

**Supplemental Materials to
“Computational Prediction of DNA Methylation Landscape in the Human Genome”**

Rajdeep Das¹, Nevenka Dimitrova², Zhenyu¹ Xuan, Robert A. Rollins³, Fatemah G. Haghighi^{3,4},
John R. Edwards^{4,5}, Jingyue Ju^{4,5}, Timothy H. Bestor³ and Michael Q. Zhang^{1,6}

¹Cold Spring Harbor Laboratory, Cold Spring Harbor NY, USA

²Philips Research, 345 Scarborough Rd, Briarcliff Manor NY, 10510, USA.

³ Department of Genetics and Development, College of Physicians and Surgeons of Columbia University, New York NY, USA

⁴Columbia Genome Center, Columbia University, New York NY, USA

⁵Department of Chemical Engineering, Columbia University, New York NY, USA

⁶Author for Correspondence:

Dr. M.Q.Zhang

Cold Spring Harbor Laboratory

1 Bungtown Road

Cold Spring Harbor, NY 11724

USA

Tel: 516 367 8393

Fax: 516 367 8461

mzhang@cshl.edu

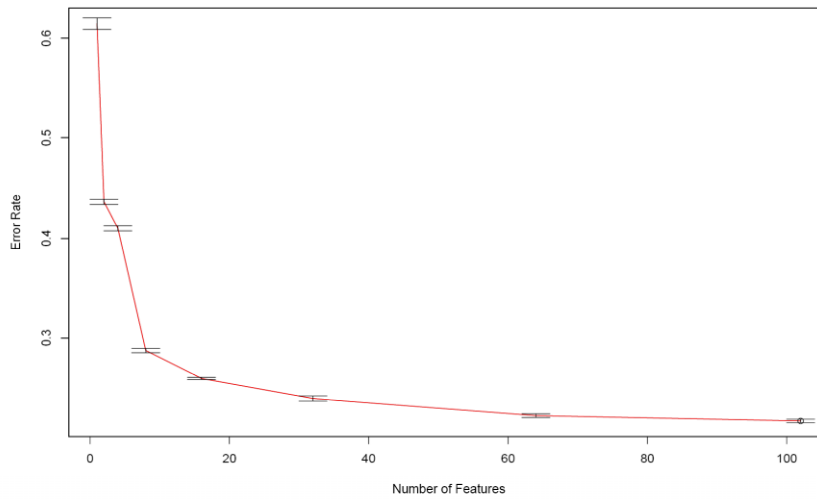
Supplemental Materials

Figure S1. Recursive feature elimination: accuracy vs. number of features.

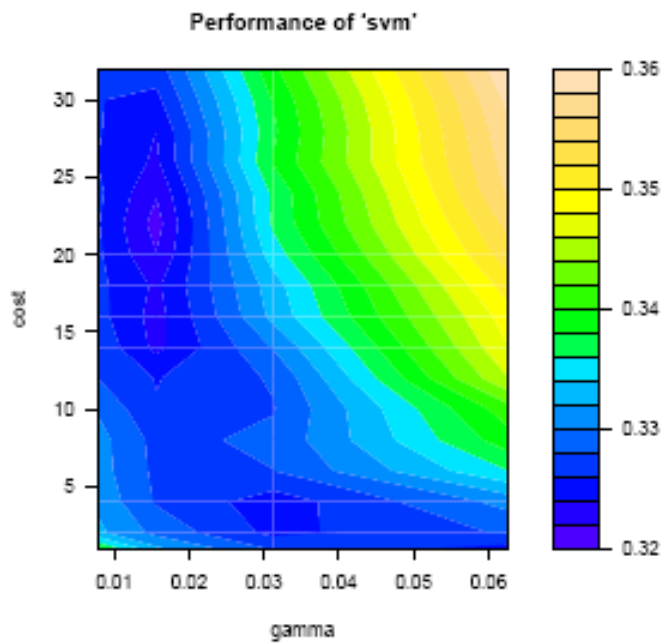


Figure S2. Grid search results for the performance of SVM classifier.

