TABLES & FIGURES

Table 1: Known Imprinted Genes and CpG Islands

Imprinted Gene(s)	EnsEMBL Gene ID	Chromosomal Location	DMR
Nnat	ENSMUSG00000027648	2.158434932-158435545	M
Gnas, Gnasx1, Nesp, Nespas	ENSMUSG00000027523	2.175526919-175528515	P
-	ENSMUSG00000027523	2.175537603-175542938	M
-	ENSMUSG00000027523	2.175569444-175573436	U
Copg2, Copg2as, Copg2as2	ENSMUSG00000025607	6.31047005-31047354	P
-	ENSMUSG00000025607	6.31047674-31047873	P
-	ENSMUSG00000025607	6.31080211-31080466	U
Sgce	ENSMUSG00000004631	6.4453104-4453415	M
-	ENSMUSG00000004631	6.4460514-4460754	M
Nap1-l5	ENSMUSG00000029805	6.59357121-59357436	M
H19	ENSMUSG00000000031	7.133006782-133007370	P
Igf2, Igf2as	ENSMUSG00000000033	7.133086642-133086847	P
-	ENSMUSG00000000033	7.133091687-133097997	P
Mash2/Ascl2	ENSMUSG00000009248	7.133401068-133403152	N
Tapa1/Cd81, Tssc4	ENSMUSG00000037706, ENSMUSG00000037699	7.133489190-133489542	U
Tssc4, Kvlqt1/Kcnq1, Kvlqt1as/Lit1	ENSMUSG00000037699, ENSMUSG00000009545	7.133505316-133505877	P
Kvlqt1/Kcnq1, Kvlqt1as/Lit1	ENSMUSG00000009545	7.133543703-133544475	N
-	ENSMUSG00000009545	7.133732998-133734530	M
-	ENSMUSG00000009545	7.133842243-133842445	N
P57KIP2/Cdkn1c	ENSMUSG00000000154	7.133883008-133883764	N
Slc22a1l/Impt1, P57KIP2/Cdkn1c	ENSMUSG00000037664, ENSMUSG00000000154	7.133896782-133902620	P
Slc22a1l/Impt1, Tssc3/Ip1	ENSMUSG00000037664, ENSMUSG00000010760	7.133943302-133944789	P
Tssc3/Ipl, Nap1-l4/Nap2	ENSMUSG00000010760, ENSMUSG00000010759	7.133990276-133991044	N
Obph1/Osbpl5	ENSMUSG00000037606	7.134182532-134182744	U
Ube3a, Ube3aas	ENSMUSG00000025326	7.48955078-48955908	U

Snrpn, Snurf, Ipw, Pwcr1	ENSMUSG00000000948	7.49379542-49379959	M
Ndn, Magel2	ENSMUSG00000033585, ENSMUSG00000033574	7.51709425-51709854	B
Zfp127/Mkrn3,Zfp127as/Mkrn3as	ENSMUSG00000033564	7.51780860-51781093	M
Frat3, Zfp127	ENSMUSG00000033564, ENSMUSG00000033551	7.51824782-51825751	M
Peg3/Pw1 -	ENSMUSG00000002265 ENSMUSG00000002265	7.6089053-6089383 7.6110648-6112563	U M
Rasgrf1	ENSMUSG00000032356	9.90370716-90371012	P
Zac1, Hymai	ENSMUSG00000019817	10.12974657-12975123	M
Meg1/Grb10 -	ENSMUSG00000020176 ENSMUSG00000020176	11.11953633-11954614 11.11964394-11965591	M N
Dlk/Pref1	ENSMUSG00000040856	12.103716118-103716887	N
Meg3/Gtl2	ENSMUSG00000021268	12.103788454-103788744	P
Dio3	ENSMUSG00000040837	12.104541184-104543033	N
Slc22a3	ENSMUSG00000023828	17.11835678-11835880	U
Igf2r, Igf2ras/Air -	ENSMUSG00000023830 ENSMUSG00000023830	17.12144211-12145609 17.12172251-12173135	M P
Impact -	ENSMUSG00000024423 ENSMUSG00000024423	18.12957829-12958387 18.12989774-12991228	U M
Peg1/Mest	ENSMUSG00000029794	Un.130506604-130506818	M
Asb4	ENSMUSG00000042607	No CGIs	U
Usp29, Usp29as, Zim3	ENSMUSG00000023184	No CGIs	M (Peg3)
Zim1	ENSMUSG00000002266	No CGIs	U
Zfp264	NA	NA	U
Ins2	ENSMUSG00000000215	No CGIs	N
Dcn	ENSMUSG00000019929	No CGIs	U
U2af1-rs1	NA*	NA	M
Htr2a	ENSMUSG00000034997	No CGIs	U
Slc38a4/Ata3	NA*	NA	M
Ins1	ENSMUSG00000035804	No CGIs	U
Peg13	NA*	NA	M

