BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 5 2014, pages 636–643
doi:10.1093/bioinformatics/btt585

Sequence analysis

Advance Access publication October 9, 2013

A penalized Bayesian approach to predicting sparse protein–DNA binding landscapes

Matthew Levinson and Qing Zhou*

Department of Statistics, University of California, Los Angeles, CA 90095, USA

Associate Editor: John Hancock

ABSTRACT

Motivation: Cellular processes are controlled, directly or indirectly, by the binding of hundreds of different DNA binding factors (DBFs) to the genome. One key to deeper understanding of the cell is discovering where, when and how strongly these DBFs bind to the DNA sequence. Direct measurement of DBF binding sites (BSs; e.g. through ChIP-Chip or ChIP-Seq experiments) is expensive, noisy and not available for every DBF in every cell type. Naive and most existing computational approaches to detecting which DBFs bind in a set of genomic regions of interest often perform poorly, due to the high false discovery rates and restrictive requirements for prior knowledge.

Results: We develop SparScape, a penalized Bayesian method for identifying DBFs active in the considered regions and predicting a joint probabilistic binding landscape. Using a sparsity-inducing penalization, SparScape is able to select a small subset of DBFs with enriched BSs in a set of DNA sequences from a much larger candidate set. This substantially reduces the false positives in prediction of BSs. Analysis of ChIP-Seq data in mouse embryonic stem cells and simulated data show that SparScape dramatically outperforms the naive motif scanning method and the comparable computational approaches in terms of DBF identification and BS prediction.

Availability and implementation: SparScape is implemented in C++ with OpenMP (optional at compilation) and is freely available at 'www.stat.ucla.edu/~zhou/Software.html' for academic use.

Contact: zhou@stat.ucla.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 21, 2013; revised on September 18, 2013; accepted on October 4, 2013

1 INTRODUCTION

Many complex processes in the cell, particularly gene regulation, are controlled by the binding of various factors to the DNA sequence. A key to understanding these processes is determining where each of these DNA binding factors (DBFs), including transcription factors (TFs), nucleosomes, RNA and other proteins and protein complexes, binds in the genome in a certain cell type and set of conditions. This collection of binding sites (BSs) for all DBFs over regions of interest is sometimes called a binding landscape. More formally, we define a binding landscape as the base pair–specific probability of binding for each of a library of DBFs over a set of genomic regions.

Considering DBFs one at a time leads to many false positives, both in determining which DBFs have significantly enriched BSs in a set of genomic regions and in predicting the exact locations of BSs, and results in a limited view of the processes controlled by these DBFs. This has motivated recent work on jointly predicting binding landscapes for a set of DBFs. Currently, joint landscapes at single base pair resolution for all DBFs have only been predicted in lower eukaryotes such as yeast with fewer DBFs and much smaller genomes (Wasson and Hartemink, 2009). In higher eukaryotes, predictions have been limited to a (usually small) pre-selected set of DBFs known to bind the regions of interest (He *et al.*, 2009, 2010; Kaplan *et al.*, 2011; Laurila *et al.*, 2009; Raveh-Sadka *et al.*, 2009). Of these methods, only that of He *et al.* (2009) does not require the DBF concentrations as prior knowledge, something made possible by considering at most two DBFs at a time, and only Kaplan *et al.* (2011) used ChIP-Seq data as a source of direct information. See Arnold *et al.* (2011), Ernst *et al.* (2010), Marbach *et al.* (2012), Ramsey *et al.* (2010), Teif and Rippe (2010) and Won *et al.* (2010) for recent examples of alternate approaches to answering related questions.

Owing to computational limits, it is impossible to predict a joint base pair–specific binding landscape for all DBFs with unknown concentrations over the entire genome in higher eukaryotes with large genomes and many DBFs. We are thus limited to exploring a subset of the genome. One motivating type of genomic subset is a set of regions known to be co-bound by a small group of DBFs based on ChIP-Seq data. In such a genomic subset, we do not expect most DBFs to have a significant number of BSs. Thus, the false positive BSs in a predicted binding landscape can be substantially reduced if only the DBFs with significantly enriched binding in the regions of interest are considered. However, it is limiting to require the complete set of DBFs enriched in the considered regions to be known *a priori* as is done in existing work on similar questions in higher eukaryotes.

In this article, we develop a method that offers a principled way to select an, often small, subset of DBFs active in the regions of interest and to reduce the false-positive signal in the predicted probabilistic binding landscape, eliminating the need for prior knowledge of the set of enriched DBFs or DBF concentrations. In the motivating genomic subset, our method allows for the discovery of unknown cofactors that commonly bind near the DBFs with ChIP data (ChIP DBFs). The predicted joint binding landscape provides a global and quantitative view of the binding pattern among the DBFs. This is an initial step to the study of combinatorial regulatory logic among multiple DBFs.

*To whom correspondence should be addressed.

2 MODEL AND ESTIMATION

2.1 Overview of SparScape

Our method, SparScape, proceeds in two stages. First, from a candidate set we select the DBFs with significant binding in the considered regions. Second, we do a refined prediction of the binding landscape considering only the selected DBFs. See Supplementary Figure S1 for a schematic illustration of the method.

SparScape takes as input the previously estimated binding motifs for a set of candidate DBFs, the sequence of a set of genomic regions of interest and genome-wide binding data (e.g. ChIP-Seq) for any of the candidate DBFs if available. The set of regions could be the whole genome when examining small genomes. We consider nucleosome binding because nucleosome occupancy blocks the binding of many other DBFs, and recent studies have demonstrated the utility of nucleosome models in protein binding landscapes (Kaplan *et al*., 2009; Raveh-Sadka *et al*., 2009; Wasson and Hartemink, 2009). We model nucleosome binding preferences by a position-specific Markov model proposed by Kaplan *et al*. (2009). Details are provided in Supplementary Methods Section S1.1. The binding of non-nucleosome DBFs is modeled by position-specific weight matrices (PWMs). The background is modeled by a fifth-order Markov chain estimated from a large set of sampled or simulated regions similar to the regions of interest.

