Determining coding CpG islands by identifying regions significant for pattern statistics on Markov chains

Meromit Singer[1,*], Alexander Engström[2,*], Alexander Schönhuth[2,^]
and Lior Pachter[1,2,3,^]

Departments of [1]Electrical Engineering and Computer Science, [2]Mathematics,
and [3]Molecular and Cell Biology, University of California, Berkeley, USA.
* - these authors contributed equally to this work
^- corresponding authors. AS: A.Schoenhuth@cwi.nl LP: lpachter@math.berkeley.edu

Abstract Recent experimental and computational work confirms that CpGs can be unmethylated inside coding exons, thereby showing that codons may be subjected to both genomic and epigenomic constraint. It is therefore of interest to identify coding CpG islands (CCGIs) that are regions inside exons enriched for CpGs. The difficulty in identifying such islands is that coding exons exhibit sequence biases determined by codon usage and constraints that must be taken into account. We present a method for finding CCGIs that showcases a novel approach we have developed for identifying regions of interest that are significant (with respect to a Markov chain) for the counts of any pattern. Our method begins with the exact computation of tail probabilities for the number of CpGs in all regions contained in coding exons, and then applies a greedy algorithm for selecting islands from among the regions. We show that the greedy algorithm provably optimizes a biologically motivated criterion for selecting islands while controlling the false discovery rate.

We applied this approach to the human genome (hg18) and annotated CpG islands in coding exons. The statistical criterion we apply to evaluating islands reduces the number of false positives in existing annotations, while our approach to defining islands reveals significant numbers of undiscovered CCGIs in coding exons. Many of these appear to be examples of functional epigenetic specialization in coding exons.

1 Introduction

In many genomes, the majority of cytosines that are followed by guanines (CpG sites) are methylated and occurrences of unmethylated CpGs tend to cluster together, in so-called unmethylated islands. The extent of methylation at such islands has been shown to be associated with the expression levels of nearby genes [49], and in many cases aberrant methylation of these islands is associated with diseases such as cancer [24; 42]. Computational methods based on evolutionary principles have been used to annotate these functional regions and are based on the fact that deamination of a methylated cystosine in a CpG dinucleotide results in a TpG (or, in the complementary strand, a CpA dinucleotide). Since only methylated cytosines are deaminated in such a manner, and in many organisms, in particular in mammals, cytosines are methylated at CpG sites, one observes an overall depletion of CGs and frequently observed C $\rightarrow$ T mutations in alignments of related species. Unmethylation in the germline and selection are seen as the dominating mechanisms which shield functional regions from CpG depletion [8; 40]. In the case of coding CpGs, investigation of the extent of conservation to measure methylation (*pro-epigenetic selection*) has only recently been undertaken [8; 30] with results suggesting extensive unmethylation of CpGs in developmentally related genes.

Due to the connection between CpG depletion and methylation, it is possible to identify unmethylated genomic regions based on sequence criteria alone. Interest in sequence-defining features of unmethylated regions which are over-represented in CpG dinucleotides, known as *CpG islands (CGIs)*, is also motivated by the need for CGIs as ground sets in a variety of more specific studies on the functional impacts of methylation (e.g. [20; 43; 49]). While recent advances in high-throughput sequencing have been used to probe methylation states of individual nucleotides at genome-wide resolution [18; 27; 41], providing tools for unpreceded progress in the study of regulatory methylation, such methods cannot annotate a complete set of regions which are unmethylated at some developmental stage or in some tissue. This is particularly relevant in intragenic regions, where the functionality of unmethylated islands is likely to be different from that of unmethylated islands overlapping promotors [21]. While CGIs at promotors are observed to be consistently unmethylated across different cell types and conditions [21], this is not the case for islands in intragenic regions [29], and it is hypothesized that methylation states of CGIs in intragenic regions are highly tissue specific and are mostly involved in the “fine tuning” of expression [21].

The identification of CpG islands requires a formal definition, and a variety of recent epigenomic studies [12; 20; 41; 43; 49; 51] have pointed out that the popular *Gardiner-Garden-Frommer (GF)* criteria [14] that define CpG islands based on thresholds ($\geq$ 200bp, GC % $\geq$ 50 and number observed CpGs/number of expected CpGs $\geq$ 0.6) and also the later, more stringent definition due to Takai and Jones [45] have to be adjusted.

As a consequence, a variety of novel statistical approaches to genome-wide prediction of CGIs have been proposed, leading to improvements in several relevant aspects and allowing genome-wide computation of CGIs in a variety of genomes (Table 1). It is noteworthy that the GF definition depends heavily on the precise implementation of the search algorithm. Since there is no canonical choice for a search algorithm many alternative options have been suggested (see e.g. [26; 35; 48; 45] for different sliding-window algorithms). However, it has recently been shown that these differences can lead to nearly arbitrary differences in the sets of CGIs selected [19; 21]. Recently, [19; 50] have proposed statistical definitions of CGIs and their methods also provide statistical approaches to identifying non-overlapping islands. However the sophisticated statistical techniques underlying these methods make it difficult to succinctly describe the characterizing properties of the CpG islands they predict, and it is perhaps for this reason that researchers, despite the advantages of these statistical approaches, continue to use the Gardiner-Garden-Frommer based CpG islands available from the UCSC genome browser.

In the case of coding regions, the problems discussed above are compounded by the selective pressures on nucleotides resulting from the translation of codons into amino acids that affect nucleotide composition [8; 29; 40]. The problem of modifying heuristics such as the GF definition in order to account for coding constraints is non-trivial. In principle, current state-of-the-art statistical approaches to predicting CpG islands could be adapted to explicitly incorporate the characteristics of coding regions, but we do not know of any method that has been proposed for doing this.

Our first main contribution in this paper is a direct and simple criterion for defining CpG islands—arguably the simplest criterion possible. The simplicity of our definition lies in the fact that it only requires specification of a (Markov based) null model which is a canonical choice for encoding k-mer statistics of genomic regions. We define CpG islands to be regions where the number of CpGs significantly deviates from this null model according to a p-value threshold. We demonstrate the utility of the p-value of a region in assessing the functional significance of a CpG island by showing that p-values can be used to prioritize existing CpG island predictions according to functional significance.

	Year	Island model needed	Parameters: statistical / ad hoc	Applicability of method	# Parameters (learned : specified)	Min length requirement	MHT correction
Our approach	2011	No	Statistical	Easy	17 (16:1)	No	Yes
Wu et al. [50]	2010	No	Both	Medium	($\sim 10^7$:1)	No	No
Hsieh et al. [19]	2009	Yes	Both	Weak	3 (2:1)	No	No
Straussmanet et al. [43]	2009	Yes	Both	Weak	12 (11:1)	Yes	No
Sujua et al. [44]	2008	No	Ad hoc	Weak	5 (0:5)	Yes	No
Glass et al. [13]	2007	No	Both	Easy	5 (2:3)	No	No
Hackenberg et al. [17]	2006	No	Ad hoc	Easy	5 (3:2)	No	No
UCSC Genome Browser [25]	2004	No	Ad hoc	Weak	5 (0:5)	Yes	No
Wang et al. [48]	2004	No	Ad hoc	Weak	7 (1:6)	No	No
Takai et al. [45]	2002	No	Ad hoc	Weak	6 (0:6)	Yes	No
Ponger et al. [35]	2002	No	Ad hoc	Weak	4 (0:4)	Yes	No
Gardiner-Garden et al. [14]	1987	No	Ad hoc	Weak	4 (0:4)	Yes	No

Table 1. Comparison of different methods for CpG island annotation. "Applicability of method" - relates to how easily the method can be used for different sequences (such as different organisms, genomic regions etc.). "MHT" - multiple hypothesis testing. The large number of parameters learned for [50] is due to inferring 'smooth deviations' of CG counts, individually for every genomic segment of length 16.

Our second main contribution is the development of an algorithm for selecting a set of non-overlapping islands from among all possible islands which is optimal for biologically motivated criteria. This result is proved in Theorem 1. Our algorithm is greedy, and has a stopping criterion based on the Benjamini-Hochberg theory for controlling the false discovery rate. Coupled with a dynamic programming procedure which allows to efficiently perform the hypothesis tests required, this algorithm yields a powerful and utmost flexible new approach to finding CpG islands: it can easily be applied to different genomes, or to regions subject to selective pressures that affect nucleotide composition. The only parameter to be specified in our approach is the false discovery rate — the model parameters are estimated directly from the data.

