Assessing phylogenetic motif models for predicting transcription factor binding sites

John Hawkins1,∗, Charles Grant2, William Stafford Noble2,3 and Timothy L. Bailey1
1Institute for Molecular Bioscience, University of Queensland, Qld 4072, Australia, 2Department of Genome Sciences and 3Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA

ABSTRACT
Motivation: A variety of algorithms have been developed to predict transcription factor binding sites (TFBSs) within the genome by exploiting the evolutionary information implicit in multiple alignments of the genomes of related species. One such approach uses an extension of the standard position-specific motif model that incorporates phylogenetic information via a phylogenetic tree and a model of evolution. However, these phylogenetic motif models (PMMs) have never been rigorously benchmarked in order to determine whether they lead to better prediction of TFBSs than obtained using simple position weight matrix scanning.

Results: We evaluate three PMM-based prediction algorithms, each of which uses a different treatment of gapped alignments, and we compare their prediction accuracy with that of a non-phylogenetic motif scanning approach. Surprisingly, all of these algorithms appear to be inferior to simple motif scanning, when accuracy is measured using a gold standard of validated yeast TFBSs. However, the PMM scanners perform much better than simple motif scanning when we abandon the gold standard and consider the number of statistically significant sites predicted, using column-shuffled ‘random’ motifs to measure significance. These results suggest that the common practice of measuring the accuracy of binding site predictors using collections of known sites may be misleading since such collections may be missing ‘weak’ sites, which are exactly the type of sites needed to discriminate among predictors. We then extend our previous theoretical model of the statistical power of PMM-based prediction algorithms to allow for loss of binding sites during evolution, and show that it gives a more accurate upper bound on scanner accuracy. Finally, utilizing our theoretical model, we introduce a new method for predicting the number of real binding sites in a genome. The results suggest that the number of true sites for a yeast TF is in general several times greater than the number of known sites listed in the Saccharomyces cerevisiae Database (SCPD). Among the three scanning algorithms that we test, the MONKEY algorithm has the highest accuracy for predicting yeast TFBSs.

Contact: j.hawkins@imb.uq.edu.au

1 INTRODUCTION
An important goal in current molecular biology research is to understand the cellular systems that regulate gene expression. One part of this goal lies in identifying the binding locations of transcription factors (TFs)—proteins that bind to DNA segments and regulate the expression of nearby genes. Due to the difficulty in accurately assaying protein–DNA interactions on a large scale, we have only a small amount of high-quality experimental data regarding the location of these binding sites. Hence, we require good computational tools for identifying the remaining sites.

Unfortunately, attempts to construct accurate binding site predictors have been frustrated by the small size of the binding sites, which means that DNA regions identical to functional binding sites occur by chance with high frequency. As a result, standard probabilistic motif search using a position weight matrix (PWM) suffers from overwhelming numbers of false positive predictions (Wasserman and Sandelin, 2004). In order to produce better predictors, researchers have turned to other sources of information.

The biological reality is that many components of the cellular system cooperate to produce a protein–DNA binding event. For example, TFs tend to bind in groups to form macromolecular complexes, and the binding is thus stabilized by the presence of other TFs (Levine and Tjian, 2003). Furthermore, there is emerging evidence that a range of epigenetic factors contribute to the binding of TFs (Giaccione et al., 2006; Kouzarides, 2007; Liu et al., 2006).

Computational tools have been constructed that make use of TF protein complex information (Berman et al., 2002) and epigenetic DNA modifications (Natikar et al., 2007; Whittington et al., 2008).

An alternative source of information, which is not exploited by the cell itself, comes from comparative genomics. We expect that important features in biological sequences tend to evolve more slowly than the neutral rate. This assumption has been exploited to identify functional regions by comparative genomics techniques such as phylogenetic footprinting (GuhaThakurta, 2006; Stormo, 2000). The assumption of evolutionary conservation has also been utilized by a class of motif scanning algorithms based on what we call phylogenetic motif models (PMMs). These models extend the popular PWM motif models used in computational biology to represent and identify sequence features such as TF binding sites (TFBSs), splice junctions and binding domains in DNA, RNA and protein molecules, respectively (Ghulathakurta, 2006; Stormo, 2000). Whereas a PWM computes the probability of given region in a single genome being an instance of a motif, a PMM computes the probability of an ungapped region in a multiple alignment being comprised of instances of the motif that evolved independently from an ancestral instance.

