Statistics and Its Interface Volume 8 (2015) 463–476

motifDiverge: a model for assessing the statistical significance of gene regulatory motif divergence between two DNA sequences

Dennis Kostka*, Tara Friedrich,
Alisha K. Holloway, and Katherine S. Pollard

Next-generation sequencing technology enables the identification of thousands of gene regulatory sequences in many cell types and organisms. We consider the problem of testing if two such sequences differ in their number of binding site motifs for a given transcription factor (TF) protein. Binding site motifs impart regulatory function by providing TFs the opportunity to bind to genomic elements and thereby affect the expression of nearby genes. Evolutionary changes to such functional DNA are hypothesized to be major contributors to phenotypic diversity within and between species; but despite the importance of TF motifs for gene expression, no method exists to test for motif loss or gain. Assuming that motif counts are Binomially distributed, and allowing for dependencies between motif instances in evolutionarily related sequences, we derive the probability mass function of the difference in motif counts between two nucleotide sequences. We provide a method to numerically estimate this distribution from genomic data and show through simulations that our estimator is accurate. Finally, we introduce the R package `motifDiverge` that implements our methodology and illustrate its application to gene regulatory enhancers identified by a mouse developmental time course experiment. While this study was motivated by analysis of regulatory motifs, our results can be applied to any problem involving two correlated Bernoulli trials.

Keywords and phrases: Testing, Gene regulation, Motif, ChIP-seq, Binomial, Transcription factor, Regulatory evolution.

1. INTRODUCTION

Next-generation sequencing increasingly provides insight into the locations of regulatory regions in the genomes of many organisms, and it gives information about the cell types and developmental stages in which these regulatory elements are active [1]. RNA sequencing (RNA-seq, [2, 3]) enables accurate quantification of gene expression, and techniques such as DNase sequencing (DNase-seq, [4]) and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq, [5]) pinpoint which parts of a genome are in open chromatin and therefore may be associated with regulatory activity in a given cell type. These methods can be coupled with chromatin immunoprecipitation followed by sequencing (ChIP-seq, [6]) for histone modifications, transcription factors (TFs) and co-factors to further refine predictions of regulatory elements, such as promoters, enhancers, repressors, and insulators [7]. Gene expression levels are different between cell types and dynamic during development as the result of regulatory elements that are specifically active in some cells but not in others [8, 9]. Therefore, identification of functional regulatory elements and the TFs that recognize them is a key step to characterizing any type of cell. This information also sheds light on transitions between different cell types, such as in the progression to cancer or during cellular differentiation.

Regulatory genomic elements typically contain multiple motifs for one or more TFs. The TF proteins bind to these motif sequences to combinatorially modulate the expression of nearby genes [10]. TF motifs are to some extent degenerate (i.e., mutations away from the consensus sequence are tolerated), and therefore they are typically represented as probability distributions over nucleotides (A, C, G, and T) at each position in the motif [11]. For each TF, this distribution can be represented as position specific probability matrix (PSPM). While TF binding depends on more than just the target DNA sequence (TF concentration, open chromatin, etc.), and even though the binding affinity of a TF towards a stretch of nucleotides is quantitative rather than binary, the presence or absence of TF motifs can be represented as a binary event by scoring how well a sequence matches a TF's PSPM (details below). Because sequence changes can alter how well DNA matches a PSPM, mutations and substitutions can create or destroy motif instances. It is challenging to predict the effect of a single motif loss or gain on the function of a regulatory region, because a loss may be compensated for by a nearby gain. However, a large cumulative change in the number of motifs across a regulatory region can alter expression of nearby genes, potentially resulting in differences in organismal traits, such as disease susceptibility.

arXiv: 1402.0042
*Corresponding author.

To the best of our knowledge there are no existing methods for quantifying divergence between DNA sequences based on differences in motif counts. The primary challenge is that in most biologically meaningful settings the sequences are related through evolution (i.e., they are homologous) or functional constraints, and therefore motif instances are correlated. This is the problem we address in this paper: We derive the joint distribution of the number of motifs in the two sequences, and the marginal distribution of the difference in numbers of motifs between the two sequences. From the latter distribution, we show how p–values can be computed for testing the null hypothesis of no systematic difference in motif counts between two sequences. We validate our methodology through simulations and apply it to ChIP-seq and RNA-seq data from a developmental time course.

2. A MODEL FOR REGULATORY MOTIF DIVERGENCE

We propose a probabilistic model and test for assessing the statistical significance of the difference in number of motifs for a single TF between two DNA sequences. While the core of our approach is independent of the specifics regarding TF motif modeling, we also provide methodology to estimate the distribution of our test statistic for any TF that has a motif model in the form of a PSPM. The sequences may be homologous or not, because our approach does not require (but can make use of) a sequence alignment that enables a parameter estimation scheme based on evolutionary models (see Section 2.2.5 and the Appendix). In both cases, the two sequences can be short sequence elements (e.g., pairs of orthologous gene promoter sequences) or concatenations of multiple short sequence elements that share some property (e.g., promoters of multiple genes). For the non-homologous case, any two sequences or sets of sequences can be compared. For example, one might be interested in TFs with significantly different numbers of motifs in promoters of genes that are up-regulated versus down-regulated in a cancer RNA-seq experiment, or in comparing gene promoters versus distal enhancers. For the homologous case, one might compare two genotypes present within a single species, such as a disease-associated versus healthy genotype of a gene promoter. The homologous case can also be used across species, for instance, to quantify the regulatory divergence of pairs of homologous regulatory sequences identified via ChIP-seq. We recently took this approach to compare human and fish developmental gene regulation, and we showed that TF motif differences capture functional changes in enhancer sequences better than do standard measures of sequence divergence [12].

2.1 Background: predicting TF motifs

A typical approach to identify TF motifs in DNA sequences is to scan a sequence one position at a time using a PSPM and predict a motif at any position where the likelihood of a motif-length sub-sequence under the PSPM model is significantly higher than under a background distribution (see below for details) [13]. In this context, the PSPM and background distribution are thought of as generative models. Let M be a PSPM of length l (typically about 7 to 10 bp) over the DNA nucleotide alphabet $\{A, C, G, T\}$, where M_{ij} is the probability of observing nucleotide i at position j in the motif. Let B_i be the probability of observing nucleotide i (at any position) under a background model. Such a background model can, for example, be estimated from the whole genome or from any reasonably long sequence from the species of interest. Then $L_{ij} := \log(M_{ij}/B_i)$ is the log odds for nucleotide i at position j and $T(x) = \sum_{j=1}^{l} L_{x_j}$ is the log odds score for a sequence $x = x[1, \ldots, l]$ of length l. The distribution of T can obtained numerically, and a log odds score threshold for predicting motif instances can be found in such a way that Type I error, Type II error, or a balance between the two (balanced cutoff) are controlled [13]. Alternatives to Type I error control are commonly employed, because false negatives can be important in this application; TFs frequently bind to sequences that are weak matches to their motif (i.e., would be missed with strict Type I error control), and in some cases this weak binding is functional.

We note that PSPM based log odds scores do not account for dependencies between motif positions, despite the fact that these are known to exist for TF motifs. More sophisticated methods for motif annotation that take relationships between nucleotide positions into account have been developed [14, 15, 16]. However, standard PSPM scoring is commonly used, computationally convenient, and has recently been observed to perform well [17]. The model we describe in this paper can in principle be applied together with any method for motif prediction.

To scan a sequence x of length $k \geq l$ for motifs, a sliding window approach is typically used. Starting at the first nucleotide x_1, compute $T(x_{1\to l}) := T(x[1, \ldots, l])$. Then, slide the window one nucleotide to start at position x_2 and compute $T(x_{2\to l+1})$. Continue computing $T(x_{i\to i+l-1})$ until the last test statistic $T(x_{k-l+1\to k})$ is computed. A motif is predicted at position i if $T(x_{i\to i+l-1}) > t$ for a log odds score threshold t (see above). Note that subsequent test statistics are not independent, because their underlying sequences overlap. This "in-sequence" dependency is often not accounted for, but there are methods that take it into account [18]. Our model does not explicitly include in-sequence dependency. However, based on the fact that our method performs well on simulated data with in-sequence correlation (see Section 4.2), and that other methods with similar assumptions perform well in practice [17], we believe that this is a reasonable approach. Also, we show that there is a relationship between in-sequence dependence and the dependence between motif counts in two homologous sequences (Appendix). Because of this relationship, our model is able to indirectly account for some in-sequence dependence via its parameter for between-sequence correlation (Section 2.2.5).

