Computation identification of transcription factor binding using DNase-seq

by

Tatsunori B Hashimoto

Submitted to the Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2014

Author ....
Department of Electrical Engineering and Computer Science
Janurary 31, 2014

Certified by ....
Professor David Gifford
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by ....
Professor Tommi Jaakkola
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Professor Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Theses

Computation identification of transcription factor binding using DNase-seq

by

Tatsunori B Hashimoto

Submitted to the Department of Electrical Engineering and Computer Science
on Janurary 31, 2014, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Here we describe Protein Interaction Quantitation (PIQ), a computational method that models the magnitude and shape of genome-wide DNase profiles to facilitate the identification of transcription factor (TF) binding sites. Through the use of machine learning techniques, PIQ identified binding sites for >700 TFs from one DNase-seq experiment with accuracy comparable to ChIP-seq for motif-associated TFs (median AUC=0.93 across 303 TFs). We applied PIQ to analyze DNase-seq data from mouse embryonic stem cells differentiating into pre-pancreatic and intestinal endoderm. We identified (n=120) and experimentally validated eight 'pioneer' TF families that dynamically open chromatin, enabling other TFs to bind to adjacent DNA. Four pioneer TF families only open chromatin in one direction from their motifs. Furthermore, we identified a class of 'settler' TFs whose genomic binding is principally governed by proximity to open chromatin. Our results support a model of hierarchical TF binding in which directional and non-directional pioneer activity shapes the chromatin landscape for population by settler TFs.

Substational parts of this thesis are taken from our publication on PIQ currently in press at Nature biotechnology.

Thesis Supervisor: Professor David Gifford
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Professor Tommi Jaakkola
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I'd like to thank my collaborators Richard Sherwood (for collecting primary data, running validation assays, and writing the first draft of the nature biotech paper), and Charles W. O'Donnel (for help with basic bioinformatics and sequence data preprocessing). I'd also like to thank my advisors David Gifford and Tommi Jaakkola for their valuable insights throughtout the research process. Finally, I am very thankful of my friends, family and particularly Victoria for her support throughout my research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Highly accurate genome-wide methods have been developed to localize the condition-specific binding of TFs to the genome, facilitating the elucidation of genome regulatory elements and gene regulatory networks [18, 3]. Chromatin immunoprecipitation of selected protein-DNA complexes followed by high-throughput sequencing and mapping of the immunoprecipitated DNA (ChIP-seq)[8] has become a valued method for TF location analysis and can reliably identify where TFs bind genome-wide within 10 bp [6, 5]. Each ChIP-seq experiment profiles a single TF and requires either an antibody specific to the TF or the incorporation of a tag into the TF being profiled. DNase-seq[1] is an assay that takes advantage of the preferential cutting of DNase I in open chromatin[16] and the steric blockage of DNase I by tightly-bound TFs that protect associated genomic DNA sequences[17]. After deep sequencing of DNase-digested genomic DNA from intact nuclei, genome-wide data on chromatin accessibility as well as TF-specific DNase-protection profiles revealing the genomic binding locations of a majority of TFs are obtained[2, 11, 10, 13]. These TF signature DNase profiles reflect the TFs effect on DNA shape and local chromatin architecture, extending hundreds of base pairs from a TF binding site, and they are centered on DNase footprints at the binding motif itself that reflect the biophysics of protein-DNA binding[10, 14, 1]. As DNase-seq experiments are TF-independent and do not require antibodies, it is

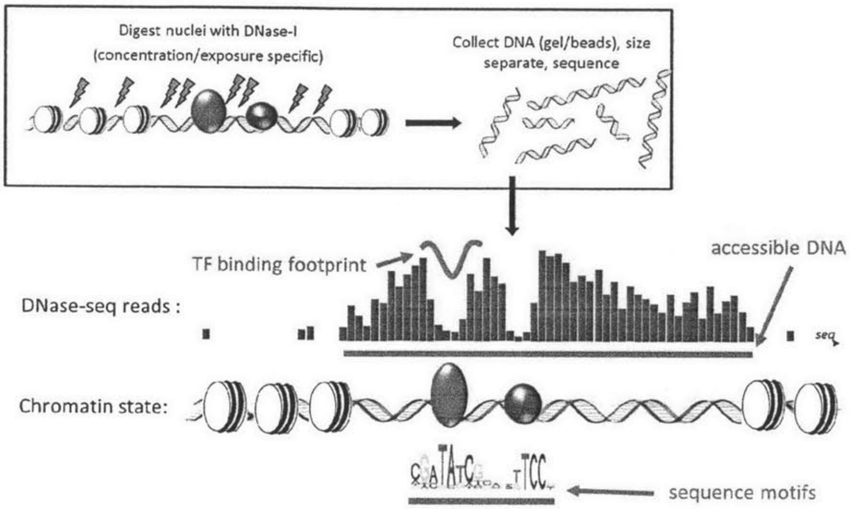
Digest nuclei with DNase-I
(concentration/exposure specific)
Collect DNA (gel/beads), size separate, sequence
TF binding footprint
accessible DNA
DNase-seq reads :
seq
Chromatin state:
sequence motifs

Figure 1-1: Schematic depiction of DNase-seq protocol resulting in characteristic patterns of accessiblity and transcription factor induced protection

possible to predict the binding of hundreds of different TFs to their genomic motifs from a single DNase-seq experiment.