We model ChIP data as a set of binary windows, where a window of base pairs around the center of a ChIP peak is called a ChIP window for that DBF. In this work, we use ChIP windows of 50 bp on either side of a ChIP-Seq peak. Our method estimates the probability that a BS for a DBF with ChIP data is entirely within one of its own ChIP windows. This is in the same spirit as the use of DNase I sensitivity measures to create an informative prior distribution for TF binding (Kaplan *et al*., 2011; Narlikar *et al*., 2007), though SparScape can exploit binding data for any of the candidate DBFs or the nucleosome. Moreover, ChIP windows are included as part of a generative probabilistic model with parameters related to the accuracy and sensitivity of ChIP peaks, resulting in a more principled and flexible utilization of ChIP data.

Estimating concentrations is a unique feature of SparScape. This makes it impossible to exactly calculate the binding landscape through forward–backward summation. Instead, we explore the posterior distribution through a penalized iterative sampling approach, simultaneously selecting DBFs, estimating model parameters and predicting the binding landscape. Note that concentration here is not related to the physical concentration of the DBF, but rather summarizes the enrichment of BSs in the considered regions and gives the probability of a BS beginning at any location in the sequence. One could use physical concentration measures such as gene expression to build an informative prior on the concentration vector, but we have not explored this possibility.

Jointly estimating the concentrations introduces a risk of excessive false positives, especially for DBFs with less informative motifs where we expect non-functional matches to occur frequently in the genome. To avoid this, we use a penalty on the predicted site counts in each iteration, penalizing in proportion to the expected number of false-positive sites estimated with control regions given the current parameter values. Intuitively, this removes the expected false-positive sites from the sampled sites. The level of penalization is controlled by a tuning parameter, chosen in the first stage through 10-fold cross-validation. With a proper level of penalization, concentrations of many DBFs will be estimated as exactly zero, achieving the goal of DBF selection. The final binding landscape, considering only the selected DBFs, is predicted in the second stage.

When selection and prediction are completed, SparScape reports a binding landscape, which gives the probability of binding at each base pair by each selected DBF, and the estimated parameter values, including the concentrations for the DBFs. It also reports the binding configuration and the parameter values sampled at each iteration, allowing, for example, construction of credible intervals for the parameters and examination of high-order interactions between binding at different sites.

2.2 The SparScape model

Consider the sequence S of a set of genomic regions with total length $|S|$, and the set of ChIP windows D in these regions for all ChIP DBFs. Let K be the number of candidate DBFs, and Θ denote the set of binding model parameters for all K DBFs, including the nucleosome, and the background model. Under the standard steric hindrance constraint, we define a binding configuration as a partition of the sequence S into unbound background sites and BSs for the K DBFs. Denote a configuration by $A = (a_1, a_2, \ldots, a_{|A|})$, where a_i is the index of one of the $K+1$ models and represents a subsequence of base pairs bound by a DBF (or is a single unbound base pair) in the current configuration. More specifically, it represents a single unbound site covering $L_0 = 1$ bp when $a_i = 0$, a nucleosome covering $L_1 = 147$ bp when $a_i = 1$, and a non-nucleosome DBF from the candidate library covering L_k bp when $a_i = k \in \{2, \ldots, K\}$, where L_k is the length of the motif for the k^{th} DBF. Figure 1a illustrates an example configuration.

Let ϕ be the probability that a BS for a ChIP DBF is entirely within one of its ChIP windows and γ be the probability that an unbound base pair is not covered by any of the ChIP windows. Write $\Phi = (\phi, \gamma)$. Denote by $T = (\tau_0, \ldots, \tau_K)$ the probabilities of initiating a background site ($k = 0$) and other DBF sites ($k = 1, \ldots, K$) at a given location $\left(\sum_k \tau_k = 1, \tau_k \geq 0\right)$. This can be thought of as a vector of local concentrations (local to the regions considered).

We wish to jointly estimate the binding configuration A, the concentrations T and the ChIP parameters Φ. Given binding configuration A, we consider the sequence S and the ChIP windows D as independent sources of information. We further assume independent priors on T and Φ. Under these model assumptions, the joint posterior distribution is

$$P(\Phi, T, A|S, D, \Theta) \propto P(S, D, A|\Theta, \Phi, T)\pi(\Phi, T)$$

$$= P(S|A, \Theta)P(D|A, \Phi)P(A|T)\pi(\Phi)\pi(T), \quad (1)$$

where $P(S, D, A|\Theta, \Phi, T)$ is the complete-data likelihood, regarding A as missing data. The posterior distribution of A gives the predicted binding landscape. Our primary goal is to select the DBFs with motif models in Θ that have a significant number of BSs in S and predict the likely binding configurations.

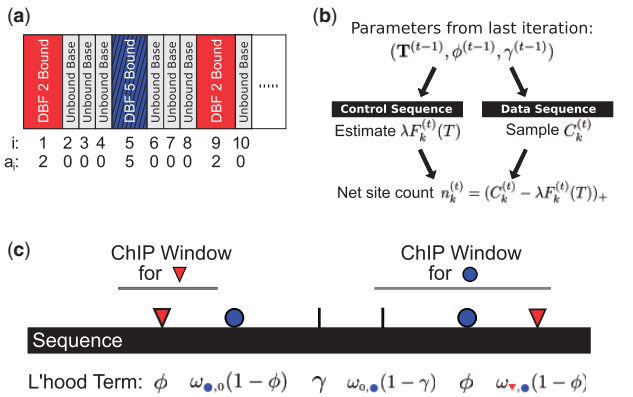

Fig. 1. Elements of the SparScape model. (**a**) Vector A for a particular binding configuration. (**b**) Procedure for calculating the net site counts and penalties. (**c**) Terms contributed to the likelihood by unbound base pairs and binding of DBFs in different types of windows. The triangle and circle represent BSs for two ChIP DBFs. The vertical bar represents an unbound site or a BS for a DBF with no ChIP data

We achieve this by assigning a statistical weight based on Equation (1) to every binding configuration and searching the space of possible configurations through Monte Carlo sampling. Note that different DBFs can bind to the same base pairs in different configurations, and thus, our sampling approach allows relatively large posterior binding probabilities for multiple DBFs for the same base pair if BSs do overlap.