PMM scanning algorithms are an extension of simple PWM scanning algorithms. Instead of scanning a single sequence, however, a PMM algorithm scans a multiple alignment of orthologous sequences, usually produced by a tool such as ClustaW (Chenna et al., 2003), MAFFT (Katoh et al., 2002), MULTIZ (Blanchette et al., 2004) or Multi-LAGAN (Bludno et al., 2003). The PMM scanning approach also requires an explicit model of
nucleic acid substitution and a phylogenetic tree describing the relationship and evolutionary distances among the species from which the orthologous sequences are taken. Several TFBS prediction algorithms have been implemented that use PMMs. These include the Monkey algorithm (Moses et al., 2004), a variant called rMonkey (Moses et al., 2006), and Morthp, which is available as part of the MEME Suite (http://meme.nbcr.net).

Surprisingly, earlier attempts to demonstrate that these PMM scanning algorithms are superior to simple PWM scanners failed to show any significant difference (Hawkins and Bailey, 2008). In that work, we evaluated the algorithms using TF motifs, multiple alignments and known TFBSs from yeast (Saccharomyces cerevisiae). In addition to indicating that simple PWM scanning might be as accurate (at least in yeast) as PMM scanning, the accuracy of PMM scanners at predicting the known binding sites was orders of magnitude less than one might therefore theoretically be, assuming that binding sites are not lost and that the multiple alignments are reasonably accurate.

One explanation for why the accuracy of PMM scanners appeared no better than that of simple PWM scans in previous work is that the gold standard might be biased against more sensitive scanning methods. One way in which this could happen is if the experimental methods used in creating the sets of known TFBSs systematically leave out weakly binding sites. Such weak sites will tend to have low PWM scores, and be missed by PWM scanners. However, weak sites are known to be important biologically (Gertz et al., 2008), so they will tend to be conserved, causing PMM scanners to detect them. If weak sites are systematically more likely to be missing from the gold standard, PMM scanners will appear to be less accurate (due to larger numbers of apparent false positive predictions—the weak-but-real sites) than simple PWM scanners.

Missing functional sites in the gold standard would also explain some, but not all, of the discrepancy between the theoretical and empirical accuracy of PMM scanners. Our earlier simplifying assumption that binding sites have not been lost in some of the orthologous species in the multiple alignment is clearly false (Moses et al., 2006). Relaxing this assumption in the theoretical model would lower the theoretically achievable prediction accuracy because loss-of-site events make real sites harder to distinguish from background sequence. Errors in the multiple alignments of orthologous species would also account for some of the discrepancy in accuracy, and the assumption that the multiple alignments correctly align orthologous positions is clearly overly optimistic.

The contributions of the present work are 4-fold. First, we extend our previous study of the empirical accuracy of PMM-based scanners using a set of known sites (a ‘gold standard’), confirming the apparent paradox noted above. Secondly, we show that the paradox disappears when we adopt an alternative evaluation approach that uses column-shuffled (‘random’) motifs, rather than a set of known sites, to measure prediction accuracy. Under this metric, PMM scanners are more accurate than a simple PWM scan, and heuristics for locally correcting errors in the multiple alignments are beneficial. Thirdly, we extend our theoretical model of PMM scanner accuracy to allow for site loss and and show that it more accurately fits the available data for yeast. Finally, we show how theoretical models of prediction accuracy can be used to estimate the number of functional binding sites of a TF in a genome, and provide such estimates for 21 yeast TFs.

2 MATERIALS AND METHODS

2.1 TFBS prediction algorithms

All the binding site prediction algorithms in this study use a matrix representation of a motif, $M$, in which the rows represent each of the four possible nucleic acids and the columns correspond to positions within the motif. Each matrix entry is the probability of seeing the given residue in the specified position. The algorithms all make use of a background distribution of nucleic acids, $B$, a single vector that contains the overall frequencies of the nucleic acids in the target genome. The PMM algorithms also require a phylogenetic tree with distances (in substitutions per site), relating the species in the alignments of orthologous regions that they scan.

Scoring a site using a PMM involves computing a log-likelihood ratio of an alignment column $\sigma$ of $N$ sequences given evolutionary models of the motif and background $\theta_B$ and $\theta_T$, respectively. Under the assumption that the columns of the motif model are independent, the log-likelihood scores are additive; hence, the scoring function generalizes easily to the score for an alignment of length $L$ by summing the scores of the individual columns. When aligned with the $j$th position in the motif, the log-likelihood score for this column is written as

$$S(t) = \log \frac{P(\sigma \mid \theta_M, T)}{P(\sigma \mid \theta_B, T)}$$

where $\theta_M$ refers to the motif model for the $j$th position in the motif and $T$ is a phylogenetic tree containing the evolutionary distances between the species. The two models $\theta_B$ and $\theta_T$ incorporate the frequencies in the position-specific probability matrix of the motif, $M$, the background frequencies of the residues, $B$, different substitution rates for the two models $K_B$ and $K_T$, respectively, and an evolutionary model for calculating the substitution probabilities. We use the HKY (Hasegawa et al., 1985) model with the Halpern–Bruno modification (Halpern and Bruno, 1998) (HKY+HB) with all PMM scanning algorithms in this study.