1.2 DNase-seq protocol

In this assay (Figure 1-1), DNA is cleaved by the enzyme DNase I, and the ends of the resulting fragments are identified by high-throughput sequencing. Mapping the sequencing reads to the genome produces a genome-wide DNase profile in which the presence of mapped reads indicates that a genomic region is sensitive to DNase I cleavage and therefore accessible to proteins. Transcription factor binding protects DNA from cleavage, resulting in distinctive changes, or footprints, in the DNase profile. The identity of the proteins bound to the protected sequences can be determined de novo3 or by reference to previous knowledge of the DNA sequence motifs recognizied by transcription factors [15, 18].

The combination of DNase-I digestion followed by high-throughtput sequencing is referred to as DNase-seq throughtout this thesis.

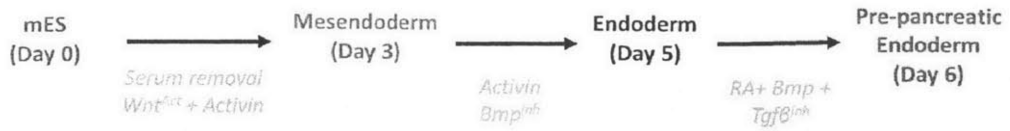

Figure 1-2: Cell identities explored within our pancreatic differentiation system

1.3 Stem cell differentiation system

We collected DNase-seq data from a developmental lineage paradigm that involves the stepwise differentiation of mouse embryonic stem cells (mESC) to pre-pancreatic and intestinal endoderm referred to hereafter as PancE and IntE27. We induced PancE and IntE differentiation by treating mESC for six days with an in vitro growth factor and small molecule treatment protocol (Figure 1-2). We collected DNase-seq data at two intermediate stages along this stepwise differentiation pathway, mesendoderm (day 3) and endoderm (day 5). This experimental structure yielded a total of five cell states (Figure 1-2) all of which were generated with >90% efficiency, providing relatively homogenous populations.

1.4 Computational identification of binding sites with DNase-seq

The binding of a transcription factor causes a characteristic footprint pattern in the observed DNase-seq read pattern. We can use this fact to determine whether or not some base in the genome is occupied by a transcription factor by comparing the local DNase-seq reads to the footprint. In aggregate, these DNase-seq footprints are distinct and serves to distinguish transcription factors (Figure 1-3).

However, while we can generate transcription factor footprints from ChIP-seq data, it is difficult to generate binding predictions from DNase-seq data alone, due to the large number of possible footprint shapes and candidate binding sites.

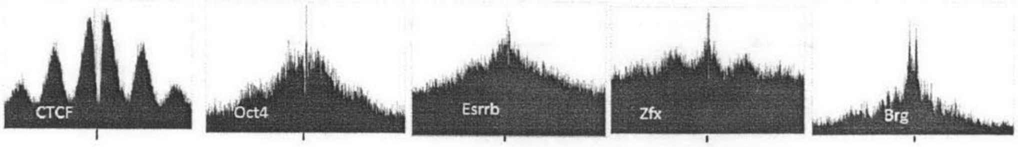

Figure 1-3: Average protection patterns across several thousand ChIP-seq binding sites for several transcription factors

1.5 Prior computational methods

Previous approaches to identification of transcription factor binding through DNase-seq can be classified into two groups: motif-free and motif-based. Motif-free approaches such as DHS footprinting used in the ENCODE phase 2 project can identify local protection patterns resembling TF binding and does not require prior knowledge of transcription factor binding motifs. Motif-based approaches such as CENTIPEDE method use transcription factor binding sites as a way to narrow the set of candidate binding sites. PIQ uses motif information as a critical part of its binding calls. We justify this decision by noting that while dependence on motif information is undesireable, it is almost unavoidable since even using motif-free methods most downstream analysis of DNase based TF binding calls use motifs to disambiguate which factor is binding at a binding site.

Our algorithm utilizes recent advances in time-series models and approximate inference to automatically correct for the experiment-level biases of DNase-seq. First we use expectation-propagation to fit a billion element Gaussian process model within minutes by exploiting stationarity. Second, we use sparse inverse covariance matrix based methods to share strenght across experiments. Lastly we use motif-association arguments to construct robust decision rules for whether a candidate binding site is bound by a trasncription factor or not.