We showcase this approach by predicting *coding* CpG islands in the human genome. To our knowledge, we are the first to predict such CGIs utilizing a statistical model that appropriately accounts for the excess of CpGs expected due to coding constraints. We show that in coding regions different rules should be applied and illustrate this by showing that previous methods miss coding CGIs in some exons while predicting non-significant islands in others.

1.1 Related Work

As sketched above, our null hypothesis will be that DNA sequences are generated by Markov chains of higher order and we will make "CpG island calls" if CpG counts significantly deviate from the background model assumption. Short range dependencies among nucleotides have been reported early on [1; 5; 34] where [5] even suggests that higher-order Markov chains are, to a certain degree, "realistic" genomic sequence models.

In this work we center on 5-th order Markov models due to the benefits of such higher order models in analyzing exonic sequences. This approach has been used before in prominent gene finding programs. One example is GENSCAN [10], which uses "phased" 5-th order models that were in turn inspired by earlier work on the topic [7; 15]. Use of "phased" 3-periodic inhomogeneous Markov chains necessitates a significant increase in the number of model parameters, which we have opted to avoid here and leave as interesting future work. There are different methods by which model order can be chosen, such as the Bayesian or Akaike Information Criterion (BIC and AIC), as suggested in [33], or by means of Chi-square tests, as suggested in [38]. Here, also following [38, p. 2], we agree that model order "from a practical point of view ... also depends on the composition of the biological sequence one wants to take into account.", in that the 5-mer previous to a nucleotide in a coding region always contains the entire codon of the previous amino acid.

Modeling exonic sequences by more involved methods such as employing hidden Markov models [11], variable length Markov models [9], drifting Markov models [46] or, as mentioned above, 3-periodic inhomogeneous (phased) Markov chains, may further reduce the number of false-positives. We leave it as future work to explore the use of such methods for this purpose, and to overcome the associated technical difficulties, such as parameter estimation and the rapid and efficient computation of p-values.

2 Methods

2.1 Notation

In the following, we denote the set of nucleotides by $\Sigma := \{A,C,G,T\}$. Let Σ^k be the set of strings over Σ of length k (k-mers) and let Σ^* be the set of all strings of finite length (oligomers of arbitrary length). We write $v \in \Sigma^k, w \in \Sigma^l$ for strings and $v \cdot w \in \Sigma^{k+l}$ (or simply vw) for their concatenation. For $B \subset \Sigma^{n-k}$ and $v \in \Sigma^k$ we write $Bv \subset \Sigma^n$ for the set of strings from B extended by the string v. In this language, genomes, or parts of the genome are collections of strings over Σ.

We write $\mathbf{G}_s^e$ for the substring of the string $\mathbf{G}$ which stretches from position s to e, including the positions s and e. Let $v, w \in \Sigma^*$ be strings and define $\#(v,w)$ to be the number of (overlapping) occurrences of the string w as a substring in v. For example, $\#(CGACG, CG) = 2$.

Let $\mathbf{P}_n$ be some probability distribution over strings of length n, $w = CG$ and $v \in \Sigma^n$ be a random string drawn according to $\mathbf{P}_n$. We denote by

$$p_{n,m} := \mathbf{P}_n(\#(v,w) \geq m) \tag{1}$$

the tail probability that an n-mer randomly drawn according to P_n contains at least m occurrences of CG. A small value for $p_{n,m}$ indicates that a substring with m CGs contains an unusually high number of CGs.

Let P_n and w be as before and let $u \in \Sigma^n$ be a random string drawn according to P_n that begins and ends with a CG. We denote by

$$p_{n,m}^{cg\to cg} := \mathbf{P}_n(\#(u,w) \geq m \mid u \in CG\Sigma^{n-4}CG) \tag{2}$$

the tail probability that an n-mer randomly drawn according to P_n *that begins and ends with CG* contains at least m occurrences of CGs.

For a given substring $\mathbf{G}_s^e$ where $n = e - s + 1$ and $m = \#(\mathbf{G}_s^e, CG)$ we denote $p(\mathbf{G}_s^e) := p_{n,m}$ and $p^{cg\to cg}(\mathbf{G}_s^e) := p_{n,m}^{cg\to cg}$.

2.2 Coding CpG Islands

As discussed above, CpG islands may be defined in terms of a region's functionality, but in this study we propose a purely sequence based definition. Sequence based definitions have the advantage of being applicable to any sequenced genome without the need for further experimental information, and can pinpoint CpG islands that are unmethylated only in particular tissues (being maintained as CGIs due to selective pressure), and therefore hard to find by experimental mapping of DNA methylation in a subset of tissues.

We begin by giving a statistical definition of CpG islands, and return to questions about association with function in the Results section. Due to the higher mutation rate of methylated cytosines to thymines, and the methylation of cytosines in mammals being mostly at CpG sites, regions that are methylated in the germline and at which there is no selection to maintain CpG sites will be sparser in CpG sites than regions at which the previous conditions do not apply. We postulate, as in [14], that CpG islands should be defined in terms of the number of CpGs. Specifically, we define CpG islands (CGIs) and coding CpG islands (CCGIs) as regions with a significantly high number of CpGs.

DEFINITION 1. *Given a threshold for significance q, a region G_s^e is a CpG island iff $p_{(e-s+1),m} < q$, where $m = \#(G_s^e, CG)$.*

DEFINITION 2. *Given a threshold for significance, q, a region G_s^e is a coding CpG island iff $p_{(e-s+1),m} < q$, where $m = \#(G_s^e, CG)$, and G_s^e is completely contained within a coding exon.*

There are two immediate advantages to using our method for direct annotation of coding CpG islands. The first being that coding exons are under strong selective pressure resulting in different mutation rates than in the noncoding sequence and therefore in different underlying null models. In the Results section we show the importance of choosing an accurate null model. The second being that by limiting ourselves to a small fraction of the genome we test less hypotheses, reducing the extent of the multiple hypothesis testing problem.

2.3 A Markov chain model

The definition of CpG islands depends on the probability distributions $\mathbf{P}_n$, representing the null model. In the case of CGIs we use the genomewide mutation rates to construct these distributions. In the case of CCGIs, we choose these distributions to reflect the coding mutation rate (CMR) present in coding exons (rather then genomewide mutation rates).

Our model needs to reflect k-mer statistics on genomic regions, and it is therefore convenient to assume that sequences are generated by a (stationary) $k-1$-th order Markov chain, denoted as $(X_t)_{t\in\mathbb{N}}$. A k-th order Markov chain emitting symbols from Σ is parametrized by a transition probability matrix

$$M = (m_{uv})_{u\in\Sigma^k, v\in\Sigma} \quad \text{where} \quad m_{uv} = \mathbf{P}(u \to v) := \mathbf{P}(X_t = v \mid X_{t-k} \ldots X_{t-1} = u)$$

and an initial probability distribution $\pi \in [0,1]^{\Sigma^k}$ that is defined as $\pi(v) = \mathbf{P}(X_1...X_k = v)$, where π is chosen such that the resulting stochastic process is stationary (in case of $k = 1$ this translates to $\pi^T M = \pi^T$). According to the rules of Markov chains one computes probabilities as

$$\mathbf{P}(v = v_1...v_n) = \mathbf{P}(X_1...X_n = v_1...v_n) = \pi_{v_1...v_k} \cdot m_{v_1...v_k,v_{k+1}} \cdot ... \cdot m_{v_{n-k}...v_{n-1},v_n}.$$

Our objective is to use this way of modeling sequences to compute $p_{n,m}$ values for coding regions, and throughout this section we explain our approach to doing so. We first describe how we have chosen the order of the Markov chain to use and what the parameters of our model are. We continue by describing how we efficiently compute, for sequences generated by a k-th order Markov chain, the probability of a sequence of length n to have $\geq m$ occurrences CG, and to end with a sequence $v \in \Sigma^k$. Lastly, we show that by using this computation and summing over all possibilities for v, we can compute $p_{n,m}$. In this section we also explain how such recursive computations are used to compute $p_{n,m}^{cg\to cg}$, which is used in the algorithm described in the next section.

In the coding regions, nucleotide sequences are organized into codons, each codon consisting of three nucleotides. Since the different codon positions (first, second and third) have different mutation rates, there are different transition rates between nucleotides at different positions in the codon [8; 40]. We therefore opted for 6-mer statistics (a 5th order Markov model), in which the entire codon of the previous amino acid is taken into account when determining the probability of transition, giving indication of the position of the codon being predicted.

We would like the parameters of our model to reflect the CMR. In our model, we would like $\pi(u)$ to be the probability that the k-mer u is randomly drawn from a coding region and m_{uv} to be the probability that in the coding regions a k-mer u is followed by the nucleotide v. The resulting Markov chain encodes the $k+1$-mer statistics of the coding regions. We denote by $\mathbf{P}_{CMR}(v)$ the probability of observing a string v in a coding region under the Markov chain model described.