The PMM scanning algorithms that we test differ primarily in how they handle gaps. The Morthp algorithm simply ignores regions with gaps. The MONKEY (Moses et al., 2004) and rMonkey (Moses et al., 2006) algorithms treat gaps as mistakes in the multiple alignment. MONKEY removes gaps locally, creating an ungapped local alignment for each position in the reference species. Each of these re-alignments can only use positions that are close to each other in the original multiple alignment. The rMonkey algorithm uses a slightly different heuristic for fixing gaps. Using the PWM model, it identifies the highest scoring site in the reference sequence, and then realigns it (without gaps) to the other sequences. Only regions that overlap the reference match by at least one base pair in the original multiple alignment are considered in the realignment. This process is repeated for the remaining binding-site free intervals in the reference sequence until no single species match remains that passes a significance threshold.

2.2 Measuring prediction accuracy

We measure the accuracy of TFBS prediction methods using false discovery rate (FDR), which expresses the proportion of the predictions that are false positives. We report accuracy as either FDR at a given sensitivity level or at a given number of predictions. In general, we are interested in knowing what the FDR is when we choose a particular score as a threshold for deciding which positions in the genome are predicted to be binding sites. Without loss of generality, assume that a prediction algorithm assigns a score, $s$, to a position in a genome, and that larger values of $s$ indicate higher confidence that the position is a binding site.

When we evaluate prediction accuracy using a gold standard set of known sites, we define the FDR for each score threshold $s$ such that there is at least one position with score $s$ or larger as

$$\text{FDR}(s) = \frac{\sum_{|s| \geq r} T}{\sum_{|s| \geq r} T + \sum_{|s| < r} N}$$

where $\sum_{|s| \geq r} T$ is the total number of positions in the genome whose score $s$, is at least $r$, and $\sum_{|s| < r} N$ is the number of ‘negatives’ (non-binding positions) in the genome with score at least $r$. 

![Page i340](https://example.com/image)
When we evaluate prediction accuracy without reference to a set of known sites, we replace the numerator in Equation (2) with the expected number of negative predictions with score \( r \) or greater. This is simply \( np \), where \( n \) is the total number positions scored, and \( p \) is the FDR value of the score threshold.

We estimate the FDR of score \( r \) using the empirical null score distribution we construct as described below in Section 2.4. This gives an estimated FDR of

\[
\text{FDR}(r) = \frac{np}{\sum_{s \geq r} p(s)} \tag{3}
\]

We also make use of the \( q \)-value, a quantity closely related to FDR. The \( q \)-value is defined as the minimum FDR at which a score is deemed significant (Storey, 2002). Assuming, as before, that higher scores are more significant, the \( q \)-value of score \( r \) is therefore

\[
q_r = \text{argmin}_{s \geq r} \text{FDR}(r). \tag{4}
\]

2.3 Estimating prediction accuracy using known sites

The most straightforward way to compare binding site prediction algorithms is by estimating their ability to discriminate a known set of binding sites from non-binding sites. The *Saccharomyces cerevisiae* Database (SCPDB) (Zhu and Zhang, 1999) contains sets of known binding sites in the *S.cerevisiae* genome for a number of TFs. We use only the 21 TFs whose sets contain at least five known sites. We build motifs for each of these TFs to measure the accuracy (FDR) of the scanning algorithms in a cross-validation setup. Using the set of sites for a single TF, we build a set of motifs in which, for each motif, exactly one of the sites is removed. Using each motif independently, we scan multiple alignments of all intergenic regions of four species of yeast, ignoring non-binding sites. The Kalos et al. (2001) approach allows us to estimate the null score distribution of the real motif, as described above.

We use the set of 100 shuffled motifs as input to the prediction algorithm to scan the yeast intergenic region multiple alignments. We also scan using the real motif. Using the scan results, we filter out all candidate control motifs that do not yield a number of ‘hits’ within \( \pm 20\% \) of the real motif, where a hit is defined as a score with \( p \)-value at most 0.001.

