Supplementary Materials

Supplementary Materials and Methods

Plasmids and Cloning Strategies

The sense strand sequences of the oligonucleotides used to generate the promoter mutants spanned the region from –50 to +14 and were:

WT:

CGCCACTGCGGTTCCCGGTTCTAAACTCTCCACCCACCCGGCTCTGCTCAGCTTCTCCCCAGA

ATG^{-6}:

CGCCACTGCGGTTCCCGGTTCTAAACTCTCCACCCACCCGACT**ATG**CTCAGCTTCTCCCCAGA

ATG^{-6}Control:

CGCCACTGCGGTTCCCGGTTCTAAACTCTCCACCCACCCGACT**TAG**CTCAGCTTCTCCCCAGA

ATG^{-24}:

CGCCACTGCGGTTCCCGGTTCTACC**ATG**GCCACCCACCCGGCTCTGCTCAGCTTCTCCCCAGA

ATG^{-24}Control:

CGCCACTGCGGTTCCCGGTTCTACC**TAG**GCCACCCACCCGGCTCTGCTCAGCTTCTCCCCAGA

ATG^{-42}:

CGCCACT**ATG**GTTCCCGGTTCTAAACTCTCCACCCACCCGGCTCTGCTCAGCTTCTCCCCAGA

ATG^{-42}Control:

CGCCACT**TAG**GTTCCCGGTTCTAAACTCTCCACCCACCCGGCTCTGCTCAGCTTCTCCCCAGA

Isolation of RNA and 5’RACE Analysis of Transcription Start Sites

The *PD1* transgenic mouse strain used for in vivo start site analysis, B10.*PD1*, was described previously (Singer et al. 1982). Spleen cells were isolated from B10.*PD1* mice and cultured at 1x10^{6} cells/ml in complete RPMI 1640 supplemented with 10% FBS, 2 mM L-glutamine, 55 uM β-mercaptoethanol, 100 uM minimal essential amino acids, 1 mM sodium pyruvate, 20 mM HEPES pH 7.2, and gentamicin sulfate (10 ug/ml) for 48 hours. Total RNA was isolated using RNA STAT-60 (TEL-TEST, Inc.). Transcription start sites were determined with the SMART RACE cDNA Amplification Kit (Clontech) using 1 mg of total splenic RNA; RACE-Ready cDNA was subsequently PCR-amplified using the SMARTII and *PD1* exon 1 (GGTGGTATATCCAGTGATTTTTTTCTCCAT) or murine H2-K^b (CCGCCCTGGCTCCGACTCAGACCCGC) gene specific primers. Radiolabeled ^{32}P-αdCTP (Amersham) was incorporated during the last five cycles; radiolabeled products were resolved on a 8% denaturing acrylamide gel.

Genome-Wide Analysis for ATG deserts and for TSS

All programs were written in Perl and statistical analyses were performed using R and S plus packages**.** Genomic sequences were downloaded from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/, ftp://ftp.ncbi.nih.gov/genomes/M_musculus/, ftp://ftp.ncbi.nih.gov/genomes/R_norvegicus/, and ftp://ftp.ncbi.nih.gov/genomes/Saccharomyces_cerevisiae/. CpG island mapping was from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/maps/mapview/BUILD.34/ seq_cpg_islands.md. Two TATAA search criteria were used, TATAAT in the 200 bp upstream or TATA(A/T)(A/T) in the 40 bp upstream sequences. ATG counts were computed in the sliding window of 100 bases. -1 indicates the first 100 bases upstream the transcript start site (TSS) where +1 marks the 100 bases downstream from the TSS. TSS were as defined by the annotation for each gene in the database. Permutation analyses for ATG distribution drew a sample from the all genes equal in size to that of the TATAA promoter set and evaluated the frequency with which ATG codons were observed. The analyses were performed by resampling 1000 times. The genes and sequences for the TSS analysis were downloaded from http://dbtss.hgc.jp/. The definition of multiple TSS was based on the method described in Suzuki et al., (Suzuki et al. 2004) as follows. Let x be the offset vector that contains the list of distances between each clone and refseq. Any outliers in x outside the interval [m-4*sd, m+4*sd] are removed, then the standard deviation (sd) is computed. Here, x is the vector defined above and m is the median of x. The decision to call a multiple TSS gene or a single TSS gene is based on the sd. A gene is classified as having mutiple TSS if sd>=5. Otherwise, the gene is classified as having a single TSS.