2.2 Modeling differences in the number of TF motifs between two sequences

Consider two sequences x and y of lengths k_x and k_y (possibly not equal). For a given TF, let a random variable X_i be the indicator for the presence of a motif at position i in x, and let Y_i be the corresponding random variable for y. We assume the prediction of a motif in a sequence is the result of a Bernoulli trial with a homogeneous success probability along the sequence. Then, the joint distribution of (X_i, Y_i) does not depend on i. Next, we define random variables $N_x = \sum_i X_i$ and $N_y = \sum_i Y_i$ for the total number of motifs in each sequence. Marginally N_x and N_y have Binomial distributions. However, note that X_i and Y_i (and therefore N_x and N_y) are not necessarily independent, because the sequences x and y are potentially related, for example due to sequence homology or shared regulatory constraints. The problem we address here is how to define and estimate the distribution of the difference in the number of motifs between the two sequences $N_{xy} = N_x - N_y$ under dependence of X_i and Y_i. Our approach is based on the two underlying, correlated Binomial trials. We note that assuming homogeneous success probabilities implies that we are neglecting effects stemming from in-sequence dependence between motif hits. We believe this is a reasonable approach for the reasons given above and in Section 2.2.5.

2.2.1 Equal length sequences

First consider the case of equal length sequences ($k := k_x = k_y$), which simplifies the model because there is a corresponding Bernoulli trial in x for each trial in y. Let N_{10} be the number of pairs (X_i, Y_i) with $X_i = 1$ and $Y_i = 0$, and let N_{01} be the number of pairs with $X_i = 0$ and $Y_i = 1$. Then $N_{xy} = N_{10} - N_{01}$. To derive the distribution of N_{xy}, we first consider the joint distribution of N_{10} and N_{01}, which is multinomial:

$$P(N_{10} = n_{10}, N_{01} = n_{01}) = (n_{10}, n_{01}, n - n_{10} - n_{01})! \times p_{10}^{n_{10}}\, p_{01}^{n_{01}} (1 - p_{10} - p_{01})^{n - n_{10} - n_{01}}, \tag{1}$$

where $(\cdot,\cdot,\cdot)!$ is the multinomial coefficient, $n = k - l + 1$ is the number of windows tested for a motif of length l, $p_{00} = Pr(X_i = 0, Y_i = 0)$, $p_{01} = Pr(X_i = 0, Y_i = 1)$, and so on.

Notably the joint distribution of (N_{10}, N_{01}) is independent of p_{00} and p_{11} and only depends on the probabilities for a motif in one sequence and not the other: p_{01} and p_{10}. Because $n_{xy} = n_{10} - n_{01}$ can be realized in $\lfloor \frac{n - n_{xy}}{2} \rfloor$ different ways, the distribution of N_{xy} is:

$$P(N_{xy} = n_{xy}) = \begin{cases} \sum_{j=0}^{\lfloor \frac{n-n_{xy}}{2} \rfloor} P(N_{10} = n_{xy} + j, N_{01} = j) & \text{for } n_{xy} \geq 0 \\ \sum_{j=1}^{\lfloor \frac{n-n_{xy}}{2} \rfloor} P(N_{10} = j, N_{01} = |n_{xy}| + j) & \text{for } n_{xy} < 0. \end{cases} \tag{2}$$

Identifying the sums in Equation (2) as hypergeometric series (Appendix), we can rewrite them in terms of the Gaussian hypergeometric function ${}_2\mathrm{F}_1$ [19]:

$$\sum_{j=0}^{\lfloor \frac{n-n_{xy}}{2} \rfloor} P(N_{10} = n_{xy} + j, N_{01} = j) = \binom{n}{n_{xy}} p_{10}^{n_{xy}} (1 - p_{10} - p_{01})^{n - n_{xy}} \times {}_2\mathrm{F}_1\left(\frac{n_{xy} - n}{2}; \frac{n_{xy} + 1 - n}{2}; n_{xy} + 1; \frac{4 p_{10} p_{01}}{(1 - p_{10} - p_{01})^2} \right), \tag{3}$$

with similar results for the other sum. Since ${}_2\mathrm{F}_1(a; b; c; 0) = 1$, N_{xy} follows a Binomial distribution with parameters p_{10} and n when $p_{01} \to 0$. This is as expected, because in this case $P(N_{10} = n_{10}, N_{01} = 0)$ is Binomial, and there is only one term contributing to the sums in Equation (2). Similarly, for $p_{10} \to 0$ the distribution $P(N_{10} = 0, N_{01} = n_{01})$ is a Binomial with parameters p_{01} and n, and N_{xy} has the same Binomial distribution, just mirrored at $n_{xy} = 0$.

Finally, we can obtain the mean and variance of N_{xy} from the multinomial distribution of N_{10} and N_{01} (Equation (1)):

$$\begin{aligned} \mathrm{E}[N_{xy}] &= n(p_{01} - p_{10}) \\ \mathrm{Var}[N_{xy}] &= n\left(p_{10}(1 - p_{10}) + p_{01}(1 - p_{01}) + 2 p_{10} p_{01}\right). \end{aligned} \tag{4}$$

2.2.2 Alternative parametrization

Instead of the parameters $(p_{11}, p_{10}, p_{01}, p_{00})$ we can use the success probabilities of the Bernoulli trials X_i and Y_i, plus their correlation. Define $p := p_{11} + p_{10}$ and $q := p_{11} + p_{01}$, and let the correlation between the two trials be $\rho := Cov[X_i, Y_i] / \sqrt{Var[X_i] Var[Y_i]}$. In this parameterization admissible values of ρ depend on p and q. Intuitively, it is clear that not all correlation coefficients can be admissible. For instance, if the trials have different success probabilities they cannot at the same time be perfectly correlated. If we assume $0 \leq p \leq q \leq \frac{1}{2}$ then $\rho_- \leq \rho \leq \rho_+$ with

$$\begin{aligned} \rho_- &= -pq / \sqrt{p(1-p)q(1-q)} \\ \rho_+ &= (1-q)p / \sqrt{p(1-p)q(1-q)}, \end{aligned} \tag{5}$$

so that our model can fully be specified by the success probabilities of the Bernoulli trials and an admissible correlation coefficient. We note that the variance of N_{xy} is maximal at $\rho = \rho_-$ (i.e., $p_{11} = 0$), not at $\rho = 0$ (i.e., $p_{11} = pq$), which corresponds to independent trials. Further, the variance of N_{xy} is minimal at $\rho = \rho_+$ (i.e., $p_{11} = \min(p, q)$).

2.2.3 Different length sequences

In most situations, even with homologous sequences, the lengths of x and y will not be identical. Suppose without loss of generality that x is the longer sequence so that $k_x \geq k_y$. Our strategy for modifying $P(N_{xy} = n_{xy})$ to account for the length difference is to treat k_y nucleotides as in Equation (2)

and to derive the distribution for the number of motifs in the remaining nucleotides of x. Recall that we model the difference in motif hits N_{xy} without conditioning on specific alignments or configurations of hit-pairs $\mathscr{A} = \{(x_i, y_i)\}$ between the two sequences. In fact, the sums in Equation 2 are equivalent to summing over all configurations consistent with observing n_{xy}: $\sum_{\mathscr{A}} P(\mathscr{A}) I(n_{xy}|\mathscr{A})$, where $I(n_{xy}|\mathscr{A})$ is one if $\mathscr{A}$ is consistent with observing n_{xy} and zero otherwise. Likewise, our approach for different-length sequences is also equivalent to averaging over all possible configurations. To that end, note that $N_{xy} = N_1 + N_2$, where N_1 is a random variable representing the number of motifs in the first $k_y - l + 1$ nucleotides of x minus the number of motifs in the corresponding nucleotides of y, and N_2 represents the number of motifs in the remaining $k_x - k_y$ possible motif start positions in x. Again, N_1 and N_2 are marginalized quantities in the sense that they average over all configurations of hit-pairs between the two sequences. Then, N_1 has the distribution defined in Equation (2) with length parameter k_y (i.e., $n = k_y - l + 1$). It is easy to see that N_2 only depends on x and is Binomially distributed with success probability $p_{10} + p_{11}$ and $k_x - k_y$ trials, as expected for the remaining Bernoulli trials. If $k_y > k_x$, we leave the definition of N_1 unchanged, but instead treat the excess trials in x as negative counts of motifs that are subtracted from the count for the same-length segment of length k_y. In this case, N_2 is Binomially distributed with success probability $p_{01} + p_{11}$ and $k_y - k_x$ trials. Thus, for different length sequences the difference in numbers of motifs is distributed as the convolution of the distributions for N_1 and N_2:

$$P(N_{xy} = n_{xy}) = \begin{cases} \sum_{j=0}^{k_x - k_y} P_s(N_1 = n_{xy} - j) Bin(N_2 = j) & \text{for} \quad k_x \geq k_y \\ \sum_{j=0}^{k_y - k_x} P_s(N_1 = n_{xy} + j) Bin(N_2 = j) & \text{for} \quad k_x < k_y, \end{cases} \tag{6}$$

where $Bin(\cdot)$ is the probability mass function of the Binomial distribution with parameters given above, and P_s denotes the probability mass function of N_{xy} in the case of equal-length sequences (Equation (2)). We get the mean and variance of N_{xy} for different length sequences from Equation (4) and the Binomial distribution:

$$\begin{aligned} \mathrm{E}[N_{xy}] &= k_y(p_{10} - p_{01}) + (k_x - k_y)p \\ \mathrm{Var}[N_{xy}] &= k_y\big(p_{10}(1-p_{10}) + p_{01}(1-p_{01}) + 2p_{10}p_{01}\big) + \\ &\quad (k_x - k_y)p(1-p), \end{aligned} \tag{7}$$

where again $k_x \geq k_y$ without loss of generality. Unlike Equation (2), which depends only on p_{10} and p_{01}, the distribution of N_{xy} for unequal length sequences (Equation (6)) depends on p_{11} as well (via $p = p_{11} + p_{10}$) and makes full use of the parametrization of (X_i, Y_i).

