA methyltransferases (DNMTs). Downstream of the TSS, the DNA is mostly CpG-depleted and is predominantly methylated in repetitive elements and in gene bodies. CGIs, which are sometimes located in gene bodies, mostly remain unmethylated but occasionally acquire 5-methylcytosine (5mC) in a tissue-specific manner (not shown). Transcription elongation, unlike initiation, is not blocked by gene body methylation, and variable methylation may be involved in controlling splicing. Gene bodies are preferential sites of methylation in the context CHG (where H is A, C or T) in embryonic stem cells[5], but the function is not understood (not shown). DNA methylation is maintained by DNMT1 and also by DNMT3A and/or DNMT3B, which are bound to nucleosomes containing methylated DNA[99]. Enhancers tend to be CpG-poor and show incomplete methylation, suggesting a dynamic process of methylation or demethylation occurs, perhaps owing to the presence of ten-eleven translocation (TET) proteins in these regions, although this remains to be shown. They also have NDRs, and the flanking nucleosomes have the signature H3K4me1 mark and also the histone variant H2A.Z[32,100]. The binding of proteins such as CTCF to

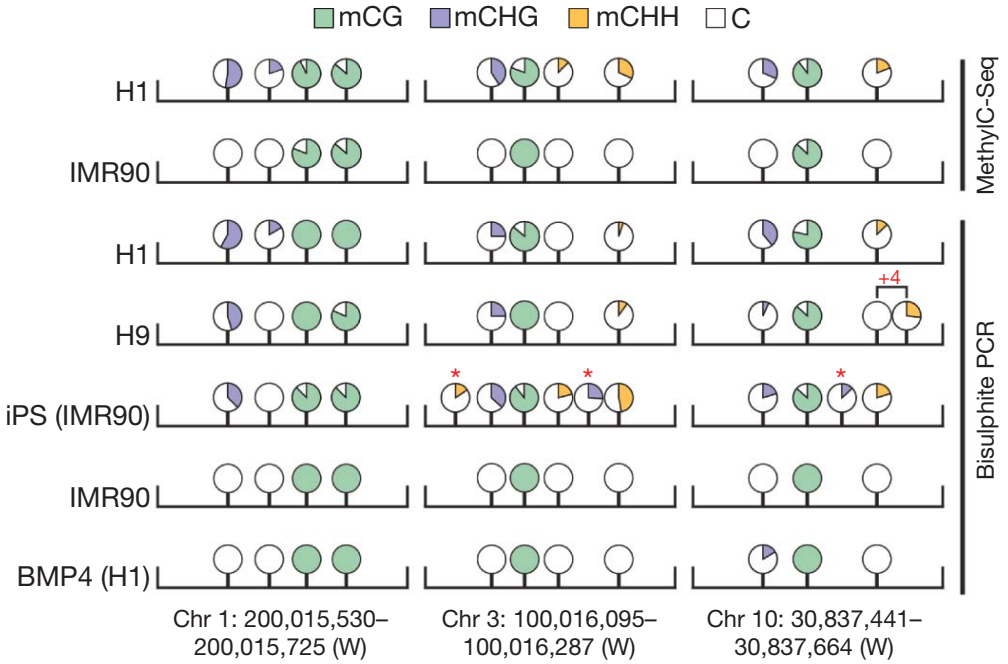
Jones *Nat Rev Genet* 13 :484 (2012)

Méthylation et cellules souches

→ méthylation de cytosine aussi à de sites CH ($\mathtt{H} = \{\mathtt{A}, \mathtt{C}, \mathtt{T}\}$)

→ change lors de la différentiation de cellule





(H1 = cellules souches ; IMR90=lung fibroblast)

Lister, Pelizzola & al *Nature* 462 :315 (2009)

Détection de méthylation

Table 1 | **Main principles of DNA methylation analysis**

Pretreatment	Analytical step			
	Locus-specific analysis	**Gel-based analysis**	**Array-based analysis**	**NGS-based analysis**
Enzyme digestion	• *Hpa*II-PCR	• Southern blot • RLGS • MS-AP-PCR • AIMS	• DMH • MCAM • HELP • MethylScope • CHARM • MMASS	• Methyl–seq • MCA–seq • HELP–seq • MSCC
Affinity enrichment	• MeDIP-PCR		• MeDIP • mDIP • mCIP • MIRA	• MeDIP–seq • MIRA–seq
Sodium bisulphite	• MethyLight • EpiTYPER • Pyrosequencing	• Sanger BS • MSP • MS-SNuPE • COBRA	• BiMP • GoldenGate • Infinium	• RRBS • BC–seq • BSPP • WGSBS