Chapter 2

Methods

We developed a novel method for detecting transcription factor binding events from DNAse hypersensitivity data. The statistical model and inference framework shown in Figure 2-1 are the natural result of several design goals which we outline first.

1. **Resistance to low-coverage** Share strength across neighboring bases by modeling reads as arising from a Gaussian Process.

2. **Integrate multiple experiments** Learn the cross-experiment structure as a Gaussian graphical model using L_1 regularization.

3. **High spatial accuracy** Use motifs to inform base-pair level positions rather than de-novo footprinting.

4. **Robust worst case behavior** Use priors that guarantee monotonicity with respect to motif score and read coverage.

5. **Scalability to thousands of factors genome wide** Fast approximate inference strategies and use of Amazon ec2.

The five design goals correspond to the major subcomponents of the algorithm and will be covered below.

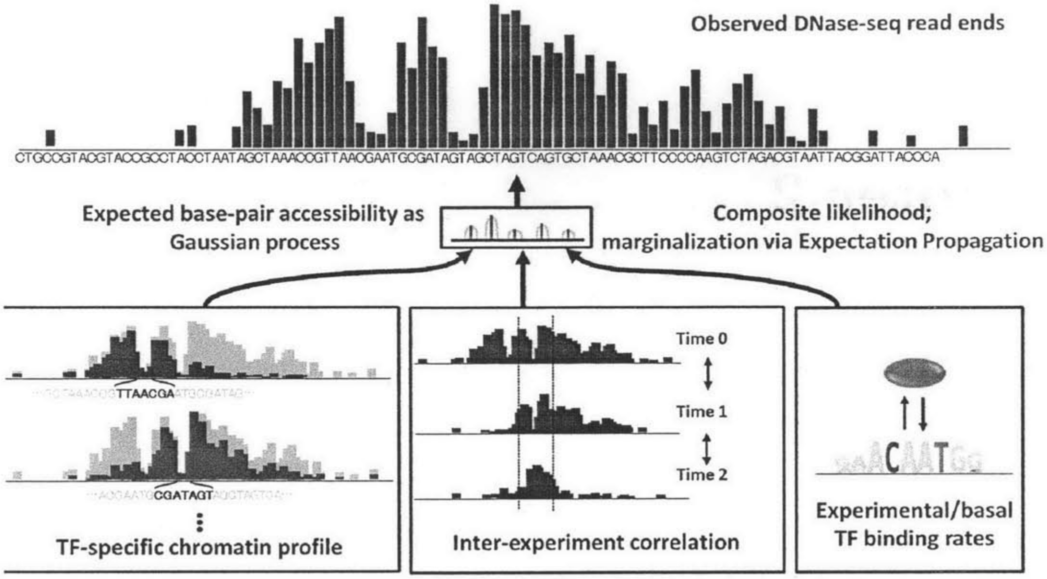
Observed DNase-seq read ends
CTGCCGTACGTACCGCCTACCTAATAGCTAAACCGTTAACGAATGCGATAGTAGCTAGTCAGTGCTAAACGCTTCCCCAAGTCTAGACGTAATTACGGATTACCCA
Expected base-pair accessibility as Gaussian process
Composite likelihood; marginalization via Expectation Propagation
TTAACGA
CGATAGT
TF-specific chromatin profile
Time 0
Time 1
Time 2
Inter-experiment correlation
ACAATG
Experimental/basal TF binding rates

Figure 2-1: Overview of the PIQ generative model. The Gaussian process ties together the TF and inter-experiment effects, generating a correlated latent state from which reads are drawn

2.1 Covariance correction

2.1.1 Generative model for reads

We model the generative model of reads in the single-experiment, single-strand, no factor binding case as a two step process. First we generate the underlying per-base accessibility of the genome to DNase as a Gaussian Process, which is a distribution over functions of a particular level of smoothness. The Gaussian Process is parametrized by μ_0, the average log-read rate per base, σ_0 the deviation in log-read rates, and $k_{|i-j|}$, the correlation between neighboring bases.

$$\mu_i \sim N(\mu_0, \Sigma)$$

$$\Sigma_{i,j} = \text{Covariance}(\mu_i, \mu_j) = k_{|i-j|}$$

Given the per-base rates, μ_i we define the read per base x_i as being distributed Poisson with log-rate equal to μ_i

$$x_i \sim \text{Poisson}(\exp(\mu_i))$$

Intuitively, μ_0, Σ model the overdispersion of read counts relative to a Poisson, while $k_{|i-j|}$ defines the degree and type of smoothness we have across the genome, allowing us to share information across adjacent bases. In the multi-experiment case, we estimate the parameters $(\mu_0, \Sigma, k_{|i-j|})$ for each experiment.