2.2.4 Computing $P(N_{xy} = n_{xy})$ and $P(N_{nx} \geq n_{xy})$

Our main application is to compute a p–value for an observed difference in motifs ($N_{xy} = n_{xy}$) between two sequences x and y. Thus, we are interested in computing a tail probability of the probability mass function of N_{xy} (Equation (6)). To test if n_{xy} is significantly larger compared to what we expect under a null hypothesis we need to obtain $P(N_{xy} \geq n_{xy})$. Similarly, we need $P(N_{xy} \leq n_{xy})$ to test for significantly fewer motifs in x compared to y.

To numerically evaluate $P(N_{xy} = n_{xy})$, we perform the convolution in Equation (6) using the fast Fourier transform. A prerequisite for this is the probability mass function $P_s(N_{xy} = n_{xy})$ for the symmetric case ($k_x = k_y$), which we get from Equation (2) and evaluate up to a pre-specified error $\epsilon \geq 0$. More specifically, let $P_s(N_{xy} = n_{xy}) = \sum_j S_j$, where the summands S_j are taken from Equation (2). Further let $w_j := S_{j+1}/S_j$. Then there exists j_- such that for j_+ with $j_- < j_+ \leq \lfloor \frac{n - n_{xy}}{2} \rfloor$ (Appendix):

$$\begin{aligned} &P_s(N_{xy} = n_{xy}) = \sum_{j=0}^{j_+} S_j + \epsilon(j_+) \\ &\text{with} \quad 0 \leq \epsilon(j_+) < S_{j_+}\Big(\frac{1 - w_{j_+}^{\frac{n-n_{xy}}{2} - j_+}}{1 - w_{j_+}} - 1\Big). \end{aligned} \tag{8}$$

We evaluate this error bound after each additional term in the sum and stop when a desired precision has been achieved. Additionally, in order to obtain $P_s(N_{xy} = n_{xy})$ for a series of values for n_{xy} the following recurrence relation (Appendix) is useful:

$$\begin{aligned} &(n - n_{xy}) p_{10} P_s(N_{xy} = n_{xy}) = \\ &(1 - p_{10} - p_{01})(n_{xy} + 1) P_s(N_{xy} = n_{xy} + 1) + \\ &p_{01}(n + n_{xy} + 2) P_s(n_{xy} + 2). \end{aligned} \tag{9}$$

The fast Fourier transform evaluates $P(N_{xy} = n_{xy})$ over an entire range of values for n_{xy}, which enables us to compute tail probabilities $P(N_{xy} \geq n_{xy})$, and thereby p–values, by direct summation.

2.2.5 Estimating model parameters

Up to this point, we have treated the model parameters (p_{10}, p_{01}, p_{11}), or alternatively (p, q, ρ), as known. In practice they must be estimated from data before one can compute p–values for an observed difference n_{xy} in the number of motif hits between two sequences. The process of predicting TF motifs (Section 2.1) suggests several properties that could influence the shape of the probability mass function of N_{xy}:

(*i*) *Sequence length.* More predicted motifs can be expected in longer sequences. Also, the larger the length-difference between two sequences, the larger the difference in motifs is expected to be. Both of these effects are explicitly included in our model (via k_x and k_y), and we assume that these sequence lengths are known.

(*ii*) *Motif information content.* Low information content (i.e., weak or uninformative) PSPMs can lead to more predicted motif instances compared to high information content PSPMs. This effect can be taken into account via the choice of the log odds score threshold t (Section 2.1). For example, selecting a value of t for each TF that controls the Type I error will make motif counts comparable across TFs.

(*iii*) *Threshold for predicting motifs.* A loose threshold t for predicting motifs will result in more motif predictions. In our model, the expected number of motifs will be reflected in the parameters p and q.

(*iv*) *Sequence composition.* For a given background distribution, the probability of a motif prediction will depend on the similarity of the nucleotides favored in the PSPM compared to the nucleotide composition of the sequence. For instance, for a GC-rich motif we expect more motifs in a GC-rich sequence compared to an AT-rich sequence. The parameters p and q account for the sequence composition of x and y, respectively. While effects of sequence composition can be further mitigated by using sequence-dependent prediction thresholds $\{t_{xy}\}$ (e.g., corresponding to sequence-dependent background distributions B_i), this is not desirable if a consistent threshold is sought for a collection of jointly analyzed sequences.

(*v*) *Relationship of the two sequences.* If the two sequences are homologous, we may expect fewer differences in motifs compared to the case of two independent sequences. As described above, we model the relationship between x and y via a correlation parameter ρ, which allows us to accommodate both correlated ($\rho > 0$) and uncorrelated ($\rho = 0$) sequences.

Taking these issues into account, we propose the following approaches to parameter estimation.

Independent sequences: Assume x and y are independent and that motifs are equally likely in both sequences. Then, we can estimate $\hat{p} = \hat{q} := (n_x + n_y)/(k_x + k_y)$ (which implies $\hat{p}_{10} = \hat{p}_{01}$). With respect to the correlation parameter ρ we have two options. First, we can choose $\hat{\rho} = 0$, reflecting the independence of X_i and Y_i. In this case, our model is fully specified. A second alternative for independent sequences leverages a relationship between in-sequence dependence and between-sequence dependence (Appendix) to account for correlation (or anti-correlation) between motif instances within each sequence. Specifically, assume the $\{X_i\}$ and $\{Y_i\}$ are realizations of two independent Markov chains. Then $\lambda_x := P(X_i = 1|X_{i-1} = 1)$ may be different from $P(X_i = 1|X_{i-1} = 0)$, and such a correlation ($\lambda_x \neq p$) influences the variance of N_x [20]. A similar effect holds for $\lambda_y := P(Y_i = 1|Y_{i-1} = 1) \neq q$ and the variance of N_y, while the expectations stay the same as in the original model. Numerical estimates for λ_x and λ_y can be obtained, and we can we can choose ρ in a way that the variance of the model with no in-sequence dependence matches the variance of this more general model. Let $\hat{\lambda}_x$ and $\hat{\lambda}_y$ be estimates for the conditional success probabilities. Then this approach leads to:

(10)
$$\hat{\rho} = \frac{-(\hat{p}(1-\hat{p})\hat{q}(1-\hat{q}))^{-\frac{1}{2}}}{2\min(k_x, k_y)}\Big(A(\hat{p}, \hat{\lambda}_x, k_x) + A(\hat{q}, \hat{\lambda}_y, k_y)\Big),$$

where $A(\cdot)$ quantifies the effect of the in-sequence dependence on the variance of N_x and N_y (Appendix, [20]). This parameter choice enables us to include some of the effects due to in-sequence dependence into our model when x and y are independent.

Dependent sequences: If x and y are homologous sequences, we propose to estimate model parameters using an evolutionary model that quantifies the probability of nucleotide changes between x and y. We will focus on evolutionary models for cross–species data based upon continuous time Markov chains (CTMCs), but population genetics models for genotypes within species could also be used.

Like in the case of independent sequences we estimate $\hat{p} = \hat{q} := (n_x + n_y)/(k_x + k_y)$. But we estimate the between-sequence correlation ρ via an estimate for p_{11} derived from the evolutionary model. More specifically, suppose there is a motif at position i in x (i.e., $X_i = 1$). Consider the probability $p_{1\to1}$ that the congruent, homologous sub-sequence of y also contains a motif. We then obtain a numerical estimate $\hat{p}_{1\to1}$ based on the sequence composition of x and y, an evolutionary model, the PSPM, the background model and the score cutoff t used to predict motifs (see Appendix for details). Finally, an estimate of the probability of a motif in both sequences is $\hat{p}_{11} = \hat{p}\hat{p}_{1\to1}$, and the resulting estimator of ρ takes the form:

(11)
$$\hat{\rho} = \big(\hat{p}_{11} - \hat{p}^2\big)/\big(\hat{p}(1-\hat{p})\big).$$

Note that $\hat{\rho} = 0$ for independent sequences ($\hat{p}_{11} = \hat{p}^2$), and $\hat{\rho} > 0$ for positively correlated sequences $\hat{p}_{1\to1} > \hat{p}$. Negative between-sequence correlation is typically not accounted for in evolutionary models, so for homologous sequences we have $\hat{\rho} \geq 0$.

