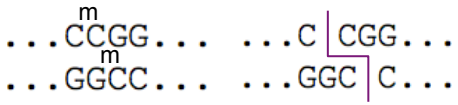
Genome-wide hCAGE-tags clusters annotation based on the new CpG islands model

Yulia A Medvedeva, John AC Archer, Vladimir B Bajic; KAUST team; RIKEN OSC; FANTOM5 consortium

October 17, 2011, Yokohama, Japan

CpG islands: brief history from experimental point of view

Bird A. **CpG-rich islands and the function of DNA methylation.** Nature. 1986.



- **~1% of the human genome is sensitive to HpaII restriction enzyme**
- **Non-methylated CpGs form clusters, sufficiently distinctive to be identified by computational analysis**
- Majority of CpG islands are unique in sequence (not duplications)
- Methylated-free regions are frequently located around promoter (often cover few first exons), but also in 3' gene region, mostly of housekeeping genes, but also tissue-specific, active in this tissue
- **Methylation of CpG island in the promoter reduces the level of transcription (methylation of non-islands region does not have such effect)**
- Methylation of CpG islands could be prevented by protein binding (hypothesis)
- Methylation of CpG island affects chromatin structure (hypothesis)

CpG islands' function is to distinguish regions of the genome that are available for interaction with nuclear components from the unavailable remainder of the genome

CpG islands: brief history from computational point of view

Gardiner-Garden M, Frommer M. **CpG islands in vertebrate genomes.** J Mol Biol. 1987.

CpG islands criteria:

- > 200 bp in length
- G+C > 0.5
- $Obs_{CpG}/Exp_{CpG} = (N_{CpG} \times L) / (N_C \times N_G) > 0.6$

Features:

- Clusters of GC-boxes (CCGCCC, GGGCGG) upstream and downstream of TSS
- Lack of TATA-box in promoters of widely expressed genes with CGIs, but not in tissue-specific ones

UCSC CpG islands: *de facto* standard

CpG islands are predicted by searching the sequence one base at a time, scoring each dinucleotide (+17 for CpG and -1 for others) and identifying maximally scoring segments. Each segment is then evaluated for criteria in previous slide.

Repetitive elements in the genome sequence were masked prior CpC island search.

But: 6–30% of CAGE-tags in mouse and human appear to be within repetitive elements (RE). Located near 5' of protein-coding loci RE frequently function as alternative promoters mostly for noncoding RNAs. More than ¼ of RefSeq genes contain RE with CAGE-tags mapped in their 3' UTR, with strong evidence for the reduced expression of these genes relative to RE-free genes. About 23,000 candidate regulatory regions derived from RE.

Faulkner et al. (FANTOM4) The regulated retrotransposon transcriptome of mammalian cells. Nature Genetics (2009)