ATG Desert Predictor using Multiple Species Sequence Alignments

Multiple species sequence alignments among human, mouse, and rat were derived from GoldenPath (http://hgdownload.cse.ucsc.edu/goldenPath/hg16/humor/). We extracted 2 kb of promoter sequence from the conserved regions. The ATG count in each of ten 100-bp windows was calculated for mouse and rat sequences as described for the human sequence. We set out to develop an ATG desert gene algorithm using conserved features among the three species. Two related algorithms were developed. One used two window ranges from w1 to w2 and w3 to w4 in the predictor. The predictor yp is defined as following:

Let Mj be the mean ATG count cross windows from w1 to w2 for the j-th organism. Let Nj be the mean ATG count cross windows from w3 to w4 for the j-th organism. For each organism, we define a predictor Yj based on Mj and the threshold x0: Yj=1 if Mj > x0, otherwise Yj=0. When Yj is 1, the j-th organism is not an ATG desert in [w1,w2]. For each organism, we define a predictor Zj based on Nj and the threshold z0: Zj=1 if Nj < z0, otherwise Zj=0. When Zj is 1, the j-th organism is not rich in ATG count in [w3,w4]. Define yp=0 if (Y1+Y2+Y3 < sum.cut1) or (Z1+Z2+Z3 > sum.cut1.b), otherwise yp=1. Y1+Y2+Y3 is the number of species for which [w1,w2] is not predicted to be an ATG desert. Z1+Z2+Z3 is the number of species in which [w3,w4] is not enriched in ATG codons. When sum.cut1=2 and sum.cut1.b=1, yp = 0 if [w1,w2] is an ATG desert in at least 2 organisms or [w3,w4] is not enriched in the ATG count in at least 2 organisms, otherwise yp=1. Denote cut1=x0 and cut1.b=z0. We found that w1= -6 w2= -1 w3= -10 w4= -7cut1= 1.4 sum.cut1= 2 cut1.b= 1.6 and sum.cut1.b= 1 gave statistically significant results.

In the other algorithm, the predictor is a two-stage threshold function. Let x0 and y0 be two thresholds. x0 is a number between 0.3 and 1.5 and y0=1, 2, or 3. Given a window range from w1 to w2, let Mj be the mean ATG count across windows from w1 to w2, for the j-th organism. For each organism, we define a predictor Yj based on Mj and the threshold x0: Yj=0 if Mj < x0, otherwise Yj=1. Based on the three predictors Y1, Y2 and Y3, we define a combined predictor yp: let xp=Y1+Y2+Y3, yp = 0 if xp < y0, otherwise yp=1. When x0=1.5 and y0=3, yp=0 means at least in one of organisims [w1,w2] is an ATG desert. Select predictors should also satisfy the following conditions: p.value < 0.05, specificity > 0.8, sensitivity > 0.15 and odds.ratio > 2.5. The parameters of w1 = -10 w2 = -1 x0 = 1.5 y0 = 3 gave statistically significant

Supplementary Figure Legends

Supplementary Figure 1. MHC class I genes contain an ATG desert spanning the TSS region. (A) Schematic of the *PD1* class I gene, highlighting distal (-1090 bp to –454 bp) and proximal (-454 bp to +31 bp) promoter regions and coding sequences (not to scale). Transcription factor binding elements USF, Enh A, ISRE and CRE in the proximal promoter region are indicated by gray rectangles. The initiator methionine is indicated by the red vertical line (arrow) and coding exons I-VIII are represented as black rectangles. The vertical blue lines indicate upstream ATGs in the proximal and distal promoter regions. The horizontal gold line indicates the ATG desert spanning the proximal promoter region and extending into the coding sequences. ATG codon frequencies in the distal and proximal promoter regions are shown at the bottom **(B)** Analysis of the ATG codon frequencies upstream and downstream of transcription initiation (indicated by the red arrow at position 0) of human class Ia genes. Each point represents the ATG frequency within a 100 bp window either upstream or downstream of transcription initiation. The shaded area represents the ATG desert.

Supplementary Figure 2. An ATG desert is specifically correlated with the absence of a TATAA box. ATG codon frequencies were analyzed in the region 2 Kb upstream and 2 Kb downstream from transcription initiation using a dataset of 16,544 unique genes. The major transcript initiation site for any given gene is located at position 0 on the X axis. Each point represents the frequency of observed ATG codons within a 100 bp window; 95% confidence intervals are indicated by the triangles. In this analysis, TATAA was defined as the sequence TATA(A/T)(A/T) occurring within 40 bp of transcription initiation. CGI were identified as sequences with a CpG content >0.55, a length of >500 bp and observed/expected ratio >0,65. The groups are designated **(A)** CGI; TATA(A/T)(A/T), **(B)** No CGI; TATA(A/T)(A/T), **(C)** CGI; No TATA(A/T)(A/T), **(D)** No CGI; No TATA(A/T)(A/T). The number of genes represented in each group is in parenthesis.

Supplementary Figure 3. ATG deserts are associated with promoters in the yeast genome. ATG codon frequencies were analyzed in the regions 4 Kb upstream and downstream of the major TSS. In this analysis, TATAA was defined as the sequence TATA(A/T)(A/T) occurring within 40 bp of transcription initiation. The major transcript initiation site for any given gene is located at position 0 on the X axis. Each point represents the frequency of observed ATG codons within a 100 bp window; 95% confidence intervals are indicated by the triangles.