Given our assumption that the underlying probability distribution follows a 5-order Markov model, we would like to compute $p_{n,m}^{cg\to cg}$ and $p_{n,m}$ for different values of n and m. Pattern counting statistics have been extensively studied [2; 28; 31; 33; 37], and the asymptotic cases of probabilities such as (1,2) have been considered in large deviations theory. We used a dynamic programming approach to efficiently compute (see section 2.5 for a further discussion on efficiency) the exact probabilities from (1,2) for Markov chains of order k.

We denote the set of n-mers that end with the string v of length k by $\Sigma^{n-k}v$. Let $u \in \Sigma^n$ be a random string generated by the Markov chain (X_t), $w = CG$, and $v \in \Sigma^k$ be some fixed string. Define

$$\pi_{n,m}(v) := \mathbf{P}_X(\{\#(u,w) \geq m\} \cap \{u = \Sigma^{n-k}v\})$$

as the probability that the Markov chain (X_t) generates a string of length n that contains at least m CGs and ends with the k-mer v. Define

$$\pi_{n,m} := [\pi_{n,m}(v)]_{v \in \Sigma^k}$$

to be the vector of values $\pi_{n,m}(v)$ for all $v \in \Sigma^k$.

We recall that a k-th order Markov chain (X_t) where $k > 1$ can be transformed into a 1-st order Markov chain (Y_t) which acts on the alphabet Σ^k rather than Σ by defining

$$\mathbf{P}_Y(v_1 ... v_k \to w_1 ... w_k) := \begin{cases} \mathbf{P}_X(v_1 ... v_k \to w_k) & v_2 ... v_k = w_1 ... w_{k-1} \\ 0 & \text{otherwise} \end{cases}. \quad (3)$$

This generates a transition probability matrix for the new 1-st order model by using M - the transition matrix of the k-th order model. We denote this new transition probability matrix by M_Y.

We recall that (Y_t) generates symbols from Σ^k rather than Σ. Let $w = w_1 ... w_k \in \Sigma^k$ and $I^w := [m_{vw}]_{v \in \Sigma^k}$ be the w-column in M_Y. We define

$$M_Y^w := (\mathbf{0}, ..., \mathbf{0}, I^w, \mathbf{0}, ..., \mathbf{0}), \quad M_Y^{CG} := \sum_{w \text{ s.t. } w_{k-1}w_k = CG} M_Y^w \quad \text{and} \quad M_Y^{\neq CG} := \sum_{w \text{ s.t. } w_{k-1}w_k \neq CG} M_Y^w$$

that is, M_Y^w results from putting all entries in M_Y to zero apart from those in the w-column I^w and M_Y^{CG} ($M_Y^{\neq CG}$) results from summing over all matrices M_Y^w where w ends (does not end) with CG. Note that $M_Y = M_Y^{CG} + M_Y^{\neq CG}$. We then obtain that

$$(\pi_{n,m})^T = (\pi_{n-1,m-1})^T \cdot M_Y^{CG} + (\pi_{n-1,m})^T \cdot M_Y^{\neq CG}. \quad (4)$$

Given $\pi_{n,m}$, one can now compute $p_{n,m}$ by the identity $p_{n,m} := (\pi_{n,m})^T \mathbf{1}$. We can use this to compute the probability that a random string of length n generated by $\mathbf{P}_X$ has at least m CGs and ends with a CG by summing over the appropriate v instances of $\pi_{n,m}$, and we can compute $p_{n,m}^{cg \to cg}$ in the case of the CMR 5-order model by initializing with

$$\hat{\pi}_{5,m}(v) = \begin{cases} \pi_{5,m}(v) & v \in CG\Sigma^3 \\ 0 & \text{otherwise,} \end{cases}$$

and summing over the appropriate v instances of $\pi_{n,m}$.

Example Suppose $\Sigma = \{G, C\}$ is the alphabet at hand, and that we use a 2-nd order Markov chain. Suppose that the parameters for our model are the following transition matrix and initial probability distribution:

$$M = \begin{array}{c} \\ CC \\ CG \\ GC \\ GG \end{array} \begin{array}{c} \begin{array}{cc} C & G \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \\ 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix} \end{array} \qquad \pi = \begin{pmatrix} \pi(CC) \\ \pi(CG) \\ \pi(GC) \\ \pi(GG) \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{pmatrix}.$$

For example, $m_{GC,C} = 0.9$. We build M_Y from M for the recursive computation using (3) and get:

$$M_Y = \begin{array}{c} \\ CC \\ CG \\ GC \\ GG \end{array} \begin{array}{c} \begin{array}{cccc} CC & CG & GC & GG \end{array} \\ \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0.9 & 0.1 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 \end{pmatrix} \end{array}.$$

We have that $(\pi_{2,0})^T = (0.1, 0, 0.3, 0.4)$ and $(\pi_{2,1})^T = (0, 0.2, 0, 0)$ and that

$$M_Y^{CG} = \begin{pmatrix} 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ and } M_Y^{\neq CG} = \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 \end{pmatrix}.$$

We can now compute $p_{n,m}$ recursively. For example,

$$(\pi_{3,1})^T = (\pi_{2,0})^T \cdot M_Y^{CG} + (\pi_{2,1})^T \cdot M_Y^{\neq CG} = (0.05, 0, 0.07, 0.16),$$

and $p_{3,1} = (\pi_{3,1})^T \mathbf{1} = 0.28$.

2.4 Determining non-overlapping CCGIs: the algorithm

Our algorithm for determining CCGIs, given a set of coding exons as input, begins with the estimation of the null model $\mathbf{P}_{CMR}$. Maximum likelihood estimates for the Markov chain parameters are determined simply by counting k-mer transitions in the set of coding exons at hand. The algorithm takes as input these maximum likelihood estimates and a threshold, q, for determining significance and proceeds as follows:

1. Compute probability tables $(p_{n,m})_{1 \leq n \leq N, 0 \leq m \leq n/2}$ and $(p_{n,m}^{cg \to cg})_{1 \leq n \leq N, 0 \leq m \leq n/2}$ for a reasonably large N, the maximum possible length for an island.

2. Order all exonic substrings $\mathbf{G}_s^e$ which are bounded by CG from both ends by their tail probabilities $p^{cg\to cg}(\mathbf{G}_s^e)$. We define p_i to be the i-th smallest among these tail probabilities and $G_i = \mathbf{G}_{s_i}^{e_i}$ the corresponding substring. That is

$$\begin{aligned} p_1 = p^{cg\to cg}(G_1) \le p_2 = p^{cg\to cg}(G_2) \le \cdots \\ \le p_{K-1} = p^{cg\to cg}(G_{K-1}) \le p_K = p^{cg\to cg}(G_K) \end{aligned}$$

where K is the number of all exonic substrings that are bounded by CGs from both ends.

3. Set a threshold $k^* = \max\{i : p_i \le \frac{i}{K}q\}$ and choose all G_i s.t. $i \le k^*$ to construct a set of candidate islands. Discard all other substrings from further analysis.

4. Order the set of candidate islands by their tail probabilities $p(\mathbf{G}_s^e)$, s.t. p_i is the i-th smallest among these tail probabilities and G_i is the matching candidate island.

5. Select a set of non-overlapping islands from the candidate island set by running:
GREEDYCCGI
CCGI $\leftarrow \emptyset$, CAND $\leftarrow \{G_i, 1 \le i \le k^*\}$
while CAND $\neq \emptyset$ **do**
  $i^* \leftarrow \underset{i}{\operatorname{argmin}} \{p(G_i) \,|\, G_i \in \text{CAND}\}$
  CCGI $\leftarrow$ CCGI $\cup \{G_{i^*}\}$, CAND $\leftarrow$ CAND $\setminus \mathcal{N}(G_{i^*})$
Output CCGI as the set of coding CpG islands.
where $\mathcal{N}(G_i)$ is the set of candidate islands that overlap G_i.

Figure 1 presents an overview of the algorithm. Details about the steps are below:

Step 1 [Computation of probabilities]

$\mathbf{P}_{CMR}$ is constructed from the maximum likelihood estimates, and two tables are constructed using the dynamic programming approach described in section 2.3.