2.4 Estimating prediction accuracy using column-shuffled motifs

An alternative way to estimate the prediction accuracy of a PWM or PMM scanning algorithm, that does not require a set of known sites, is to estimate the score distribution of the ‘negatives’ (non-binding sites). Using this null score distribution, we can compute the \( p \)-value of any score threshold \( r \),

\[
p_h(r) = P(r \geq t),
\]

which we then use to compute the FDR at that score threshold [Equation (3)].

To estimate the null distribution of scores for a given motif, we extend an approach that was developed for measuring the prediction accuracy of regular expression motifs (Kheradpour et al., 2007). The basic idea is to use a set of ‘random’ (control) motifs as input to the scanning algorithm and count the number of times each score occurs. The estimate of the probability of each score is then gotten by normalizing these counts so that the total probability sums to one. Scanning the probabilities of all scores greater than or equal to \( t \) gives the \( p \)-value, \( p_h(r) \).

The challenge in this approach is to create control motifs whose score distributions are similar to the true null distribution we wish to estimate. In a nutshell, we create a set of 20 control motifs by first creating 100 column-shuffled versions of the original motif, and then selecting 20 motifs that are least similar to known yeast TF motifs, and are as dissimilar from each other as possible. In selecting the motifs, we also enforce a few additional constraints, as discussed below.

The process begins with the generation of 100 candidate control motifs by shuffling the columns of the target motif. In generating our shuffled motifs, we use the additional constraint that columns within the motif can only be exchanged if the information content of the two columns differs by less than a certain threshold, which we set at 0.4, exactly one-fifth of the maximal information content of a column.

We use the set of 100 shuffled motifs as input to the prediction algorithm to scan the yeast intergenic region multiple alignments. We also scan using the real motif. Using the scan results, we filter out all candidate control motifs that do not yield a number of ‘hits’ within \( \pm 20\% \) of the real motif, where a hit is defined as a score with \( p \)-value at most 0.001.

2.5 Theoretical accuracy of PMM scanning allowing site loss

In order to quantify the predictive power of PMM scanning algorithms, we previously developed a method to estimate the distributions of PMM scores of binding sites and non-binding sites (Hawkins and Bailey, 2008). This approach allows us to estimate the probability of any score, \( s \), under either the motif model.

\[
P_{PB}(s) = P(s | \phi \in \Phi_B, T).
\]

or under the background model,

\[
P_{PB}(s) = P(s | \phi \in \Phi_B, T).
\]

We can thus give estimates of true positive rates and false positive rates for any PMM score.

Our previous work was based on the simplifying assumption that binding sites are not lost during evolution. Since it is well known that binding sites are frequently lost in one or more lineages (Doniger and Fay, 2007), we now present a method for calculating PMM score distributions under a theoretical model that allows sites to be lost independently in any lineage. Note that we do not permit a site to be regained once it has been lost.

We assume that the instantaneous probability of the loss of a site is independent of how long it has been conserved and does not change over time. Hence, the appropriate probability density function is a continuous exponential distribution of the form \( f(t) = \lambda e^{-\lambda t} \) where \( \lambda \) is the single parameter that governs the rate of loss. To compute the cumulative probability that a given site is lost in time \( t \), we integrate \( f(t) \) over the interval \([0, t]\) that gives the cumulative probability distribution

\[
F(t) = 1 - e^{-\lambda t}.
\]

Once a site loss occurs, the evolutionary model switches permanently from the site model (\( \Phi_B \)) to the background model (\( \Phi_B \)).

For our theoretical models, we use a phylogenetic tree with star topology and equal branch lengths. D. Following Eddy (2005), we place the target (first) genome in the center of the star. This placement allows a dynamic programming solution that computes the score probability distribution in

\[
\text{Phylogenetic motif models for TFBS prediction}
\]
Two assumptions of independence simplify the process of calculating the probability distributions required in computing the distribution of the log-likelihood scores, \( \xi \). First, the assumption of a phylogenetic star with the target genome in the center means that each genome evolves from the target independently; hence, the probability of \( N \) genomes is the probability of the first \( N-1 \) genomes times the probability of seeing the \( N \)-th genome. Second, the assumption of independence between the positions within the motif implies that the probability distribution for the score considering only the \( m \)-th column.

These assumptions allow us to apply dynamic programming to calculate a discretized approximation to the score probability distributions (Staden, 1990). We calculate the distribution under both the assumption that we are dealing with a conserved motif, and under the assumption that we are dealing with a neutral sequence. We are then able to generate the cumulative distributions under each model and determine the false positive and false negative rates at each possible score.