AIMS, amplification of inter-methylated sites; BC–seq, bisulphite conversion followed by capture and sequencing; BiMP, bisulphite methylation profiling; BS, bisulphite sequencing; BSPP, bisulphite padlock probes; CHARM, comprehensive high-throughput arrays for relative methylation; COBRA, combined bisulphite restriction analysis; DMH, differential methylation hybridization; HELP, *Hpa*II tiny fragment enrichment by ligation-mediated PCR; MCA, methylated CpG island amplification; MCAM, MCA with microarray hybridization; MeDIP, mDIP and mCIP, methylated DNA immunoprecipitation; MIRA, methylated CpG island recovery assay; MMASS, microarray-based methylation assessment of single samples; MS-AP-PCR, methylation-sensitive arbitrarily primed PCR; MSCC, methylation-sensitive cut counting; MSP, methylation-specific PCR; MS-SNuPE, methylation-sensitive single nucleotide primer extension; NGS, next-generation sequencing; RLGS, restriction landmark genome scanning; RRBS, reduced representation bisulphite sequencing; –seq, followed by sequencing; WGSBS, whole-genome shotgun bisulphite sequencing.

Laird *Nat Rev Genet* 11 :191 (2010)

Séquençage bisulphite

Watson >>AC^{m}GTTCGCTTGAG>>
Crick <<TGC^{m}AAGCGAACTC<<

C^{m} methylated
C Un-methylated

1) Denaturation

Watson >>AC^{m}GTTCGCTTGAG>> Crick <<TGC^{m}AAGCGAACTC<<

2) Bisulfite Treatment

BSW >>AC^{m}GTTUGUTTGAG>> BSC <<TGC^{m}AAGUGAAUTU<<

3) PCR Amplification

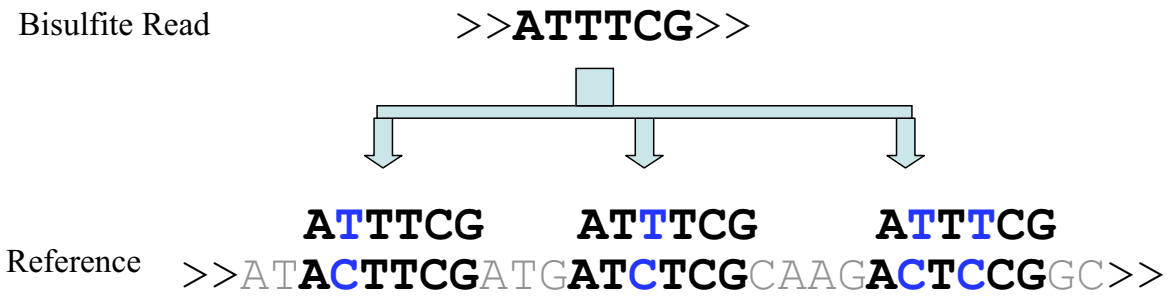
BSW >>AC^{m}GTTTGTTTGAG>> BSC <<TGC^{m}AAGTGAATTT<<
BSWR <<TG CAAACAAACTC<< BSCR >>ACG TTCACTTAAA>>

Xi & Li *BMC Bioinformatics* 10 :232 (2009)

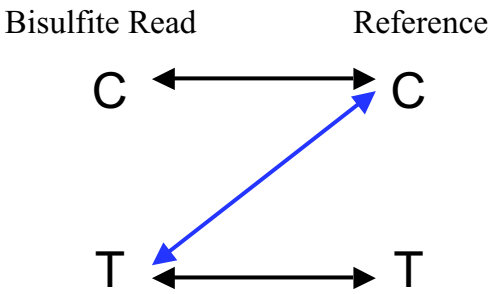
Séquençage bisulphite 2

Alignement asymétrique . . .