The first part of the statistical weight is the ratio between the likelihood of the sequence given a configuration A and the likelihood given the null configuration A_0 where no DBFs are bound. Let S_j be the base at position j and $S_{\text{Start}(i):\text{End}(i)}$ be the subsequence in S from the first to last base pairs covered by element a_i in the A. Define the single-element sequence likelihood ratio as follows:

$$H_k\left(S_{\text{Start}(i):\text{End}(i)}\right) = \frac{P\left(S_{\text{Start}(i):\text{End}(i)}|\theta_k\right)}{P\left(S_{\text{Start}(i):\text{End}(i)}|\theta_0\right)} \tag{2}$$

for $a_i = k \in \{1, \ldots, K\}$, where θ_k is the parameter (e.g. PWM) for the kth binding model. See Supplementary Methods Section S1.2 for the detailed formulation. By definition $H_0 \equiv 1$.

The second part of the statistical weight is the ratio between the likelihood of the ChIP windows given a configuration A and the likelihood given the null configuration A_0. The background window is defined as the set of base pairs not covered by any of these windows. Thus, if we have M ChIP DBFs, we will have $M+1$ sets or types of windows, where windows of type 0 are the background windows. Let $d_{\text{Start}(i):\text{End}(i)}$ be the type of the window covering the element a_i. When a_i indicates a ChIP DBF, we define

$$P(d_{\text{Start}(i):\text{End}(i)} = k|a_i = k, \phi) = \phi, \tag{3}$$

$$P(d_{\text{Start}(i):\text{End}(i)} = j|a_i = k, \phi) = \omega_{k,j}(1-\phi), \tag{4}$$

where $\omega_{k,j}$ $(j \neq k)$ is proportional to the total length of all windows of type j and $\sum_{j\neq k}\omega_{k,j} = 1$. For background sites $(a_i = k = 0)$, ϕ is replaced by γ in (3) and (4). The model for a DBF without ChIP-Seq data is identical to that for background sites. See Figure 1c for an illustration and Supplementary Methods Section S1.3 for more technical details. If most of the ChIP DBF BSs are covered by a corresponding ChIP window, the parameter ϕ will be close to one. The value of γ is determined mostly by the percentage of the base pairs not covered by any ChIP windows. We have found that when running SparScape ignoring the ChIP data, the percentage of predicted BSs within what would have been ChIP windows had they been considered tends to be similar across DBFs. Thus, we assume a single parameter ϕ shared among all ChIP DBFs in the current work. See Kaplan *et al.* (2011) and Kharchenko *et al.* (2008) for other models of ChIP data in motif finding.

With all the terms in Equation (1) defined we can compute the full likelihood ratio. Define $B(k,\ell) = P(d_{\ell:\ell'}|a_i = k, \Phi)$ with $\ell' = \ell + L_k - 1$. Then the full likelihood ratio for element $a_i = k$ in configuration A, starting at sequence position ℓ and covered by a window of type $d_{\ell:\ell'}$, is

$$\mathcal{L}(k,\ell) = (\tau_k/\tau_0) \quad H_k(S_{\ell:(\ell+L_k-1)}) \quad B(k,\ell)/B(0,\ell). \tag{5}$$

Note that when $a_i = k = 0$, i.e. the ith element is an unbound base pair, by definition we have $\mathcal{L}(0,\ell) = 1$ for all ℓ. The product of $\mathcal{L}$ over a_i defines the complete-date likelihood ratio,

$$\frac{P(S,D,A|\Theta,\Phi,T)}{P(S,D,A_0|\Theta,\Phi,T)} = \prod_{i=1}^{|A|}\mathcal{L}(a_i, \text{Start}(i)). \tag{6}$$

2.3 Sparsity through penalization

The total number of candidate DBFs K is often large. We seek DBF selection because we expect that BSs for a large majority of candidate DBFs are not enriched in the considered genomic regions. Considering DBFs that are not truly enriched when predicting the final landscape increases false positive predictions, sometimes dramatically.

One way to achieve DBF selection is to estimate many concentrations τ_k as exactly zero, as τ_k is the probability of initiating a BS for DBF k. It can be seen from (6) that the log-likelihood for T given A is $\sum_{k=0}^{K} C_k \log \tau_k$, where C_k is the number of BSs of DBF k in A. If we take the conjugate Dirichlet prior on T with prior counts $\alpha_k > 0$, for $k = 0, \ldots, K$, the conditional posterior distribution for T is a Dirichlet distribution $Dir(C_0 + \alpha_0, \ldots, C_K + \alpha_K)$. A sample from this distribution always has positive components, based on which we cannot construct a sparse estimation of T. Thus, we run our algorithm in the selection stage for a burn-in period with $\alpha_k = 1$, and then set $\alpha_k = 0$. If $C_k = \alpha_k = 0$ at some iteration, then the conditional posterior distribution has a point mass at $\tau_k = 0$. This allows us to achieve sparsity in the sense that some $\tau_k = 0$ with a positive probability. DBFs that hit $\tau_k = 0$ at any sampling iteration are selected out.

Unfortunately, for many DBFs we expect relatively strong non-functional motif matches to occur randomly, leading to a non-negligible number of false positive predicted sites such that τ_k almost never hits zero. We counteract this false positive signal with penalty terms on the parameters T and Φ, leading to a penalized complete-data log-likelihood of the form

$$\log P(S,D,A|\Theta,\Phi,T) - \lambda\sum_{k=2}^{K} F_k \log \tau_k - \rho(\Phi), \tag{7}$$

where $F_k \geq 0$ is the expected count of false positive BSs for DBF k, $\lambda \geq 0$ is a tuning parameter and $\rho(\Phi)$ denotes the penalty for Φ. Given the current concentrations, F_k is estimated empirically from a set of control sequences, possibly simulated, with no (known) true sites (Supplementary Methods Section S1.4). See Figure 1b for a schematic of the penalty count and net site count calculation. We did not penalize the nucleosome concentration (τ_1) in the results presented here, but SparScape supports such penalization.