45 CGIs were identified in the gene and 50kb upstream sequences of 41 EnsEMBL gene loci associated with known imprinted mouse genes. Genes are listed here with EnsEMBL Gene IDs and associated CGIs identified by cpgplot (http://www.emboss.org). CGIs are categorized according to the methylated allele (P = paternal, M = maternal, U = unknown, N = Neither, B = both). The 27 CGIs that are differentially-methylated (M or P) are DMR-CGIs, and the remaining 18 (U or N or B) are UMR-CGIs. This annotated database of imprinted genes was generously provided to us by Smith and Kelsey.
* excluded from the original analysis because they were not present in the EnsEMBL mus_musculus_core_9_3 database, but were analyzed subsequently (see Discussion).

Table 2: Number of Significant *k*-mers for UMR & DMR vs Control CGIs

k	α	DMR-CGIs	UMR-CGIs	Expected
5	10^{-2}	294	60	10
5	10^{-3}	154	22	1
5	10^{-4}	92	10	0
5	10^{-5}	62	6	0
6	10^{-2}	580	119	41
6	10^{-3}	234	34	4
6	10^{-4}	128	10	0
6	10^{-5}	72	2	0
7	10^{-2}	1208	372	164
7	10^{-3}	372	68	16
7	10^{-4}	150	16	2
7	10^{-5}	82	4	0

The number of *k*-mers with significantly different frequencies in DMR-CGIs and UMR-CGIs relative to Control CGIs are shown for each combination of word length *k* and significance level α used. In all cases, many significant differences were observed between DMR and Control CGIs. Relatively fewer significant differences were seen between UMR and Control CGIs, indicating that the majority of significant compositional differences between DMR and Control CGIs are related to differential methylation. The number of 'false positives' expected to arise merely by chance based on the significance level and number of *k*-mers tested is shown for comparison.

Figure 1: Clustering Analysis of Significant Heptamers

CLUSTER 1A	LOD
CCGCCGC..	-1.98
.CGCCGCG.	-1.78
GCGCCGC..	-1.62
.CGCCGCC.	-1.92
..GCCGCCG	-2.09
..GCCGCCT	-2.8
CLUSTER 1B	**LOD**
.TGCCGCA	1.89
TTGCCGC.	2.01
.TGCCGCG	1.79

