CpG counts are approximately Poisson distributed

Lemma 1: *The number of different ways to pick n non-consecutive integers from 1 to L is* $\binom{L-n+1}{n}$.

Proof: Assume there are n white balls and $L-n$ black balls. The number in question is the same as the number of different ways to align all the L balls, with no adjacent white balls. This is equivalent to inserting white balls into to the spaces between black balls, when one space can accommodate no more than one white ball. Since there are $L-n+1$ spaces between the black balls, the number of ways is $\binom{L-n+1}{n}$. $\square$

Definition: For a sequence of L success/failure experiments, assume the marginal success rate for each experiment is a constant p, and one can not have two consecutive successes. Let S be the total number of successes. We denote the distribution of S as a "non-consecutive binomial" with parameters L and p, and use the shorthand $NCBin(L,p)$.

Lemma 2: *The probability mass function (pmf) for* $N \sim NCBin(L,p)$ *is*
$\Pr(N=n) = \binom{L-n}{n} p^n (1-p)^{L-2n} + \binom{L-n}{n-1} p^n (1-p)^{L-2n+1}$.

Proof: For a given particular sequence $\{Z_t\}_{t=1,\ldots,L}$ let S be the set of indices where Z equals 1, i.e. $Z_i = 1$ for $\forall i \in S$, $Z_i = 0$ for $\forall i \notin S$. If $N=n$ then S contains n elements. Now consider the probability of this sequence for two possibilities.

1. When $Z_L = 0$, we have

$$\begin{aligned}\Pr(\{Z_t\}_{t=1,\ldots,L}) &= \left[\prod_{i\in S} \Pr(Z_i = 1, Z_{i+1} = 0)\right] \times \left[\prod_{i\notin S, (i-1)\notin S} \Pr(Z_i = 0)\right] \\ &= p^n (1-p)^{L-2n}.\end{aligned}$$

   Because knowing $Z_L = 0$ is equivalent to picking n non-consecutive integers from 1 to $L-1$, by lemma 1, the number of ways to pick such a sequence is $\binom{L-n}{n}$.

2. Define $S_1 = S \backslash \{L\}$. When $Z_L = 1$, S_1 contains $n-1$ elements. Then, we have

$$\begin{aligned}\Pr(\{Z_t\}_{t=1,\ldots,L}) &= \left[\prod_{i\in S_1} \Pr(Z_i = 1, Z_{i+1} = 0)\right] \times \left[\prod_{i\notin S, (i-1)\notin S_1} \Pr(Z_i = 0)\right] \\ &\times \Pr(Z_L = 1) = p^n (1-p)^{L-2n+1}.\end{aligned}$$

Because $Z_L = 1$ implies $Z_{L-1} = 0$, to pick such a sequence is equivalent to picking $n-1$ non-consecutive integers from 1 to $L-2$, therefore by lemma 1, the number of ways to pick such a sequence is $\binom{L-n}{n-1}$.

Putting together 1 and 2 proves the result. □

THEOREM 1: *If in a genomic segment of length L bp, the probability of a dinucleotide being CG is a constant p, $L \to \infty$, $p \to 0$, and $Lp \to \alpha$, then the number of CG dinucleotide in the segment converges in distribution to a Poisson random variable with mean α.*

Proof: Let B_t be the base at genomic location t, $t = 1, \ldots, L$. B_t takes values A,T,G,C. Define Bernoulli random variable $Z_t = \mathbb{1}(B_t = C, B_{t+1} = G)$. Let $P(Z_t = 1) = p$ for $t = 1, \ldots, L-1$. Note that Z_t cannot be 1 at consecutive locations, we have $P(Z_{t+1} = 0|Z_t = 1) = 1, P(Z_{t+1} = 1|Z_t = 1) = 0$. Let $N = \sum_{i=1}^{L-1} Z_t$ be the number of CG, N follows $NCBin(L-1, p)$. By lemma 2 the pmf for N is: $\Pr(N = n) = \binom{L-n-1}{n} p^n (1-p)^{L-2n-1} + \binom{L-n-1}{n-1} p^n (1-p)^{L-2n}$.