Together with the prior distribution $\pi(\Phi, T)$, we obtain a penalized posterior distribution. To understand this penalized posterior, consider sampling from the conditional distributions taking the penalties into account. The conditional sampling from $[A|\Phi, T, S, D, \Theta]$ is not affected by the penalization and can be implemented by forward summation and backward sampling (Gupta and Liu, 2003; Zhou and Wong, 2004). See Supplementary Methods Section S1.5 for details.

Consider penalized sampling from $[\Phi, T|A, S, D, \Theta]$. Let $n_k = C_k$ for $k = 0, 1$ and $n_k = (C_k - \lambda F_k)_+$ for $k \geq 2$, where $x_+ = \max(0, x)$. We think of n_k as the net site count after discounting for false positives λF_k. Then the penalized complete-data log-likelihood for the concentrations T is $\sum_{k=0}^{K} n_k \log \tau_k$. If the penalty $\lambda F_k \geq C_k$ for some iteration, then the net site count $n_k = 0$ and the conditional posterior distribution of T (with $\alpha_k = 0$) has a point mass at $\tau_k = 0$. Thereafter, we drop DBF k from further consideration. See Supplementary Methods Section S1.6 for more details about this conditional sampling step. Figure 2 shows penalized and unpenalized sample paths for the concentrations of a true and a false DBF in a simulated dataset, with more examples in Supplementary Figure S2. Figure 2a demonstrates that moderate penalization can eliminate a false-positive DBF even when its concentration would otherwise stabilize around a highly inflated value. While τ_k may hit zero for some false-positive DBFs even with $\lambda = 0$, as shown in these figures, there are many cases where DBF selection is achieved by using a non-zero penalty $\lambda > 0$. Setting the prior counts $\alpha_k = 0$ is mainly for mathematical rigor so that τ_k may hit zero exactly, as any positive value of α_k will never give $\tau_k = 0$ in any sampling iteration. In addition, this choice also avoids the use of a threshold value on the estimated τ_k for DBF selection, which would be necessary if α_k were positive.

Penalization on the parameters in Φ is similar. The conditional posterior distributions for ϕ and γ given A are beta distributions. The counts in both are related to the site counts inside and outside the relevant set of ChIP windows. We penalize the counts for these beta distributions in the same manner as described above. Suppose we have sampled C_k sites for DBF k and $C_k^{(\text{in})}$ is the count of these sites within a ChIP window of type k. Then the net site count $n_k^{(\text{in})} = n_k(C_k^{(\text{in})}/C_k)$ for the inside count, and likewise for the outside count (Supplementary Methods Section S1.7).

The penalty parameter λ controls the level of penalization on sampled site counts in each iteration of our algorithm. We developed a method for choosing λ similar to the work by Fu and Zhou (2013) via 10-fold cross-validation, with details in Supplementary Methods Section S1.8. The DBFs with a non-zero concentration in at least 5 out of 10 training sets for the chosen λ are regarded as selected DBFs.

2.4 Landscape prediction

After enriched DBFs are selected, the binding landscape is predicted considering only the selected DBFs with prior counts $\alpha_k = 1$. Recall that a final run of our iterative sampling algorithm is necessary because with unknown concentrations calculating binding probabilities exactly through forward-backward summation is impossible. For an outline of the full algorithm, see Supplementary Methods Section S1.9.

In the DBF selection stage we penalize the BS counts to encourage sparsity in concentration estimation. Overall, the penalty biases DBF concentration estimates downward and biases the predicted binding landscape toward higher specificity and lower sensitivity. One can increase the sensitivity of BS prediction in the final stage by not penalizing. One may also consider reestimating PWMs, using previously published PWMs as prior information. This can increase sensitivity but may result in minor modes with some concentrations highly inflated. In such cases the landscape prediction with penalized site counts and fixed PWMs must be used despite the downward bias.

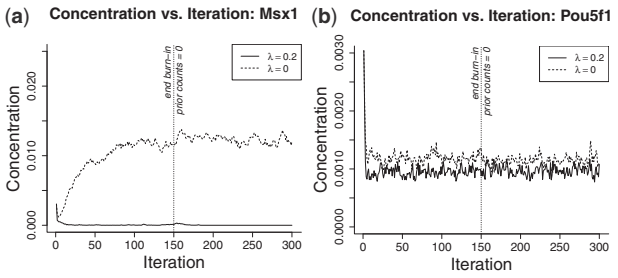

Fig. 2. Sample paths for the concentration of a DBF in the selection stage in a simulated dataset for a penalized run ($\lambda = 0.2$) and an unpenalized run ($\lambda = 0$). The dotted vertical line indicates the iteration at which the prior counts (α_k) were reduced from 1 to 0. (**a**) A DBF with no true sites. With no penalty ($\lambda = 0$), the concentration estimate is massively inflated. Twenty percent penalization pushed the concentration estimate to zero and eliminated this false positive DBF. (**b**) The sample path for a DBF with true BSs, illustrating the fact that concentration estimates for true positive DBFs are generally stable under non-excessive penalization

3 RESULTS

3.1 Applications to mouse embryonic stem cell and simulated data

We investigate two mouse datasets derived from multiple TF loci (MTL) regions defined in Chen *et al.* (2008). In that article, 12 TFs known to play a key role in maintenance of embryonic stem cells (ESCs) were studied with ChIP-Seq experiments. These 12 TFs are Stat3, Nanog, Klf4, Pou5f1/Oct4, Esrrb, Sox2, cMyc, Smad1, nMyc, E2f1, Tcfcp2l1 and Zfx. Two main clusters of BSs were found, one centered around Oct4 (Nanog, Sox2, Oct4, Smad1, Stat3) and the other around cMyc (cMyc, nMyc, E2f1, Zfx). We interrogate 1553 MTL regions built around the Oct4 group and 1178 regions built around the cMyc group. ChIP windows were built around ChIP peaks in these regions. Mouse ESC gene expression data from Zhou *et al.* (2007) were used to cull the considered DBFs to 170 with non-negligible expression. The PWMs of the 170 DBFs were extracted from the

TRANSFAC database (Matys *et al*., 2003). We use SparScape to nominate novel possible co-factors that act in concert with these known core TFs in mouse ESCs.

