9-11-2009

REDEFINING CpG ISLANDS USING A HIDEEN MARKOV MODEL

Hao Wu
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Brain Caffo
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Harris A. Jaffee
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Andrew P. Feinberg
The Johns Hopkins University, School of Medicine, Molecular Medicine

Rafael A. Irizarry
Johns Hopkins University, Bloomberg School of Public Health, Department of Biostatistics, ririzarr@jhsph.edu

Suggested Citation

Wu, Hao; Caffo, Brain; Jaffee, Harris A.; Feinberg, Andrew P.; and Irizarry, Rafael A., "REDEFINING CpG ISLANDS USING A HIDEEN MARKOV MODEL" (September 2009). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 199.
http://biostats.bepress.com/jhubiostat/paper199

Redefining CpG Islands Using a Hidden Markov Models

Hao Wu, Brian Caffo, Harris A. Jaffee, Andrew P. Feinberg, and Rafael A. Irizarry[*]

June 10, 2009

Abstract

The DNA of most vertebrates is depleted in *CpG* dinucleotides; C followed by a G in the 5' to 3' direction. CpGs are the target for DNA methylation, a chemical modification of cytosine (C) heritable during cell division and the most well characterized epigenetic mechanism. The remaining CpGs tend to cluster in regions referred to as CpG islands (CGI). Knowing CGI locations is important because they mark functionally relevant epigenetic loci in development and disease. For various mammals, including human, a readily available and widely used list of CGI is available from the UCSC Genome Browser. This list was derived using algorithms that search for regions satisfying a definition of CGI proposed by Gardiner-Garden and Frommer more than 20 years ago. Recent findings, enabled by advances in technology that permit direct measurement of epigenetic endpoints at a whole-genome scale, motive the need to adapt the current CGI definition. In this paper we propose a procedure, guided by hidden Markov models, that permits an extensible approach to detecting CGI. The main advantage of our approach over others is that it summarized the evidence for CGI status as probability scores. This provides flexibility in the definition of a CGI and facilitates the creation of CGI lists for other species. The utility of this approach is demonstrated by generating the first CGI lists for invertebrates, and the fact that we

[*] To whom correspondence should be addressed. E-mail: ririzarr@jhsph.edu

can create CGI lists that substantially increases overlap with recently discovered epigenetic marks. A CGI list and the probability scores, as a function of genome location, for each specie are available at `http://www.rafalab.org`.

1. INTRODUCTION

DNA methylation is a type of chemical modification of DNA that can be inherited without changing the DNA sequence. This type of heritable mechanism is referred to as *epigenetic inheritance*. DNA methylation involves the addition of a methyl group to DNA and typically occurs at a C followed, in the 5' to 3' direction, by a G. Biologist refer to this dinucleotide as a *CpG*, where the p implies the 5' to 3' direction. Figure 1A is a simplified illustration of how DNA methylation is maintained during cell division. DNA methylation is of particular interest because it is involved in gene regulation. It affects the transcription of genes in two ways. First, the methylation of DNA can impede the binding of transcriptional proteins to the gene, thus blocking transcription. Second, methylated DNA may be bound by proteins that start a series of chemical events that result in the formation of compact DNA that readers it inactive. Note that although two cell types in an organism have the same genome, their methylation pattern can be different (Figure 1B). The fact that DNA Methylation is heritable makes it the most prominent mechanism used by differentiated cells to pass tissue specific transcription patterns to daughter cells in cell division. Therefore, DNA methylation is regarded as the "fifth base" of the genome and is of great interest to biologists.

[FIGURE 1 AROUND HERE]

The DNA of most vertebrates is depleted in CpG dinucleotides. The remaining CpGs tend to cluster in regions referred to as CpG islands (CGI) (Figure 2). Interest in CGI grew when it was demonstrated that, in vertebrates, they are enriched in regions of the genome involved in gene transcription referred to as *promoters* (Bird 1986). Furthermore, many investigators have observed altered DNA methylation of CGI in development and cancer (Feinberg 2007).

Potash, Sabunciyan and Feinberg (2008) recently demonstrated that *CGI shores*, defined as regions within 2,000 bases pairs but not inside CGI, are useful predictors for genomic locations that are differentially methylated across different tissues and between cancer and normal samples.

[FIGURE 2 AROUND HERE]

A specific example of the need for knowledge of CGI locations is their use in the construction of high-throughput assays. The traditional technique for measuring DNA methylation, bisulfite modification-based sequencing, is labor intensive and not suitable for genome wide studies. New molecular biology techniques, along with the use of microarrays or second generation sequencing technologies, has made high throughput unbiased methylation profiling feasible. However, whole-genome assays are too costly for most research groups. Knowledge of CGI locations provide manufacturers a way to construct cost-effective products that focus on regions known to be associated with important epigenetic events (Agilent 2008; Meissner et al. 2008).

Although existing CGI lists have been widely used, comprehensive measurements of methylation, enabled by recent advances in technology, demonstrate that the current definition needs to be improved. Furthermore, the current definition was developed for humans and interest in measuring DNA methylation in other organisms motivate the need for a more general definition. In Section 2 we describe existing approaches to detecting CGI and point out their limitations. In Section 3 we motivate the need for a new approach and the statistical model that we use to redefine the concept of a CGI. In Section 4 we present the model. In Section 5 we describe improvements over existing approaches obtained from fitting our model. Finally, in Section 6 we summarize our findings.

2. PREVIOUS WORK

A formal definition of a CGI was provided by Gardiner-Garden and Frommer (1987). A CGI is defined as a region of at least 200 base pairs, with the proportion of Gs or Cs, referred to as *GC-content*, greater than 50%, and observed to expected CpG ratio (O/E) greater than 0.6. The observed to expected ratio is calculated by dividing the proportion of CpG dinucleotides in the region by what is expected by chance when bases are assumed to be independent outcomes of a multinomial distribution. The formula currently used is

$$\text{O/E} = \frac{\#CpG/N}{\#C/N \times \#G/N}$$

with N the number of base pairs in the segment under consideration. Various computer algorithms have been developed that efficiently scan the genome for regions satisfying the definition. The most widely used CGI list is based on this definition and is hosted by the UCSC Genome Browser (Kent et al. 2002). However, this definition is somewhat arbitrary, because the choice of the cut-offs has a great influence on what is considered an island. The cut-off choice was likely derived from exploratory data analysis (Gardiner-Garden and Frommer 1987, Figure 1), but neither a biological argument nor a formal statistical motivation was used.

Alternative algorithmic definitions have been proposed. For example, Takai and Jones (2002) considered slightly different cut-offs and demonstrated that by using a minimum length of 500 base pairs (bp), a minimum GC-content of 55%, and a minimum O/E of 0.65, the enrichment for promoter regions of genes was largely not affected. However, most undesirable Alu-repetitive elements were excluded from the UCSC Genome Browser's CGI list. Repetitive elements are sequences that appear over and over again on the genome. The Alu sequences appears more than 1,000,000 times. These are not associated with Epigenetic marks but some satisfy the CGI definition. Therefore, we rather exclude them. However, Alu sequences are easily identified and can be filtered without altering the CGI definition.

Glass et al. (2007) described a completely different algorithm. For every CpG they recorded the length of a segment needed to cover the nearest 27 CpGs. They then observed that, for certain species, a histogram of these lengths shows a bimodal distribution. The histogram was used to select a cut-off and regions associated with the first mode are defined as CpG "clusters" (their terminology for CGI). However, both these alternative definitions also depend on cut-offs based on a difficult to interpret scale.

Because we assume that the underlying structure of the genome includes unobserved states (CGI and baseline), which are presumed to be correlated along the genome (see for example Figure 2), Hidden Markov Models (HMMs) are a natural method to consider. Churchill (1989) introduced the use of HMM to sequence analysis. More recently, HMMs have been effectively used to partition genomes into segments of similar stochastic structure (Muri 1998; Nicolas, Bize, Muri, Hoebeke, Rodolphe, Ehrlich, Prum and Bessières 2002; Boys and Henderson 2004, for example). In these approaches, the hidden state is assumed to be a homogeneous first order Markov chain. The distribution of the observed base at location t, conditioned on the hidden state, is a heterogeneous first order Markov chain. States are then inferred from the base-to-base transitions observed in the genome in question. In the examples cited above, this approach is effectively used to discover heterogeneities in the genome of bacteria (Nicolas et al. 2002) and to segment these genomes (Boys and Henderson 2004).