1) Multiple Mapping



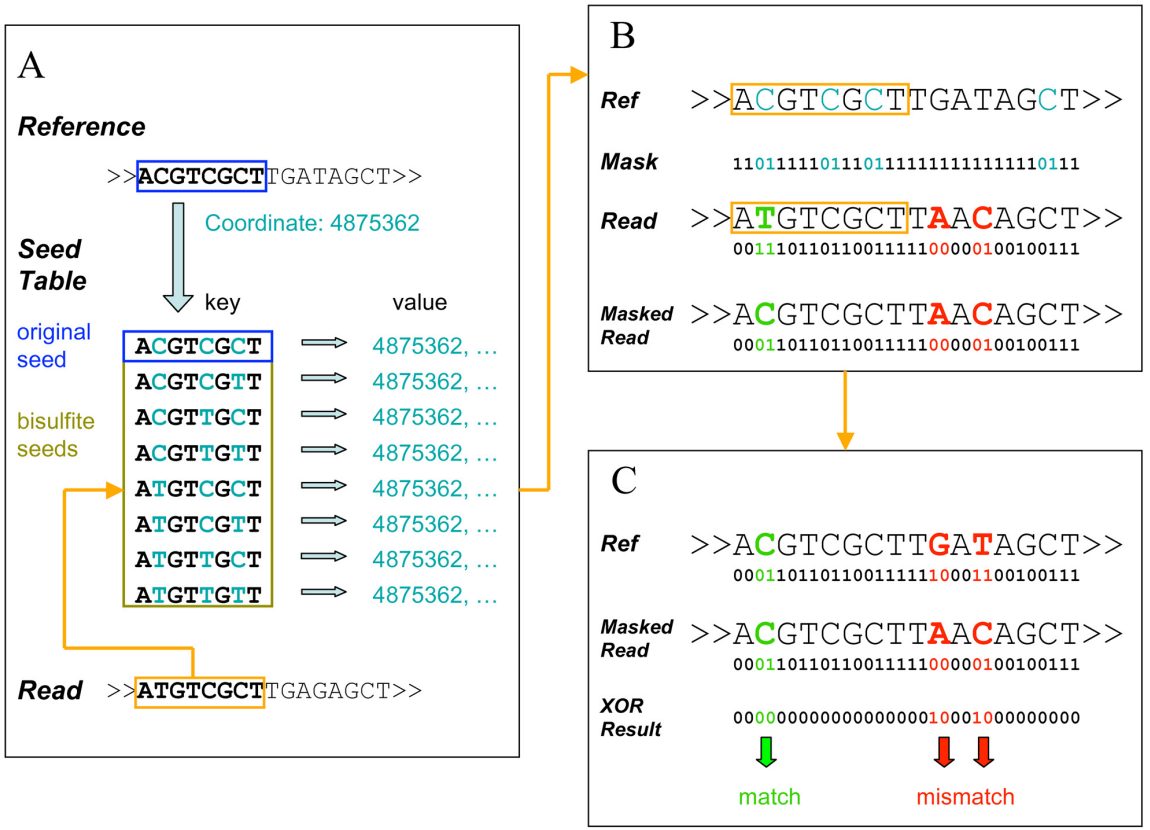
2) Mapping Asymmetry



Xi & Li *BMC Bioinformatics* 10 :232 (2009)

BSMAP

hachage + manipulation de bits pour détecter match/mismatch (stocker un masque pour la référence)



Xi & Li *BMC Bioinformatics* 10 :232 (2009)

Pénalisation modifiée

On peut aussi trouver une pénalisation propre (LODS) à la conversion bisulfite :

Table 1. Score matrix for aligning bisulfite-converted DNA reads to a reference genome sequence

	a	c	g	t
a	6	−18	−18	−18
c	−18	6	−18	3
g	−18	−18	6	−18
t	−18	−18	−18	3

Columns refer to bases in the read, and rows refer to bases in the genome.