Two simulated datasets, each composed of 1000 simulated regions, were designed to mimic, respectively, the two mouse datasets, with a detailed description in Supplementary Methods Section S1.10. To evaluate performance across a range of sample sizes, for both the mouse and simulated data, the entire Oct4 group was analyzed together, but the cMyc group was randomly divided into subsets of 100 regions. DBFs were selected using only the first random subset and binding landscapes were predicted over all subsets with the selected tuning parameters and DBFs.

The most comparable method with available software is COMPETE (Wasson and Hartemink, 2009). COMPETE requires a fixed concentration vector and a pre-selected set of DBFs as input. We chose the concentration vector following the tuning procedure outlined in their article. Since the primary source of information for both SparScape and COMPETE is the sequence, we also compared against a naive approach considering the raw binding score at each locus. This raw score is the ratio of the PWM score over the background model score for a given w-mer (2).

3.2 Comparing DBF selection

Both COMPETE and the raw score method use only sequence information and do not use ChIP data like SparScape. To make a fair comparison, we first show that SparScape outperforms COMPETE and the raw score method when we ignore the available ChIP data and only use the sequence data. This illustrates the value of penalization and joint concentration estimation in SparScape. When the ChIP data are used, SparScape outperforms the competitors more dramatically.

To select DBFs using COMPETE, we ranked the DBFs by the total predicted binding probabilities over all base pairs and chose rank cutoffs for comparison with other methods. For the raw score method, when considered separately by locus, the number of scores for a DBF that exceed a chosen threshold (we used 1000 and 2000) approximately follows a Poisson distribution when there are no true sites. An expected false-positive count of scores over this threshold was estimated from control regions and used as the rate parameter of the Poisson distribution to find a P-value for the hypothesis that the DBF had no true sites in the examined regions. We considered three approaches to controlling for multiple comparisons. Our most conservative method was the ranking method where we ranked all the DBFs by their P-values and considered only the top N, where N is the number of DBFs selected by SparScape. The next most conservative approach was the use of the standard Bonferroni multiple testing adjustment to control the family-wise error rate at 5%. The least conservative approach was controlling the false discovery rate (FDR) at 5%.

Table 1 summarizes DBF selection results for the three methods. Even in the best case for the alternate methods, SparScape provides more powerful and more accurate DBF selection, in the large (Oct4 group) and small (cMyc group) datasets and in the real and simulated data, using ChIP data or not. To achieve similar sensitivity for the factors we know are enriched, the other methods nominate between 4 and 10 times as many possible co-factors, fewer of which are plausible compared with those nominated by SparScape. SparScape also did a better job selecting the DBFs around which the examined regions were built, selecting three of the five members of the Oct4 group and all four members of the cMyc group. When choosing the same number of DBFs selected by SparScape, COMPETE or the raw score method selected at best one member of the Oct4 or cMyc groups in the respective sets of MTL regions.

Table 1. DBF selection results for the raw score (RS) method, COMPETE (CO), SparScape with no ChIP data (SS-NC) and SparScape (SS)

		Cutoff	RS 1000	RS 2000	CO	SS-NC CV	SS CV
(A)	*Oct4*	*Rank*	2/11	4/11	0/11	10/11	11/13
	Group	*Bonf.*	10/140	10/127	11/127		
		FDR	10/148	10/134	11/134		
	cMyc	*Rank*	6/12	7/12	1/12	9/12	9/13
	Group	*Bonf.*	7/19	8/14	2/14		
		FDR	8/26	9/22	4/22		
(B)	*Oct4*	*Rank*	2/9	3/9	1/9	4/9	8/12
	Group	*Bonf.*	6/32	4/21	3/21		
		FDR	7/48	6/32	3/32		
	cMyc	*Rank*	0/5	0/5	0/5	2/5	5/11
	Group	*Bonf.*	7/78	7/54	2/54		
		FDR	9/100	8/84	6/84		

Note: (A) Simulated data and control. (B) Mouse data and sampled control. In (A), T/N represents T true DBFs out of N DBFs selected. In (B), T/N represents T ChIP DBFs out of N DBFs selected. For COMPETE, the number of DBFs to select from the ranked list was chosen to match that of SparScape (the *Rank* row) or the raw score method (the *Bonf* and *FDR* rows). The *Rank* row gives ranked results for the raw score method and COMPETE that match the number selected by SS-NC. The *Bonf* row gives selection results using a Bonferroni-corrected P-value threshold of 0.05. The *FDR* row gives selection results using an FDR of 5%. CV indicates that the DBFs were selected via cross-validation by SparScape.

When regions are analyzed to select target DBFs for lab-based follow-up experiments, it is of critical importance to reduce the number of false leads suggested by computational analysis. A major reduction in false positives in selecting DBFs is a key advantage of our new method. The substantial improvement over COMPETE highlights the critical roles of penalizing false positive counts and estimating concentrations in SparScape.

The joint DBF selection is a key factor in the improved performance achieved by SparScape. To demonstrate this point, we compared against an individual selection approach under the same framework of SparSpace but considering one candidate DBF at a time with the nucleosome. We applied this individual approach to the cMyc group mouse data with ChIP windows and ran SparScape 10 times for each DBF with the same λ as chosen in our joint run on that dataset. This individual approach selected 112 DBFs, 97 of them in all 10 runs, including 9 of the 12 ChIP DBFs. As expected, this result is close to that of the raw score method (with FDR control). As shown in Table 1, the joint run selected only 11 total DBFs including 5 ChIP DBFs. This highlights the huge reduction in the false positives for DBF selection that results from considering all the DBFs together.

In the Oct4 regions, SparScape selected Zfp219, Egr1/Krox24, Sp1 and Nr6a1/GCNF besides the ChIP DBFs. All four have been identified in the literature as having some association with

differentiation, ESCs or being key regulators in maintenance of pluripotency. Zfp219 and Sp1 have been identified as members of protein interaction networks for pluripotency with Oct4 and Nanog in mouse ESCs (Kim *et al*., 2008; Wang *et al*., 2006). Zfp281 has also been identified as a Nanog interacting protein required for proper cell differentiation (Fidalgo *et al*., 2011). The binding motifs for Zfp219 and Zfp281 are similar, so it is possible we are picking up sites for both TFs. Sp1 was also found to be a significant co-factor in this same dataset by He *et al*. (2009). The Egr and Sox families have been shown to interact in Schwann cells (Jessen and Mirsky, 2008). Nr6a1 is required to suppress Oct4 and recruit co-factors to affect DNA methylation and histone modifications in the Oct4 promoter during differentiation (Gu *et al*., 2011).