In general, HMMs have been extensively used in sequence analysis to discover functional elements in various genomes. In a seminal book on the topic, Durbin (1998) proposes the use of HMMs for the task of detecting CGIs. Specifically, eight states are assumed: the four nucleotides in each of the two states (CGI and baseline). Regions for which the state (HMM or baseline) is predetermined (using the current definition) are used to estimate the transition probabilities. With the transition probabilities in place and a sequence of dinucleotides, CGI and baseline states can be predicted by fitting an HMM.

3. MOTIVATION

Recent advances in technology have enabled high-throughput measurement of epigenetic events, such as differentially methylated regions (DMRs) across tissue types. Newly available data has motivated the need for a more flexible CGI definition. For example, we examined data published by Irizarry et al. (2008) and found many DMRs not associated with CGIs but that were nevertheless in the shores of CpG-enriched sequences. For example, one DMR reported by Irizarry et al. (2008) was within $1,000$ bp of a CpG cluster not currently defined as a CGI (Figure 3a). Furthermore, this region coincides with a gene promoter. Despite coinciding with two functional elements associated with CGI, this region meets only two of the three criteria of the formal definition: O/E is only 0.5. Therefore this region is not in the Genome Browser list of CGI. Visual inspection of the base composition around other DMRs not associated with CGI demonstrated that this was a general problem (data not shown).

[FIGURE 3 AROUND HERE]

None of the existing competing algorithms solve this problem. By focusing only on promoters of known genes, we find that the definition proposed by Takai and Jones (2002), although successfully filters out more undesirable repetitive regions, results in even less sensitivity for functional epigenetic elements. Furthermore, the Genome Browser list was filtered to remove repeats, which is a viable solution that does not involve changing to a more restrictive definition. The algorithm described by Glass et al. (2007) has limitations as well. A specific problems is that several smaller clusters agglomarate into larger ones (Figure 3b shows an example). As a consequence, relatively long stretches of CpG depleted regions are included in the CGI. Furthermore, the 27 CpG requirement results in a list that leaves out many shorter CpG clusters that are associated with DMRs. For example, the CGI described above (Figure 3a), is excluded.

Similarly, more statistically based approached have limitations. Although the model proposed by Durbin (1998) serves as an elegant illustration, implementing the approach has not yielded a practical method for genome-wide identification of CGI. To elaborate, note that the typical HMM approach in sequence analysis models the transitions between bases directly. When applied to CGIs, the fundamental difference between the two states must therefore be the transition from C to G, with islands having a bigger transition probability. However, below we demonstrate that for this approach to fit the data, we would have to include much more than two hidden states, due to the variability in base composition observed in most genomes. Moreover, in our experience, the level of complexity required by an HMM, applied to the individual bases, impedes the development of a procedure useful for the creation of CGI lists.

If CGIs are simply a cluster of CpGs, then a procedure that scans through the genome searching for regions with larger than expected CpG rates would suffice. However, the evolutionary theory for CpG islands motivates a more sophisticated approach. Briefly, the human genome is depleted of CpG because the mutation rate for this specific dinucleotide is higher than others (Venter et al. 2001). CGIs are believed to be the result of certain segments of the genome being somewhat protected from the mechanism that leads to this mutation. This is a possible explanation for the association of CGI and locations relevant to development. This evolutionary argument, based on differing mutation rates, suggest that the fundamental property that defines a CGI is not the CpG density per se but the CpG density conditioned on GC-content. This is because regions that originally had high GC-content had more CpG dinucleotides which, even unprotected, resulted in relatively high CpG counts. Gardiner-Garden and Frommer's definition, based on the observed to expected ratio as opposed to just the number of CpG, agrees with the above described theory. Our data exploration, described below, supports and builds on this approach.

We divided the human genome into non-overlapping segments of length 256 bases after removing the Alu-repetitive-elements. Figure 4 shows a histogram of the CpG rates of these segments. This figure does not provide a clear cut-off, based on CpG rates, for distinguishing CGI from baseline. However, if we stratify the segments by GC-content (Figure 5), distinct bi-modal distributions of CpG rates are observed. The two modes support the existence of two states: CGI and baseline. The fact that the center of the two modes increases with GC-content implies that we should consider rates of CpG counts to the GC content of the bin. That is, we consider the number of CpGs relative to a quantity measuring the number of opportunities for CpGs, similar to considering the number of events is relative to the size of the risk set in survival analysis. Note that the O/E concept of Gardiner-Garden and Frommer, is a clever and simple method for adhering to this principal.

Our data exploration revealed another interesting characteristic of the genome. Figure 6A shows GC-content for a representative region of the genome (with no repetitive elements). Note that there appears two states for GC-content as well. Figure 6B shows a density plot of GC-content for the entire genome, with mixture components obtained from fitting a HMM, described in detail in Section 4. In Section 4 we describe the relevance of this characteristic in our approach to defining islands.

Figures 2 and 5 support the claim that CpGs are clustered and that there are two states of O/E. Therefore, a two-state hidden Markov model is a natural method to consider. However, modeling the emission probability at a single location is complicated because GC-content, needed to compute O/E, varies widely across the genome, as seen in Figure 6. Another complication is that the distribution of CpG counts at a single location is somewhat complicated, because outcomes from consecutive locations are not independent. For example, it is impossible to have two consecutive CpGs. In Section 4 we described a procedure, motivated by hidden Markov models, that overcomes the described problems of existing approaches and the difficulties of modeling sequence data directly. By modeling CpG counts in small bins instead of base-to-base transitions, the complexity of the emission model is greatly reduced. The models are therefore relatively simple and can be fitted without cut-off choices which facilitated the extension to species for which CpG islands have never been reported.

4. MODEL

For any given genome, we assumed that each chromosome is divided into three states: Alu-repetitive elements, baseline and CGI. Because the locations of the Alu repetitive elements are well characterized, they are inherently not of interest for the current statistical problem and therefore such regions were removed. Hence we characterize the problem as that of a semi-Hidden Markov Model, with a known state for Alu repetitive elements. Our analysis then considers the two-state chain conditional on being in a non-Alu repetitive state.

We followed the basic statistical concepts first used by Churchill (1989), described by

Durbin (1998) and used by bioinformatic tools such as MEME (Bailey, Williams, Misleh and Li 2006), MAST (Bailey 1998), and BLAST (Altschul, Gish, Miller, Myers and Lipman 1990). The foundation of these tools is the stochastic modeling of bases in the genome. We denoted $B(t)$ the base at genomic location t, $p_b(t)$ the probability of $B(t) = b$ for $b = A, T, G, C$, and $p_{CG}(t)$ the probability of being CpG at location t. The depletion of CpG implies that the probability of a C at time t followed by a G is less likely than would be predicted by chance under independence: $p_{CG}(t) < p_C(t) \times p_G(t+1)$. We have argued that a useful model for detection of CGI needs two states to describe changes in $p_C(t), p_G(t)$, and $p_{CG}(t)$. However, we have specified three parameters for each genomic location t, resulting in an over-determined system. Placing parsimonious modeling assumptions on the chain of bases that imply in a two-state stochastic process for the chain of CpGs would result in undue complexity. Instead, we describe and motivate simple assumptions that permitted the derivation of a useful model from the general model described above.

We first divided the non-Alu regions into non-overlapping segments of length L bp. For the results shown here we used $L = 16$. This choice is justified in Section 4.2. We denoted $N_C(s)$, $N_G(s)$, and $N_{CG}(s)$ as the number of C, G, and CpG in segment s, and $Y(s)$ the hidden state for segment s with states: $Y(s) = 1$ as CGI and $Y(s) = 0$ as baseline.