Supplementary Figure 4. ATG deserts are correlated with non-TATAA promoters and multiple TSS. The DBTSS database (http://dbtss.hgc.jp/) was utilized to evaluate the TSS pattern of 1,019 genes relative to the presence or absence of a TATAA box. Single versus multiple TSS were as defined by Suzuki et al. (Suzuki et al. 2001). ATG codon frequencies were analyzed in the region 2 Kb upstream and 2 Kb downstream from transcription initiation. The major transcript initiation site for any given gene is located at position 0 on the X axis. Each point represents the frequency of observed ATG codons within a 100 bp window; 95% confidence intervals are indicated by the triangles. In this analysis, TATAA was defined as the sequence TATA(A/T)(A/T) occurring within 40 bp of transcription_initiation. The groups are designated **(A)** TATA; Multiple TSS, **(B)** TATA; Single TSS, **(C)** No TATA; Multiple TSS, **(D)** No TATA; Single TSS.

Supplementary Figure 5. ATG deserts are associated with promoters with multiple TSS. The DBTSS database (http://dbtss.hgc.jp/) was utilized to identify the TSS pattern of 1,019 genes. Single versus multiple TSS were as defined by Suzuki et al. (Suzuki et al. 2001). For those promoters for which alignments with rat and mouse genes were possible, the second ATG desert algorithm described in Methods was applied separately to the sets of multiple TSS and single TSS. (**A)** human promoters with multiple TSS; (**B)** human promoters with single TSS; (**C)** mouse promoters with multiple TSS; (**D)** mouse promoters with single TSS; (**E)** rat promoters with multiple TSS; (**F)** rat promoters with single TSS. The major TSS for any given gene is located at position 0 on the X axis. Each point represents the frequency of observed ATG codons within a 100 bp window; 95% confidence intervals are indicated by the triangles.

Supplementary Table I. Characterization of CGI in MHC Class I genes

Gene	**Class**	**Expression**	**CGI**	**TATAA**	**Upstream of Met^{+1}**	**Total Length**	**G+C%**	**CpG0/e**
Pig								
PD1	Ia	U	Y	N	586	1860	0.638	0.742
*PD1*4	Ia	U	Y	Y	169	1585	.666	0.770
*PD*6	Ib	B	Y	N	174	1703	0.569	0.648
Human								
HLA-A	Ia	U	Y	N	384	1802	0.640	0.739
HLA-B	Ia	U	Y	N	279	1604	0.647	0.739
HLA-C	Ia	U	Y	N	295	1670	0.648	0.708
HLA-E	Ib	U	Y	Y	253	1213	0.625	0.743
HLA-F	Ib	R	Y	N	194	1555	0.636	0.713
HLA-G	Ib	R	Y	N	231	1398	0.647	0.676
Mouse								
H2K	Ia	U	Y	Y	401	1674	0.639	0.790
H2D	Ia	U	Y	Y	397	1584	0.647	0.744
Q1	Ib	R	Y	Y	185	1106	0.632	0.679
Q2	Ib	R	Y	Y	193	1383	0.647	0.727
Q4	Ib	U	Y	Y	238	1496	0.640	0.700
Q5	Ib	R	Y	Y	225	1282	0.630	0.746
Q8/9	Ib	B	Y	N	333	1447	0.642	0.707
Q10	Ib	R	Y	Y	316	1523	0.650	0.716
T3	Ib	R	N	N	-	-	-	-
T10	Ib	U	N	Y	-	-	-	-
T22	Ib	U	N	Y	-	-	-	-
T24	Ib	U	N	N	-	-	-	-
M3	Ib	U	Y	N	35	514	0.555	0.608
M9	Ib	NE	N	Y	-	-	-	-
M10	Ib	R	N	N	-	-	-	-

Key: U-ubiquitous, R-restricted, B-broad expression, NE-not known to be expressed

Among the mouse genes, the classical class I genes (class Ia) that encode transplantation antigens, as well as the non-classical (class Ib) Qa genes also occur within CGI. In this gene family, the presence of a CGI around the promoter is not correlated with the absence of a canonical TATAA box, since all but one of the mouse class I genes has a TATAA element in the promoter. It is also not correlated with in vivo patterns of expression, since both tissue restricted and ubiquitously expressed class I genes contain CGI. In contrast, the non-classical class Ib genes located within the T and M regions of the murine MHC, which are highly divergent from the classical class I genes and are expressed in a tissue restricted fashion, are not contained within CGI. Therefore, there is no clear correlation between the presence of a CGI and any single attribute of class I promoter structure or pattern of expression. The only correlation observed relates to the position of the mouse class I genes relative to the centromere: the presence of a CGI in the class I promoter erodes with increasing distance of the gene from the centromere (data not shown).