3. SOFTWARE PACKAGE

We implemented statistical tests for differences in the number of motifs between two sequences in an open source software package, called `motifDiverge`, which is written in the R programming language. The package includes functions for predicting motifs in sequences and computing p–values based on an estimate of the distribution of motif differences between two sequences. The difference distribution and p–value account for sequence lengths, nucleotide composition of the sequences and the motif, the total number of motifs, and the similarity of the two sequences. The `motifDiverge` package is freely available by request from the first author or can be downloaded from http://www.kostkalab.net/software.

4. SIMULATION STUDY

We performed a study on simulated data to assess whether the model in Equation (6) describes differences in the number of annotated motifs between two sequences well. In order to assess the model and our proposed heuristics for parameter estimation, we compare the shape of estimated histograms for $P(N_{xy} = n_{xy})$ to the true distribution under different scenarios. We also assess the distribution of p–values obtained from data simulated under the null hypothesis. These analyses make use of generative phylogenetic models for pairs of DNA sequences. We simulate independent sequence pairs (x, y), as well as correlated sequences where transitions between corresponding nucleotides in x and y are modeled by a continous time Markov chain (CTMC).

4.1 Simulation approach

We use a phylogenetic hidden Markov model (phyloHMM) [21] to generate pairs of sequences (x, y). Let τ denote the evolutionary time separating x and y. When τ is small, x and y are correlated (e.g., homologous), while $\tau \to \infty$ generates independent sequences. To simulate motif instances, our phyloHMM consists of three states: a background (BG) and two motif states (M_1, M_2, which are reverse complements of each other). The transition probabilities between these states are $1-\zeta$ for BG to BG, M_1 to BG, or M_2 to BG, and $\zeta/2$ for BG to M_1, BG to M_2 or between M_1 and M_2 (Appendix). The parameter ζ encodes motif prevalence. The background state consists of a CTMC with a strand-symmetric and time-reversible rate matrix estimated from neutrally evolving sites in primate genomes (46-way Conservation track from the UCSC Genome Browser, http://genome.ucsc.edu). It emits two corresponding nucleotides (one in sequence x and one in sequence y) separated by evolutionary distance τ (i.e., there are τ expected substitutions between x and y per nucleotide, also some times denoted K or D). The motif state consists of a similar CTMC except that the equilibrium probabilities of each position equal the probability distribution given by the TF's PSPM (or its reverse complement). Each motif state emits two sequences of motif-length (one for x and one for y).

We repeatedly generated sequence pairs (x, y) and predicted motifs for the transcription factor Nkx2-5 using a log odds score threshold t with a false positive rate (Type I error, see section 2.1) for motif hits of 1%. Sequence pairs were generated with different lengths (k_x, k_y), different between sequence divergence parameters τ, and different motif-prevalence parameters ζ. To simulate $k_x \neq k_y$, we generate two sequences of the longer length and then delete the excess nucleotides from the shorter sequence. In most simulations, the motif prevalence is the same in x and y, so that we are simulating data reflecting $P(N_{xy} = n_{xy})$ under the null hypothesis of no motif differences between x and y.

For each simulation scenario we generated 100,000 sequence pairs, counted motif-hit differences, and then computed three estimates of $P(N_{xy} = n_{xy})$ based on the simulated data: (i) Maximum likelihood estimation given our model, where we find the parameters that maximize the likelihood according to Equation (6); (ii) A Gaussian distribution with mean and variance estimated from the simulated data; and (iii) the same Gaussian distribution with continuity correction that accounts N_{xy} being an integer. We also estimated p–values using different estimation schemes for the model parameters, which we describe in Section 2.2.5. These cover independent versus homologous sequences and count-based versus phyloHMM-based estimates.

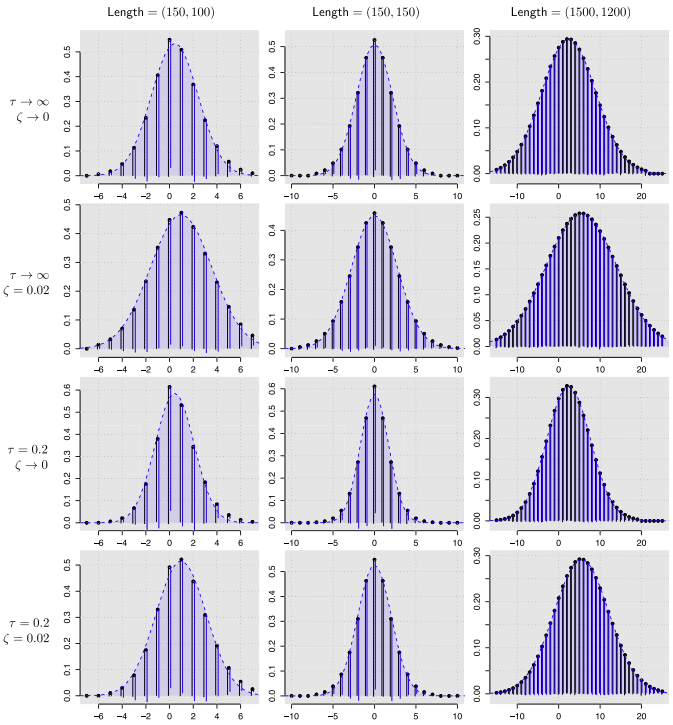

Figure 1. $P(N_{xy} = n_{xy})$ describes differences in motif hits well. The rows show different between-sequence dependence, the columns different sequence lengths.

4.2 Simulation results

First, we show that the proposed estimators of $P(N_{xy} = n_{xy})$ describe differences in motif hits well. Figure 1 shows results for three combinations of (k_x, k_y) (columns) and four combinations of (τ, ζ) (rows). For each scenario, we simulated 100,000 data sets. Each plot shows a hanging rootogram [22] of the differences in the number of observed Nkx2-5 motifs. That is, the vertical axis denotes the square root of the probability, and the horizontal axis the difference in motif counts. The solid circles correspond to the maximum likelihood fit of $P(N_{xy} = n_{xy})$ to the simulated data. The blue dashed lines correspond to a Gaussian approximation with the estimated mean and variance, and the blue vertical bars are the corresponding Gaussian values with continuity correction. These should be compared to the lengths of the black vertical bars, which cor-

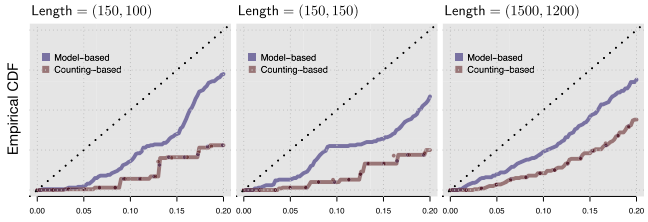

Figure 2. Partial empirical CDF of 1,000 p–values computed using different parameter estimates for data simulated under the null hypothesis. Three panels show different sequence lengths.

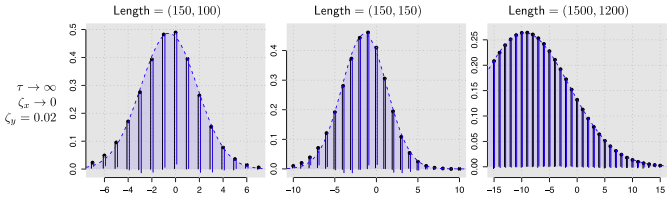

Figure 3. $P(N_{xy} = n_{xy})$ for TF motif differences for sequences with different motif prevalence (ζ_x vs. ζ_y).

respond to the true frequencies of n_{xy} in the simulation. The first two rows show simulations for independent sequences ($\tau \to \infty$) for different values of ζ, while in the second two rows x and y are related ($\tau = 0.2$ expected substitutions per nucleotide). Across these different scenarios, we find that all three estimators of $P(N_{xy} = n_{xy})$ very accurately capture the observed distribution of motif-count differences in our simulations. In other words, the black vertical bars nearly all end at zero; the blue bars are often similar in length to the black bars, and the dotted blue density in general matches the other three distributions fairly closely.

Next, we looked at the accuracy of our estimated p–values. We simulated 1,000 sequence pairs with $\tau = 0.02$, $\zeta = 0.02$, and three combinations of sequence lengths (k_x, k_y). Figure 2 summarizes the results. Each panel shows the (partial) empirical cumulative distribution function (CDF) of p–values obtained from different parameter estimates. The blue lines represent model-based estimates, whereas the red lines represent count-based estimates (see Appendix for definitions of different parameter estimates). The solid lines treat the sequence-pairs as homologous (which is how the data were generated), whereas the dotted lines assume independence between x and y. We find that our p–values are mostly conservative, and that for longer sequences they become approximately uniformly distributed for smaller p. Interestingly, when the simulated sequence pairs are uncorrelated, the estimates are very similar for count-based and for model-based parameter estimates. In light of the greater computational effort for model-based estimates this may suggest the usage of count-based estimates for non-homologous sequences.

Finally, to assess the model fit of $P(N_{xy} = n_{xy})$ when motif prevalence is different between x and y, we simulated 100,000 sequence pairs in the following way. Sequence x was simulated from a phyloHMM with $\zeta_x \to 0$ and sequence y from a model with $\zeta_y = 0.02$. Taking single sequences from two different phyloHMMs corresponds to $\tau \to \infty$. Figure 3 is analogous to Figure 1 and shows the result. We find that even when motif prevalence is different, our estimators of $P(N_{xy} = n_{xy})$ accurately capture the properties of the true, simulated distribution of N_{xy}.