We observe that, using the notation as in the previous section,

$$\begin{aligned} (\pi_{n,m})^T - (\pi_{n-1,m})^T = [(\pi_{n-1,m-1})^T - (\pi_{n-2,m-1})^T] \cdot M_Y^{CG} \\ +[(\pi_{n-1,m})^T - (\pi_{n-2,m})^T] \cdot M_Y^{\neq CG} \end{aligned}$$

which by induction on n and m gives the intuitive result

$$\pi_{n,m} \ge \pi_{n-1,m} \quad \text{for all} \quad v \in \Sigma^k \qquad \text{and therefore} \qquad p_{n,m} \ge p_{n-1,m} \quad \forall n,m \ge 0.$$

LEMMA 1. *Let $v \in \Sigma^n, w \in \Sigma^k, k \le n$ be two sequences with equal content of CGs, that is $m := \#(v,CG) = \#(w,CG)$. Then $p_{k,m} \le p_{n,m}$ and $p_{k,m}^{cg\to cg} \le p_{n,m}^{cg\to cg}$, i.e., the shorter one is more significant.*

Step 2 [Ordering of p-values]

$p_{n,m}^{cg \to cg}$ values are computed for each substring within an exon that begins and ends with a CG, setting the stage for step 3. We can restrict ourselves to those substrings because Lemma 1 guarantees that only such substrings will be selected in step 5 of the algorithm. This results in a large decrease in the number of hypotheses tested, significantly reducing both the multiple hypothesis testing problem and the algorithm's running time. We make use here of $p_{n,m}^{cg \to cg}$ because when applying an FDR correction to a set of hypothesis tests the distribution of the p-values for the hypotheses generated by the null model needs to be uniform.

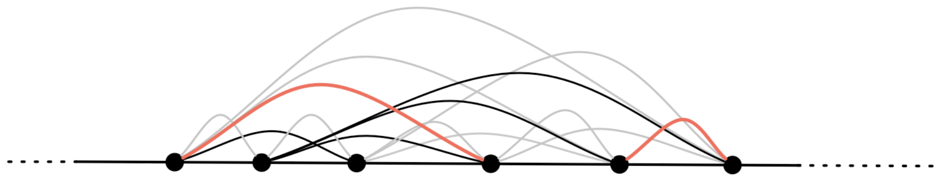

Figure 1. Illustration of the result of the CCGI algorithm. Nodes marked along the genome correspond to CpG sites. An edge between every two nodes corresponds to an interval, and represents a possible CCGI. A p-value is associated with every edge (computed by dynamic programming). Light grey edges are intervals with p-values that do not pass the FDR threshold. Red edges denote non-overlapping intervals chosen by the algorithm to maximize the product of the p-values subject to the constraints in Theorem 1.

Step 3 [Benjamini-Hochberg correction]

In this step a set of candidate islands is chosen by using the Benjamini-Hochberg correction [3; 4] with the significance threshold q. This guarantees that under our null model the expected proportion of false-positives in our set of candidate islands is less than q. Note that our hypotheses are dependent, for example in the case of overlap between substrings. For our case the positive regression dependency condition required for multiple testing under dependency [4] translates to that CG content in substrings which overlap with one another is positively correlated—which is an obvious observation.

Step 4 [Reordering of p-values]

The candidate islands are sorted by their p-values, not conditioning on having a CG at the beginning and end (which solely served to ensure uniformity of the p-value distribution for step 3). This step is a preparatory for step 5.

Step 5 [Greedy algorithm]

The algorithm GREEDYCCGI receives a collection of exonic substrings $\mathbf{G}_s^e$ and iteratively selects regions $\mathbf{G}_s^e$ with minimal $p(\mathbf{G}_s^e) = p_{n,m}$ (where $n = e - s + 1$ and $m = \#(\mathbf{G}_s^e, CG)$), removing from the candidate set any substrings overlapping the chosen one. Theorem 1 shows that the output is of maximal significance under a biologically reasonable constraint. Before we present the theorem, we give some notation.

Let $\mathbf{H}$ be a superstring (in the following an exon) and $\{H_l\}_{l=1}^L$ be a set of non-overlapping substrings of $\mathbf{H}$, ordered by position in $\mathbf{H}$: $H_1 = \mathbf{H}_{s_1}^{e_1} < H_2 = \mathbf{H}_{s_2}^{e_2} < \cdots < H_L = \mathbf{H}_{s_L}^{e_L}$, where ordering translates to $s_1 \le e_1 < s_2 \le e_2 < \cdots < s_L \le e_L$. Let $S_1 < S_2 < \cdots < S_L < S_{L+1}$ be the strings in between, that is

$$S_1 \cdot H_1 \cdot S_2 \cdot H_2 \cdots S_l \cdot H_l \cdot S_{l+1} \cdots S_L \cdot H_L \cdot S_{L+1} = \mathbf{H}$$

is the entire superstring. Based on this notation, we write

$$\begin{aligned} & \overline{H_l H_k} := H_l \cdot S_{l+1} \cdot H_{l+1} \cdots S_k \cdot H_k \\ \text{and} \quad & \overleftrightarrow{H_l H_k} := S_l \cdot \overline{H_l H_k} \cdot S_{k+1}. \end{aligned}$$

THEOREM 1.

Let $\{G_i\}_{i=1}^K$ be a set of possibly overlapping substrings of a superstring $\mathbf{G}$ and $\mathbf{S}(G_i) := \log p(G_i)$ be a score for a substring G_i. Applying GREEDYCCGI *to $\{G_i\}_{i=1}^K$ selects L non-overlapping substrings $\{H_l\}_{l=1}^L$ from among the substrings $\{G_i\}_{i=1}^K$ such that $\sum_{l=1}^L \mathbf{S}(H_l)$ is minimized among all choices of L non-overlapping substrings, subject to the constraints*

$$S(G) \ge \min_{l \le j \le k} \mathbf{S}(H_j) \textit{ for all } 1 \le l,k \le L, \textit{ and } G \subset \overleftrightarrow{H_l H_k}. \tag{5}$$

In the setting of this paper, the constraints in (5) translate to the property that the region between any two CpG islands H_l, H_k does not contain a sub-region with a p-value better than the p-values of the islands in between H_l, H_k.

Proof. By design of GREEDYCCGI, its output clearly satisfies the constraints. Assume that there is a set of non-overlapping substrings $\{J_l\}_{l=1}^L$ which satisfy the constraints in (5) such that

$$\sum_{l=1}^L \mathbf{S}(J_l) < \sum_{l=1}^L \mathbf{S}(H_l)$$

where $\{H_l\}_{l=1}^L$ are the islands returned by the GREEDYCCGI. Let $\mathbf{S}^* = \min_i\{\mathbf{S}(G_i)\}$, and $W = \{G_i \mid \mathbf{S}(G_i) = \mathbf{S}^*\}$ be the set of substrings with the lowest score. We prove the result by induction on L. We show in Lemma 2 that there are no erroneous overlaps among the islands of W [1]. We distinguish between the following two cases:

- $L > |W|$. In this case $\{H_l\}_{l=1}^{L}$ will contain all substrings of W.
  - If one of the substrings from $\{J_l\}_{l=1}^{L}$ is in W, upon removal of that substring from $\{J_l\}_{l=1}^{L}$ and $\{G_i\}_{i=1}^{K}$, and removal of their overlapping neighbors from $\{G_i\}_{i=1}^{K}$ we obtain a new set of substrings $\{\tilde{G}_i\}_{i=1}^{\tilde{K}}$. By induction, GREEDYCCGI will return a set of islands from $\{\tilde{G}_i\}_{i=1}^{\tilde{K}}$ which is optimal, resulting in a contradiction.
  - If non of the substrings from $\{J_l\}_{l=1}^{L}$ are in W, let $\hat{J}$ be some island in $\{J_l\}_{l=1}^{L}$. When considering the entire superstring $\mathbf{G}$ condition (5) is violated w.r.t. $\hat{J}$, in contradiction.
- $L \leq |W|$. $\{H_l\}_{l=1}^{L}$ consists entirely of elements of W, achieving the minimum score.

□

LEMMA 2. *Let $\{H_i\}_{i=1}^{K}$ be a set of substrings of a set of superstrings $\mathbf{H}$, $p^* = \min_i p(H_i)$ and $W = \{H_i \mid p(H_i) = p^*\}$. Then there are no two islands in W that overlap and are not contained in each other, i.e., $H_{s_1}^{e_1}, H_{s_2}^{e_2}$ s.t. $s_1 < s_2 < e_1 < e_2$.*

Proof. Assume that there are two islands $H_{s_1}^{e_1}, H_{s_2}^{e_2} \in W$ s.t. $s_1 < s_2 < e_1 < e_2$. We will show that in this case $p(H_{s_1}^{e_2}) < p(H_{s_1}^{e_1}) = p(H_{s_2}^{e_2})$, in contradiction.