We base the data generating process with a hierarchical model that we subsequently fit using direct estimates in a iterative stepped approach rather than a complex joint numerical evaluation with MCMC or equivalent. The most complex portion of the model involves a model for the CG-content counts, $N_C(s) + N_G(s)$. We require a model that adheres to the following: *i*. it must account for jumps in CG content of roughly the same height that define CGI, *ii*. slowly varying trends must also be accounted for *iii*. fitting must be reasonably fast and able to accommodate the large size of the data.

The lowest level of the hierarchy specifies that the proportion of GC content in segment s, $\{N_C(s) + N_G(s)\}/L$, follows a Hidden Markov model with a non-zero conditional mean, $p(s)$, and latent Markov process, $X(s)$ representing the hidden state for segment s with states: $X(s) = 1$ as high GC-count regions and $X(s) = 0$ as baseline. We presume that

$X(s)$ is a stationary first order Markov chain with invariant probabilities $\pi_i = \Pr\{X(s) = i\}$, say, and two by two transition matrix P. For this approach we use a normal approximation and do not force a binomial variance. This gives us added flexibility in the model though requires $\{N_C(s) + N_G(s)\}/L$ to lie away from the 0 and 1 boundaries for the distributional assumptions to be valid. However, this is well indicated by the data.

Let $\{S_j\}$ be the collection of segments defined by a constant latent state. That is: $S_1 = \{1, \ldots, M_1\}$ where M_1 is the smallest index so that $X(M_1) \neq X(M_1 - 1)$, $S_2 = \{M_1 + 1, \ldots, M_2\}$ where $M_2 > M_1$ is the smallest index so that $X(M_2) \neq X(M_2 - 1)$ and so on. This process divides the segments into regions of low or high GC-content.

The hidden Markov models accounts for auto-correlation and fast variation in the chain of GC content. However, there is clearly a component of slow variation in the GC content within segments of similar type (Figure 6) that must be accounted for. We presume the following model on the conditional mean $p(s)$

$$p(s) \mid s \in S_j \text{ and } X(M_j) = i \sim \text{Normal}\{c_i + f(s), \sigma^2\}$$

where $\int_s f(s) = 0$ represents smooth deviations while the additive constant c_i represents jumps in the CG content defined by the HMM.

Finally, we assumed, conditioned on $\{p(s)\}$ and $Y(s) = i$ we assume a HMM model on $N_{CG}(s)$ with Poisson emission probabilities with conditional means

$$a_i \times L \times p_C(s) \times p_G(s) = a_i \times L \times \frac{1}{4} p(s)^2.$$

Here we are making the parsimony assumption that $p_C(s) = p_G(s) = \frac{1}{2}p(s)$. This assumption, though perhaps aggressive if the bin sizes are small, is biologically well motivated. Further, the Poisson assumption is motivated in the next section. Note that the parameters a_1 and a_0 can be interpreted as the O/E for the CGI and baseline regions respectively.

4.1 Motivation for Poisson Model

An important model assumption is that the number of CpG occurrences in a segment of the genome approximately follows a Poisson distribution. Note that the counts are not binomial because one can not have two CpGs in a row. We termed the distribution non-consecutive binomial and proved that, asymptotically, we obtain the same results as if the counts were based on independent Bernoulli trials. Detailed proof can be found at supplemental materials.

We examined the small sample properties of a our random variable using simulations. Figure 8 shows the mass function of a non-consecutive binomial and Poisson are similar for different L and p values.

[FIGURE 8 AROUND HERE]

4.2 Choosing the segment length

The Poisson approximation, described in Section 4.1, requires L to be "large". However, there is a trade-off in that smaller values of L provide better resolution for the edges of CGI. In this section we present a simulation and data-motivated rationale for choosing this parameter.

Our simulations showed that the approximation was appropriate for length larger than $L = 8$ (Figure 8). We further assessed the performance on real data by creating CGI lists as described in Section 5 for the human genome using segment lengths of $L = 8, 16$, and 32. The resulting lists were similar: 96% of the bases in the $L = 8$ CGIs were in the $L = 16$ CGIs, 98% of bases in the $L = 16$ CGIs were in $L = 8$ CGIs, and 93% of bases in the $L = 32$ CGIs were in the $L = 16$ CGIs. However, only 83% of the bases in the $L = 16$ CGIs were in $L = 32$ CGIs. Visual inspection revealed that the reason for this were various instances where smaller proximal $L = 16$ CGIs were engulfed into a larger $L = 32$ CGI. Finally, we created validation plots based on the association of CGI with epigenetic marks for each length as described in Section 5 for each length; $L = 16$ showed the best performance. Therefore $L = 16$ was used throughout this manuscript and we recommend its use in practice. However, we emphasize that for future applications, because of the computational shortcuts proposed, performing a similar sensitivity analysis on this parameter can be easily done.

4.3 Parameter estimation

We used an iterative stepped approach to fit the posited hierarchical model. The benefits of this strategy are many and most notably include the ability to use existing software for fitting, as well as making the computational problem of fitting the model feasible. Moreover, by fitting the model in stages, we thus obtain values based on the most direct evidence. This provides some robustness against model misspecification. However, this approach comes at the cost of theoretical continuity and perhaps leads to understating uncertainty in parameter estimates.

A difficult problem is the assumption of a non-zero conditional for the HMM on $\{N_C(s) + N_G(s)\}/L$. Typically HMM algorithms presume a detrended signal. To address this concern we use an iterative algorithm. To start the iterative algorithm we assume $f(s) = 0$. The standard forward-backward algorithm, as described by Rabiner (1989), was applied to the GC-content data: $\{N_C(s) + N_G(s)\}/L$. This algorithms provides estimates for the for conditional means for each states, i.e. c_0 and c_1, as well as posterior probabilities for each state for each segment s. The posterior probabilities were thresholded to obtain a binary (0 or 1) estimate $\hat{X}(s)$ of $X(s)$. Then for each segment we subtract the means from observed values to obtain the residuals:

$$r(s) = \{N_C(s) + N_G(s)\}/L - c_0^{1-\hat{X}(s)} c_1^{\hat{X}(s)}$$

We then estimate $f(s)$ by applying a smoother to $r(s)$. Specifically, we used a moving weighted average with weights obtained from Tukey's biweight kernel with a window size of 5 segments (80 bases). We then iterated the process. Namely we subtract the smooth estimate, say $\hat{f}(s)$, from the observed GC-content and apply the forward-backward algorithm to $\{N_C(s) + N_G(s)\}/L - \hat{f}(s)$ and repeat the above procedure until convergence.

The use of HMMs and this iterated scheme, as opposed to a complete maximum likelihood solution, for example, is motivated by HMMs established applicability to sequence data, the availability of robust fitting algorithms and the satisfactory performance we have seen on the data. Moreover, as stated above, we have placed a high emphasis utilizing methods that can be easily implemented and use the most direct information available. Convergence is usually obtained quickly, in five iteration or so.

The result of this algorithm is a smoothed estimate of $p(s)$ that accommodates change points from regions of high CG content and a slowly varying trend. By iterating these steps, we mirror a blocked maximization procedure, such as is common in back-fitting and related procedures. At convergence a smoothed estimate of f is obtained as well as estimates for the c_i terms, which represent local constant increases or decreases in CG content.

With the estimate of $p(s)$ in hand, estimating the HMM on N_{CG} is much simpler. Since we assume

$$N_{CG}(s)|Y(s) = i \sim Poission(a_i \times L \times \frac{1}{4}p(s)^2),$$

the HMM can be fitted with standard forward-backward algorithm with EM. The result will give estimates for a_1, a_0 and posterior probabilities for $Y(s)$. Now we have state probabilities for the two latent Markov chains, one defining ares of high CG content, and one defining areas of high CpG content. Here, the areas of CpG content correctly accounts for the number of opportunities for CpG, rather than looking at the raw number in isolation. We estimate the posterior probabilities of being a CGI state, i.e. $Y(s) = 1$, for each segment s. We also obtain the posterior probabilities of being in a high GC-content state, i.e. $X(s) = 1$, for each segment s. Because the forward-backward algorithm calculates these quantities, they are readily available. We can then estimate the states for X and Y using these posteriors.