In the cMyc group, we proposed six co-factors, Atf5, Erf, Rfxap, Nrf1, Sp1 and Zbtb7b/cKrox/ThPok. Atf4 (with a nearly identical motif as Atf5) has been identified in a co-expression and regulation network centered around cMyc (Furuya *et al*., 2008). Erf is key in cell differentiation mediated by cMyc repression (Verykokakis *et al*., 2007). Nrf1 has been shown to interact with cMyc in regulating apoptosis and implicated as a key actor in pluripotency maintenance (Mason *et al*., 2009; Morrish *et al*., 2002). Sp1 interacts with cMyc (Herkert and Eilers, 2010). Zbtb7b directs the CD4 and CD8 T cell differentiation, while related TFs such as Miz-1 are known to cooperate with cMyc (Kerosuo *et al*., 2008).

Taken together, we see that SparScape was able to select DBFs known to bind in the regions of interest and nominate a small group of likely co-factors for both the Oct4 and cMyc groups. This can provide investigators with a higher return on experimental validation and follow-up studies.

3.3 The binding landscape and concentration

In the second stage, we predict the binding landscape and estimate concentrations for the DBFs selected in the first stage. The concentration estimates for the selected DBFs (SS in Table 1) in the mouse and simulated Oct4 group regions are shown in Figure 3. For the simulated dataset, the ratio between an estimated concentration and the true value ranged from 0.71 to 1.46 for the non-nucleosome DBFs in the penalized run ($\lambda = 0.2$), while the concentrations were generally overestimated without penalization ($\lambda = 0$) (Fig. 3a). Particularly, the unpenalized estimate for Nanog does not appear in the figure, as it was massively overestimated, indicating an enormous number of false positive predicted BSs. This is a general danger of jointly estimating the concentrations that can be avoided with proper penalization. This also explains, at least partially, why the estimated nucleosome concentration was lower in the unpenalized run than in the penalized run. With $\lambda = 0$, many nucleosome BSs were crowded out by the false positive sites for Nanog and other DBFs due to the competing nature between DBF binding in our model. Furthermore, nucleosome concentration τ_1 was not penalized even when $\lambda > 0$ [see Equation (7)]. Figure 3b shows the concentration estimates in the mouse Oct4 group data from a penalized run with fixed PWMs and an unpenalized run with reestimated PWMs. Similar patterns are observed as those in the simulated data.

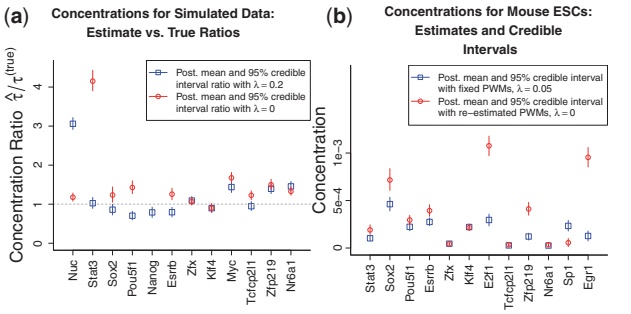

Fig. 3. Concentration estimates. (**a**) Ratios of the posterior mean concentration estimates and 95% credible intervals over the corresponding true concentrations for the correctly selected DBFs in the Oct4 group simulated data. (**b**) Posterior means and 95% credible intervals for the selected DBFs in the Oct4 group mouse data

Concentration estimates are stable over different runs of SparScape. Supplemental Figure S3 shows a box plot of the coefficients of variation (standard deviation divided by the mean) for the posterior mean estimates of concentrations over 10 independent runs of predicting the final binding landscape for the cMyc group data. These ranged from ~zero to just >0.02 for the 11 non-nucleosome DBFs selected by our two-stage run of SparScape and was <0.05 for the nucleosome.

Examples of predicted binding landscapes are shown in Supplementary Figure S4. The posterior binding probabilities summarized there can be used directly in further analyses, but we demonstrate the effectiveness of our method by choosing a posterior probability cutoff for predicting sites. A summary of results for SparScape and raw score predictions in the simulated datasets is given in Table 2. COMPETE results are not included in the table because the highest sensitivity achieved (<0.1) was so low that the results are not comparable in this format. We only present results for the DBFs with true sites, ignoring in the FDR calculations the large number of false-positive DBFs selected by the raw score method and those selected by SparScape.

For each category of DBFs, we chose the raw score cutoff that gave a similar sensitivity to that achieved by our method at the given posterior probability cutoff and compared the FDRs. SparScape reduces the FDR compared with the raw score method for every category of DBFs across both datasets, achieving a reduction of 21–63% in the Oct4 group data and 13–33% in the cMyc group. In the Oct4 simulation, we achieved an overall sensitivity of 0.74 with an FDR of 0.23 as compared with an FDR of 0.42 for the raw score method with similar sensitivity. In the cMyc simulation, we achieved a sensitivity of 0.7 and an FDR of 0.3, compared with an FDR of 0.4 for the raw score method. The raw score method performed relatively better on the DBFs with no ChIP data because those DBFs have, on average, stronger and more informative motifs. Moreover, SparScape uses ChIP data in a principled way, which led to more substantial improvement over the raw score method on the ChIP DBFs. Both methods predicted nucleosome sites with an FDR slightly below and above 50%, respectively, across a range of sensitivities. See Supplementary Figure S5 for more results with a wide range of posterior probability cutoffs.