CLUSTER 2	LOD
AAAATTT.	2.67
.AAATTTG	3.26
CAAATTT.	3.26
.AAATTTT	2.67

CLUSTER 3	LOD
.ATGTGAC.	3.25
CATGTGA..	2.93
..TGTCACA	2.72
..TGTGACA	2.72

CLUSTER 4	LOD
..CGGCTCG..	1.41
....GCTCGGC	1.8
...GGCTCGG.	1.57
CTCGGCT....	2.22
.TCGGCTC...	2.02
.TCGGCTG...	1.68

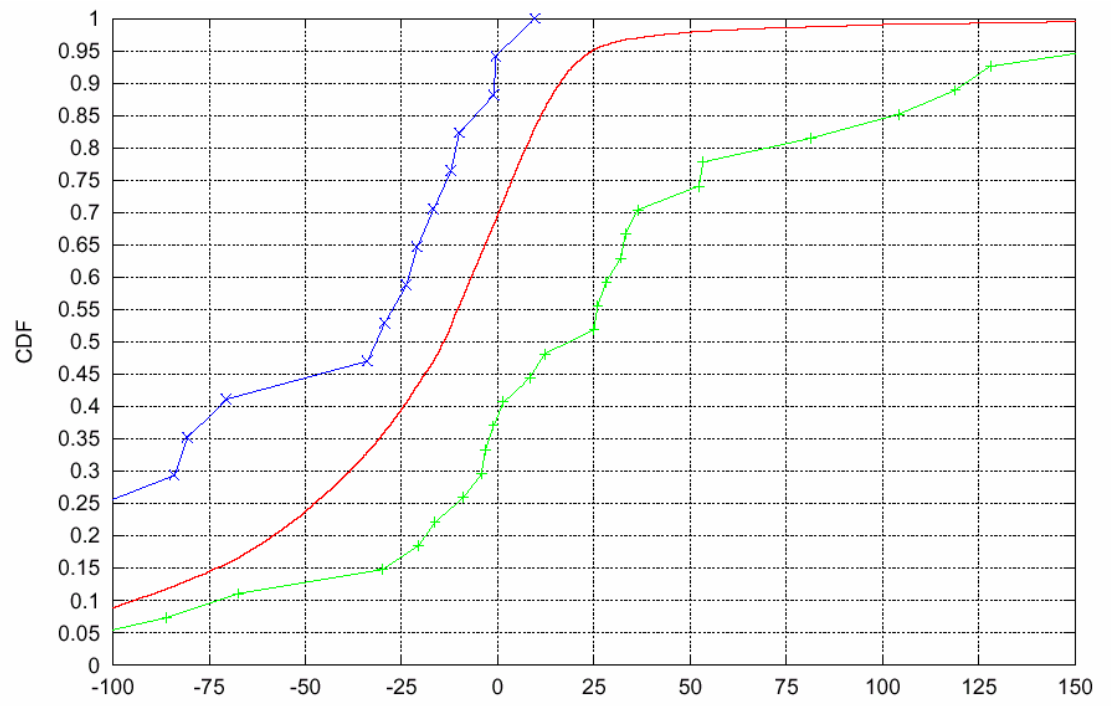
All heptamers present at significantly different ($\alpha = 10^{-5}$) frequencies in DMR and Control CGIs were clustered based on sequence similarity, and four unique clusters with more than three members each were identified. Clustered heptamers and the log-odds ratio in bits (LOD) of their frequencies in DMR and Control CGIs are shown as well as Pictogram representations of motifs for each cluster. Clusters 1A/1B correspond to *CTCF*-binding sites.

Table 3: Results of CGI Scoring Analysis

k	α	Control Avg.	DMR Avg.	>95% Control	*p*-value
5	10^{-2}	-42.4	-24.6	0.19 (5)	8.0E-02
5	10^{-3}	-48.1	-41.2	0.15 (4)	2.6E-01
5	10^{-4}	-43.1	-40.6	0.11 (3)	3.8E-01
5	10^{-5}	-43.1	-40.5	0.07 (2)	3.0E-01
6	10^{-2}	-44.7	-12.3	0.26 (7)	6.0E-04
6	10^{-3}	-36.7	-16.2	0.30 (8)	3.9E-03
6	10^{-4}	-28.6	-10.0	0.26 (7)	7.5E-04
6	10^{-5}	-23.7	-9.6	0.22 (6)	1.7E-03
7	10^{-2}	-26.9	27.8	0.52 (14)	3.7E-06
7	10^{-3}	-16.0	16.6	0.44 (12)	7.0E-07
7	10^{-4}	-8.0	13.5	0.41 (11)	7.5E-05
7	10^{-5}	-5.6	11.0	0.52 (14)	1.6E-05