Let n_1, n_2 and n_3 be the lengths of $H_{s_1}^{e_1}$, $H_{s_2}^{e_2}$ and $H_{s_1}^{e_2}$, respectively, and let f_s^e be the frequency of CG sites (the number of CGs divided by $e-s+1$) in the substring starting at s and ending at e, inclusive. Suppose w.l.g. that $n_1 \leq n_2$. Since the two substrings have the same p-values, $f_{s_2}^{e_2} \leq f_{s_1}^{e_1}$ (a shorter substring must have a higher frequency of CGs to reach the same significance), and therefore $f_{e_1+1}^{e_2} \leq f_{s_1}^{s_2-1}$. By concatenating both $H_{s_1}^{s_2-1}$ and $H_{e_1+1}^{e_2}$ to $H_{s_2}^{e_1}$ we concatenate two strings to $H_{s_2}^{e_1}$ which together are longer and have more (or equal) frequency of CGs than concatenating just $H_{e_1+1}^{e_2}$. Therefore, since $f_{s_2}^{e_2} \leq f_{s_1}^{e_2}$ and $n_2 < n_3$, $p(H_{s_1}^{e_2}) < p(H_{s_2}^{e_2})$, in contradiction.

□

2.5 Running time and implementation

We compute two tables holding values of $p_{n,m}$ and $p_{n,m}^{cg \to cg}$ for $1 \leq n \leq N := 10000$ and $0 \leq m \leq N/2$ where N is an upper bound on the maximum length of a CpG island in an exon. Using dynamic programming in the implementation of (4),

[1] For any two islands in W that overlap, one island is strictly contained in the other. In case the algorithm encounters overlaps, it chooses the shorter island and it can be easily seen that the proof holds for such cases.

the running time is $O(|\Sigma|^{k+1} \cdot N^2)$ for alphabets Σ and model order k in general where here $\Sigma = \{A,C,G,T\}$ and $k = 5$. Note that each recursive iteration $(\pi_{n,m})^T = (\pi_{n-1,m-1})^T \cdot M_Y^{CG} + (\pi_{n-1,m})^T \cdot M_Y^{\neq CG}$, due to the sparsity of M_Y^{CG} and $M_Y^{\neq CG}$, requires only $O(|\Sigma| \cdot |\Sigma|^k)$ runtime. Storage of tables requires $O(N^2)$ space. Note that [32] suggests a routine for fast computation of a single $p_{n,m}$ which requires $O(|\Sigma| \cdot L \cdot n \cdot m)$ where $L = O(|\Sigma|^k)$ depends on the model order k which coincides with the runtime of our approach.

A different algorithm presented in [39] requires a runtime of $O((|\Sigma|^k)^2 \cdot N \log N)$ which yields advantages for small model orders in particular. In our case, due to evaluating powers of matrices which yields the factor $\log N$, one would have to call the respective routine $O(N)$ times, yielding an overall runtime of $O((|\Sigma|^k)^2 \cdot N^2 \log N)$ which equally establishes no advantage.

Subsequently, we traverse each coding region and assign both $p_{n,m}$ and $p_{n,m}^{cg \to cg}$ to each substring which is bounded by CGs. In human, this results in $10.8M$ substrings with $p_{n,m}^{cg \to cg} \leq 0.01$. In order to allow a rapid FDR treatment, we designed and implemented a routine which yields an estimation of the FDR threshold that is stricter than the exact FDR (guaranteeing our expected FDR to be less than or equal to the specified α). The estimated threshold is obtained by storing for each substring G_i, $\lfloor \log p^{cg \to cg}(G_i) \rfloor$, assigning

$$i^* := \underset{i}{\operatorname{argmax}} \, \{\lfloor \log p^{cg \to cg}(G_i) \rfloor + 1 \leq \log(i\frac{\alpha}{K})\},$$

and declaring any G_i with $\log p^{cg \to cg}(G_i) < \lfloor \log p^{cg \to cg}(G_{i^*}) \rfloor$ as a candidate CCGI. Since $\lfloor \log(p_i) \rfloor \leq \log(p_i) \leq \lfloor \log(p_i) \rfloor + 1$, we are guaranteed to be taking a threshold which is at least as stringent as that attained from the Benjamini-Hochberg procedure. Then, the greedy algorithm picks a final list of CCGIs from the candidates.

The methods were implemented in C and the running time on a standard MacPro desktop was 7 hours for generating the p-value tables and 30 minutes for finding the CCGIs for all of the coding exons in the human genome. The software is available upon request from the authors.

Our coding CpG islands Genome Browser track for hg18 is available at

`http://bio.math.berkeley.edu/CCGI/`

3 Results

3.1 The p-value statistic

We began by investigating whether the p-values are indeed a good measure for determining the extent of unmethylated CpGs, by comparing the p-value measure to the epigenetic score as defined in [6] and to the observed/expected number of CpGs used in [14] ($\frac{\#CpG}{\#C \times \#G} \times N$, where N is the length of the island).

Because genome-wide methylation has not yet been characterized for a significant number of developmental stages or cell types, we used two different datasets that are believed to be correlated with functional unmethylated regions to determine our true-positive sets (whether a region is unmethylated or not). The first dataset consisted of regions experimentally determined by the ENCODE project as open chromatin regions in non-malignant cell types [2] [16]. Overlap of an examined island with an open-chromatin region determined the island at hand as functional. The second dataset consisted of genomewide site-specific methylation measurements for two cell types [27]. A methylation score of between 0 and 1 was computed for a region as the mean of the methylation scores of the CpG sites within it, and the region was considered as unmethylated if the score was ≤ 0.3 (this is an arbitrary threshold but since the island scores are well partitioned it does not affect our results). Only sites with coverage of at least 8 reads and islands with at least 5 such sites were considered. An island was considered functional if it was unmethylated in both cell types.

To test whether our *p*-value statistic is informative we computed the *p*-values for the set of bona fide CpG islands annotated in [6] and compared how well they indicate functionality compared to the islands' epigenetic scores and the islands' observed/expected scores. We used the set of bone fide islands for this comparison because they are annotated along with scores, allowing us to compare the ordering of the assigned *p*-values to a different scoring method.

The *p*-values were computed using a first order Markov model estimated from statistics on the *entire genome* (and not only on coding regions). A receiver operating characteristics (ROC) curve for the *p*-values, the epigenetic scores and the observed/expected score, using the ENCODE open chromatin data to asses functionality (Figure 2.A), shows that the *p*-value statistic achieves better overall performance - area under curve (AUC) of 0.76 - than the two other scores (AUC of 0.73 and 0.754 for epigenetic score and observed/expected score, respectively). When using the methylation scores from [27] to asses functionality (Figure 2.B), we again observe better overall performance of the *p*-values: AUC of 0.865 as opposed to 0.852 and 0.844 for the epigenetic score and observed/expected scores, respectively. This is surprising since the epigenetic scores incorporate functional information such as the presence of promoter activity (not from the ENCODE project), whereas our *p*-values are computed from sequence alone.

3.2 Choice of null model

In order to asses the importance of the null model choice on the *p*-value scores we examined the effect of estimating the null model parameters using only coding sequence as opposed to the entire genome sequence. We learned parameters for

[2] One file of open chromatin was compiled from: ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeChromatinMap/ using the files: wgEncodeUncFAIREseqPeaks{H1hesc,Nhek,Gm12878V2,Huvec,Panislets}.narrowPeak

a 5-th order Markov chain for the two types of sequences, and calculated the log ratio of the p-values from the two models for the $194,586$ unique, but possibly overlapping, coding exons from [36] (Figure 3). For all exons the p-value from the genome null model was lower than that from the coding null model. This was to be expected, given the higher frequency of guanine and cytosine residues in coding regions.

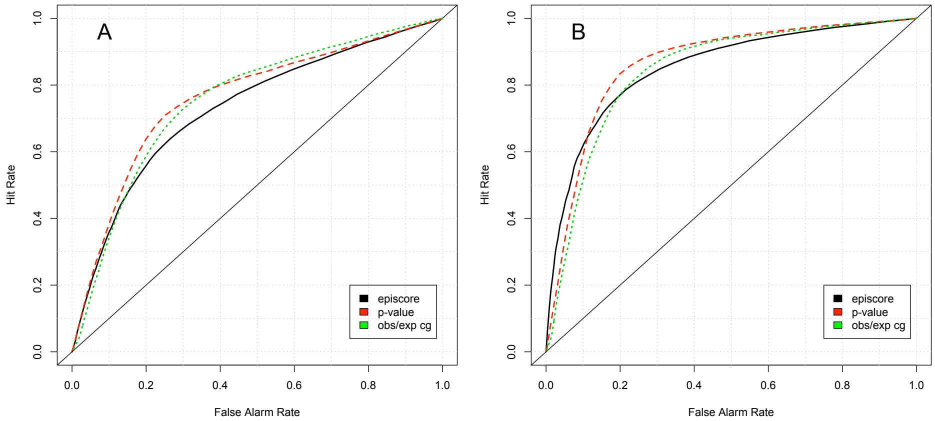

Figure 2. ROC plots showing correlation of different scoring schemes (epigenetic score, p-value, and observed over expected CG) with (A) FAIRE-Seq open chromatin regions and (B) low methylation in H1 and IMR90 cell types.