5. RESULTS

Our main motivation for the development of a new CGI definition was the fact that recently discovered epigenetic marks were not associated with CGI based on the current definition but were associated with CpG-enriched regions. Specifically, many DMRs not associated with existing CGI lists, were in CpG shores. Below we describe how CGI lists based on the results of fitting the HMMs, described in Section 4, improve coverage of these locations. We compare our list, which we refer to as the model-based CGI, to CGI lists provided by the UCSC Genome Browser (Kent et al. 2002), denoted as *Genome Browser* CGI, and the *Glass* *et al.* CGI (Glass et al. 2007).

We created a CGI list by considering regions of locations with posterior probability greater than 0.5. We also found that the CGIs that coincided with regions of baseline GC-content were not associated with epigenetic marks (data not shown) and therefore we filtered these regions. Table 1 shows the joint distribution of the observed posteriors for X and Y. Note, that the majority of locations with evidence of CGI state, occur when the genome is in the high GC-content state.

[TABLE 1 AROUND HERE]

This CGI list covered 95% of the DMRs reported by Irizarry et al. (2008). This is a dramatic increase from the 81% covered by the Genome Browser CGIs and the 86% covered by the Glass et al. CGIs. This improvement was made possible by the flexibility to control specificity. Note that the number of CGIs produced with a posterior probability cut-off of 0.50 was 144,228 and the number in the Genome Browser list is 28,226. To compare lists of similar specificity we created model-based CGI lists with posterior probability cut-offs ranging from 0.50 to 0.999 for the human and mouse genomes. We compared the association of each list with two functional elements: gene promoters and DMRs.

Because the Genome Browser CGIs are mostly annotated on the non-repetitive region, we filtered regions with more than 35% repetitive bases from all lists to make results comparable. To assess sensitivity we computed the percentage of DMRs within 2,000 bases of a CGI. We also performed comparisons similar to those previously used to assess CGI lists. Namely, we compared the percent of gene promoters covered by CGIs for human and mouse, as done by Takai and Jones (2002) and Glass et al. (2007). To assess specificity in a comparable way for the three approaches, we computed the total number of bases covered by each CGI list. Figure 9 shows plots of sensitivity versus specificity.

Glass *et al.* CGIs overlap with a larger percentage than Genome Browser CGIs (66.6% versus 58.2%). However, to achieve this gain in sensitivity, twice as many bases are used. The ability to control specificity with the model-based CGIs demonstrates that only slight improvements over the Genome Browser CGIs are possible at the same specificity level (Figure 9). In contrast, a substantial improvement was achieved by the model-based approach in the overlap with the DMRs. Using a probability cut-off of 0.999 the total lengths of the model-based CGIs (21.3 Mbp), was comparable to the total length of the Genome Browser CGIs (21.1 Mbp), but the overlap with DMR increased from 81% to 86%. A cut-off of 0.975 made the model-based CGIs (41.7 Mbp) comparable in size to the Glass *et al.* CGIs (41.1 Mbp) but the the overlap with DMR increased from 86% to 91%.

Another advantage of our approach is that we can easily fit the HMMs to genomes of other species. We fitted the model to 12 species: *H. sapiens* (human), *P. troglodytes* (chimpanzee), *M. musculus* (mouse), *B. taurus* (cow), *C. familiaris* (dog), *G. gallus* (chicken), *A. mellifera* (bee), *D. melanogaster* (fruit fly), *C. elegans* (worm), *A. thaliana* (Arabidopsis), *E. Coli* and *S. cerevisiae* (yeast). CGI have only been reported for vertebrates. We therefore tested for the presence of CGI by computing a likelihood ratio comparing a model with two states to a model with one state. Of the 12 species we tested, only the unicellular organisms, i.e. yeast and E. Coli, did not have significant evidence in favor of the presence of CGI. We are therefore reporting the first CGI lists for bee, worm, and fruit fly. Previous approaches were not successful because the required cut-offs for these species are very different than for humans. This is demonstrated by examining the fitted a_0 and a_1 parameters. Note that these can be interpreted as the average O/E in the baseline and CGI regions respectively.

[TABLE 2 AROUND HERE]

[FIGURE 9 AROUND HERE]

6. DISCUSSION

We have proposed a procedure for building CGI lists based on HMMs. The main motivation for the development of a new approach was the observation that many DMRs were near regions of high CpG density that did not meet the current definition nor any of the alternative definitions. Our new approach greatly improved the overlap with known DMRs. The improvements achieved with our approach was mainly due to the data-driven nature of the procedure. Many of the CpG dense regions were left out by algorithmic approaches, because they did not satisfy a predetermined rule. Re-running these algorithms with different cut-offs is no easy task. However, generating CGI lists with different cut-off for the HMM-generated posteriors probabilities is trivial.

Figure 10 shows GC-content versus O/E for the model-based human CGI list. The red horizontal and vertical lines are from Gardiner-Garden and Frommer CGI definition (GC content>50%, O/E>0.6). Based on the current definition only the points above the horizontal line and to the right of the vertical line are CGIs. Various of the model based CGI do not satisfy the original definition. A histogram of the lengths of model-based CGIs shows many model-based islands are smaller than the formal definition's requirement of 200 bases (Figure 10b). These figures demonstrate how the added flexibility permits shorter regions with slightly lower O/E.

Our probability-based estimates have units that are interpretable across species. Thus, in a sense, we have transformed the problem onto a standardized scale which will facilitate discussion of thresholding definitions. Because of this, fitting the model to the genomes of other species was simple - no additional user input or algorithmic tweaking was required. To demonstrate this, we fitted the model to the genome of 12 species.

In addition to providing CGI information for these species in isolation, it led to some interesting scientific findings when compared across specifies within taxonomic and evolutionary classes. Strong evidence for the presence of CGI was found for all multi-cellular organisms. The estimated model parameters confirmed that vertebrates are CpG depleted in their baseline level. Invertebrates were not CpG depleted in their baseline levels but showed higher than expected levels in the CGI. Arabidopsis was somewhere in between. Evidence of methylation has been reported for species for which we found evidence of CGI. The fruit fly had the weakest evidence for the presence of CGI. Interestingly, only small amounts of methylation are detected for this organism (Lyko, Ramsahoye and Jaenisch 2000).

A promising application of the newly defined CGIs is the creation of efficient DNA methylation arrays or enrichment schemes for second generation sequencing. For example, we can construct microarrays that tile only CGI shores. Note that if the current Genome Browser definition will miss out on a substantial number of DMRs. Furthermore, it would be possible to construct this array for any species for which the genome has been sequenced. Furthermore, the ability to control specificity will permit us to deal with different array densities.

ACKNOWLEDGMENTS

We thank Herve Pages and Robert Gentleman for help with the Bioconductor packages. The work of R.I and H.J. was partially funded by National Institutes of Health (1R01GM083084-01 and 1R01RR021967-01A2. The work of H.W. was partially funded by T32GM074906. The work of A.F. was funded by NIH Grant P50HG003233.

REFERENCES

Agilent (2008), “http://www.chem.agilent.com/Scripts/PDS.asp?lPage=50884,”.

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990), “Basic local alignment search tool,” *J. Mol. Biol*, 215(3), 403–410.

Bailey, T. (1998), “Combining evidence using p-values: application to sequence homology searches,” *Bioinformatics*, 14(1), 48–54.

Bailey, T., Williams, N., Misleh, C., and Li, W. (2006), “MEME: discovering and analyzing DNA and protein sequence motifs,” *Nucleic Acids Research*, 34(Web Server issue), W369.

Bird, A. (1986), “CpG-rich islands and the function of DNA methylation,” *Nature*, 321(6067), 209–213.

Boys, R., and Henderson, D. (2004), “A Bayesian approach to DNA sequence segmentation,” *Biometrics*, 60(3), 573–581.