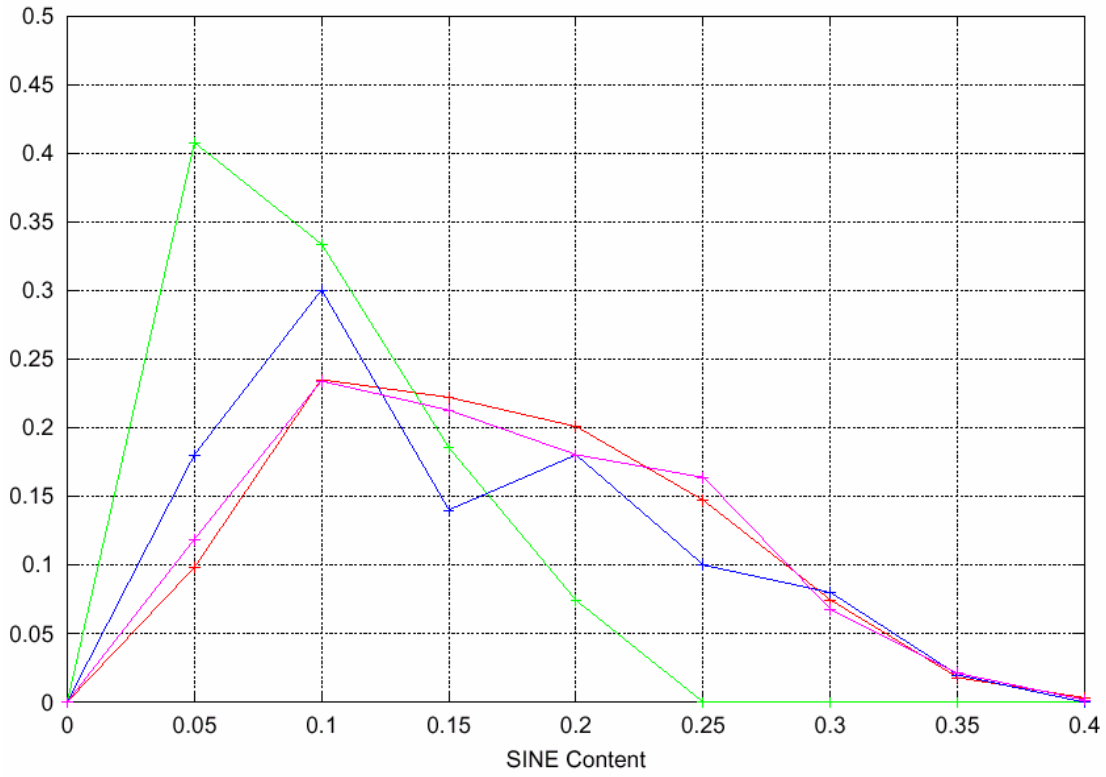
The DMR and Control CGIs were scored for all combinations of word length (k = 5,6,7) and significance level ($\alpha = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$) and many of the score distributions were significantly different for DMR-CGIs and Control CGIs (p-value determined by Wilcoxon rank-sum test). The average CGI Scores for DMR and Control CGIs are shown, along with the proportion of DMR-CGIs that scored higher than 95% of all Control CGIs. The scoring function with parameters (k = 7, $\alpha = 10^{-2}$) was optimal, as it yielded the greatest difference in mean CGI scores between Control and DMR-CGIs and the largest proportion of DMR-CGIs that scored higher than 95% of all Control CGIs.