We use the HKY (Hasegawa et al., 1985) substitution model to calculate the substitution probabilities for both the background and the motif evolutionary models. (Our analysis allows any of the standard substitution models, and our implementation incorporates the Jukes-Cantor, Kimura 2-parameter, F81, F84, HKY and Tamura-Nei models.) We use the parameter settings of the HKY model employed in MONKEY (Moses background distribution, \( B \), the first \( N \) target genome in the center means that each genome evolves from the target independently; hence, the probability of \( N \) genomes is the probability of the first \( N-1 \) genomes times the probability of seeing the \( N \)-th genome. Second, the assumption of independence between the positions within the motif implies that the probability distribution for the score considering only the \( m \)-th column.

We validated the effectiveness of this approach by empirically generating multiple alignments in which motif sites were lost with a frequency determined by our exponential loss function. The numerical approximations produced probabilities correct to three decimal places.

### 2.6 Estimating the number of TFBs in a genome

One way to estimate the number of binding sites in a genome is to choose the smallest number of sites such that our theoretical model of PWM score distributions predicts a higher (better) sensitivity (number of predictions) at all \( q \)-values than we observe when measuring empirical accuracy using column-shuffled motifs. This is motivated by the assumption that the theoretical model provides an (estimated) upper bound on prediction accuracy, so the empirical accuracy estimates should not exceed the theoretical accuracy estimates. As described in Section 3.4, we find that the requirement that the theoretical sensitivity be higher at all \( q \)-values is too strict. Therefore, we instead estimate the smallest number of sites such that the requirement is met on the \( q \)-value interval \([0, Q]\). We study the behavior of this estimate as we vary \( Q \).

To compute our estimate of the true number of binding sites, \( b \), for a given \( T \), first we use the shuffled-motif method to compute empirical estimates of, for each observed score, \( \xi \) the \( q \)-value of the score, \( Q_T(\xi) \) and the observed number of predictions with that score or better, \( C_T(x) \). We then use our theoretical model of PWM score distributions to compute the equivalent (theoretical) values \( Q_T(x) \) and \( C_T(x) \), each of which depends on an assumed number of true binding sites, \( b \). To describe our requirement that the theoretical accuracy always be better than the empirical at all \( q \)-values, we need to define functions that map \( q \)-values to scores. Because more than one score may have the same \( q \)-value, we define the inverse functions as the minimum score with a given \( q \)-value,

\[
S_T(q) = \arg \min_{q < Q_T} q \\
S_T(q, b) = \arg \min_{q < Q_T(b)} q
\]

For each value of \( Q \), \( \hat{b}(Q) \) gives an estimate of the true number of sites.

In order to get our theoretical estimates of the \( q \)-values and numbers of predictions at different scores, we proceed as follows. Given that there are \( b \) true binding sites and \( n-b \) non-binding sites, we expect that there will be \( TP_T(x, b) = bP_T(x) \) true positive predictions and \( FP_T(x, b) = (n-b)P_T(x) \) false positive predictions at a score threshold of \( x \), where \( P_T(x) \) and \( P_T(x) \) are as defined in Equations (5) and (6), respectively. The expected theoretical FDR
of score $s$ is the proportion of false positive predictions:
\[
\text{FDR}(s,b) = \frac{FP(s,b)}{FP(s,b) + TP(s,b)} = \frac{n - b \hat{P}(s)}{n - b \hat{P}(s) + \hat{P}(s)}
\]
Plugging this estimate of FDR into Equation (4) gives us $Q_s(s,b)$, our theoretical score to $q$-value mapping function. The theoretical score to number of predictions map is just the sum of the numbers of true and false positives, $\sum_j TP(s_j,b_j) + FP(s_j,b_j)$.

For comparison, we apply a previously described statistical technique which assumes that the observed score distribution is a mixture of (null distribution) and the other characterizes the background sites (alternative distribution), and the other characterizes the background sites (null distribution). The values $\tau_0$ and $\tau_1$ are defined as the fraction of scores drawn according to the null and alternative distributions, respectively. To estimate these two values, we use an implementation of the bootstrapping method described by Storey (2002). This method selects an optimal $P$-value threshold, $\hat{p}$, and then estimates $\tau_0$ as
\[
\tau_0(\hat{p}) = \frac{\hat{p} > \hat{p}}{\hat{p}}
\]
where $n$ is the total number of $P$-values. Using this approach, the estimate of the number of real sites is
\[
\hat{n}(\hat{p}) = n \tau_0(\hat{p})
\]

3 RESULTS AND DISCUSSION

3.1 PMM scanners perform worse than a simple PWM scan at predicting known TFBSs

We first show that using a gold standard of known yeast TFBSs in a cross-validated experiment indicates that there is no advantage in prediction accuracy with PMM scanners compared with a standard PWM scan. Actually, the opposite appears true, as shown in Table 1. At a sensitivity level of 50%—when half of the known binding sites are detected—the simple PWM scan has better accuracy (lower cross-validated FDR) than all three PMM scanners tested for 14 out of the 21 TF motifs. The apparent superiority of simple PWM scan persistence over all sensitivity levels (data not shown).