HMM-based CpG islands model

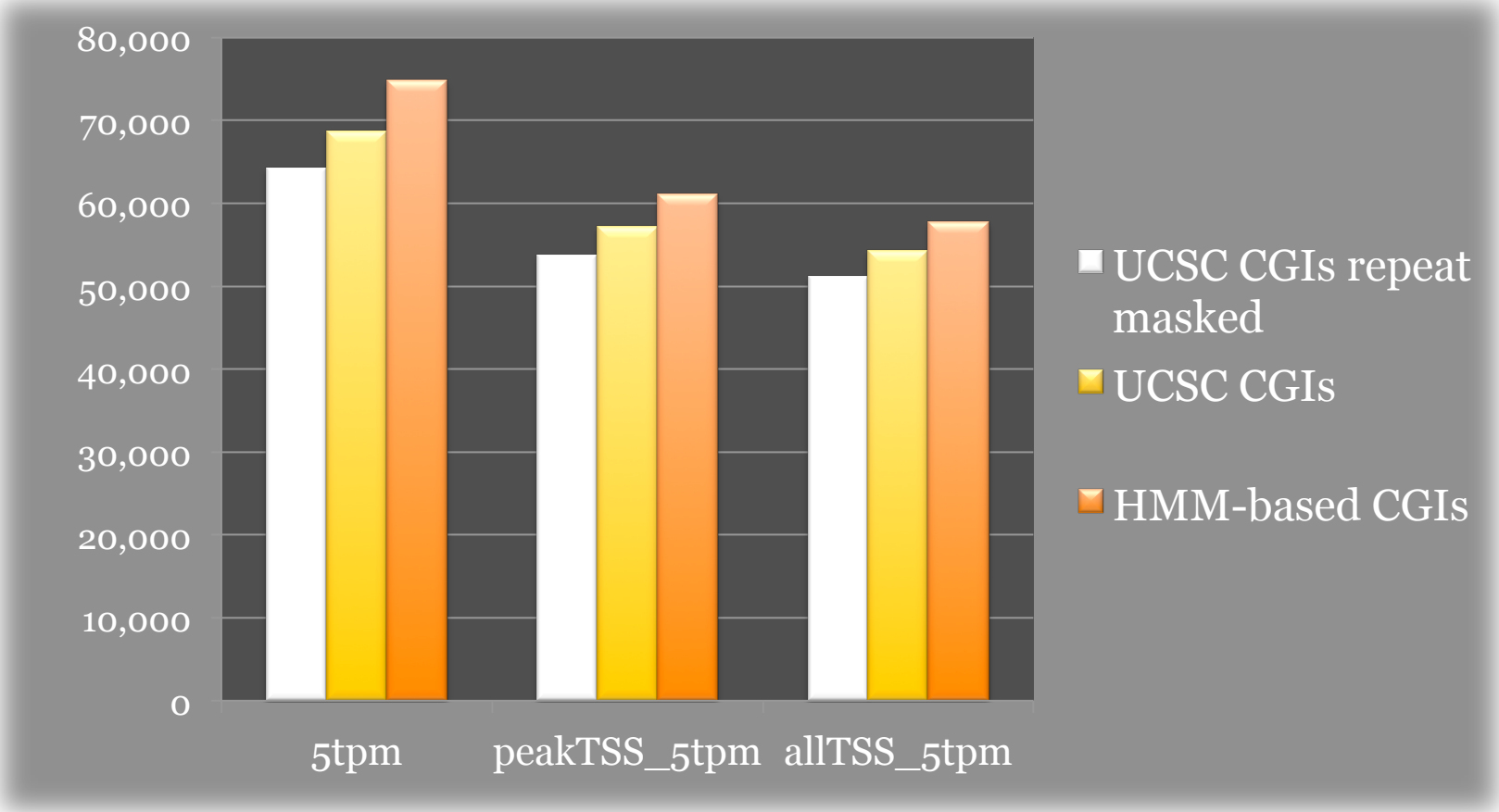
Hao Wu, et al. **Redefining CpG islands using hidden Markov models.** *Biostat (2010) 11 (3): 499-514*