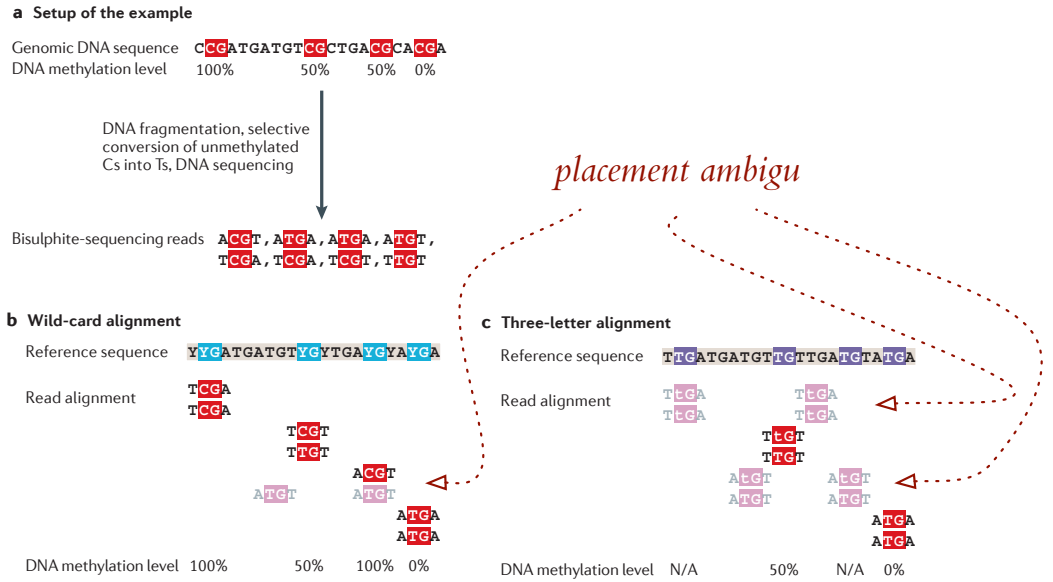
référence

match C:T ou T:T

Frith, Mori & Asai *Nucleic Acids Res* 40 :e100 (2012)

Ambiguïté

Placement est difficile : plus de matches C :T et T :T, régions de compléxité réduite (CpG)



Bock *Nat Rev Genet* 13 :705 (2012)

Inférence

On veut surtout détecter régions de méthylation différente (*differentially methylated region*, DMR) entre échantillons (100s de génomes)

a

Genomic DNA sequence		CG	…	CG	…	…	…	CG	…	CG	…	…	…	CG	…	CG	…	CG	…	CG	…	CG	…	…	CG
Cases	Sample 1	3%		6%				80%		57%				1%		0%		1%		1%		42%			78%
	Sample 2	2%		0%				50%		74%				0%		1%		0%		0%		38%			85%
	Sample 3	0%		1%				95%		86%				2%		0%		0%		0%		41%			67%
Controls	Sample 4	0%		2%				8%		1%				12%		3%		15%		8%		36%			72%
	Sample 5	1%		4%				5%		2%				15%		5%		33%		11%		39%			94%
	Sample 6	0%		2%				13%		1%				19%		2%		24%		22%		33%			92%

b

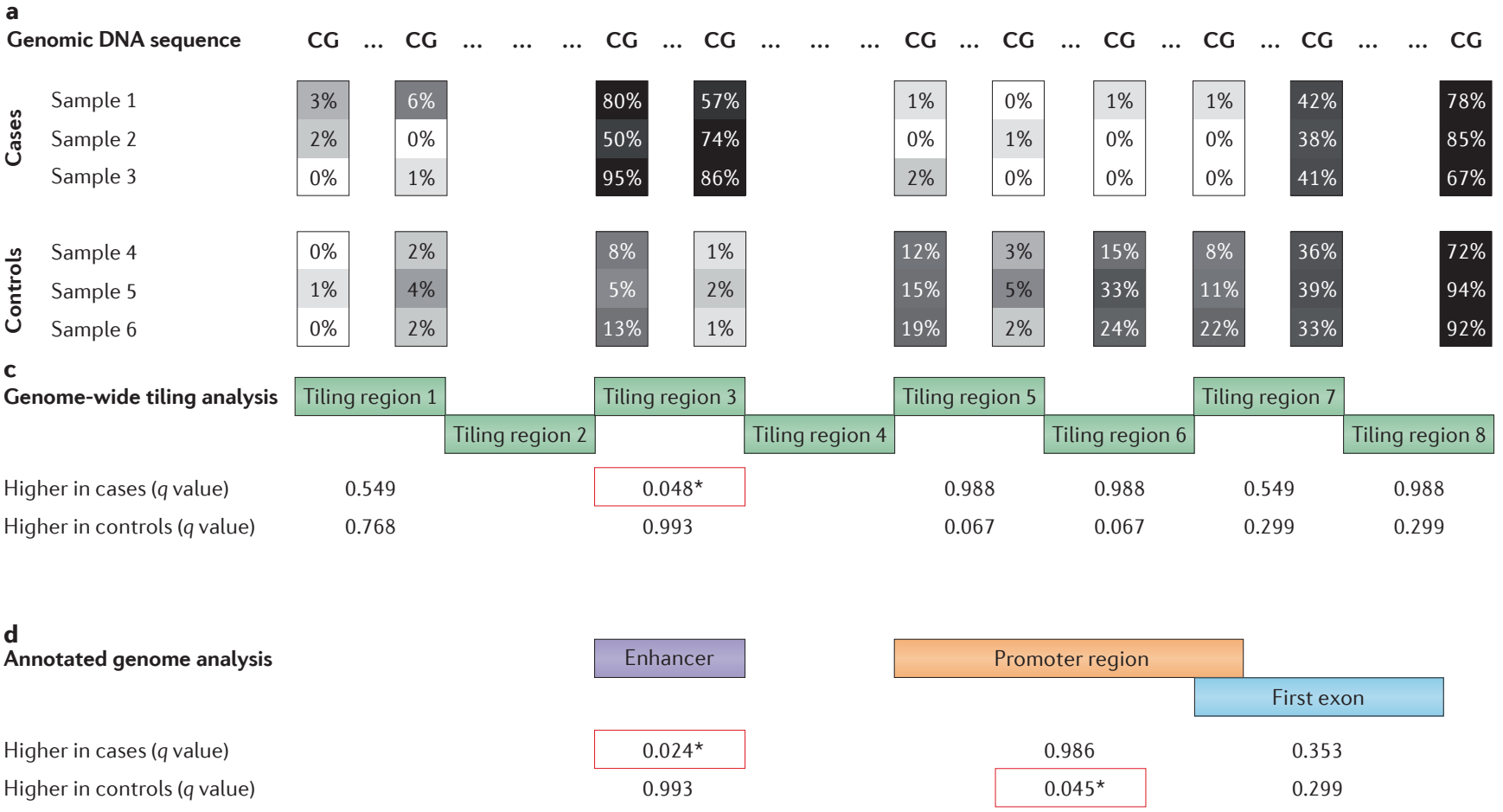
Single-CpG analysis	CG1	CG2	CG3	CG4	CG5	CG6	CG7	CG8	CG9	CG10
Higher in cases (*q* value)	0.333	0.993	0.085	0.068	0.993	0.993	0.993	0.993	0.196	0.993
Higher in controls (*q* value)	0.993	0.732	0.993	0.993	0.070	0.104	0.104	0.110	0.993	0.351