The differences in accuracy are often substantial. For one TF (PDR1), the PWM scan of just the *S.cerevisiae* genome achieves a 0% FDR, while the three PMM algorithms, which scan multiple alignments of four yeast genomes, have false discovery rates of at least 55% and up to 99%. In the few cases where the simple PWM scan does not outperform the other methods, its FDR is often quite close to that of the other algorithms. The one exception is the TF ROX1, where the phylogenetic algorithms achieve much lower FDR than the single sequence scan (20% versus 87%). In general, however, this experiment fails to demonstrate any substantial benefit to using phylogenetic motif scanning.

Several explanations are possible for the apparent failure of the phylogenetic motif scanners in this test. Errors in the multiple alignments of the intergenic regions could be causing conserved sites to receive poor scores. This explanation is supported by the observation that the rMonkey algorithm, which performs extensive re-alignment of the sequences during scanning, does slightly better than Motiph, which simply ignores gapped regions. Re-alignment strategies used by Monkey and rMonkey can also recover sites to receive poor scores. This explanation is supported by the alignments of the intergenic regions could be causing conserved phylogenetic motif scanners in this test. Errors in the multiple alignments of four yeast genomes, have false discovery rates of at least 55% and up to 99%. In the few cases where the simple PWM scan does not outperform the other methods, its FDR is often quite close to that of the other algorithms. The one exception is the TF ROX1, where the phylogenetic algorithms achieve much lower FDR than the single sequence scan (20% versus 87%). In general, however, this experiment fails to demonstrate any substantial benefit to using phylogenetic motif scanning.

Several explanations are possible for the apparent failure of the phylogenetic motif scanners in this test. Errors in the multiple alignments of the intergenic regions could be causing conserved sites to receive poor scores. This explanation is supported by the observation that the rMonkey algorithm, which performs extensive re-alignment of the sequences during scanning, does slightly better than Motiph, which simply ignores gapped regions. Re-alignment strategies used by Monkey and rMonkey can also recover sites that have ‘drifted’–the original site is not conserved but a new site has emerged nearby. Both algorithms restrict re-alignments to very narrow windows, however, and neither can correct for sites that have drifted to different stripe. These limitations may partially explain the fact that Monkey and rMonkey do not perform substantially better than Motiph, which has no re-alignment strategy. (Careful examination of Table 1 shows that Motiph has lower FDR than both Monkey and rMonkey in 8 out of 21 cases.)

An alternative explanation for the apparent failure of the phylogenetic motif scanners is that errors within the gold standard are systematically biased against phylogenetic-based scanning. We expect that the gold standard is missing some, and possibly many, functional sites. These missing sites will be flagged as false positives, and can cause an algorithm that correctly detects them to have a larger (apparent) FDR. The sites missing from the gold standard are likely to be minor variants of the motif with slightly lower binding affinity, such that the experimental methods for identifying protein to DNA binding sites overlook them. If these sites are nevertheless highly conserved, then the phylogenetic scanners will score them much higher and appear to have a larger number of false positive predictions. This effect may be contributing to the apparent failure of the phylogenetic motif scanners to improve upon single sequence scanning. We provide evidence for this hypothesis in the next section.
3.2 PMM scanners are more accurate than PWM scanners

When we measure prediction accuracy without reference to gold standard sets of known sites, it becomes clear that PMM scanners are more accurate than simple PWM scans. The \( q \)-value (minimum FDR) of each of the scanners, estimated using the column-shuffled motif approach, is shown for the yeast TFBs in Table 2. The table gives the prediction accuracy at different levels of sensitivity—20 and 50 total predictions. When making 50 binding site predictions, PMM scans have lower minimum FDR than PWM scans with 16 of the 21 yeast TF motifs used in this study. When 20 binding sites are predicted, PMM scans have lower minimum FDR for 17 out of 21 motifs. The improvement over the simple PWM scan of each of the PMM scanners is statistically significant according to a signed rank test (Motiph, \( p = 0.013 \); Monkey, \( p = 0.001 \); and rMonkey, \( p = 0.008 \)).

The systematic improvement of all phylogenetic motif scanners over the PWM scanner strongly supports the hypothesis that the gold standard data are systematically biased against the phylogenetic motif scanners. Therefore, these results suggest that not only do all of the models perform better than the gold standard indicates, but in many cases the PMMs provide a considerable advantage over PWM scanning.