Model is based on 2 HMMs: for G+C content and for CpG dinucleotides

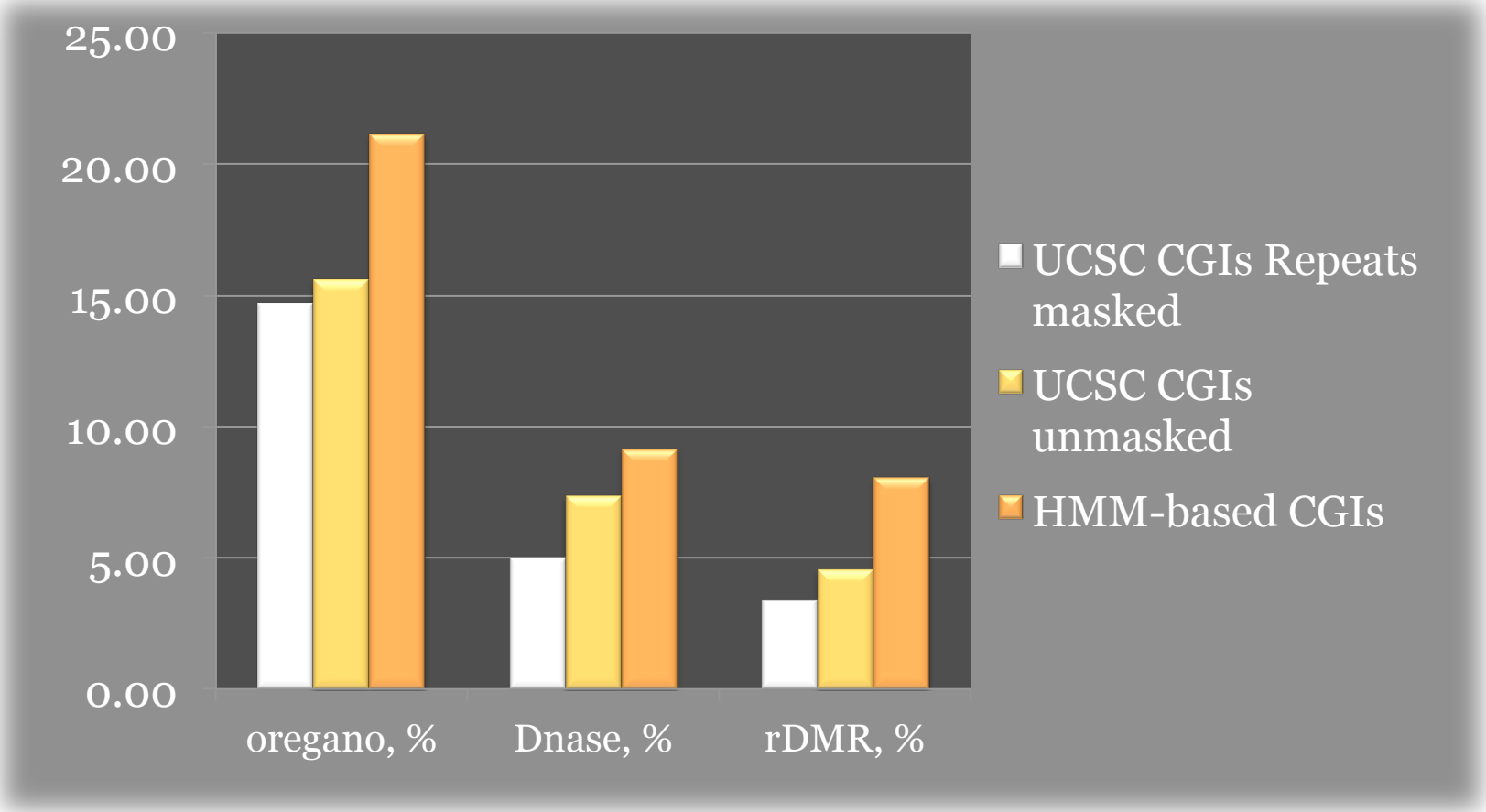
Basics statistics

CGIs	#CGIs	Total genome coverage, %	Length, bp (median)	G+C content (median)	Obs/Exp_{CpG} (median)
UCSC CGIs Repeat masked	28,690	0.70	559	0.66	0.85
UCSC CGIs unmasked	50,889	0.97	316	0.64	0.86
HMM-based CGIs	114,893	1.43	216	0.60	0.79