Churchill, G. A. (1989), “Stochastic models for heterogeneous DNA sequences,” *Bull Math Biol*, 51(1), 79–94.

Durbin, R. (1998), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge: Cambridge University Press.

Feinberg, A. P. (2007), “Phenotypic plasticity and the epigenetics of human disease,” *Nature*, 447(7143), 433–440.

Gardiner-Garden, M., and Frommer, M. (1987), “CpG islands in vertebrate genomes.,” *J Mol Biol*, 196(2), 261–282.

Glass, J. et al. (2007), “CG dinucleotide clustering is a species-specific property of the genome.,” *Nucleic Acids Research*, 35(20), 6798.

Irizarry, R., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J., Sabunciyan, S., and Feinberg, A. (2008), “Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores,” *Nature Genetics*, . Available online.

Kent, W. J. et al. (2002), “The human genome browser at UCSC,” *Genome Res*, 12(6), 996–1006.

Lyko, F., Ramsahoye, B., and Jaenisch, R. (2000), “Development: DNA methylation in Drosophila melanogaster,” *Nature*, 408(6812), 538–540.

Meissner, A. et al. (2008), “Genome-scale DNA methylation maps of pluripotent and differentiated cells,” *Nature*, 454(7205), 766–770.

Muri, F. (1998), Modelling bacterial genomes using hidden Markov models,, in *COMPSTAT ’98 Proceedings in Computational Statistics*, eds. R. Payne, and P. J. Green, Physica-Verlag, Heidelberg, pp. 89–100.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., Prum, B., and Bessières, P. (2002), "Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models," *Nucleic Acids Res*, 30(6), 1418–1426.

Rabiner, L. (1989), "A tutorial on hidden Markov models and selected applications inspeech recognition," *Proceedings of the IEEE*, 77(2), 257–286.

Takai, D., and Jones, P. (2002), "Comprehensive analysis of CpG islands in human chromosomes 21 and 22," *Proceedings of the National Academy of Sciences*, 99(6), 3740.

Venter, J. et al. (2001), "Initial sequencing and analysis of the human genome," *Nature*, 409, 860–921.

Wu, H., Jaffee, H., Feinberg, A., and Irizarry, R. (2008), "A Species-Generalized Probabilistic Model-Based Definition of CpG Islands," , . Submitted to Nature Methods.

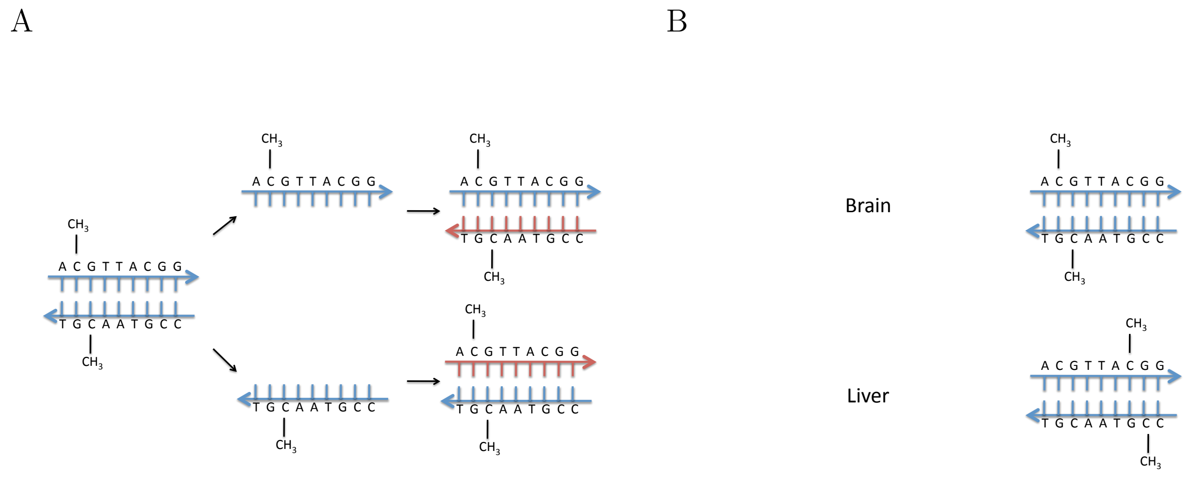

Figure 1: Cartoon illustrating how DNA methylation is inherited in cell division on how it could be involved in tissue differentiation. A) The fact that the complement of a CpG is also a CpG facilitates the the inheritance mechanism. The cartoon illustrates how, during Mitotic cell division, DNA methylation is inherited. B) This cartoon illustrates how two cells can have the same genomic sequence but a different methylation pattern.

Table 1: Joint distribution of posterior probabilities for X (GC content) and Y (CpG rate) on Human hg18 genome. Numbers in each cell are the percentages of bins with posterior probabilities fall in a category. For example, there are 64.3% bins with both probabilites between 0 and 0.1.

	Post. prob. for CpG rate				
Post. prob. for GC content	(0,0.1]	(0.1,0.5]	(0.5,0.9]	(0.9,1]	total
(0,0.1]	64.3	2.5	0.7	0.4	67.9
(0.1,0.5]	1.6	0.1	0.0	0.0	1.7
(0.5,0.9]	1.6	0.1	0.0	0.0	1.7
(0.9,1]	23.0	1.9	1.2	2.6	28.7
total	90.5	4.6	1.9	3.0	100

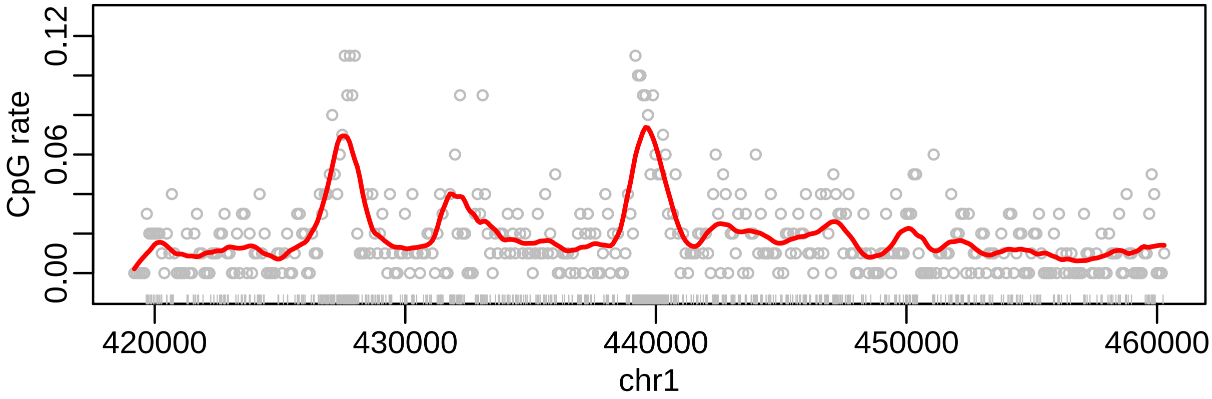

Figure 2: A genomic region of 40,000 bases from chromosome 1 is shown. The ticks on the x-axis represent CpG locations. The points represent CpG rates in segments of length 256 bases The curve is the results of a kernel smoother of the points. Approximately 20% of the genome are Cs and 20% are Gs. Thus we expect about 4% of dinucleotides to be CpG. However, most points are well below rates 4% with two clusters well above 4%. The latter are CpG islands.