5. MOTIF DIVERGENCE IN GENE REGULATORY ENHANCERS DURING CARDIAC DEVELOPMENT

To illustrate the use of `motifDiverge` on genome sequence data, we analyze a collection of gene regulatory elements identified via ChIP-seq for the active enhancer-marking histone modification histone 3 lysine 27 acetylation (H3K27ac) by Wamstad *et al.* [9]. This study identified genomic sequences marked by H3K27ac in mouse embryonic stem cells (ESCs) and at several subsequent developmental time points along the differentiation of ESCs into cardiomyocytes (CMs), which are beating heart cells. Our analysis uses these cell type specific enhancer sequences to illustrate applications of `motifDiverge` to both non-homologous and homologous sequences. We also leverage RNA-seq gene expression measurements from the same ESC and CM samples [9] to identify expressed TFs. Tissue development is a useful system for illustrating our approach, because active regulatory elements and TFs that are important for regulating gene expression differ across cell types and between species.

5.1 Motif divergence between mouse and human enhancer sequences

We first explored the use of `motifDiverge` to quantify motif differences between homologous sequences. For each of the 8,376 H3K27ac-marked enhancers from mouse CMs, we identified the homologous human sequence (if any) using the whole-genome, 30-way vertebrate multiple sequence alignments available from the UCSC Genome Browser (http://genome.ucsc.edu), which are based on the hg18 and mm9 genome assemblies. It is interesting to compare CM gene regulation between these two species, because there are a number of structural and electrophysiological differences between their hearts. We identified 1,617 orthologous human-mouse sequence pairs that were at least 20 nucleotides long. For each enhancer pair, we estimated the number of motif hits in the human and mouse sequence with JASPAR PSPMs (http://jaspar.genereg.net) for all 53 TFs expressed in mouse CMs (defined as those that have at least 10 sequence fragments per kilobase of sequence in the gene per million fragments aligned to the genome: RPKM > 10). We set the log odds score threshold to achieve a Type I error rate of 5%. Our findings were fairly robust to this thresh-

Table 1. Transcription factors with the most enhancers showing significant divergence in motif counts between human and mouse sequences, excluding those with more than 2% of enhancers showing discordant results between model-based and count-based parameter estimation methods

Transcription factors with more motifs in mouse	
TF	**Proportion of CM enhancers with more motifs in mouse**
Egr1	0.137
Mycn	0.064
Fhl1	0.033
Pbx1	0.032
Jdp2	0.012
Transcription factors with more motifs in human	
TF	**Proportion of CM enhancers with more motifs in human**
Mafk	0.275
Mycn	0.035
Creb3l2	0.033
Trp53	0.025
Jdp2	0.023
Srebf1	0.020
Fhl1	0.019
Gabpa	0.012
Deaf1	0.0093

old choice (Appendix; Figure 4). Then we tested for TFs with significant differences in motif counts between human and mouse in each CM enhancer region using count-based parameter estimation. Model-based estimation produced p–values that were highly correlated with those from the count-based analysis (Appendix; Figure 5).

After adjusting for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) controlling procedure [23], we found that a large percentage of enhancers (82%) show evidence of significant differences in motif counts for at least one TF (FDR $< 5\%$; count-based parameter estimation). About two thirds of CM enhancers (66%) have significant differences in motif counts for multiple TFs, and several have significant differences for ten or more TFs. Conversely, most TFs only have significant differences in counts between human and mouse for a small percentage of CM enhancers. This suggests that differences in the motif composition of ESC and CM enhancers is driven mostly by a few TFs. The TFs with the largest percentage of enhancers showing significant differences are listed in Table (1). These TFs are promising candidates for understanding differences in CM gene regulation between humans and mice. Interestingly, Mycn, Jdp2 and Fhl1 have some enhancers with significantly more motifs in human and some enhancers with more motifs in mouse, suggesting that these TFs may target somewhat different sets of enhancers—and potentially different genes—in the two species.

5.2 Differences in motifs between enhancers active in different cell types

Next, we used `motifDiverge` to compare motif counts between non-homologous sequence pairs. This application also illustrates how `motifDiverge` can be applied to perform a single test to compare two sets of sequences. We concatenated the sequences of the 10,580 H3K27ac-marked regions in CMs to create a single, long sequence containing all the active enhancers for this cell type. Then, we generated a similar concatenation of all 7,159 enhancers from ESCs. Any genome sequence marked by H3K27ac in both ESCs and CMs was removed from both data sets, so that the resulting two ESC and CM enhancer sequences were non-overlapping. We predicted motifs in the ESC and CM sequences as described above with PSPMs for all 73 TFs expressed in either cell type. Then we tested for TFs with significant differences in motif counts between the combined enhancer regions of the two cell types. At FDR $< 5\%$, we found 40 TFs with significantly more motifs in ESC enhancers and 27 TFs with significantly more motifs in CM enhancers.

To better understand the biological meaning of these results, we used the Wamstad *et al.* RNA-seq data to quantify the expression of each TF in ESCs and CMs. Several TFs are only highly expressed in one cell type, while others are expressed in both ESCs and CMs. The TFs with the most significant motif divergence included many that were highly expressed in the cell type with more motifs, but also some with low–though potentially biologically significant–expression levels (Table 2). This is not surprising, since TFs can function at low expression levels. Expression levels of some TFs were much higher in the cell type with more motifs compared to the other cell type (e.g., Gbx2 and Sox15 in ESCs, Egr1 in CMs), but in many cases expression was similar or even higher in the cell type with fewer motifs (e.g., Nkx2-5). This suggests that RNA-seq data might also be useful for filtering out significant motif differences that are not biologically meaningful; Nkx2-5 is not expressed in ESCs, making it unlikely that the additional motifs affect ESC gene regulation. More likely, these motifs reflect similarity in the Nkx2-5 motif to other TF's motifs or usage of the ESC regulatory regions in other cell types where Nkx2-5 is expressed, a hypothesis that could be tested as RNA-seq data from more cell types becomes available. Finally, Nkx2-5 and many other TFs have multiple different motif models (PSPMs) in different databases, and results should also be compared across PSPMs for the same TF, which can be quite different from one another. In the case of Nkx2-5, enrichment in ESCs is not recapitulated with some of the alternative PSPMs, further supporting the idea that the ESC motif hits are not biologically meaningful. These analyses show how `motifDiverge` can be used to analyze data from ChIP-seq experiments and how RNA-seq data can be used to filter and interpret `motifDiverge` findings, leading to robust conclusions about the role of sequence differences in gene regulation.

Table 2. Transcription factors with the most significant differences in TF motif counts between ESCs and CMs. Expression values are reads per kilobase per million fragments sequenced (RPKM)

Transcription factors with more motifs in ESC			
TF	**FDR adjusted p–value**	**ESC Expression**	**CM Expression**
Rhox6	<1e-300	11.890	0.129
Pou3f1	<1e-300	15.081	0.380
Zfp187	<1e-300	8.215	13.816
Sox2	<1e-300	212.888	0.120
Hmbox1	<1e-300	4.799	12.281
Pou2f1	<1e-300	12.103	7.669
Sox12	<1e-300	22.933	23.458
Foxd3	<1e-300	20.746	0.038
Zfp105	<1e-300	10.949	4.216
Srf	<1e-300	33.402	42.569
Sox13	<1e-300	14.345	2.693
Tbp	<1e-300	15.609	5.676
Hbp1	<1e-300	17.552	28.762
Arid3a	<1e-300	5.055	15.516
Sox4	<1e-300	17.788	41.915
Pbx1	<1e-300	6.593	43.487
Gata6	<1e-300	0.152	75.887
Mafk	<1e-300	3.695	18.851
Pou5f1	<1e-300	669.960	0.043
Yap1	9.5e-294	51.1	57.6
Cebpb	5.3e-271	6.2	16.7
Gbx2	7.7e-258	22.7	0.01
Zfp652	5.1e-156	8.0	11.9
Dbp	4.1e-132	3.5	28.5
Elf3	1.1e-130	24.3	0.6
Zbtb12	3.2e-84	23.5	25.8
Tcf7	2.3e-75	16.3	7.6
Fhl1	4.7e-63	26.6	29.7
Nkx2-5	3.0e-60	0.9	177.6
Sox15	1.0e-42	14.1	0.1
Transcription factors with more motifs in CM			
TF	**FDR adjusted p–value**	**ESC Expression**	**CM Expression**
Tcfap2c	<1e-300	23.884	0.055
Zic2	5.5e-248	26.8	0.1
Srebf1	1.1e-185	15.5	27.5
Esrrb	7.4e-98	85.2	2.4
Zic3	4.3e-92	38.6	0.07
Creb3l2	2.0e-69	1.3	50.1
Stat3	1.1e-62	8.4	21.8
Tgif1	5.6e-61	43.3	6.4
Smad3	1.8e-56	5.3	21.1
Myc	2.9e-53	32.0	4.2
Glis2	1.6e-49	12.5	12.9
Mlx	2.8e-35	18.3	7.03
Tcf3	1.2e-24	54.7	21.0
Mycn	2.6e-20	125.5	10.8
Xbp1	2.0e-18	18.8	20.5
Egr1	2.5e-18	19.9	192.2
Atf1	1.4e-17	26.3	7.2
Zbtb7b	2.7e-17	3.4	10.8

6. CONCLUSION

In this paper, we propose a new model for the difference in counts between two correlated Bernoulli trials representing numbers of TF motifs in a pair of DNA sequences. Our major results are the model derivation, accurate methods for parameter estimation, and a software package called `motifDiverge` that can be used to predict TF motifs and to perform tests comparing motif counts in two sequences. We illustrate the use of `motifDiverge` to discover TFs with significant differences in motifs (i) between two species, or (ii) between two cell types. These applications demonstrate the power of our methodology for discovering specific genes and regulatory mechanisms involved in species divergence and tissue development through careful analysis of ChIP-seq data.