2.1.2 Cross-experiment and cross-strand model

Cross-experiment and cross-strand effects for DNase affinity are treated identically. Let $\mu_{i,k}$ be the read rate at base i in experiment or strand $k \in \{1 \ldots K\}$, then we model the distribution over different experiment as a multivariate Gaussian parametrized by a cross-experiment correlation matrix $\hat{\Sigma}$ subject to a L_1 penalty prior with parameter λ,

$$\{\mu_{i,1} \ldots \mu_{i,K}\} \sim \text{Multivariate Normal}(\mu_0, \hat{\Sigma})$$

$$\log(P(\hat{\Sigma})) \propto -\lambda|\hat{\Sigma^{-1}}|$$

Parameterizing the cross-experiment correlation by a matrix $\hat{\Sigma}$ is natural, since the single experiment rates μ_i are already Gaussian. The L_1 penalty induces sparsity over the precision matrix $(\hat{\Sigma^{-1}})$ which has the effect of preventing loosely related experiments from sharing information. In our experimental design, this penalty is particularly important, since the differentiation protocol results in a highly structured cross-experiment correlation structure.

2.2 Binding call classifier

We will first cover the single-experiment, single-factor case since the generalization to multiexperiment and multifactor are straightforward.

PIQ represents a transcription factor as a motif (shared cross-experiment) and a DNase footprint parameter β, not shared across either experiments or factors. A particular binding site is represented as a pair of variables, indicating the binding site location and whether the site is bound, (y_j, I_j).

Given the covariance matrix Σ, binding calls can be determined by simply clustering the set of binding sites after de-correlating the input counts. We propose two such methods: a mixture model based approach which gives PIQ a probabilistic interpretation, as well as a SVM based approach which more directly optimizes our target objective function. All results were generated with the older mixture model based classifier, but we have found that the SVM approach is faster and has slightly better worst case performance.

2.2.1 Mixture model based classifier

Given a motif for some factor, we call a base a binding site candidate if its score passes some threshold (in all analysis, we used any position occurring with less than 1e-5 frequency with respect to background sequence). For the binding site candidate indexed by j, let y_j be the base-pair representing the midpoint of such a motif match. Then we define the binding-adjusted read rates for the two strands $(\hat{\mu}_i^+, \hat{\mu}_i^-)$ in terms of the binding indicator I_j which is one if a factor is bound, and a DNase-footprint parameters for each strand, $\beta^+ = \{\beta^+_{-M} \dots \beta^+_0 \dots \beta^+_M\}$ and an analogous β^-.

$$\hat{\mu}_i^+ = \mu_i^+ + \begin{cases} \beta^+_{i-j} & : \ |i-j| \leq M \text{ and } I_j = 1 \\ 0 & : \ \text{otherwise} \end{cases}$$

In the multi-experiment case, each experiment and factor receives its own footprint $(\beta^+, \beta^-)_k$, and in the multi-factor case we simply sum over all matching β.

2.2.2 SVM based classifier

In the SVM approach we consider the overall objective of DNase-seq binding call to find a set of candidate sites whose footprint patterns are significantly enriched in PWM match sites compared to background sites. This goal can be cast as a straightforward classification problem, with the 'positive class' being drawn from PWM matches, and 'negative class' drawn from sites uniformly at randomly offset at least 1kb but less than 100kb away from each motif match.

An advantage of the SVM method is that even if our covariance correction fails to completely remove basewise correlation, the SVM applied to variance stabilized data will remove the residual covariance terms to find the optimal linear decision bound. We have found that this property is provides substantial benefits over using a simpler approach such as LDA (linear discriminant analysis) and does not suffer from local minima as the mixture model does.

Any classifier more complicated than a simple linear decision bound has runtime and interpretability problems, and the SVM approach is likely to be the best practical classifier based technique for binding calls. Using a online-gradient optimizer such as PEGASOS that exploits the sparsity underlying the problem, it takes only a couple minutes to process millions of candidate binding sites.

2.3 Hypothesis testing for binding

Identifying the significance of binding sites through the use of background is a core innovation in PIQ and guarantees that binding sites discovered by PIQ are biologically relevant. We separate the model fitting and significance testing components such that even if our fitted model is incorrect, the significances are not.

For every candidate motif match site, we generate a background binding site by randomly selecting a coordinate at least 1kb away but within 100kb of the candidate site. Assuming that the probability of randomly selecting a non sequence specific transcription factor binding site is negligible, this gives us a confident set of examples drawn from the set of non-bound region.

Running the classifier on this background site gives us the null distribution of scores expected from unbound transcription factor sites. While this distribution can be used directly to construct a set of positive binding sites with a given p-value, we use this distribution in order to find the set of confidently bound sites.