We then tested whether the difference in the p-values has an effect on the set of coding exons considered significant. At an FDR threshold of 0.05, 54,596 coding exons were determined significant when using the genome model, while 9,639 were determined significant when using the coding model. At an FDR of 0.01, 37,597 and 7,050 coding exons were determined as significant for the genome model and coding model, respectively. The drastic reduction in the number of coding exons predicted as significant with respect to their CpG content shows the importance of the specified null distribution.

3.3 Coding CpG islands

Using our algorithm on the coding regions of the human genome with FDR threshold 0.01, we determined a Coding CpG island (CCGI) set for the human coding exons, that consisted of 12,445 islands.

As a first validation of the sensitivity of the CCGIs we examined a region rich in HOX genes in chr7 (Figure 4.a). This region was studied in [8] where it was shown that CpGs in protein coding exons of HOX genes are subject to pro-epigenetic selection. Our results not only define CpG islands in every one of these

HOX genes (with the exception of HOXA10 in which the $5'$ exon consists of only 2 amino acids), but also provide quantitative information about the extent of over-representation of CpGs via p-values.

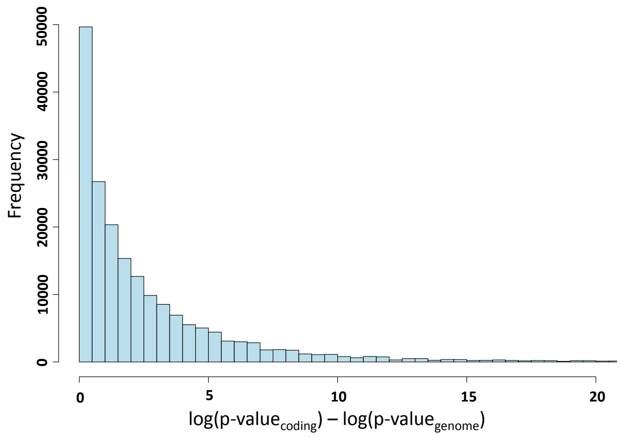

Figure 3. Fold change in the log p-value when computing p-values of exons using a *genome* null model versus a *coding* null model. 82,505 exons have p-values that differ by more than a factor of 2.

We compared our CCGIs to two alternative CpG island sets: the Gardiner-Garden-Frommer based CpG island set [14] available from the UCSC genome browser (GFCGIs) and islands predicted by the method in [22] (HMMCGIs). We chose to compare to GFCGIs, because despite the arbitrary thresholds that are problematic in the method, this set of islands is very popular and widely used. The HMMCGIs represent recent work applying state-of-the-art statistical methods to CpG island prediction. Testing the specificity of the CCGI set is not trivial because one cannot validate that a region determined as a CCGI is not unmethylated and functional in some developmental state or cell type. We therefore chose to test the extent to which the CCGIs were enriched for known functional regions (Table 2).

In a comparison to the GFCGIs, we found 3,000 CCGIs that did not overlap any GFCGI. In contrast, of the 1,143 CGIs that were completely within some exon (and therefore could have been detected by our algorithm) only 34 did not overlap a CCGI. Although many of our CCGIs are located in the first exon of genes (44%), of the set not overlapping GFCGIs we found only 14% in first exons. This suggests that the GFCGIs are missing coding CpG islands in exons not located adjacent to large CpG islands in promoters. To better understand why the GFCGIs were missing in many exons, we examined the properties of the CCGIs that they did not overlap. We found that these CCGIs were mostly shorter than the 200bp cutoff for

	Total	Open Chromatin	17-Cons Track	First Exon	Near A-TSS
CCGIs	12445	2734(0.21)	11041(0.88)	5539(0.44)	7248(0.58)
CCGIs \ GFCGIs	3000	189(0.06)	2687(0.89)	433(0.14)	1248(0.41)
CCGIs \ HMMCGIs	802	25 (0.03)	706 (0.87)	82 (0.10)	286 (0.35)

Table 2. Overlap of CCGI sets with different functional regions. The proportion for each computation is in parentheses. "Open Chromatin" - The same regions used in section 3.1. "17-Cons Track" - Regions of the UCSC 17-way Conservation track. "First Exon" - First exons of RefSeq genes. "Near A-TSS" - Regions of ±500 bps from alternative transcription start sites [47].

	Total	H1 (r1)	H1 (r2)	IMR90 (r1)	IMR90 (r2)
CCGIs	12445	0.075 (73/961)	0.108 (439/4047)	0.168 (663/3912)	0.164 (565/3429)
GFCGIs in exons	1143	0.053 (14/262)	0.052 (43/817)	0.072 (60/826)	0.065 (48/737)
HMMCGIs in exons	3617	0.029 (26/895)	0.023 (55/2312)	0.039 (86/2177)	0.037 (77/2066)

Table 3. Proportion of islands completely within exons determined as unmethylated from the data in [27]. The absolute counts in parentheses are the number of unmethylated islands over the total number of islands for which there was sufficient data do determine the methylation state.

GFCGIs (2,689 out of the 3,000). Moreover, the constant threshold for accepting an island as a GFCGI (#CpG observed / expected > 0.6) implicitly leads to more stringent requirements for longer regions to be declared islands, and can lead to a higher rate of false-negatives for the longer lengths than for the shorter lengths (for example, an island of length 1,500 bps with #CpG observed / expected ratio of 0.5 is much more likely to be an unmethylated island than a 200 bps island with the same ratio, but the constant threshold does not account for this). To verify this, we calculated the correlation between the log of the p-values of GFCGIs and their length, and found it to be -0.78 (with a fitted line of slope -5.164).

The CCGIs we found not overlapping GFCGIs did mostly overlap HMMCGIs. In fact, of our 12,445 CCGI predictions, only 802 did not overlap HMMCGIs. However the comprehensive coverage of HMMCGIs comes at a price of specificity. We examined the set of 1,636 HMMCGIs that are completely within some exon (and therefore could have been predicted by our method) but do not overlap with a CCGI, and found that their average p-value was 0.105. This surprisingly large number suggests that many of these islands are being predicted on the basis of an incorrect null model.

Differential Methylation. In order to further investigate the characteristics of the CCGIs we used the dataset from [27], in which methylation was measured for two different cell types - H1 (human embryonic stem cells) and IMR90 (fetal lung fibroblasts). We computed a methylation score for each CCGI and determined unmethylated CCGIs in the same manner as described in section 3.1. In all experiments (two replicated for each cell type) the large majority of the CCGIs was found to be methylated (Table 3). To validate that this is a general phenomenon for CpG islands inside exons we examined the GFCGIs and HMMCGIs that were completely within exons and found that the large majority of those CpG islands were methylated (Table 3). This supports the hypothesis that CpG islands present within exons play an important role in tissue specific regulatory mechanisms. Interestingly, in all island sets tested, the percentage of unmethylated CpG islands in IMR90 was consistently larger than the percent of unmethylated CpG islands in H1 cells, suggesting that in the stages *after* differentiation the intragenic CpG islands have a greater regulatory role. We compared the methylation status from the two different cell types (using replicate 2 of H1 and replicate 1 from IMR90 since they produced data for a larger number of islands) and found that of the 2,956 CCGIs for which there was sufficient data in both experiments, 249 islands were unmethylated in both samples and 103 were differentially methylated (38 were unmethylated in H1 and 65 were unmethylated in IMR90). Figure 4.b presents an interesting example of such a CCGI which was not detected as a GFCGI or an HMMCGI.

Alternatively Spliced Exons As an exon-specific phenomenon, we examined the overlap of the different island sets with *alternatively spliced exons* where the length of the exon varies due to splicing events.