For the mouse datasets, true BSs are not annotated, and therefore, we compared different methods based on the numbers of

Table 2. Sensitivity and FDR in BS prediction in the simulated data

		SparScape			Raw score		
	DBF subset	*PP*	*Sensitivity*	*FDR*	*Cutoff*	*Sensitivity*	*FDR*
Oct4	*All*	0.5	0.74	0.23	900	0.74	0.42
Group	*Non-Nuc*	0.5	0.71	0.16	1250	0.71	0.35
	ChIP	0.5	0.62	0.20	700	0.63	0.54
	No ChIP	0.5	0.82	0.11	5000	0.82	0.14
cMyc	*All*	0.6	0.70	0.30	4000	0.70	0.40
Group	*Non-Nuc*	0.6	0.67	0.26	5000	0.68	0.33
	ChIP	0.6	0.72	0.21	5000	0.76	0.33
	No ChIP	0.6	0.60	0.32	4000	0.60	0.37

Note: Non-Nuc is all considered DBFs except for the nucleosome. ChIP is the set of DBFs for which ChIP data were available. No ChIP is the set of DBFs for which ChIP data were not available. In the cMyc group simulation, there were 10 small datasets. Combined results are reported here but the predictions were made separately in each set. PP stands for posterior probability cutoff.

Table 3. BS prediction inside and outside matching ChIP windows

	SparScape			COMPETE		Raw score		
	PP	*In*	*Out*	*In*	*Out*	*Cutoff*	*In*	*Out*
Oct4	0.25	3103	56	148	584	500	1168	3485
Group	0.4	2150	39	3	7	1000	662	1789
	0.6	1074	25	0	0	2000	438	1110
	0.8	317	9	0	0	4000	254	583
cMyc	0.25	1328	387	12	540	500	1311	2930
Group	0.4	861	181	0	49	1000	979	2136
	0.6	497	61	0	0	2000	707	1500
	0.8	256	13	0	0	4000	442	951

predicted BSs for a ChIP DBF inside and outside its ChIP windows. The MTL regions were chosen by requiring multiple ChIP peaks from a small set of TFs to occur close to each other, making it much less likely that peaks in these regions are false positives. Likewise, we expect few true BSs further than 50 bp from a ChIP-Seq peak (the coverage of our ChIP windows) but within a few hundred base pairs. So we expect that the ChIP windows capture a high percentage of true BSs in these regions.

As reported in Table 3, for the selected ChIP DBFs, SparScape predicted a high percentage of BSs in a corresponding ChIP window for different cutoffs on the posterior binding probabilities. For example, in the Oct4 group data, we predicted 317–3103 BSs in a matching ChIP window for the eight selected ChIP DBFs, with only 9–56 sites outside. Since the performance of the raw score method and COMPETE in DBF selection was unsatisfactory, we report their results on the sites predicted for all 12 ChIP DBFs (Table 3). One clearly sees a much higher percentage of BSs predicted outside ChIP windows. In the most sensitive cases, SparScape predicted 77–97% of sites for ChIP DBFs within a corresponding ChIP window despite not restricting site prediction to within ChIP windows, while only 25–30% of sites predicted by the raw score method and 20% of sites predicted by COMPETE fell within a corresponding ChIP window. This suggests that a high percentage ($>70\%$) of the BSs predicted by COMPETE and the raw score method are false positives, demonstrating that for the questions we consider, our method is more sensitive and vastly more specific than the alternatives.

Although the joint estimation of DBF concentrations and binding landscape requires computationally expensive iterative sampling, SparScape is reasonably fast. SparScape is implemented in C++ with OpenMP, so runs most quickly on boards with more CPUs. We tested computation time with 10 independent runs on a subset of the cMyc mouse data consisted of just >52 000 bp in 100 regions with the 11 selected DBFs (Table 1) in addition to the nucleosome. A total of 1250 iterations took on average 37 min (with minimal variation) on a MacBook Pro running OS X 10.6.8 with 2.53 GHz Intel core 2 duo processors and 4 GB of RAM.

3.4 Application to gene promoters

SparScape is not designed to carry out DBF selection on entire chromosomes or the whole genome. The enrichment of BSs for any DBF other than the nucleosome is too thin and selection results may be too sparse or inconsistent. For prediction of a binding landscape over an entire chromosome or genome, we recommend running SparScape on the entire DBF library without selection and with a non-zero but small value of λ to prevent concentration inflation of DBFs with low-information or GC-rich motifs. SparScape can, however, select DBFs effectively on data less specialized than co-bound regions such as the Oct4 and cMyc groups explored above.

We demonstrate this by randomly sampling 2000 mouse genes and running SparScape on the upstream 1000 bp of these genes, using upstream $4001 - 5000$ bp as the control regions. The candidate library consisted of all 203 DBFs in our library with unique PWMs. We ran 10-fold cross-validation to select $\lambda = 0$, and selected 45 DBFs with $\lambda = 0$ and 41 with $\lambda = 0.05$ (the value used in the Oct4 and the cMyc mouse data). This represents 22% of the DBFs considered, a very reasonable number to be enriched in a random sample of 10% of gene promoters from the mouse genome. Supplementary Table S1 shows the DBFs chosen with these two values of λ ranked by posterior mean concentration. One indication that SparScape selected truly enriched proteins is the presence of TATA box binding protein in both lists (ranked ninth with the selected $\lambda = 0$), since the TATA box is known to be close to the transcription start site.

4 DISCUSSION

Generating a base-pair–specific binding landscape for all known DBFs over the entire genome of higher organisms with large genomes, without prior knowledge of DBF concentrations, is currently prohibitively computationally expensive. When one then considers only certain genomic regions, it is expected that a large proportion of DBFs will not have true BSs, and performance suffers if all DBFs are considered. Requiring prior knowledge of the set of DBFs active in the regions of interest and their concentrations introduces the need for (often ad hoc) user-dependent prior or iterative analysis or a need for additional experimental data. We contribute a new method, SparScape, that eliminates the need for this prior information and takes advantage of binding data where available while outperforming alternate methods. One of the key features of our method is the

inclusion of penalization in Bayesian inference based on the expected false positive site counts. This significantly reduces false positive results both in DBF selection and in BS prediction. A similar idea has been used in the contrast motif finder (Mason *et al*., 2010). The unsatisfactory performance of COMPETE for the problems we investigate suggests that tuning prior DBF concentrations when they are not given is difficult in practice and using an improper vector of concentrations can be risky for joint prediction of landscapes. Moreover, both SparScape and the raw score method select DBFs by comparing against some control regions, but COMPETE does not have this critical component. We want to stress that our comparison against COMPETE is mostly for demonstration purpose as the method is not targeted at DBF selection or predicting sparse binding landscapes.