CAGE-tags clusters associated with CGIs



Features associated with CGIs



De novo motif family search

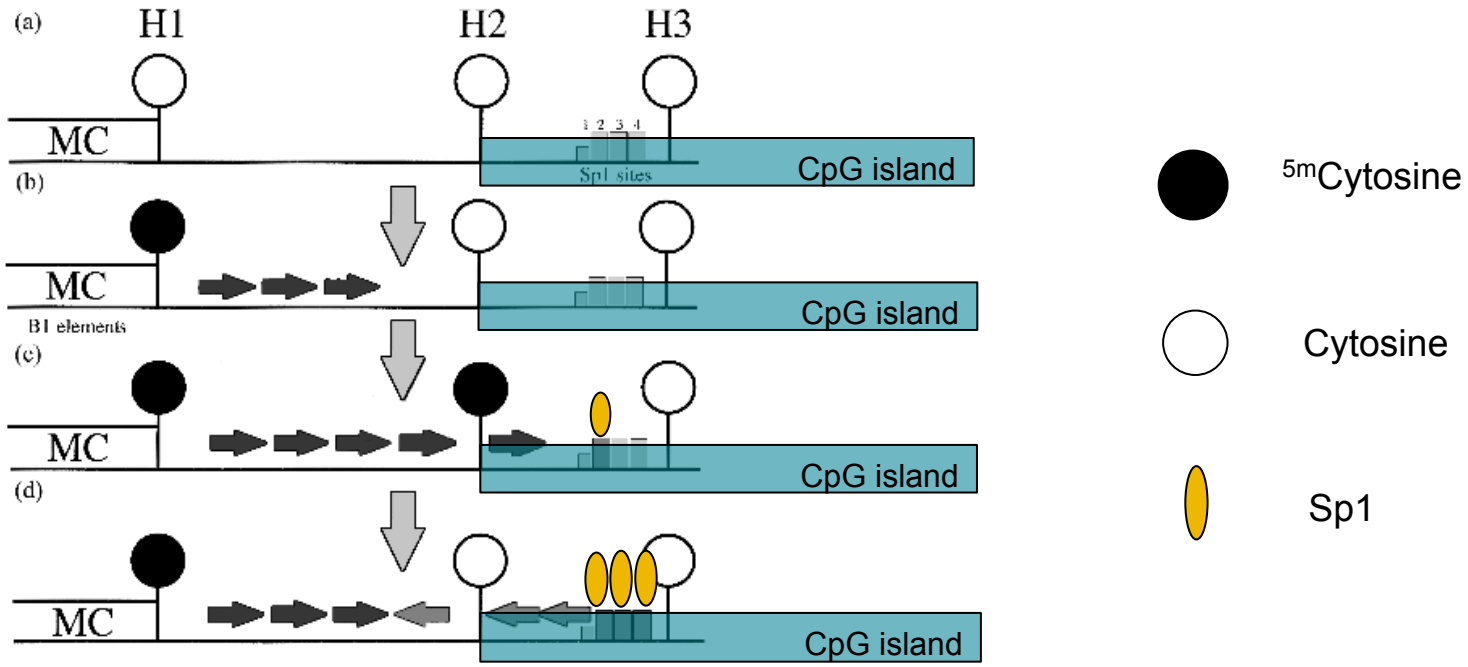
Motif families of the length: 8-12 bp

Total number: 300

Statistically significant, comparing to shuffled sequence: 268

Association of the motif families with known transcriptional factors (HOCOMOCO, etc…)

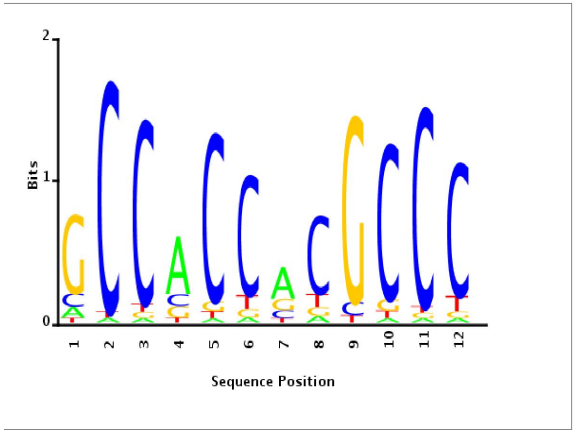
Mouse aprt gene



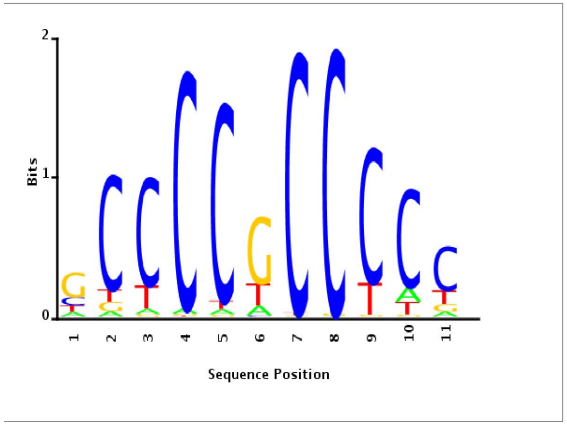
Turker MS. **The establishment and maintenance of DNA methylation patterns in mouse somatic cells.** Semin Cancer Biol. 1999 Oct;9(5):329-37 (modified).

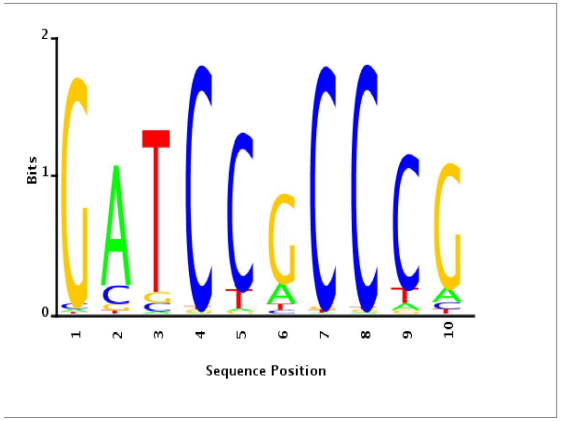
What motif families are located in the “core” or “boundaries” of CpG islands (-50..+50 bp around boundaries)?