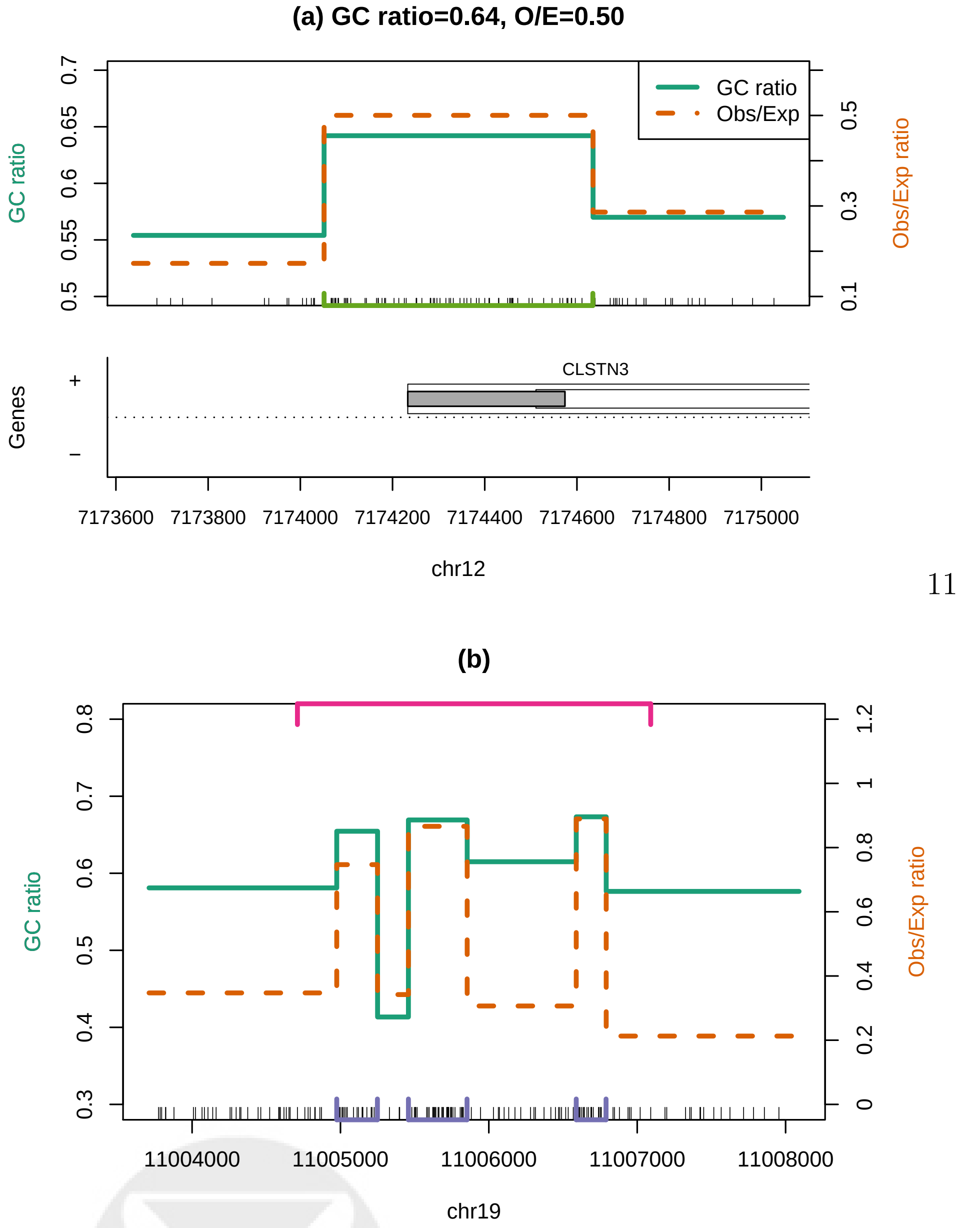

Figure 3: The observed to expected ratio (green) and percentage of G+C (orange) are shown for two regions of the human genome. CpG clusters are denoted with bars along the bottom or top of the plot. A) For a region covering the 5' end of CLSTN3 a CpG dense region that is not in current CGI list is denoted by the lime green bar at the bottom. B) The top (pink) bar denotes one of Glass *et al.* CGI that engulfs three Genome Browser CGIs (denoted with purple bars at end). The regions between the Genome Browser CGI have low observed to expected ratio.

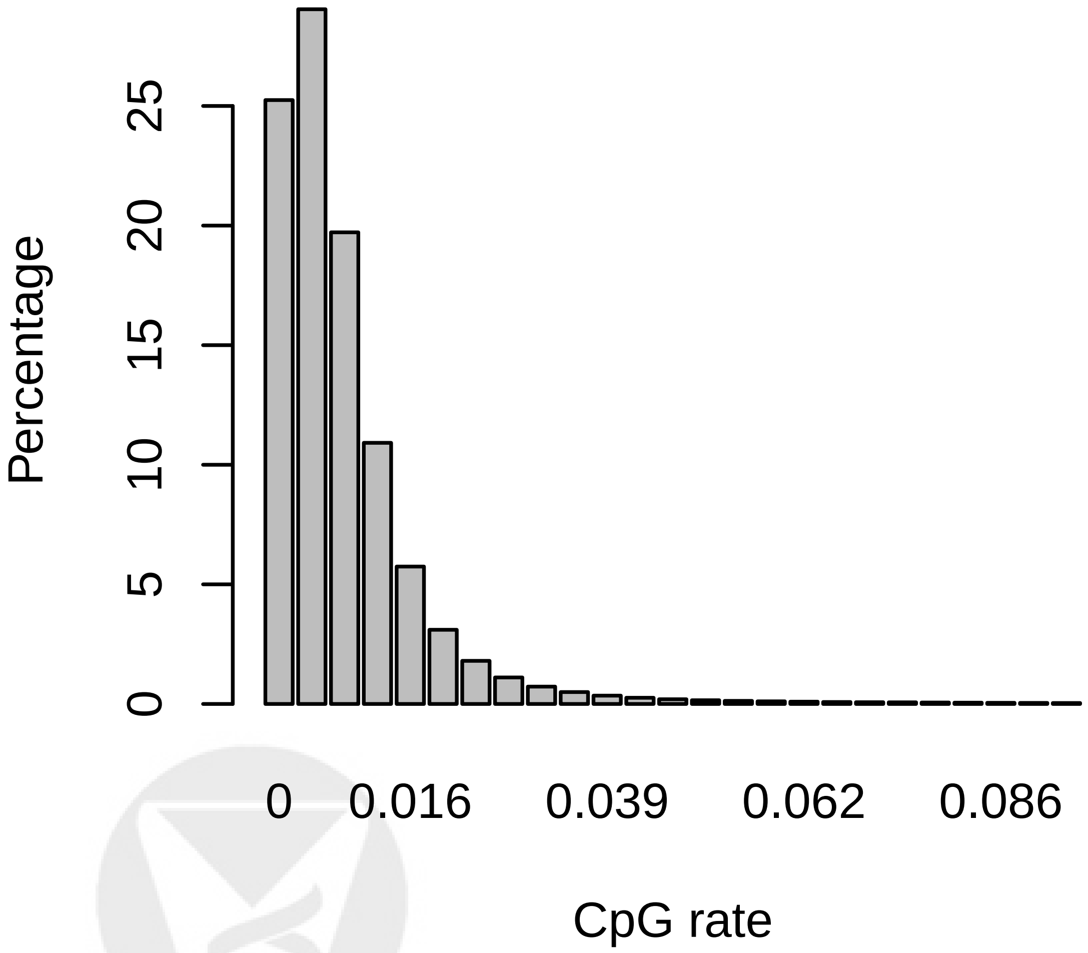

Figure 4: Histogram of CpG rates in non-overlapping genomic segments of length 256 bases.