As demonstrated by the results in this article, with the default method for choosing the level of penalization, a standard usage of SparScape usually works well. But in fact, SparScape is highly flexible. A user could independently choose less stringent DBF selection by setting a smaller penalty value and perform a single selection run, instead of 10-fold cross-validation. This option is particularly useful when the goal is to predict sites for all DBFs that could at all plausibly have BSs in the considered regions. The default choice in the final post-selection landscape prediction is to predict with the same penalty chosen in the selection stage and with fixed PWMs. A user may choose to make the predicted landscape less sparse by running the final prediction with no penalty, reestimated PWMs or both.

Our modeling framework can easily be extended to include other types of information. Some measure of absolute binding affinity, as opposed to the relative binding affinity information in a normalized PWM, is available from protein binding microarrays (Berger *et al*., 2006) and could be included as energy models (Djordjevic *et al*., 2003; Foat *et al*., 2006; Zhou, 2010). Further work integrating ChIP peak strength in the scoring of binding in ChIP windows could be fruitful, especially for nucleosomes. Other location-specific information on DBF binding to the sequence could also be used. We plan further work using SparScape's joint binding landscape to predict gene expression and offer insights into the underlying regulatory network.

Funding: NSF grants (DMS-1055286 and DMS-1308376 to Q.Z.) and Burroughs Wellcome Fund fellowship (to M.L.).

Conflict of Interest: none declared.

REFERENCES

Arnold,P. *et al*. (2011) MotEvo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics*, **28**, 487–494.

Berger,M. *et al*. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol*., **24**, 1429–1435.

Chen,X. *et al*. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Djordjevic,M. *et al*. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res*., **13**, 2381–2390.

Ernst,J. *et al*. (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res*., **20**, 526–536.

Fidalgo,M. *et al*. (2011) Zfp281 functions as a transcriptional repressor for pluripotency of mouse embryonic stem cells. *Stem Cells*, **29**, 1705–1716.

Foat,B. *et al*. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, **22**, e141–e149.

Fu,F. and Zhou,Q. (2013) Learning sparse causal gaussian networks with experimental intervention: regularization and coordinate descent. *J. Am. Stat. Assoc*., **108**, 288–300.

Furuya,S. *et al*. (2008) Inactivation of the 3-phosphoglycerate dehydrogenase gene in mice: changes in gene expression and associated regulatory networks resulting from serine deficiency. *Funct. Integr. Genomics*, **8**, 235–249.

Gu,P. *et al*. (2011) Differential recruitment of methyl CpG-binding domain factors and dna methyltransferases by the orphan receptor germ cell nuclear factor initiates the repression and silencing of oct4. *Stem Cells*, **29**, 1041–1051.

Gupta,M. and Liu,J. (2003) Discovery of conserved sequence patterns using a stochastics dictionary model. *J. Am. Stat. Assoc*., **98**, 55–66.

He,X. *et al*. (2009) A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS One*, **4**, e8155.

He,X. *et al*. (2010) Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput. Biol*., **6**, e1000935.

Herkert,B. and Eilers,M. (2010) Transcriptional repression: the dark side of myc. *Genes Cancer*, **1**, 580–586.

Jessen,K. and Mirsky,R. (2008) Negative regulation of myelination: relevance for development, injury, and demyelinating disease. *Glia*, **56**, 1552–1565.

Kaplan,N. *et al*. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.

Kaplan,T. *et al*. (2011) Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early drosophila development. *PLoS Genet*., **7**, e1001290.

Kerosuo,L. *et al*. (2008) Myc increases self-renewal in neural progenitor cells through miz-1. *J. Cell Sci*., **121**, 3941–3950.

Kharchenko,P. *et al*. (2008) Design and analysis of chip-seq experiments for dna-binding proteins. *Nat. Biotechnol*., **26**, 1351–1359.

Kim,J. *et al*. (2008) An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**, 1049–1061.

Laurila,K. *et al*. (2009) A protein-protein interaction guided method for competitive transcription factor binding improves target predictions. *Nucleic Acids Res*., **37**, e146.

Marbach,D. *et al*. (2012) Predictive regulatory models in *Drosophila* melanogaster by integrative inference of transcriptional networks. *Genome Res*., **22**, 1334–1349.

Mason,M. *et al*. (2009) Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics*, **10**, 327.

Mason,M. *et al*. (2010) Identification of context-dependent motifs by contrasting chip binding data. *Bioinformatics*, **26**, 2826–2832.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.

Morrish,F. *et al*. (2002) c-MYC apoptotic function is mediated by NRF-1 target genes. *Gene Dev*., **17**, 240–255.

Narlikar,L. *et al*. (2007) A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol*., **3**, e215.

Ramsey,S. *et al*. (2010) Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, **26**, 2071–2075.

Raveh-Sadka,T. *et al*. (2009) Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res*., **19**, 1480–1496.

Teif,V. and Rippe,K. (2010) Statistical-mechanical lattice models for protein-DNA binding in chromatin. *J. Phys: Condens. Matter*, **22**, 414105–414118.

Verykokakis,M. *et al*. (2007) The RAS-dependent erf control of cell proliferation and differentiation is mediated by c-Myc repression. *J. Biol. Chem*., **282**, 30285–30294.

Wang,J. *et al*. (2006) A protein interaction network for pluripotency of embryonic stem cells. *Nature*, **444**, 364–368.

Wasson,T. and Hartemink,A. (2009) An ensemble model of competitive multi-factor binding of the genome. *Genome Res*., **19**, 2101–2112.

Won,K.-J. *et al*. (2010) Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol*., **11**, R7.

Zhou,Q. (2010) On weight matrix and free energy models for sequence motif detection. *J. Comput. Biol*., **17**, 1621–1638.

Zhou,Q. and Wong,W. (2004) CisModule: *de novo* discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–12119.

Zhou,Q. *et al*. (2007) A gene regulatory network in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **104**, 16438–443.