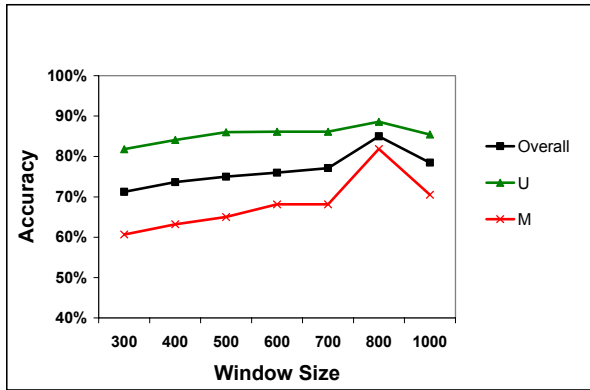


Figure S3. Accuracy of the SVM classifier as a function of window length.

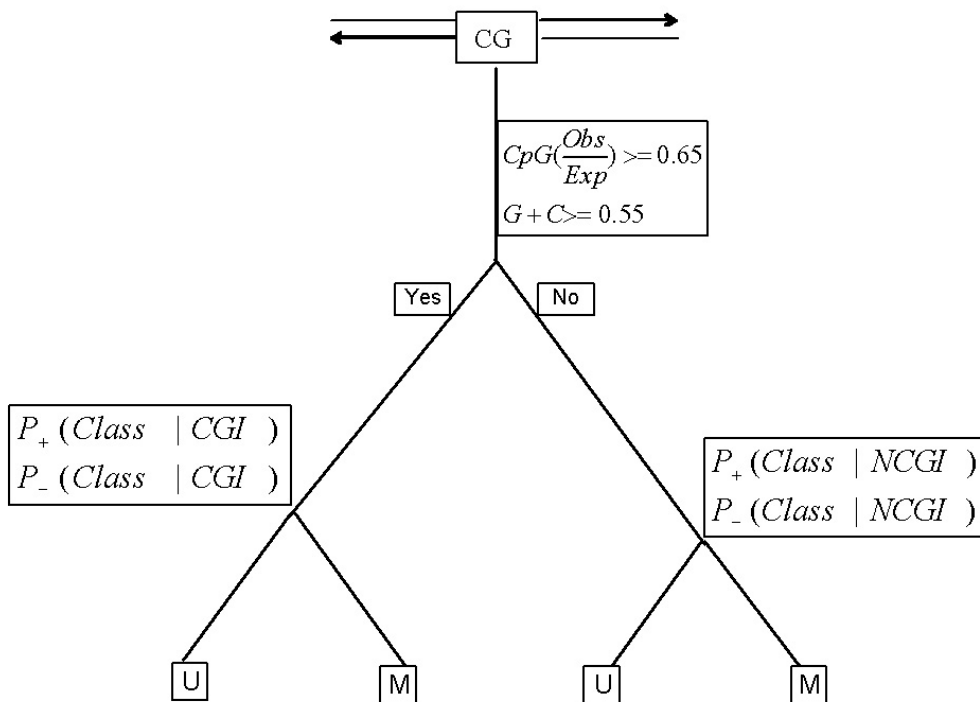


Figure S4. Methylation Prediction algorithm.

Table S1. Transcription factors, methylated vs. unmethylated set (non-CGI, non-Alu).