Sequence divergence is usually measured in numbers of DNA substitutions or model-based estimates of rates of substitutions. These measures do not account for whether or not substitutions create or destroy TF motifs and are not well suited to quantify functional divergence [12]. Our tests capture how changes to DNA sequences affect their TF motif composition, and therefore they provide a more meaningful measure of divergence for regulatory regions. Hence, our model will be useful for understanding when non-coding mutations affect or do not affect the function of regulatory sequences. This information will enable, for example, identification of causal mutations in genomic regions identified as associated with diseases or other phenotypes. Since the majority of these genome-wide association study (GWAS) hits are outside of protein-coding regions [24], `motifDiverge` has the potential to have a large impact on human genetics research.

In future work, it would be interesting to extend our approach to model the joint distribution of multiple correlated Bernoulli trails and univariate summary statistics (e.g., sums, differences) of this distribution. As with two sequences, the main challenge is modeling correlations between the sequences. The phylogenetic tree models we used here can measure relationships between multiple homologous, but not equally related, DNA sequences; therefore they could provide a natural solution to this problem. Another interesting application would be to leverage motif divergence for phylogenetic tree construction, in place of the usual metric of overall sequence divergence. This could potentially be achieved in a maximum likelihood framework after development of a tree-based version of `motifDiverge` for multiple species.

We focus on comparing counts of TF motifs in two (possibly homologous) sequences, but our model is not specific to motifs in any way. The random variables N_x and N_y could represent other features of interest in two related DNA sequences, such as counts of microRNA binding sites, repetitive elements, polymorphisms, or experimentally measured events (e.g., ChIP-seq peaks). In fact, the two Bernoulli trials do not need to measure events on sequences, and our

model could be applied to many other types of correlated count data.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Institutes of Health (#GM82901 and #HL098179), a National Science Foundation graduate fellowship, and institutional funds from the Gladstone Institutes and the University of Pittsburgh School of Medicine.

APPENDIX

Derivation of Equation (3)

$P(N_{xy} = n_{xy})$ is a hypergeometric function for $k_x = k_y$

The probability mass function of N_{xy} for equal length sequences (Equation (2)) can be written as a sum: $P(N_{xy} = n_{xy}) = \sum_j S_j$, with the summands S_j given by Equation (1). Taking the ratio of two successive summands we get:

$$S_{j+1}/S_j = \frac{(n - n_{xy} - 2j)(n - n_{xy} - 2j - 1)}{(j + n_{xy} + 1)(j + 1)} \frac{p_{10}p_{01}}{(1 - p_{10} - p_{01})^2} = \frac{(j + \frac{n_{xy}-n}{2})(j + \frac{n_{xy}+1-n}{2})}{(j + n_{xy} + 1)(j + 1)} \frac{4p_{10}p_{01}}{(1 - p_{10} - p_{01})^2}. \tag{12}$$

We note that this is a rational function in j, n_{xy} and n and identifies the arguments $(n_{xy} - n)/2$, $(n_{xy} - n + 1)/2$ and $(n_{xy} + 1)$ of the Gaussian hypergeometric function in Equation (3) [19].

Derivation of Equation (8)

Error bound for evaluating $P(N_{xy} = n_{xy})$ for $k_x = k_y$

Let $w_j := S_{j+1}/S_j$. From Equation (12), we get that increasing j decreases the numerator S_{j+1} and increases the denominator S_j, so that w_j is decreasing in j. Therefore, there exists j_-, with $w_{j_-} < 1$ (i.e., the summands S_j are decreasing for $j \geq i_-$). The error $\epsilon(j_+)$ of truncating the sum over j at $j_+ \geq j_-$ is then:

$$\epsilon(j_+) = \sum_{j=j_++1}^{\lfloor \frac{n-n_{xy}}{2} \rfloor} S_j = \sum_{j=j_++1}^{\lfloor \frac{n-n_{xy}}{2} \rfloor} w_{j-1}w_{j-2}\ldots w_{j_+}S_{j_+} < \sum_{j=1}^{\lfloor \frac{n-n_{xy}}{2} \rfloor - j_+} w_{j_+}^j S_{j_+} = S_{j_+}\left(\frac{1 - w_{j_+}^{\lfloor \frac{n-n_{xy}}{2} \rfloor - j_+}}{1 - w_{j_+}} - 1\right), \tag{13}$$

where we have used the following: (i) $S_j = (S_j/S_{j-1})S_{j-1} = w_{j-1}S_{j-1}$, (ii) S_j are decreasing for $j \geq j_-$, (iii) $S_j \leq 1$ are non-negative multinomial probabilities (see Equation (2)), and (iv) the geometric sum. Thus, to estimate the probability mass function of N_{xy} to a desired precision ϵ, $\sum_j S_j$ an be truncated at the first $j_+ \geq j_-$ for which $\epsilon(j_+) \leq \epsilon$.

Derivation of Equation (9)

Recurrence relation for $P(N_{xy} = n_{xy})$ for $k_x = k_y$

Let $P(N_{xy} = n_{xy}) = \sum_j S(j, n_{xy}, n)$, where the summands $S(j, n_{xy}, n)$ are taken from Equation (2). Recurrence relations in n and n_{xy} can be obtained via the Zeilberger algorithm [19], for instance as implemented in the computer algebra system Maxima (http://sourceforge.net/projects/maxima). For a recurrence in n_{xy}, the Maxima code is:

```
(%1) Sj : n!/((n_{xy}+j)!*i!*(n-n_{xy}-2*j)!)
       *p10^(n_{xy}+j)*p01^j*(1-p10-p01)^(n-n_{xy}-2*j) $
(%2) load(zeilberger) $
(%3) Zeilberger(Sj,j,n_{xy});
(%o3) [[-(j*(n+n_{xy}+2)*p10)/(n_{xy}+j+1),
      [(n-n_{xy})*p10,(n_{xy}+1)*(p10+p01-1),
      -(n+n_{xy}+2)*p01]]]
```

This output defines the following quantities:

$$\begin{aligned} a_0(n_{xy}, n) &= (n - n_{xy})p_{10} \\ a_1(n_{xy}, n) &= -(n_{xy} + 1)(1 - p_{10} - p_{01}) \\ a_2(n_{xy}, n) &= -(n + n_{xy} + 2)p_{10} \\ R(j, k, n_{xy}) &= -\frac{j(n + n_{xy} + 2)p_{10}}{n_{xy} + j + 1}, \end{aligned}$$

which satisfy the recurrence relation

$$a_0(n_{xy}, n)S(j, n_{xy}, n) + a_1(n_{xy}, n)S(j, n_{xy} + 1, n) + a_2(n_{xy}, n)S(j, n_{xy} + 2, n) = R(j + 1, n_{xy}, n)S(j + 1, n_{xy}, n) - R(j, n_{xy}, n)S(j, n_{xy}, n). \tag{14}$$

Summing Equation (14) over j gives the recurrence for $P(N_{xy} = n_{xy})$. We confirm that the right hand side is zero:

$$a_0(n_{xy}, n)P(N_{xy} = n_{xy}) + a_1(n_{xy}, n)P(N_{xy} = n_{xy} + 1) + a_2(n_{xy}, n)P(N_{xy} = n_{xy} + 2) = \sum_{j=0}^{\lfloor \frac{n-n_{xy}}{2} \rfloor} (R(j + 1, n_{xy}, n)S(j + 1, n_{xy}, n) - R(j, n_{xy}, n)S(j, n_{xy}, n)) = R\left(\lfloor \frac{n - n_{xy}}{2} \rfloor + 1, n_{xy}, n\right) S\left(\lfloor \frac{n - n_{xy}}{2} \rfloor + 1, n_{xy}, n\right) - R(0, n_{xy}, n)S(0, n_{xy}, n) = 0.$$

That $R(0, n_{xy}, n) = 0$ follows straight from the definition, and that $S(\lfloor \frac{n-n_{xy}}{2} \rfloor + 1, n_{xy}, n) = 0$ follows via $S_{j+1} = (S_{j+1}/S_j)S_j$ and Equation (12).