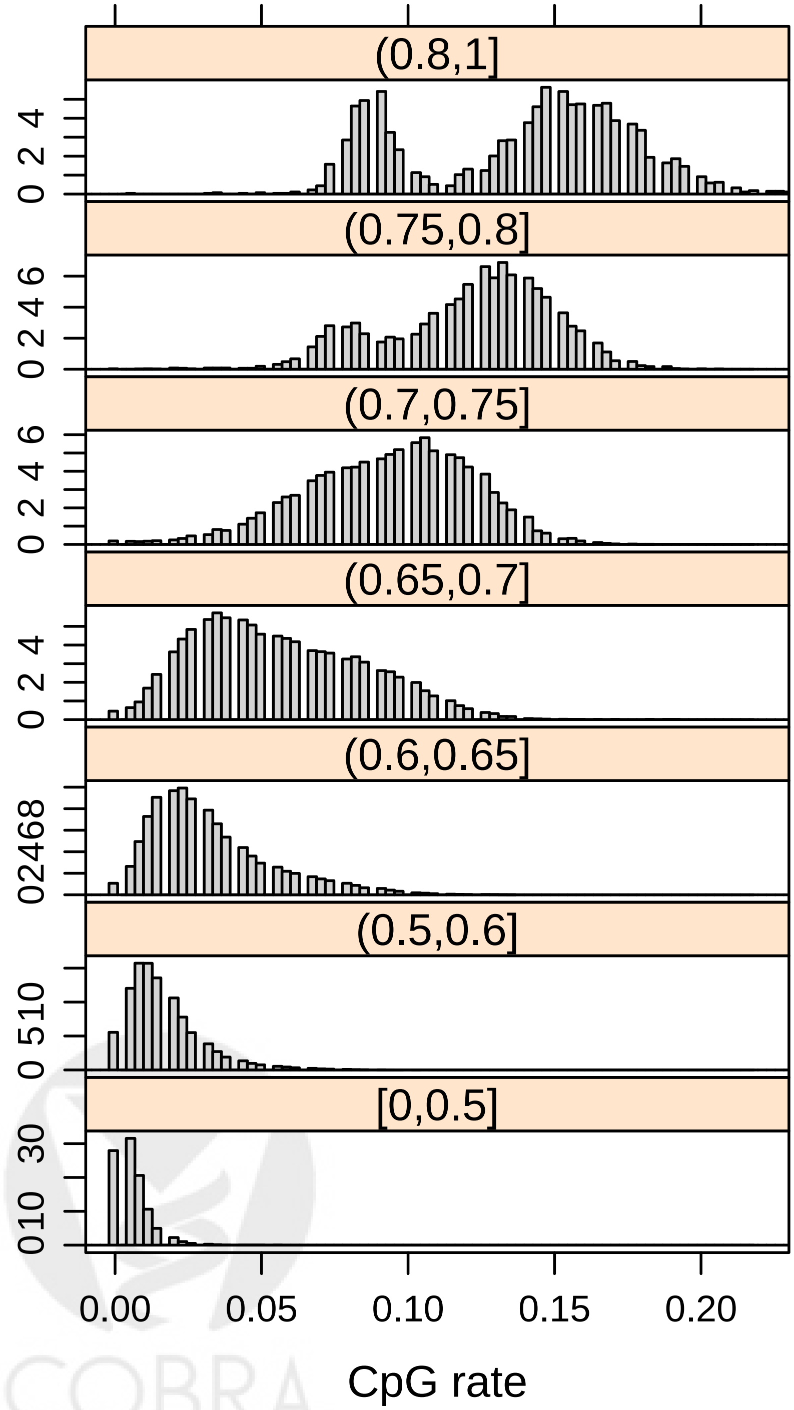

Figure 5: Histogram of CpG rates in genomic segments of length 256 bases, as in Figure 4, but stratified by GC-content. The GC-content strata is shown on top of each histogram.

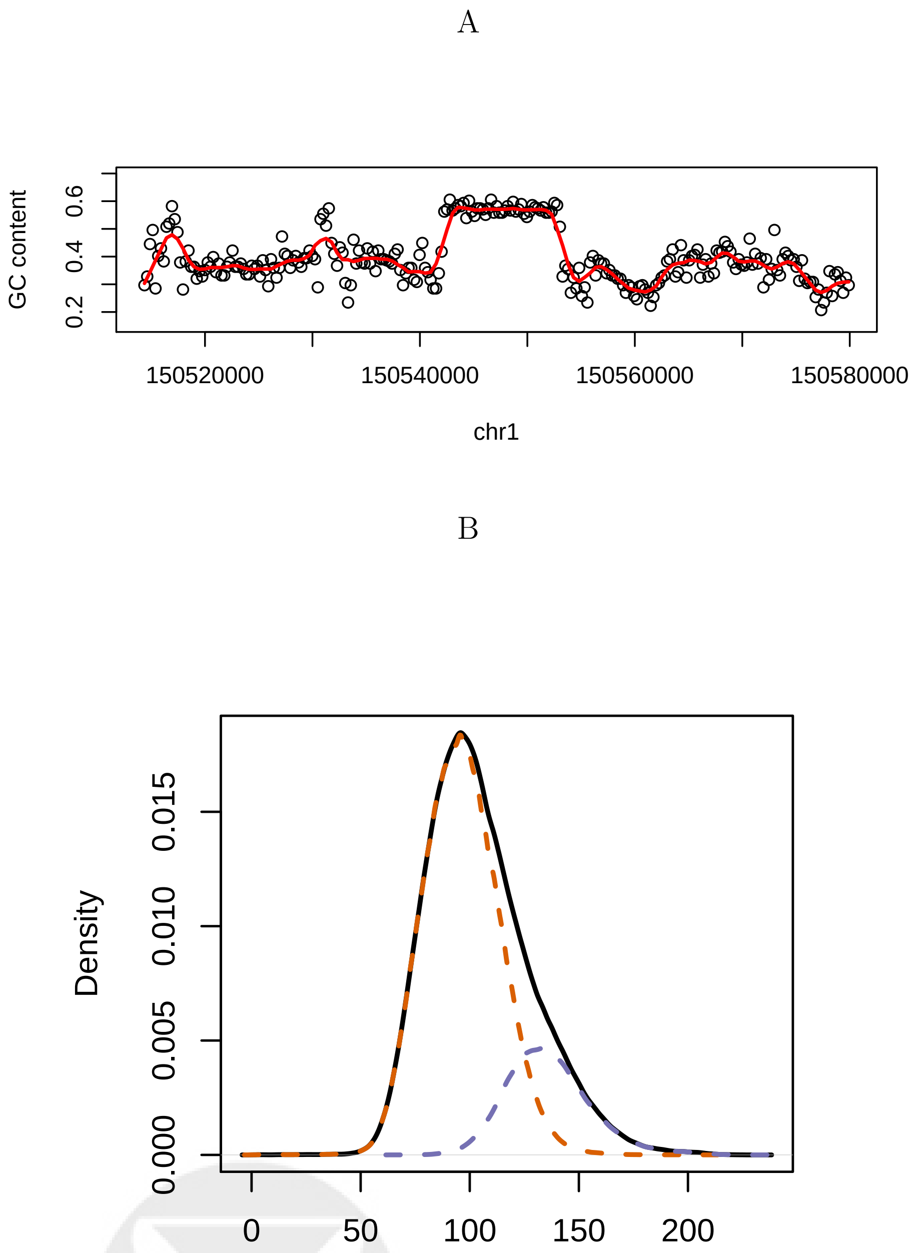

Figure 6: GC-content plots. A) A region with no Alu-repeats was divided into non-overlapping segments of length 256 bases. The points are the GC-content of each segment. The curve is the results of a kernel smoother of the points. B) The solid line is density plot for GC-content for the 256 base segments from all non-Alu-repetitive regions. A HMM (described in Section 4) was fitted to the entire genome and the dashed lines show the density plot for segments from the two states.