The results in Table 2 also indicate that the different realignment strategies used by the three PMM scanners make a large difference. The approach used by the Monkey algorithm works best with the 21 yeast TF motifs studied here. Monkey has the best accuracy among all the prediction algorithms for 11 out of the 21 motifs at a sensitivity level of 20 predictions, and for 10 motifs at a sensitivity level of 50 predictions.

3.3 Incorporating site loss improves our theoretical estimate of the statistical power of PMM scanners

The extension of our earlier model of the statistical power of PMM scanners to allow loss-of-site events causes it to agree more closely with the observed power of PMM scanners on the yeast gold standard TFBS sets. We evaluate the accuracy of the theoretical model by comparing the ROC curves (Swets, 1988) it predicts with the ROC curves generated by our cross-validation experiments with the 21 yeast TF motifs. An example of the behavior of the theoretical model under five different assumptions of the probability of site loss events, ranging from no loss to up to 51% chance of loss in any lineage, is shown in Figure 1. Assuming 51% chance of loss shifts the theoretical ROC curve about an order of magnitude closer to the observed ROC curve. This plot is typical of the plots for all 21 TF motifs (data not shown).

The extended theoretical model also agrees well with empirical estimates of statistical power based on the shuffled-motif approach. Figure 2 shows that, if we assume that there are 51 real ABF1 binding cites in the yeast intergenic regions, the theoretical estimate of the statistical power of PMM scanning fits the empirical results using...
This value acts as a reasonable upper bound on the expected amount lost on an individual branch of the tree. We observe that, for all (Tuch loss model to emulate the most pessimistic estimate in the literature allowing binding site loss, we have calibrated the parameters of the theoretical model with and without binding site loss. In the plot we report the mean and SD of the estimate of true binding sites. The complete list of \( \hat{b} \)-values estimated using the shuffled-motif approach, uses \( P \)-values that do not belong to the null distribution, hence the \( q \)-value threshold, \( Q \). The alternative method, called bootstrapping, uses \( P \)-values for all positions scanned within the genome is used to estimate the proportion, \( \pi \), of the \( P \)-values that do not belong to the null distribution, hence the number of real binding sites (Storey, 2002) (Section 2.6). We ran the bootstrapping method, with 10,000 bootstraps, 100 times, and we report the mean and SD of the estimate of true binding sites.

The plot in Figure 3 illustrates the results of our model-based method with a typical example of the relationship between the \( q \)-value threshold, \( Q \), and number of predicted real sites, \( b(\hat{Q}) \), for the motif ABF1. We include in the plot the results for the theoretical model with and without binding site loss. In the plot allowing binding site loss, we have calibrated the parameters of the loss model to emulate the most pessimistic estimate in the literature (Tuch et al., 2008), such that a site has 51% probability of being lost on an individual branch of the tree. We observe that, for all motifs, the estimated number of sites undergoes a phase transition at high \( q \)-values. Up to some value of \( Q \), the no-loss estimate tends upwards on a slight slope. However, when we use the loss model, the estimate remains much flatter throughout the first two-thirds of the plot. We also observe that, for all motifs, the estimated number

### Table 3. Estimated number of real binding sites for 21 yeast TFs

| TF Name          | SCPD          | Model-based | Bootstrapping |
|------------------|---------------|-------------|---------------|
|                  | No. of sites  | No. of sites| No. of sites  |
|                  | (SD)          | (SD)        | (SD)          |
| ABF1             | 16            | 51          | 175 (62)      |
| BAS1             | 5             | 15          | 117 (0)       |
| GAL4             | 10            | 9           | 49770 (556)   |
| GCN4             | 10            | 47          | 62587 (1957)  |
| HAPI             | 5             | 44          | 56753 (2892)  |
| HSEJHSTF         | 7             | 34          | 131 (0)       |
| MATalpha2        | 12            | 20          | 0 (0)         |
| MCM1             | 34            | 40          | 0 (0)         |
| MIG1             | 9             | 37          | 739 (23)      |
| PDR1             | 11            | 4           | 36 (0)        |
| PHO4             | 8             | 36          | 1091 (82)     |
| RAPI             | 16            | 24          | 42434 (1967)  |
| REB1             | 17            | 105         | 59299 (709)   |
| repressor_of_CAR1| 12            | 42          | 51189 (990)   |
| ROXI             | 8             | 18          | 46 (0)        |
| SW5              | 6             | 5           | 133 (0)       |
| TRP              | 10            | 118         | 427 (8)       |
| UASH             | 14            | 14          | 145 (31)      |
| UASPHR           | 15            | 25          | 93 (9)        |
| UAS              | 5             | 21          | 1144 (127)    |
| URSHH            | 13            | 48          | 176 (46)      |