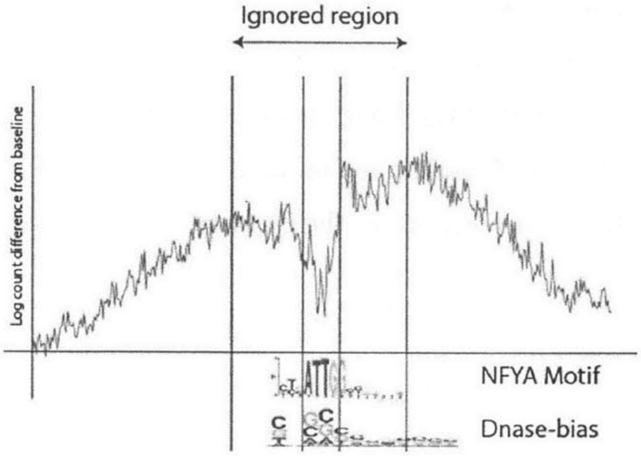

Figure 2-2: Sequence bias avoidance method: we ignore any bases involved in the motif match plus a 10-bp flanking sequence. We flag any motifs whose footprints are not statistically significant outside this possibly sequence-biased region

Each background binding site is assigned a PWM score drawn from the permuted null distribution of PWM scores. We then find a linear separator over the pair (PIQ score, PWM score) which maximizes the enrichment ratio of candidate sites to background sites. After this procedure, any binding site that passes cutoff is guaranteed to have a PIQ score significantly exceeding the expected score at its PWM score.

This nonparametric permutation based cutoff procedure ensures that when the classifier performs badly, PIQ calls fewer sites instead of calling more false sites.

2.4 Correcting sources of bias

The permutation test based binding site calls only guarantee that the binding sites called as positive are significantly correlated with sequence, but does not guarantee that this correlation is not bias due to the DNase enzyme.

In order to determine whether a particular motif's DNase footprint is due to sequence bias, we look at the footprint region outside the motif match (Figure 2-2) and compare this deviation against those of random K-mers of equal length, which we expect to have no substantial TF binding.

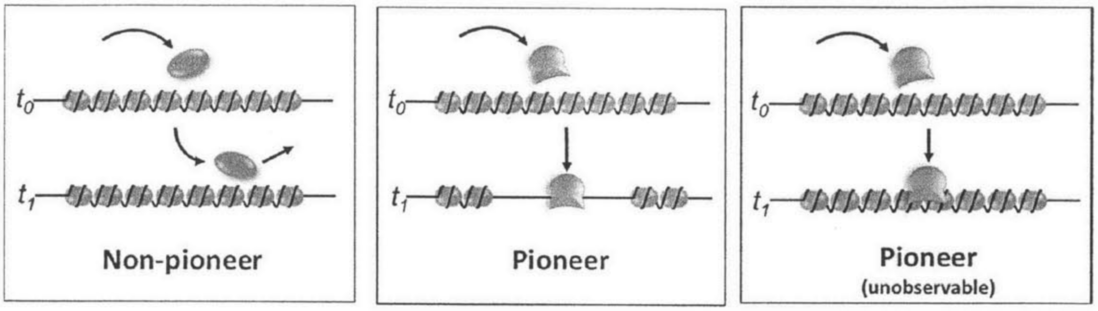

Figure 2-3: Classification of transcription factors: Non-pioneers bind only at open chromatin, Pioneers can bind regardless. We can detect pioneers that additionally open surrounding chromain.

In order to avoid falsely detecting transcription factors we significance test the classification vector. By selecting motifs whose non-motif associated footprint has deviation at least 3 standard deviations we can ensure that we do not falsely detect binding sites based upon PWM scores alone.

2.5 Detection of 'pioneer' factors

We asked whether PIQ could provide an initial understanding of the rules governing TF binding site choice. We focused first on whether some TFs act as "pioneers24," shaping the chromatin landscape and the binding of other TFs (Figure 2-3). Several reports of TFs possessing pioneer activity exist in the literature24, 26, 28-33, but these reports are empirical experimental studies that do not use standard criteria to define pioneer TF activity, are often unconfirmed functionally and to date no systematic attempts have been taken to categorize pioneer TFs. Although pioneer TFs have been defined in various ways, we chose to probe the existence of pioneer TFs capable of binding to closed chromatin and opening nearby chromatin for future occupancy by other TFs. Utilizing our time series, we designed a pioneer index to measure the expected motif-specific local increase in DNase accessibility with respect to baseline at sites whose binding changes between successive timepoints according to PIQ for each of our 733 motifs. A higher pioneer index corresponds to higher chromatin opening activity from one timepoint to the next in our developmental timecourse.

Chapter 3

Implementation

3.1 Inference

We estimate hyperparameters via a modifed expectation propagation iteration over the estimates μ_0,Σ,and σ.