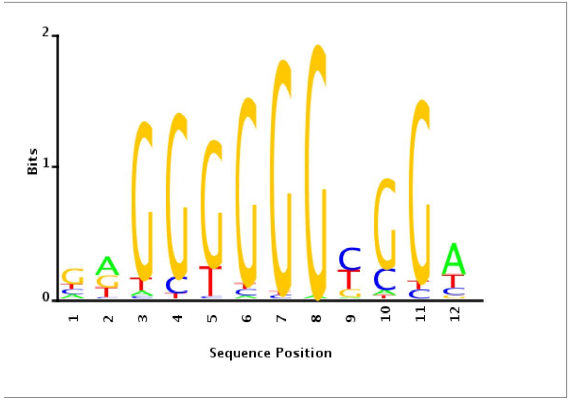
Motifs in the center regions of CGIs



GC-box,
Sp-family,
Zinc-finger
family

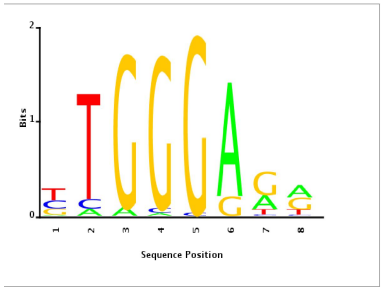




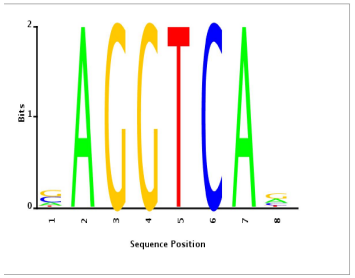


Motifs in the boundary regions

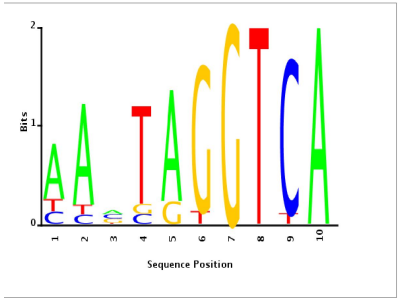
IKZF1



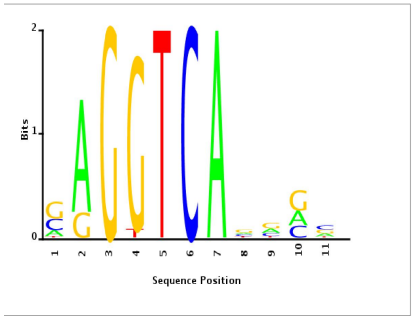
RARA



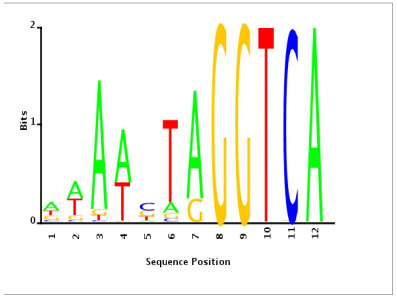
NR1D1



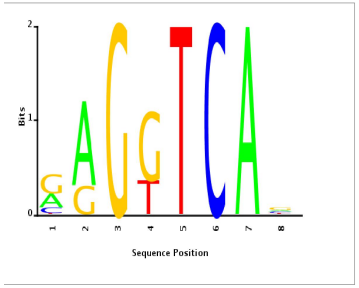
RARB



RORA



RARG



RAR as heterodimer with RXR without ligand can induce histone acetylation, chromatin condensation and transcriptional suppression. On ligand binding, the corepressors dissociate from the receptors and associate with the coactivators leading to transcriptional activation

Annotation

Genome-wide annotation with this model of CGIs is uploaded to Zenbu

Annotation of hCAGE-clusters will be provided soon

Thanks

RIKEN OSC team

KAUST team

Benoit Marchand, George Marselis, Haitham Ashoor, Wail Ba Alawi, Shariful Islam Bhuyan, Michael Schramm

VIGG team

Ivan Kulakovskiy, Vsevolod Makeev

FANTOM5 consortium