Hexamer	Matrix id	Divergence	Transcription factor	Species
		nce		
BCCCWG	M00480	0.89	LUN-1	human
CCCWGH	M00088	0.91	Ik-3 Ikaros 3	mouse
	M00316	0.95	Imperfect Hogness/Goldberg BOX	
	M00480	0.81	LUN-1	human
	M00721	0.99	CACCC-binding factor	human
	M00947	0.88	CP2/LBP-1c/LSF	mouse human
CCTGMV	M00002	0.95	E47	human
	M00073	0.86	deltaEF1	chick
	M00277	0.98	Lmo2 complex, bound to Tal-1, E2A proteins, and GATA-1, half-site 1	mouse human
	M00319	0.69	MEF-3 MEF-3	
	M00412	0.59	AREB6 AREB6 (Atpl1a1 regulatory element binding factor 6)	human
	M00531	0.85	NERF1a new ets-related factor 1a	human
	M00411	0.88	HNF-4alpha1 Hepatocyte nuclear factor 4	rat
GMCCCN	M00491	0.99	MAZR MAZ related factor	human mouse
	M00518	0.83	PPARalpha:RXR-alpha alpha:retinoid X receptor alpha)	human frog rat mouse
	M00152	0.62	SRF serum response factor	rat human cat chick
WGCCCH	M00531	0.97	NERF1new ets-related factor 1a	human
CKGSCM				

Table S2. Transcription factors, unmethylated vs. methylated set (non-CGI, non-Alu).

Hexamer	Matrix id	Divergence	Transcription factor	Species
		nce		
AAWGGR	M00410	0.91	SOX-9 SOX (SRY-related HMG box)	human
	M00745	0.83	LEF-1	frog human mouse
AAATKT	M00135	0.74	Oct-1 octamer factor 1	mouse hamster gibbon chick monkey rat frog human
ATGVAA	M00135	0.82	Oct-1 octamer factor 1	mouse hamster gibbon chick monkey rat frog human
	M00161	0.39	Oct-1 octamer-binding factor 1	mouse hamster gibbon chick monkey rat frog human

	M00195	0.77	Oct-1 octamer factor 1	mouse hamster gibbon chick monkey rat frog human
	M00210	0.59	OCT-x Octamer binding site	cat gibbon chick rat frog mouse monkey human
	M00249	0.96	CHOP:C/EBPalpha heterodimers of CHOP and C/EBPalpha	hamster rat chick human mouse frog
	M00342	0.62	Oct-1 Octamer binding factor 1	mouse hamster gibbon chick monkey rat frog human
	M00795	0.62	Octamer	hamster gibbon chick rat frog cat monkey mouse human
	M00930	0.52	Oct-1	mouse gibbon chick monkey rat frog human
TGVAAA	M00135	0.76	Oct-1 octamer factor 1	mouse hamster gibbon chick monkey rat frog human
	M00138	0.94	Oct-1 octamer factor 1	mouse hamster gibbon chick monkey rat frog human
	M00161	0.35	Oct-1 octamer-binding factor 1	mouse hamster gibbon chick monkey rat frog human
	M00195	0.67	Oct-1 octamer factor 1	mouse hamster gibbon chick monkey rat frog human
	M00210	0.52	OCT-x Octamer binding site	cat gibbon chick rat frog mouse monkey human
	M00342	0.55	Oct-1 Octamer binding factor 1	mouse hamster gibbon chick monkey rat frog human
	M00795	0.64	Octamer	hamster gibbon chick rat frog cat monkey mouse human
ATGVAA	M00930	0.51	Oct-1	mouse gibbon chick monkey rat frog human
TGVAAA	M00935	0.89	NF-AT	rat mouse human
CWGAMA	M00992	0.97	FOXP3	human mouse
AATKAA	M00099	0.97	S8 S8	mouse
	M00206	0.92	HNF-1 Hepatic nuclear factor 1	mouse human rat
	M00241	0.91	Nkx2-5 homeo domain factor Nkx-2.5/Csx, tinman homolog	mouse
	M00424	0.67	NKX6-1 NKX6-1	rat chick mouse human hamster