The probability mass function can be sandwiched by upper and lower bounds P_0 and P_1.

$$\begin{aligned}
\Pr(N = n) &> \binom{L-n-1}{n} p^n (1-p)^{L-2n-1} \equiv P_0 \\
\Pr(N = n) &< \binom{L-n-1}{n} p^n (1-p)^{L-2n-1} + \binom{L-n-1}{n-1} p^n (1-p)^{L-2n-1} \\
&= \binom{L-n}{n} p^n (1-p)^{L-2n-1} \equiv P_1.
\end{aligned}$$

Using the same arguments often used to prove a binomial random variables converge in distribution to Poisson, both P_0 and P_1 converge to $Poisson(\alpha)$. Therefore, $\Pr(N = n)$ converges to $Poisson(\alpha)$. □

Choosing the segment length L

We applied the proposed method to human hg18 genome using different segment lengths: $L = 8, 16$ and 32. Figure 1 shows the ROC curves for overlapping TSS and DMRs. It can be seen that $L = 16$ gives the best results in covering human TSS and the results for covering human DMR are similar for all choices. Therefore we chose $L = 16$ as the segment length.

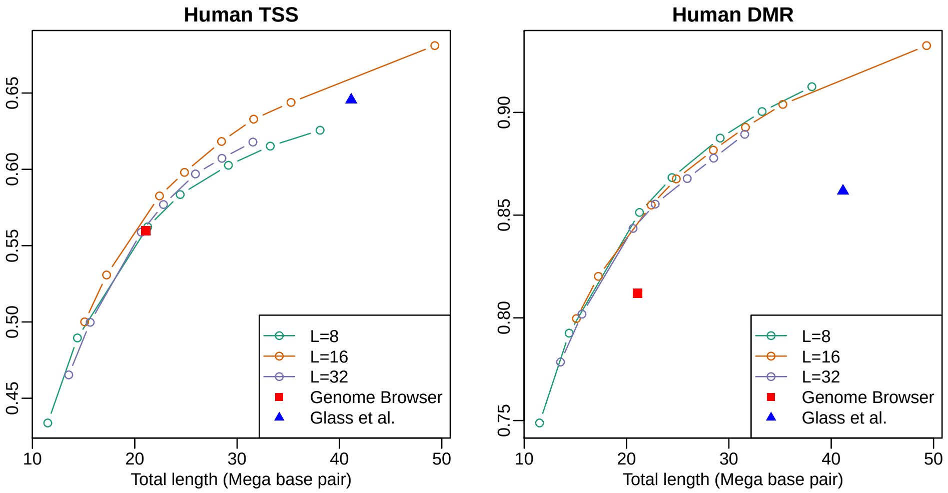

Figure 1: ROC curves of using different L. $L = 16$ gives the best overall results.

Choosing the smoothing window size

We applied the proposed method to human hg18 genome with different window sizes in smoothing $r(s)$. We tried using 80, 200, 400 and 600 base pairs and not smoothing (equivalent to infinite smoothing window). The ROC curves in Figure 2 showed that the results are actually similar, with 400 base pairs window being slightly better than others. Therefore we chose 400 base pairs as the smoothing window size in this manuscript.