Derivation of Equation (10)

In-sequence and between-sequence correlation

As mentioned in the main text, PSPM based annotation of motifs generates in-sequence dependence that is not per-se accounted for in our model. Suppose there is a first order

(Markov) dependence of X_i on X_{i-1}, quantified by the parameter λ_x (and likewise for Y_i). Under these assumptions the expected value for N_x is still $k_x p$, but for the variance we find [20]:

$$\begin{aligned}&\mathrm{Var}(N_x) = \\ &k_x p(1-p) + \frac{2p(1-p)(\lambda_x - p)}{1-\lambda_x} \times \\ &\left[(k_x - 1) - \frac{\lambda_x - p}{1-\lambda_x}\left(1 - \left[\frac{\lambda_x - p}{1-p}\right]^{k_x}\right)\right],\end{aligned} \tag{15}$$

and an equivalent expression for N_y. For $N_{xy} = N_x - N_y$ we then find (assuming no between-sequence dependence)

$$\begin{aligned}\mathrm{Var}(N_{xy}) = \; & k_x p(1-p) + A(p, \lambda_x, k_x) + \\ & k_y q(1-q) + A(q, \lambda_y, k_y)\end{aligned} \tag{16}$$

where $A(\cdot,\cdot,\cdot)$ represents the second term in the variance formula in Equation (15) and $\mathrm{Cov}(X_i, Y_i) = 0$. Comparing Equation (16) with Equation (7), substituting $p = p_{11} + p_{10}$ and $q = p_{11} + p_{01}$ we arrive at Equation (10) after some algebra. Note that a negative correlation between X_i and X_{i+1} decreases the variance in $N_x = \sum_i X_i$, and similarly for N_y. If both sequences have negative correlation between subsequent successes, the variance of N_{xy} decreases. This is the same effect a correlation between X_i and Y_i has on the variance of N_{xy}.

Parameter estimates contributing to $\hat{\rho}$ in Equations (10) and (11)

Count-based and model-based parameter estimates

Here we describe estimates for the parameters $\lambda_x = \lambda_y =: \lambda$ (for independent sequences)and $p_{1\to1}$ (for homologous sequences with alignment). These quantities reflectin-sequence and between-sequence dependencies, respectively: $\lambda = P(X_i = 1|X_{i-1} = 1)$ and $p_{1\to1} = P(Y_i = 1|X_i = 1)$, see Section 2.2.5 in the main text. We assume that the in-sequence dependence is the same in x and y, that motif gains and losses are time-reversible (i.e., $P(Y_i = 1|X_i = 1) = P(X_i = 1|Y_i = 1)$ and present count-based estimates as well as estimates based on a phylogenetic hidden Markov model (phyloHMM).

Count-based estimates: For a count-based estimate for λ, we count the number of adjacent motif hits in both x and y, and then divide it by the number of overall motif hits in both sequences. This is analogous to the estimate $\hat{p}$ for the success probability of the two Binomial trials X and Y, as described in the main text. For $\hat{p}_{1\to1}$, in turn, we count the number of congruent motif hits in x and y from alignments, and then divide the result by the overall number of motif hits in x. The advantage of both these estimates is that they do not take much effort to calculate. The downside is that typically p is small (for instance because a strict Type I error cutoff t in motif prediction, see Section 2.1). This, in turn, means that (especially for short and intermediate length sequences) not many adjacent or congruent motif hits will be observed. Therefore these count-based estimates can be very variable in those situations.

Model-based estimates: To overcome the variability in the count-based estimates described above to some extent, we assume a phyloHMM as an underlying, generative model for the two sequences x and y. For this approach we require a sequence alignment of x and y. Essentially,we fit the phyloHMMto our observation (the sequences x and y, plus the corresponding motif hits) and then derive the parameters of interest as large sample properties from the fitted model. As described in Section 4.1, the phyloHMM consists of three states: a background state (corresponding to a neutral evolutionary model), a motif state, and a state for the reverse complement of the motif. First, we model the transition probabilities to be $\zeta/2$ for background-to-motif and motif-to-motif transitions, and $1-\zeta$ for background-to-background and motif-to-background transitions. We then fix the ζ in such a way that

$$E_{\mathscr{O}}[P_{S\sim\mathscr{O}}(T(S) > t)] = \hat{p}, \tag{17}$$

where $\hat{p}$ is our count-basedestimate of the success probability, S is a nucleotide sequence of motif-length (with log odds score $T(S)$) emitted by the phyloHMM Ψ as either of the two sequences, and $\mathscr{O}$ is the state-path of motif-length generated by the Markov chain in Ψ that underlies the emission of S. Note that the left-hand side in Equation (17) depends on ζ because the probability for each state-path depends on the transition probabilities; but the LHS is independent of τ (the evolutionary time between sequences x and y), because the Ψ is time-reversible and S is a "marginalized" single sequence, not a sequence-pair. To evaluate the expectation in Equation (17) we enumerate all possible state-paths $\mathscr{O}$ and calculate (i) their Type I motif-hit error according to the PSPM and background distribution used for motif annotation (see Section 2.2), and (ii) their probability of occurrence from the equilibrium frequencies of the Markov chain. This yields an estimate $\hat{\zeta}$.

Next, to obtain an estimate for τ we maximize the likelihood of the sequences x and y:

$$\hat{\tau} = \mathrm{argmax}_\tau \; L((x,y)|\Psi(\tau, \hat{\zeta})),$$

where $L()$ denotes the likelihood of jointly observing x and y. Overall this procedure yields a fully specified (fitted) phyloHMM $\Psi(\hat{\tau}, \hat{\zeta})$.

Finally, we use this fitted phyloHMM to obtain estimates for λ and $p_{1\to1}$. To that end we generate two very long (100,000 nucleotides or longer) sequences and take (i) $\hat{\lambda}$ to be the fraction of adjacent motif hits, and (ii) $\hat{p}_{1\to1}$ to be the fraction of motif hits that is congruent between the two generated sequences. We note that it is straightforward to obtain bounds for these estimates via Binomial tail inversion [25].

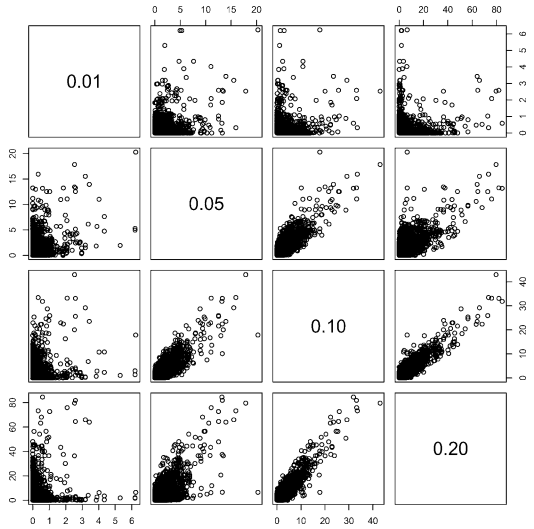

Figure 4. Scatter plots of `motifDiverge` *–log(p–values) comparing human versus mouse cardiomyocyte enhancers for all expressed TFs at different motif hit thresholds* $(0.01, 0.05, 0.1, 0.2)$.

Note that highly similar, well-aligned sequences will lead to short estimated evolutionary times $\hat{\tau}$, and therefore high values for $\hat{p}_{1\to 1}$, which will in-turn lead to large estimates of $\hat{\rho}$ in Equation (11). Conserved motif instances in both sequences will, during the estimation procedure, "vote" for larger estimates of ζ and for smaller estimates of τ; non-conserved motif hits, on the other hand, will still favor large ζ, but also large τ (in contrast to small τ). Long evolutionary times are the only way the null-model (of no difference in motif-prevalence between the two sequences) can account the creation/destruction of motif hits. The evolutionary model also accounts for the fact that some nucleotide changes are more likely than others, and that some motifs are therefore more likely to be lost and gained than others, based on their nucleotide composition. Finally, more realistic evolutionary models that explicitly include insertions and deletions could in principle be utilized in this framework.

Effects of the threshold used to identify motif hits

We investigated the sensitivity of `motifDiverge` findings to the choice of threshold used to identify motif hits in each sequence. Specifically, for the analysis of human versus mouse CM enhancers (Section 5.1), we tested the robustness of our results by varying the log odds score threshold across a range of values: $1\%, 5\%, 10\%, 20\%$. We found that the resulting `motifDiverge` p–values are correlated (Figure 4) across thresholds, with higher correlation at more similar thresholds, as expected. We also observed that our conclusions about human versus mouse enhancer motif content are not dramatically affected by the choice of threshold.