	M00510	0.55	Lhx3 LIM homeobox transcription factor 3	human mouse
	M01000	0.82	AIRE	human
AAATGV	M00059	0.89	YY1 Yin and Yang 1	human
	M00210	0.91	OCT-x Octamer binding site	cat gibbon chick rat frog mouse monkey human
	M00225	0.66	STAT3 signal transducer and activator of transcription 3	mouse human
GRAAT	M00150	0.98	Brachyury Brachyury	mouse
VAAAT	M00024	0.93	E2F E2F	mouse human
	M00135	0.84	Oct-1 octamer factor 1	mouse hamster gibbon chick monkey rat frog human
	M00161	0.46	Oct-1 octamer-binding factor 1	mouse hamster gibbon chick monkey rat frog human
	M00195	0.84	Oct-1 octamer factor 1	mouse hamster gibbon chick monkey rat frog human
	M00210	0.56	OCT-x Octamer binding site	cat gibbon chick rat frog mouse monkey human
	M00225	0.9	STAT3 signal transducer and activator of transcription 3	mouse human
	M00342	0.68	Oct-1 Octamer binding factor 1	mouse hamster gibbon chick monkey rat frog human
	M00795	0.84	Octamer	hamster gibbon chick rat frog cat monkey mouse human
	M00930	0.64	Oct-1	mouse gibbon chick monkey rat frog human
TRAATT	M00463	0.68	POU3F2	mouse rat human

Table S3. Transcription factors, unmethylated vs. methylated set (CGI., non-Alu)

Hexamer	Matrix id	Divergence	Transcription factor	Species
CCGSSC	M00034	0.93	p53 tumor suppressor	mouse human
CGSCCS	M00196	0.78	Sp1 stimulating protein 1	human rat mouse
	M00650	0.83	MTF-1	human mouse
	M00931	0.95	Sp-1	rabbit hamster human frog monkey chick rat pig mouse
VGCGGG	M00196	0.77	Sp1 stimulating protein 1	human rat mouse
	M00803	0.99	E2F	rat mouse human

	M00919	0.96	E2F	rat mouse human
	M00920	0.95	E2F	rat mouse human
	M00931	0.75	Sp-1	rabbit hamster human frog monkey chick rat pig mouse
	M00932	0.89	Sp-1	human rat
	M00933	0.91	Sp-1	human rat mouse
	M00939	0.96	E2F-1	rat mouse human
	M00982	0.93	KROX	human monkey rat mouse
TCCSSG	M00223	0.95	STATx signal transducers and activators of transcription	sheep human mouse
	M00224	0.4	STAT1 signal transducer and activator of transcription 1	human mouse
	M00225	0.36	STAT3 signal transducer and activator of transcription 3	mouse human
	M00341	0.94	GABP GA binding protein	mouse human
	M00457	0.95	STAT5A (homodimer) signal transducer and activator of transcription 5a	human mouse sheep
KCCSGC	M00634	0.99	GCM	mouse human

PCA Analysis

For first four eigenvectors for non-CpG island data are shown in Figure S5 and first four eigenvectors for CpG island data are shown in Figure S6. In both cases, the principal component for the first vector is Alu coverage. Alu gave us the strongest signal (0.99) and showed that it behaves independently of all other features. For the non-CpG island data the principal components for the second vector include AA, AT, TA, TT, GG, GC, and CC. Interestingly, GG, GC and CC are negatively correlated with AA, AT, TA, and TT. The principal components for the third vector include AA, GG, GAA, GGA, AAWGGR, ATGVAA, TGRAAT which are negatively correlated with TT and TTT. The principal components for the fourth vector include: TT, AAWGGR, ATGVAA, TGRAAT negatively correlated with AA and AAA. The principal components for the features extracted from the 800bp window have very similar values, except that GG does not appear in the second vector, and GA appears in the third vector. For the CpG island data, the principal components for the second vector are all negative and include TG, GT, TGT, GTG and CGT. The principal components for the third vector are all positive and include AC, TA, ACC, GTA, and CAC. The principal component for the fourth vector AG is negatively correlated with GC, CG, GCG and CGG.