q-value : estimation de taux de faux positives (*False Discovery Rate*, FDR)

Bock *Nat Rev Genet* 13 :705 (2012)

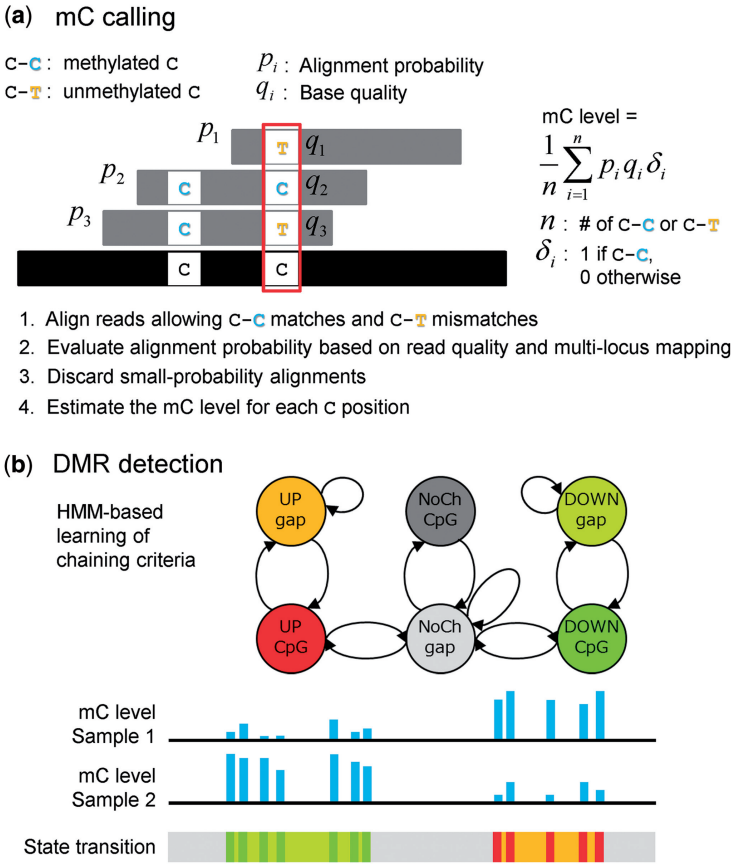
Inférence 2

(signal plus fort dans fenêtres ou selon annotation)



Bock *Nat Rev Genet* 13 :705 (2012)

Inférence avec HMM ?



→ 2 échantillons
→ émissions : restreintes à CpG seulement
distribution binomiale (m *reads* avec méthylation sur n, niveau θ spécifique à l'état)

$$\binom{m}{n}\theta^m(1-\theta)^{n-m}$$

→ duration géométrique de rester dans un état
→ LODS score d'une région :

$$\log \frac{\mathbb{P}\left\{\text{région} \,\middle|\, \text{UP}\right\}}{\mathbb{P}\left\{\text{région} \,\middle|\, \text{NoCH}\right\}}$$

Saito, Tsuji & Mituyama *Nucleic Acids Res* (2014)