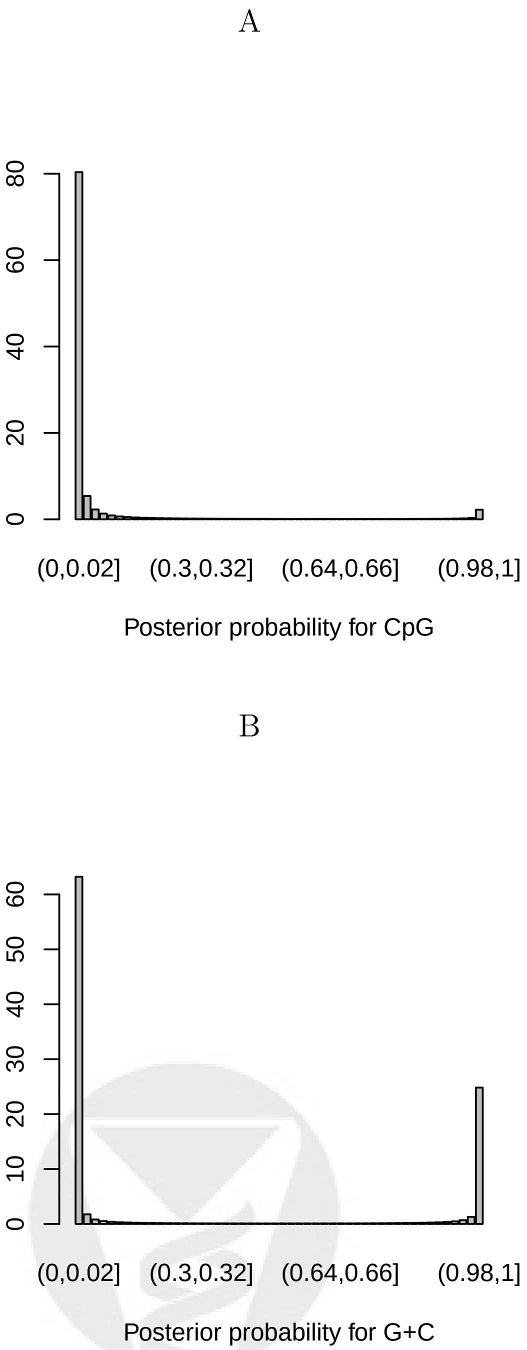

Figure 7: Histograms of posterior probabilities obtained from the hidden Markov model. A) CGI B) GC-content.

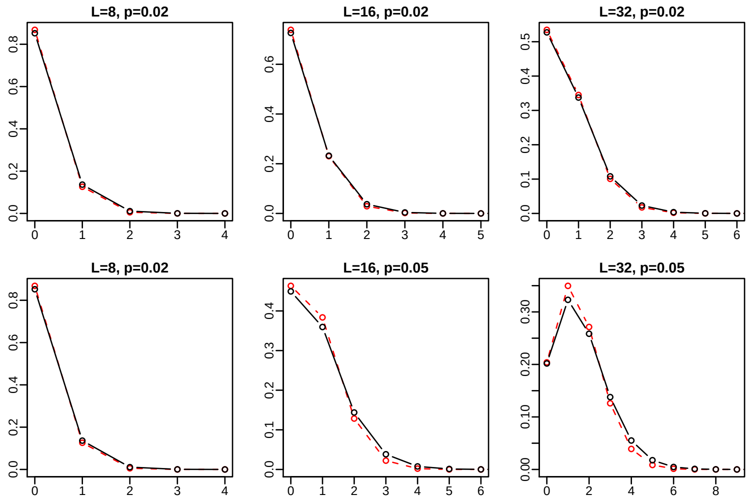

Figure 8: Pmf's for $NCBin(L, p)$ and $Poisson(Lp)$ are similar when L is large and p is small.

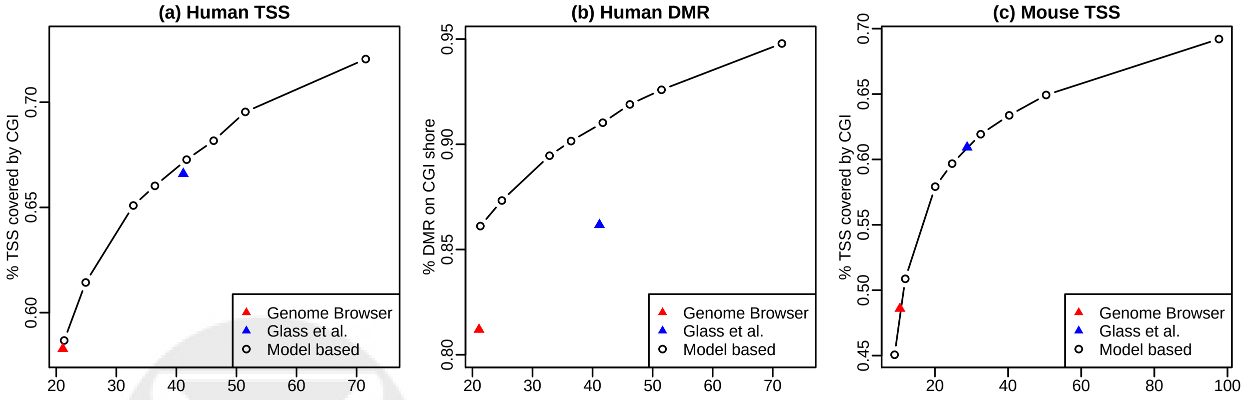

Figure 9: ROC-like plots showing the sensitivity versus total length for different CGI lists (used as a measure of specificity). The sensitivity is defined as the percentage of functional elements associated to CGI. The four figures are for different functional elements. (a) Human (HG18) transcription start sites (TSS), (b) human (HG18) unknown sequence tag found using sequencing, (c) human (HG18) differentially methylated (DMR) and (d) mouse (MM8) TSS, respectively.

Table 2: In the HMM the parameters a_0 and a_1 represent the average observed to expected ratios in the baseline and island regions. The table below shows the estimated parameters in twelve species.

	a_0	a_1
Human	0.15	0.54
Chimp	0.16	0.54
Mouse	0.14	0.44
Cow	0.17	0.49
Dog	0.16	0.62
Chicken	0.18	0.68
Bee	0.77	1.51
Fruit fly	0.84	0.90
Worm	0.83	1.28
Arabidopsis	0.49	0.92
Yeast	0.76	0.77
E. coli	1.12	1.12

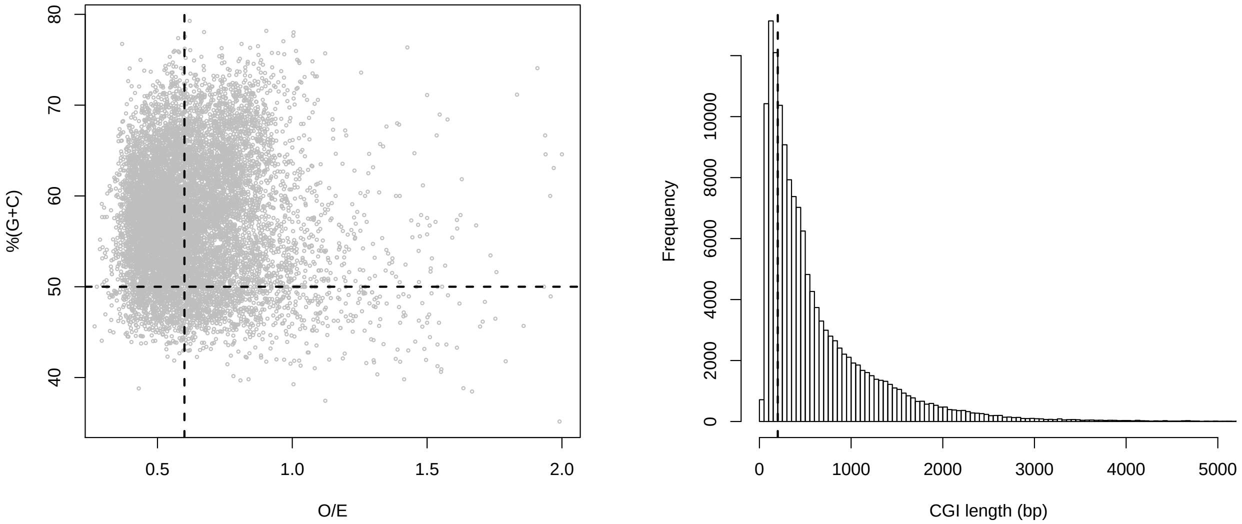

Figure 10: Statistical characteristics of model-based CGI list for Human (HG18). (A) GC content versus O/E. The red vertical and horizontal lines represent the cut-offs used by the Gardiner-Garden and Frommer definition: O/E>0.6, GC content>0.5. (B) Histogram of CGI lengths. The vertical line is at the minimum length requirement of Gardiner-Garden and Frommer CGI definition (200bp)