The estimated number of real binding sites compared with the numbers present in SCPD. Two methods are used to produce the estimates: the bootstrapping method (Storey, 2002) and our model-based method that chooses the number of sites that insures that the theoretical statistical power of the scan is an upper bound on the empirical power. For our model-based method, we report (0.5).
The flatness of the plot using the site loss model demonstrates that measuring the accuracy of binding site evolution that includes loss-of-site events. This model provides a constant estimate of the number of binding sites, indicating that it is a better model of binding site evolution.

We have demonstrated that the PWM suggests that the results are not an aberration. Among the true set of binding sites. Nevertheless, the observation that all PMMs reported in this work are a perfect reflection of the prediction on the PWM suggests that the results are not an aberration. Among the three scanning algorithms that we test, the MONKEY algorithm has the highest accuracy for predicting yeast TFBSs.

27 July 2018

ACKNOWLEDGEMENT

The authors wish to acknowledge the ARC Centre of Excellence in Bioinformatics.

Funding: Australian Research Council (grant DP0770471) (to J.H.).

NIH (grant R01 RR021692) (to T.L.B. and W.S.N.).

Conflict of Interest: none declared.

REFERENCES

Berman,B.P. et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl Acad. Sci. USA, 99, 757–762.

Blanchette,M. et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res., 14, 708–715.

Bolletli,D. et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science, 299, 1391–1394.

Bromer,A.R. et al. (2007) Divergence of transcription factor binding sites across related yeast species. Science, 317, 815–819.

Bruno,M. et al. (2003) Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res., 13, 721–731.

Chen,E. et al. (2003) Multiple sequence alignment with the chalais of series programs. Nucleic Acids Res., 31, 3497–3506.

Doniger,S.W. and Foy,J.C. (2007) Frequent gain and loss of functional transcription factor binding sites. PLoS Comput. Biol., e99.

Eddy,S.R. (2005) A model of the statistical power of comparative genome sequence analysis. PLoS Biol., e10.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol., 17, 368–376.

GertJ. (2008) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature, 457, 215–218.

Guccione,E. (2006) Myc-binding-site recognition in the human genome is determined by chromatin context. Nat. Cell Biol., 8, 764–7625.

GulahThakurta,D. (2006) Computational identification of transcriptional regulatory elements in DNA sequences. Nucleic Acids Res., 34, 3085–3096.

Gummeci,D.L. et al. (1992) Phylogenetic footprinting reveals a nuclear protein which binds to silenced sequences in the human gamma and epsilon globin genes. Mol. Cell Biol., 12, 4910–4920.

Halpern,A.L. and Bruno,W.J. (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol., 15, 910–917.

Hauzenew,M. et al. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol., 22, 160–174.

Halpern,A. et al. (2007) Multiple sequence alignment with the clustal series of programs. Nucleic Acids Res., 31, 3497–3506.

Katoh,K. et al. (2002) Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res., 30, 3859–3868.

KoSi,M. et al. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature, 423, 241–254.

Kheradpour,P. et al. (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res., 17, 1919–1931.

Kouzarides,T. (2007) Chromatin modifications and their function. Cell, 128, 693–705.

Levine,M. and Tjian,R. (2002) Transcription regulation and animal diversity. Nature, 424, 147–151.

Liu,X. (2006) Whole-genome comparison of Lox5 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. Genome Res., 16, 1517–1528.
Moses, A. M. et al. (2004) MONKEY: identifying conserved transcription factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, 5, R98.

Moses, A. M. et al. (2006) Large-scale turnover of functional transcription factor binding sites in drosophila. *PLoS Comput. Biol.*, 2, e130.

Narlikar, L. et al. (2007) A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.*, 3, e235.

Staden, R. (1990) Searching for patterns in protein and nucleic acid sequences. *Methods Enzymol.*, 183, 193–211.

Storey, J. D. (2002) A direct approach to false discovery rates.

Stormo, G. D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, 16, 16–23.

Swets, J. A. (1988) Measuring the accuracy of diagnostic systems. *Science*, 240, 1285–1293.

Tuch, B. B. et al. (2008) The evolution of combinatorial gene regulation in fungi. *PLoS Biol.*, 6, e238.

Wasserman, W. W. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5, 276–287.

Whittington, F. et al. (2008) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.*, 37, 14–25.

Zhu, J. and Zhang, M. Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, 15, 607–611.