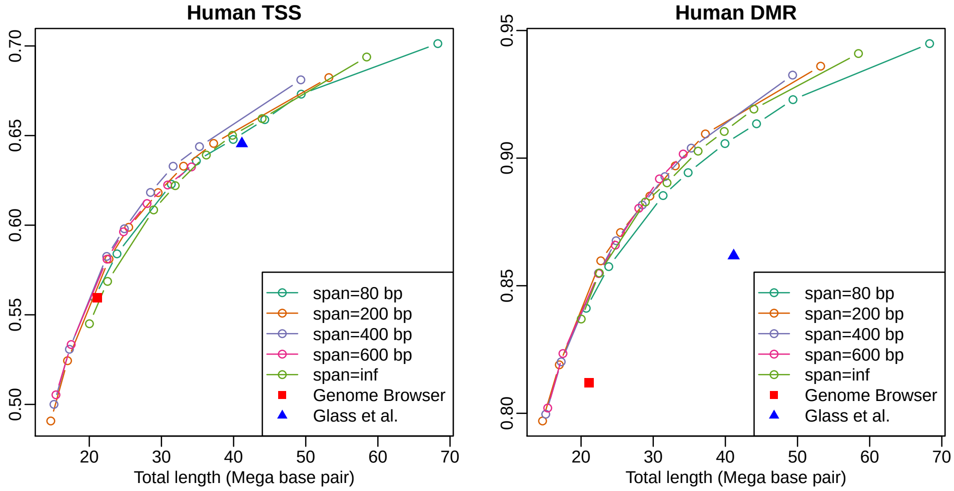

Figure 2: ROC curves of using different smoothing window. Window size of 400 base pairs gives the best results.

Example of CGIs detected from HMM

The limitations of current CGI definitions were demonstrated in Figure 3 in the manuscript. The lower panels in Figure 3 shows the result posterior probabilities from two HMMs for these two regions. It can be seen that desired CGIs can be obtained by properly thresholding the posterior probabilities.

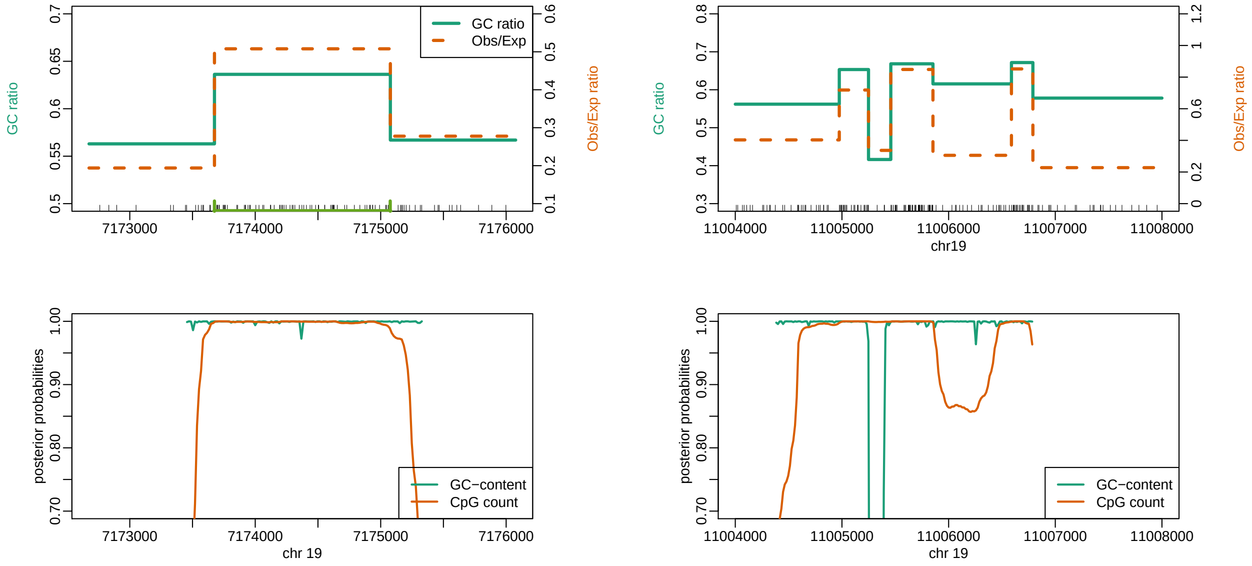

Figure 3: Posterior probabilities from two HMMs for two regions shown at Figure 3 in the manuscript.