Recall that our factorization takes the form

$$p(\lambda|C_i) \approx \prod_i N(\lambda_i|\mu(c_i), \sigma(c_i))N(\lambda|\mu_0, (\Sigma^{-1} - 1/\sigma)^{-1})$$

We can rewrite this in a form resembling expectation propagation by writing

$$p(\lambda, \Sigma, \mu_0|C_i) = \prod_i P(C_i|\lambda_i)P(\lambda, \Sigma, \mu_0) \approx \prod_i t_{C_i}(\lambda_i)P(\lambda|\Sigma, \mu_0)P(\Sigma, \mu_0) = q(\lambda, \Sigma, \mu_0)$$

Note that unlike the standard Expectation Propagation factorization for Poisson Gaussian Processes which uses N terms t_i, all bases i having the same observed count share the same approximation, t_{C_i}.

Under the standard EP update rules we obtain straightforward estimates. At time

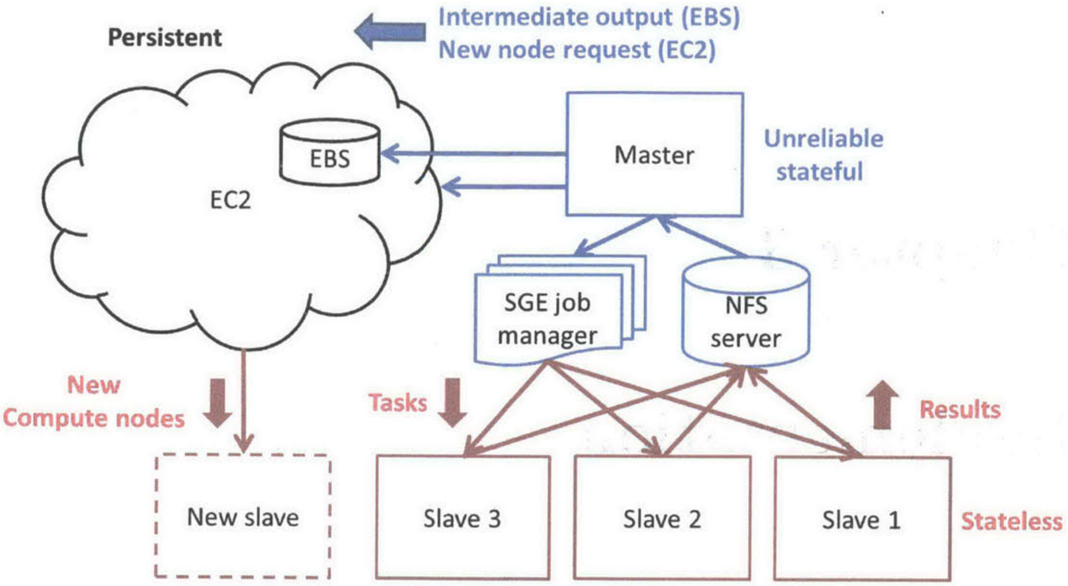

Figure 3-1: Overview of PIQ architecture on Amazon EC2. Red nodes carry almost no state acting only as compute nodes, blue nodes carry short-term state necessary to allocate compute nodes, and EBS storage is persistent and stores all fitted parameters and results

t, given some estimate q_t, σ_t, and $\sigma_t(c)$ of the approximate distribution:

$$
\begin{aligned}
\Sigma &= \mathrm{Cov}_{q_t}(\lambda|C)\\
\mu_0 &= E_{q_t}(\lambda|C)\\
\sigma_{t+1} &= (\sum_i((\mathrm{var}_{q_t}(\lambda_i|C))^{-1} - 1/\sigma_t(c_i) + 1/\sigma_t)^{-1})/M\\
\sigma_{t+1}(k) &= \sum_{\{i:C_i=k\}} \frac{1}{M}\int(\mu_0-\lambda_i)^2 P(C_i|\lambda_i)N(\lambda_i|\mu_0,\sigma_{t+1})d\lambda_i
\end{aligned}
$$

The expectation propagation updates allow us to find the minimum KL divergence approximation using σ weight and Σ,μ_0 hyperparameters.

3.2 Cloud computing

Even with fast approximate inference techniques scaling PIQ up to thousands of transcription factors is not feasible on a single computer. We estimate that a large

attribute	typical value	scaling
Number CPUs	80 CPU	M
Number Motifs	1500 Motifs	L
Number Experiments	10 Experiments	K
Window size	400 bases	W
Genome size	2.8 billion	N
Runtime	1 day	$O(NLK/M + W^3K/M + K^3)$
Memory	2Gb / CPU	$O(W^2K + K^2)$

Table 3.1: Typical problem size and asymptotic scaling for the PIQ algorithm

problem with ten experiments and 1337 motifs would take up to a cpu-year to compute (Table 3.1). In order to overcome this computational limitation we use cloud computing resources from Amazon EC2 to scale our computation capacity up.