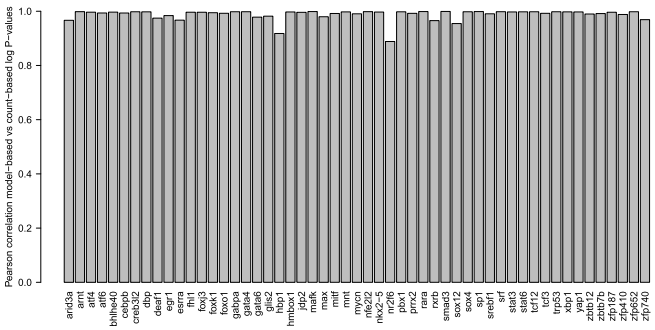

Figure 5. Correlation between `motifDiverge` p*–values from count-based versus evolutionary model-based parameter estimation. Tests are for enrichment in mouse compared to human cardiomyocyte enhancers.*

Effects of parameter estimation methods

We compared several methods of `motifDiverge` model parameter estimation in our analysis of human versus mouse CM enhancers (Section 5.1). Overall, model-based estimation identified more TFs with significantly different numbers of motifs between the two species. We found that 100% of enhancers had at least one TF with significant differences compared to 82% with count-based estimation. However, `motifDiverge` p–values from the two approaches are highly correlated (Figure 5). For many TFs, including Smad3, Atf6, Fhl1, Jdp2, and Arnt, there were almost no differences in the number of enhancers with significant losses or gains when comparing count-based and model-based estimation (<1% of enhancers discordant). On average across TFs, about 2% of enhancers produced discordant results when using a Type I error threshold of 5% for calling motif hits and FDR < 0.05 for `motifDiverge` tests. No TFs had more than 5% discordant results across enhancers. When using a stricter Type I error threshold of 1% for calling motif hits, results were more discordant between count-based and model-based estimation procedures (typically 10%−30% discordant). Thus, while our simulations indicate that model-based estimation can be more sensitive, we found in practice that for species at the divergence of human and mouse, accounting for the phylogenetic relationship between sequences does not have a big impact on `motifDiverge` results. Both options are available in the R package and can be explored by users for their particular application.

Received 1 February 2014

REFERENCES

[1] Rivera, C. M. and Ren, B., "Mapping human epigenomes," *Cell*, vol. 155, pp. 39–55, Sept. 2013.

[2] McGettigan, P. A., "Transcriptomics in the RNA-seq era," *Current Opinion in Chemical Biology*, vol. 17, pp. 4–11, Feb. 2013.

[3] Ozsolak, F. and Milos, P. M., "RNA sequencing: advances, challenges and opportunities," *Nat Rev Genet*, vol. 12, no. 2, pp. 87–98, 2011.

[4] John, S., Sabo, P. J., Canfield, T. K., Lee, K.,Vong, S., Weaver, M., Wang, H., Vierstra, J., Reynolds, A. P., Thurman, R. E., and Stamatoyannopoulos, J. A., "Genome-scale mapping of DNase I hypersensitivity," *Current protocols in molecular biology/edited by Frederick M. Ausubel ... [et al.]*, vol. Chapter 27, pp. Unit 21.27–21.27.20, July 2013.

[5] Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., and Lieb, J. D., "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin," *Genome Research*, vol. 17, pp. 877–885, June 2007.

[6] Furey, T. S., "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions," *Nat Rev Genet*, vol. 13, pp. 840–852, Dec. 2012.

[7] Wang, C., Zhang, M. Q., and Zhang, Z., "Computational identification of active enhancers in model organisms," *Genomics, Proteomics and Bioinformatics*, vol. 11, pp. 142–150, June 2013.

[8] de Laat, W. and Duboule, D., "Topology of mammalian developmental enhancers and their regulatory landscapes," *Nature*, vol. 502, pp. 499–506, Oct. 2013.

[9] Wamstad, J. A., Alexander, J. M., Truty, R. M., Shrikumar, A., Li, F., Eilertson, K. E., Ding, H., Wylie, J. N., Pico, A. R., Capra, J. A., Erwin, G., Kattman, S. J., Keller, G. M., Srivastava, D., Levine, S. S., Pollard, K. S., Holloway, A. K., Boyer, L. A., and Bruneau, B. G., "Dynamic and Coordinated Epigenetic Regulation of Developmental Transitions in the Cardiac Lineage," *Cell*, vol. 151, no. 1, pp. 206–220, 2012.

[10] Maston, G. A., Landt, S. G., Snyder, M., and Green, M. R., "Characterization of enhancer function from genome-wide analyses," *Annual Review of Genomics and Human Genetics*, vol. 13, no. 1, pp. 29–57, 2012.

[11] Stormo, G. D., "DNA binding sites: representation and discovery," *Bioinformatics*, vol. 16, pp. 16–23, Jan. 2000.

[12] Ritter, D. I., Li, Q., Kostka, D., Pollard, K. S., Guo, S., and Chuang, J. H., "The Importance of Being Cis: Evolution of Orthologous Fish and Mammalian Enhancer Activity," *Molecular Biology and Evolution*, vol. 27, no. 10, pp. 2322–2332, 2010.

[13] Rahmann, S., Müller, T., and Vingron, M., "On the power of profiles for transcription factor binding site detection," *Statistical Applications in Genetics and Molecular Biology*, vol. 2, p. Article7, 2003. MR2086500

[14] Siddharthan, R., "Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix," *PLoS ONE*, vol. 5, no. 3, p. e9722, 2010.

[15] Zhao, Y., Ruan, S., Pandey, M., and Stormo, G. D., "Improved models for transcription factor binding site identification using nonindependent interactions," *Genetics*, vol. 191, pp. 781–790, July 2012.

[16] Mathelier, A. and Wasserman, W. W., "The next generation of transcription factor binding site prediction," *PLoS Comput Biol*, vol. 9, no. 9, p. e1003214, 2013.

[17] Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., Agius, P., Arvey, A., Bucher, P., Callan, C. G., Chang, C. W., Chen, C.-Y., Chen, Y.-S., Chu, Y.-W., Grau, J., Grosse, I., Jagannathan, V., Keilwagen, J., Kiełbasa, S. M., Kinney, J. B., Klein, H., Kursa, M. B., Lahdesmaki, H., Laurila, K., Lei, C., Leslie, C., Linhart, C., Murugan, A., Myvsivckova, A., Noble, W. S., Nykter, M., Orenstein, Y., Posch, S., Ruan, J., Rudnicki, W. R., Schmid, C. D., Shamir, R., Sung, W.-K., Vingron, M., Zhang, Z., Bussemaker, H. J., Morris, Q. D., Bulyk, M. L., Stolovitzky, G., and Hughes, T. R., "Evaluation of methods for modeling transcription factor sequence specificity," *Nature Biotechnology*, vol. 31, no. 2, pp. 126–134, 2013.

[18] Pape, U. J., Rahmann, S., Sun, F., and Vingron, M., "Compound poisson approximation of the number of occurrences of a position frequency matrix (PFM) on both strands," *Journal of computational biology: a journal of computational molecular cell biology*, vol. 15, pp. 547–564, July 2008. MR2425441

[19] Petkovsek, M., Wilf, H. S., and Zeilberger, D., *A = B*. A K Peters, Ltd., 1996. MR1379802

[20] Klotz, J., "Statistical Inference in Bernoulli Trials with Dependence," *The Annals of Statistics*, vol. 1, pp. 373–379, Mar. 1973. MR0381103

[21] Hubisz, M. J., Pollard, K. S., and Siepel, A., "PHAST and RPHAST: phylogenetic analysis with space/time models," *Briefings in Bioinformatics*, vol. 12, no. 1, pp. 41–51, 2011.

[22] Tukey John, W., "Some Graphic and Semigraphic Displays," in *Statistical papers in honor of George W. Snedecor*, pp. 293–316, The Iowa State University Press, 1972. MR0448637

[23] Benjamini, Y. and Hochbert, Y., "Controlling the False Discovery Rate: A practical and powerful apporach to multiple testing," *Journal of the Royal Society B*, vol. 57, no. 1, pp. 289–300, 1995. MR1325392

[24] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A., "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 9362–9367, 2009.

[25] Kaariainen, M. and Langford, J., "A Comparison of Tight Generalization Error Bounds," in *Proceedings of the 22Nd International Conference on Machine Learning*, (New York, NY, USA), pp. 409–416, ACM, 2005.

Dennis Kostka
Department of Developmental Biology
Department of Computational & Systems Biology
University of Pittsburgh School of Medicine
530 45th Street
Pittsburgh, PA 15201
USA
E-mail address: kostka@pitt.edu

Tara Friedrich
Gladstone Institutes
Integrative Program in Quantitative Biology
University of California
1650 Owens Street
San Francisco, CA 94158
USA
E-mail address: tara.friedrich@gladstone.ucsf.edu

Alisha K. Holloway
Gladstone Institutes
Division of Biostatistics
University of California
1650 Owens Street
San Francisco, CA 94158
USA
E-mail address: alisha.holloway@gladstone.ucsf.edu

Katherine S. Pollard
Gladstone Institutes
Institute for Human Genetics
Division of Biostatistics
University of California
1650 Owens Street
San Francisco, CA 94158
USA
E-mail address: kpollard@gladstone.ucsf.edu