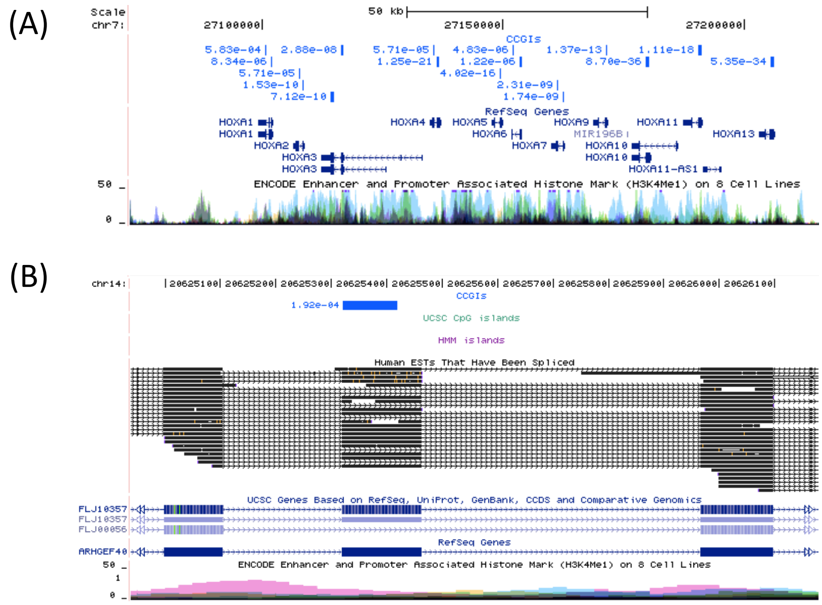

Figure 4. UCSC genome browser screen shots. (A) A HOX gene cluster showing our CCGI predictions. (B) A differentially methylated CCGI on an alternatively spliced exon was not detected as a GFCGI or a HMMCGI. Histone marks associated with promotors and enhances (bottom track) are seen mostly in the NHLF cells (seen in pink, normal human lung fibroblasts). This aligns well with the CCGI being unmethylated in IMR90 cells and methylated in H1 cells.

As displayed by Table 4 there are 190,181 regions overlapped by some exon, and 4,092, 266 and 31 regions overlapped by at least 2, 3 and 4 different, but overlapping exons, respectively. In order to rate the percentage of overlap in combination with the numbers of islands reported, we display the corresponding p-values, as determined by the hypergeometric distribution $\mathbf{P}_{N,m,n}$ for population size $N = 190,181$, the number of all exonic regions, number of successes m, the number of exonic regions that are covered by at least 2,3 or 4 exons, and number of samples drawn n. For example, in table 4,

$$1.53e-25 = \mathbf{P}_{190181,4092,12443}(X \geq 443)$$
$$6.28e-4 = \mathbf{P}_{190181,266,26020}(X \geq 56).$$

It can be seen that CCGIs outperform both HMMCGIs and GFCGIs, both in terms of precision as well as in terms of significance.

Epigenetic association. We investigated the extent of overlap of different CpG island sets with epigenetically associated regions at varying distances from the exon boundaries. For this purpose, we used the differentially methylated regions (DMRs) [23]. Due to differential methylation varying relative to distance from

Overlap	All Exons No	Coding CGIs No	Coding CGIs Prec	Coding CGIs p-value	HMM CGIs No	HMM CGIs Prec	HMM CGIs p-value	GFCGIs No	GFCGIs Prec	GFCGIs p-value
≥ 1	190,181	12,445	-	-	26,020	-	-	15,821	-	-
≥ 2	4,092	443	3.560	1.53e-25	767	2.948	3.0e-20	453	2.863	3.97e-10
≥ 3	266	41	0.329	3.22e-7	56	0.215	6.28e-4	41	0.259	9.85e-5
≥ 4	31	3	0.024	0.33	2	0.008	0.94	2	0.013	0.74

Table 4. Statistics on CpG island sets and their overlap with regions which are covered by overlapping, alternatively spliced exons. The first column indicates the overall number of regions covered by the specified number of different exons, 'Prec' (precision) holds the percentage of islands hitting such a region and 'p-value' indicates significance in terms of hypergeometric tail probabilities, putting precision and amount of islands reported into context.

non-coding regions (data not shown), we hypothesized that the predictive power of CCGIs with respect to these DMRs can be further attuned by looking inside the exons. This studies the correlation of differential methylation with the enrichment of CpGs at different exon locations (far vs. close to boundary). Moreover, this compares the performance of different CpG island definitions in this respect.

We compiled a file from the DMR dataset holding all regions which overlap an exon, extending each end by 100 bp. From each CpG island, if it overlapped an exon, we multiplied its presence in the set based on the number of exon slices it overlapped, pairing each instance with an exon region (the whole exon or an inner region of the exon). We then constructed CpG island subsets where each subset, S_x , contains all CpG islands for which the minimum distance from the island to the boundary of the exon it was assigned to is at least x (where -1 is used to denote all islands which overlap an exon and 0 denotes islands that do not cover an exon at its full length). For each such subset we computed precision (Prec) as the percentage of islands overlapping a region in the dataset investigated, recall (Rec) as the percentage of regions of the dataset investigated that overlap an island, out of the regions overlapping the exonic regions at least x bases from a non-coding region, and the F-measure (F-M), combining the evaluations of precision and recall, as the ratio of geometric over arithmetic mean of precision and recall.

Table 5 presents the comparison of CCGIs, HMMCGIs and GFCGIs. We note that these differentially methylated regions were published by the group that annotated the HMMCGIs [22; 23] and that HMMCGIs have been demonstrated to serve as excellent predictors of such regions in [22]. As we grow the distance from non-coding sequence, the islands of all sets become worse in terms of recall but compensate this to some extent by a raise in precision. The decrease in recall is drastic for the HMMCGIs and the GFCGIs. In the interior of exons the CCGI set outperforms the GFCGI set in terms of the F-M measure, for all tested distances from non-coding regions. The CCGI set outperforms the HMMCGI set in terms of the F-M measure for all distances larger than 0 up to 60, is slightly lower at 90 and is equal at 120. These results show that in the interior of exons the CCGI set outperforms the other two island calling methods with respect to the tradeoff between recall and precision, when measuring overlap with differentially methylated regions.

	Coding CGIs				HMM CGIs				GFCGIs			
Dist.	No	Rec	Prec	F-M	No	Rec	Prec	F-M	No	Rec	Prec	F-M
-1	12934	22.5	3.1	5.5	26848	53.5	3.6	6.7	16320	31.1	3.4	6.2
≥ 0	12923	22.5	3.1	5.5	11809	26.3	4.0	7.0	5465	14.2	4.7	7.0
≥ 1	12324	21.9	3.2	5.6	3785	7.3	3.4	4.7	1213	3.4	5.0	4.1
≥ 6	8705	17.0	3.4	5.7	3574	6.9	3.4	4.5	1115	3.1	4.9	3.8
≥ 12	5765	13.3	4.0	6.2	3275	6.6	3.5	4.5	998	2.9	5.0	3.7
≥ 30	2870	9.0	5.0	6.5	2561	5.8	3.6	4.5	790	2.6	5.2	3.4
≥ 60	1365	4.9	5.0	4.9	1753	5.3	4.2	4.7	602	2.3	5.3	3.2
≥ 90	859	3.9	5.4	4.5	1365	5.3	4.5	4.9	468	2.3	5.8	3.3
≥ 120	632	3.9	6.2	4.8	1098	5.0	4.6	4.8	357	2.2	6.2	3.3

Table 5. Statistics on CpG island sets, selected relative to exonic location, and their overlap with differentially methylated regions $\pm$100bp up- and downstream. See the text for definitions of precision (Prec), recall (Rec) and F-measure (F-M). The first column indicates the minimum distance of island start/end to a non-coding regions while the second/sixth/tenth column displays the number of such islands.

4 Conclusions

We have presented a novel method for annotating CpG islands in coding regions. Our method is based on a simple, statistically natural, criterion: the statistical significance of the CpG content of a genomic region. Based on this idea, we have developed an algorithm which provably optimizes a biologically motivated criterion for selecting islands while controlling the false discovery rate.

We found that we can reduce the number of false positives in existing annotations while discovering previously undetected coding CpG islands which are likely to contain unmethylated CpGs. Finally, we demonstrate the value of explicit annotation of coding CpG islands by providing evidence for a relationship between the location of islands within the exons with their type of function.

References

[1] ALMAGOR, H. A Markov analysis of DNA sequences. *J. Theor. Biol. 104* (1983), 633–645.

[2] ATTESON, K. Calculating the exact probability of language-like patterns in biomolecular sequences. In *Proceedings of the ISMB 98 Conference* (1998), pp. 17–24.

[3] BENJAMINI, Y., AND HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B 57*, 1 (1995), 289–300.

[4] BENJAMINI, Y., AND YEKUTIELI, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*, 4 (2001), 1165–1188.