We use a master-slave architecture, with a single central master node maintaining job state (Figure 3-1). Due to the nature of EC2, nodes can be terminated at any time and so we keep minimal state on transient hardware and frequently synchronize to a non-transient block storage device. The cloud computing infrastructure is outlined below in figure X.

Task priority and dependencies were accounted for using the sun grid engine (SGE) and task to parent communications were performed through binary files serialized onto a central NFS drive served from the master node and backed by a persistent EBS (elastic block store) drive.

Chapter 4

Results

4.1 Computation results

We tested sparse poisson approximation on a panel of simulated sparse data drawn from a Poisson Gaussian process. As baseline, we compare against the full expectation-propagation factorization, shown below , following analogous factorizations proposed in the literature [9].

$$p(\lambda, \Sigma, \mu_0 | C_i) = \prod_i P(C_i | \lambda_i) P(\lambda, \Sigma, \mu_0) \approx \prod_i t_i(C_i | \lambda_i) P(\lambda | \Sigma, \mu_0)$$

Simulated data was generated by using a stationary Gaussian kernel whose bandwidth is fixed at 100 with total window size varying from 50 to 2000. The mean of the Gaussian process was taken via fitting a Poisson-lognormal to real world DNase sequencing data.

In order to remove any possibility of implementation based differences, the sparse solver is implemented in native R code using no specialized BLAS packages, and no conjugate gradient type iterative solver is used. Full EP is implemented in hand optimized C++ using the Eigen linear algebra library, using fast LLT based matrix inverse subroutines.

All plots show results over 1000 replicates of simulated data.

4.1.1 Runtimes

We measured the runtime of the sparse solver and full EP under optimal conditions where all hyperparameters were set to ground truth values. In all dimensions the sparse solver performs orders of magnitude faster. In dimensions 50-200 runtime is unmeasurably quick giving the median runtime of 0 (Figure 4-1). In higher dimensions we record 100-1000 fold improvements in runtime where full EP takes 100 seconds to process a 2000 bp window, and we take less than half a second.

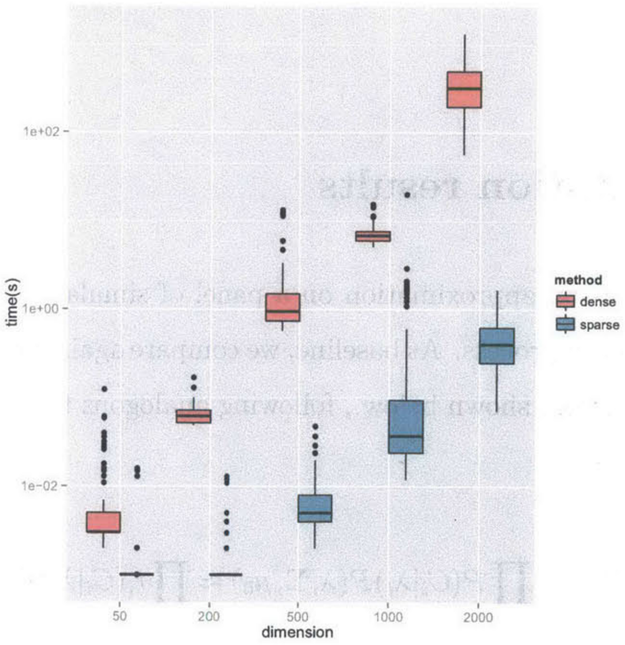

Figure 4-1: Runtimes as a function of window size. Sparse EP is substantially faster than the naive approach.

4.1.2 Parameter estimate accuracy

We also compare the accuracy of the sparse solver to full EP for estimating the expected value of the log rate λ. Once again, all algorithms are given true values for hyperparameters.

We measure the squared deviation between each algorithms' estimated value for λ and its true value over 1000 replicates.

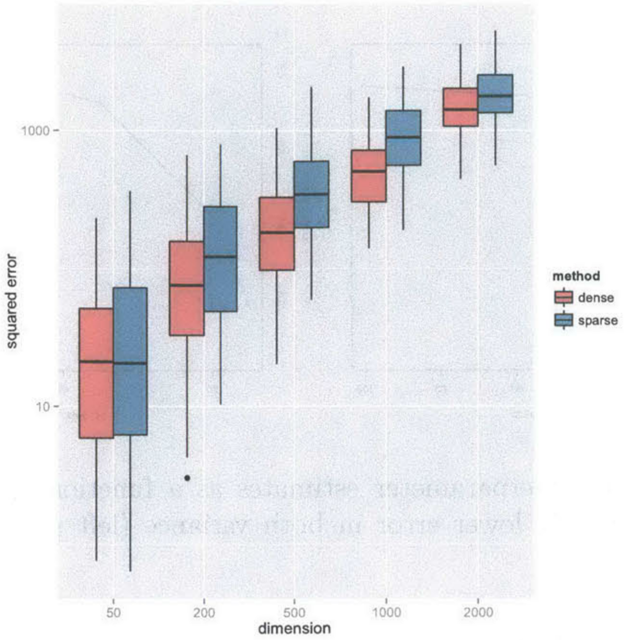

Figure 4-2: Mean estimate error as a function of size of window

The approximation error is a relatively constant factor of 1-1.5 more for sparse EP compared to its dense variant, showing that the uncertainty in the poisson Gaussian Process model dominates the sparse approximation error. (Figure 4-2)