Figure 2: Cumulative Score Distributions for DMR, UMR & Control CGIs



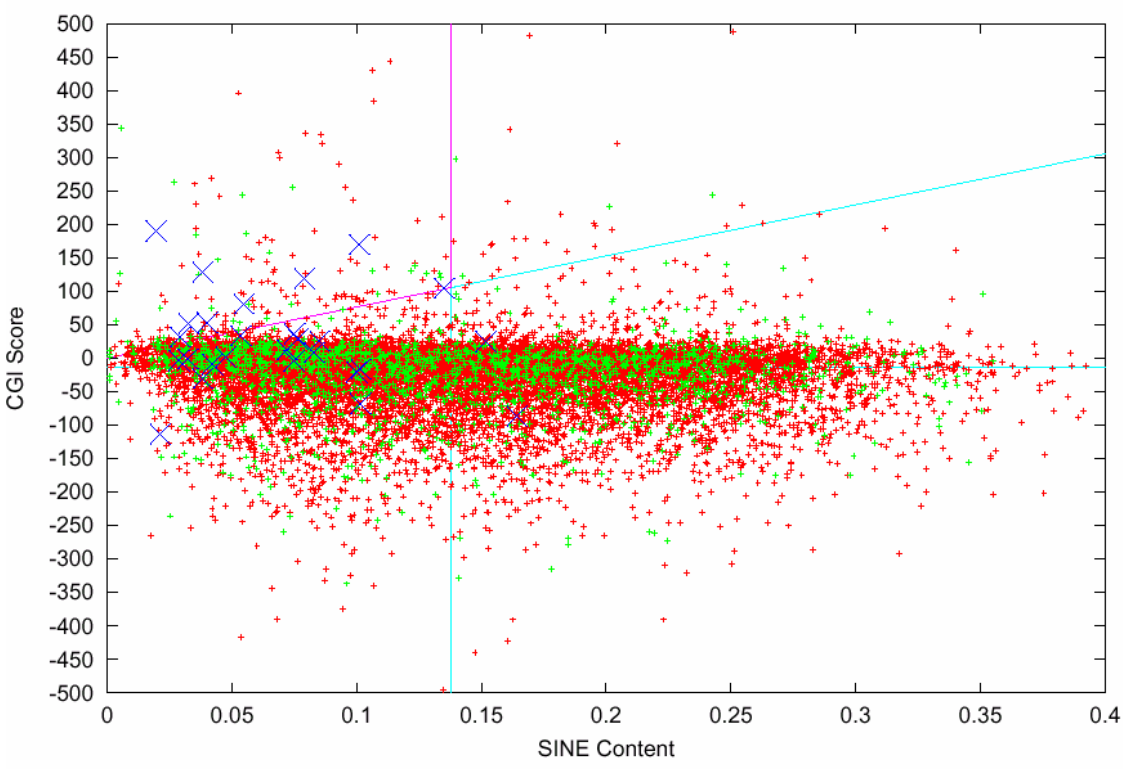
All DMR (green), UMR (blue), and Control (red) CGIs were scored according to their significant heptamer composition ($k = 7$, $\alpha = .01$) as described in methods and the cumulative distribution functions (CDF) were plotted. The average CGI score for DMR-CGIs (27.8) was significantly greater ($p < 10^{-5}$) than for Control CGIs (-26.9), demonstrating that differences in oligonucleotide composition can contribute to the prediction of imprinted loci. The average CGI score for UMR-CGIs (-62.0) was significantly lower than for both DMR-CGIs ($p < 10^{-4}$) and Control CGIs ($p < .05$), indicating that UMR-CGIs are compositionally distinct from DMR-CGIs despite their location in imprinted domains.

Figure 3: SINE Repeat Content Distributions



The SINE repeat content of the sequences and 100kb flanking regions was analyzed for DMR-CGIs (green), Control CGIs (red), High Scoring CGIs (blue) and FANTOM2 Candidate-Associated CGIs (pink). Imprinted loci display a significant reduction in SINE content compared to control regions ($p < 10^{-9}$) that is conserved between mouse and humans. The average SINE content of High Scoring CGIs (12.6%) is significantly greater than for DMR-CGIs (7.4%) and significantly less than for Control CGIs (14.4%), suggesting it contains a mixture of novel imprinted loci together with non-imprinted loci. The mean SINE content of FANTOM2 Candidate-Associated CGIs (14.2%) is not significantly different than for Control CGIs, indicating that many non-imprinted genes (some of which may be downstream regulatory targets of imprinted genes) are included in the FANTOM2-candidate set.

Figure 4: Linear Discriminant Classification of CGIs



A linear discriminant function (LDF) was developed to predict imprinted loci using CGI Scores in conjunction with regional SINE Content. Each CGI is represented as a point in the plane by its SINE Content and CGI Score: DMR-CGIs are blue X's, FCA-CGIs are green dots, and other Control CGIs are red dots. All points that fall above the diagonal pink/aqua line satisfy the LDF [0.012 × CGI Score – 9.175 × SINE Content ≥ 0] and all points to the left of the vertical pink/aqua line have SINE Content lower than the median value for all Control CGIs (0.1378). Points that lie in the upper-left region enclosed by the pink lines meet both criteria and are therefore classified as imprinted loci. 9 DMR-CGIs are correctly classified, while 48 FCA-CGIs and 170 other Control CGIs are predicted to be novel imprinted loci by this method (see Appendix 2).