[5] BLAISDELL, B. Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. *J. Mol. Evol. 21* (1985), 278–288.

[6] BOCK, C., WALTER, J., PAULSEN, M., AND LENGAUER, T. CpG island mapping by epigenome prediction. *PLoS Comput Biol 3*, 6 (June 2007), e110.

[7] BORODOVSKY, M., MCININCH, J., KOONIN, E., RUDD, K., MEDIGUE, C., AND DANCHIN, A. Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Research 23*, 17 (1995), 3554–3562.

[8] BRANCIAMORE, S., CHEN, Z.-X., RIGGS, A., AND RODIN, S. CpG island clusters and pro-epigenetic selection for CpGs in protein-coding exons of hox genes and other transcription factors. *Proc. Nat. Acad. Sci. USA 107*, 35 (2010), 15485–15490.

[9] BÜHLMANN, P., AND WYNER, A. Variable length Markov chains. *Annals of Statistics 27*, 2 (1999), 480–513.

[10] BURGE, C., AND KARLIN, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol. 268* (1997), 78–94.

[11] CHURCHILL, G. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology 51*, 1 (1989), 79–94.

[12] EDWARDS, J., O'DONNELL, A., ROLLINS, R., PECKHAM, H., LEE, C., ET AL. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Research 20* (2010), 972–980.

[13] GLASS, J.L., ET AL., CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res* (October 2007).

[14] GARDINER-GARDEN, M., AND FROMMER, M. CpG islands in vertebrate genomes. *Journal of molecular biology 196*, 2 (1987), 261–282.

[15] GELFAND, M. Prediction of function in DNA sequence analysis. *Journal of Computational Biology 2*, 1 (1995), 87–115.

[16] GIRESI, P., KIM, J., MCDANIELL, R., IYER, V., AND LIEB, J. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research 17* (2007), 877–885.

[17] HACKENBERG, M., PREVITI, C., LUQUE-ESCAMILLA, P. L., CARPENA, P., MARTÍNEZ-AROZA, J., AND OLIVER, J. L. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics 7* (2006), 446.

[18] HARRIS, R. A., WANG, T., COARFA, C., NAGARAJAN, R. P., HONG, C., ET AL. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotechnology advance online publication doi:10.1038/nbt.1682* (September 2010).

[19] HSIEH, F., CHEN, S.-C., AND POLLARD, K. A nearly exhaustive search for CpG islands on whole chromosomes. *The International Journal of Biostatistics 5*, 1 (2009), 14.

[20] ILLINGWORTH, R., KERR, A., DESOUSA, D., JØRGENSEN, H., ELLIS, P., ET AL. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol 6*, 1 (January 2008), e22.

[21] ILLINGWORTH, R. S., AND BIRD, A. P. CpG islands–'a rough guide'. *FEBS letters 583*, 11 (June 2009), 1713–1720.

[22] IRIZARRY, R., WU, H., AND FEINBERG, A. A species-generalized probabilistic model-based definition of CpG islands. *Mammalian Genome 20* (2009), 674–680.

[23] IRIZARRY, R. A., LADD-ACOSTA, C., WEN, B., WU, Z., MONTANO, C., ONYANGO, P., CUI, H., GABO, K., RONGIONE, M., WEBSTER, M., JI, H., POTASH, J. B., SABUNCIYAN, S., AND FEINBERG, A. P. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics 41*, 2 (January 2009), 178–186.

[24] JONES, P. A., AND BAYLIN, S. B. The epigenomics of cancer. *Cell 128*, 4 (February 2007), 683–692.

[25] KAROLCHIK, D., HINRICHS, A. S., FUREY, T. S., ROSKIN, K. M., SUGNET, C. W., HAUSSLER, D., AND KENT, W. J. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res 32*, Database issue (January 2004), D493–D496.

[26] KIM, K. CpG islands detector: a window-based CpG island search tool. *Genomics and Informatics 8* (2010), 58–61.

[27] Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M. M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature 462*, 7271 (Nov. 2009), 315–322.

[28] Marschall, T., and Rahmann, S. Probabilistic arithmetic automata and their application to pattern matching statistics. *Proceedings of the 19th Annual Symposium on Combinatorial Pattern Matching* (2008), 225–232.

[29] Maunakea, A., Nagarajan, R., Bilenky, M., Ballinger, T., D'Souza, C., et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature 466* (2010), 253–257.

[30] Medvedeva, T., Fridman, M., Oparina, N., Malko, D., Ermakova, E., et al. Intergenic, gene terminal, and intragenic CpG islands in the human genome. *BMC Genomics 11* (2010), 48.

[31] Nicodeme, P., Salvy, B., and Flajolet, P. Motif statistics. *Theoretical Computer Science 287* (2002), 593–617.

[32] Nuel, G. Effective p-value computations using Finite Markov Chain Imbedding (FMCI): application to local score and pattern statistics. *Algorithms in Molecular Biology 1* (2006), 5.

[33] Nuel, G. Numerical solutions for pattern statistics on Markov chains. *Statistical Applications in Genetics and Molecular Biology 5*, 1 (2006), 26.

[34] Phillips, G., Arnold, J., and Ivarie, R. The effect of codon usage on the oligonucleotide composition of the E.coli genome and identification of over- and underrepresented sequences by markov chain analysis. *Nucleic Acids Research 15*, 6 (1987), 2627–2638.

[35] Ponger, L., and Mouchiroud, D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics 18*, 4 (2002), 631–633.

[36] Pruitt, K. D., Tatusova, T., and Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res 35*, Database issue (January 2007).

[37] Regnier, M., and Szpankowski, W. On pattern frequency occurrences in a Markovian sequence. *Algorithmica 22*, 4 (1998), 631–649.

[38] Reinert, G., Schbath, S., and Waterman, M. Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology 7*, 1/2 (2000), 1–46.

[39] Ribeca, P., and Raineri, E. Faster exact Markovian probability functions for motif occurrences: a DFA-only approach. *Bioinformatics 24*, 24 (2008), 2839–2848.

[40] Saxonov, S., Berg, P., and Brutlag, D. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Nat. Acad. Sci. USA 103*, 5 (2006), 1412–1417.

[41] SINGER, M., BOFFELLI, D., DHABHI, J., SCHÖNHUTH, A., SCHROTH, G., MARTIN, D., AND PACHTER, L. Metmap enables genome-scale methyltyping for determining methylation states in populations. *PLoS Computational Biology 6*, 8 (2010), e1000888.

[42] SMIRAGLIA, D. J., RUSH, L. J., FRUHWALD, M. C., DAI, Z., HELD, W. A., COSTELLO, J. F., LANG, J. C., ENG, C., LI, B., WRIGHT, F. A., CALIGIURI, M. A., AND PLASS, C. Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies. *Hum. Mol. Genet. 10* (2001), 1413–1419.

[43] STRAUSSMAN, R., NEJMAN, D., ROBERTS, D., STEINFELD, I., BLUM, B., BENVENISTY, N., SIMON, I., YAKHINI, Z., AND CEDAR, H. Developmental programming of CpG island methylation profiles in the human genome. *Nat Struct Mol Biol 16*, 5 (2009), 564–571.

[44] SUJUAN, Y., ASAITHAMBI, A., AND LIU, Y. CpGIF: an algorithm for the identification of CpG islands. *Bioinformation 2*, 8 (2008), 335–8.

[45] TAKAI, D., AND JONES, P. A. Comprehensive analysis of CpG Islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA 99*, 6 (2002), 3740–3745.

[46] VERGNE, N. Drifting Markov models with polynomial drift and applications to DNA sequence. *Statistical Applications in Genetics and Molecular Biology 7*, 1 (2008), 6.

[47] WAKAGURI, H., YAMASHITA, R., SUZUKI, Y., SUGANO, S., AND NAKAI, K. Dbtss: database of transcription start sites, progress report 2008. *Nucleic Acids Research 36*, Database issue (2008), D97–D101.

[48] WANG, Y., AND LEUNG, F. An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics 20*, 7 (2004), 1170–1177.

[49] WEBER, M., HELLMANN, I., STADLER, M. B., RAMOS, L., PÄÄBO, S., REBHAN, M., AND SCHÜBELER, D. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics 39*, 4 (2007), 457–466.

[50] WU, H., CAFFO, B., JAFFEE, H., IRIZARRY, R., AND FEINBERG, A. Redefining CpG islands using hidden Markov models. *Biostatistics 11*, 3 (2010), 499–514.

[51] ZHANG, Y., ROHDE, C., TIERLING, S., JURKOWSKI, T. P., BOCK, C., ET AL. DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS Genet 5*, 3 (2009), e1000438.