4.1.3 Covariance structure

A advantage of the sparse solver compared to the full EP method is that since the hyperparameters can be estimated directly, rather than via an EM outer loop used in standard EP [9], the estimates of covariance Σ are significantly more accurate.

Comparing the estimated hyperparameters μ_0 and Σ for full EP and σ for our sparse solver we find that the sparse solver dominates the full EP in estimating variance when there is low correlation, and the reverse is true for correlation, where sparse solver converges to the true correlation structure for large bandwidth, while full EP systematically underestimates the overall correlation (Figure 4-3).

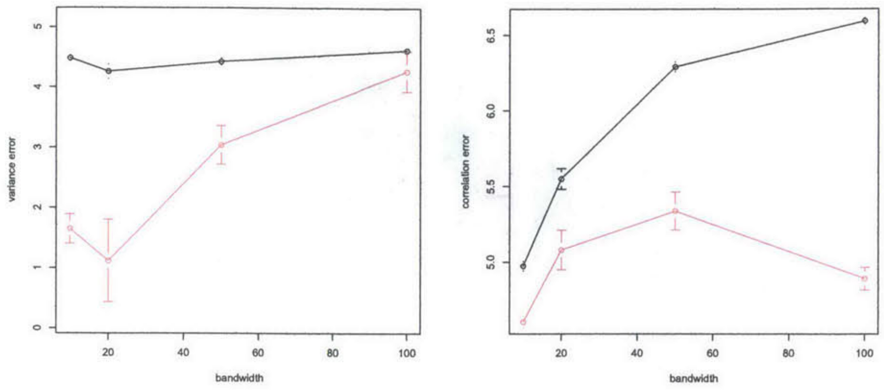

Figure 4-3: Error in hyperparameter estimates as a function of bandwidth. Red (sparse) has substantially lower error in both variance (left panel) and correlation (right panel) error.

4.2 Biological results

4.2.1 Comparison to ChIP-seq

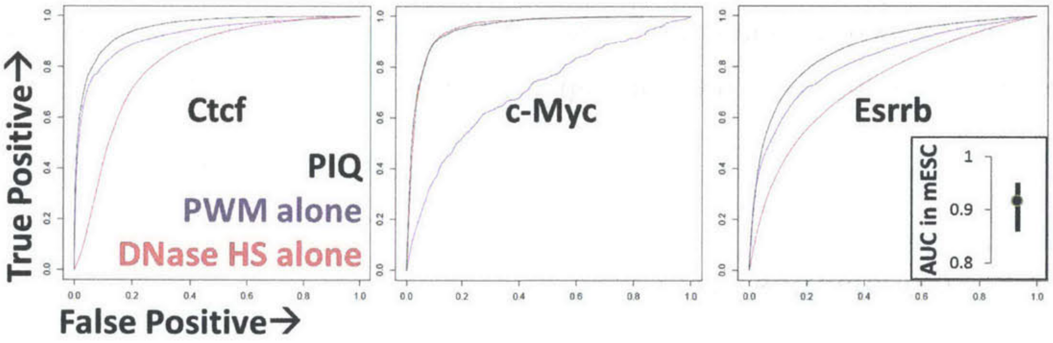

Figure 4-4: Comparison of PIQ calls with ChIP-seq shows that PIQ based DNase-seq binding calls are highly concordant with ChIP-seq, with both AUC and PPV near 0.9 across our mouse ES ChIP-seq experiments.

The high correspondence of PIQ output with ChIP-seq results suggests that PIQ provides a valuable tool for predicting protein regulatory interactions for hundreds of TFs genome wide. PIQ allows TF binding site prediction with similar accuracy to ChIP-seq for motif-supported direct protein-DNA binding events, with a median AUC of 0.93 (Figure 4-4). With a small number of replicate experiments PIQ is able to predict the binding of over 733 factors, and can do so in the absence of specific TF antibodies or tagged TFs. However, PIQ cannot detect TF motif-free binding events which are observed in ChIP-seq for certain TFs. Some motif-free ChIP-seq events may be mediated by cofactor proteins with diverse sequence specificities, and PIQ would miss these regulatory interactions, although some motif-free events may also be artifacts.

4.2.2 Detection of pioneers

4.2.3 Biological validation of pioneeers

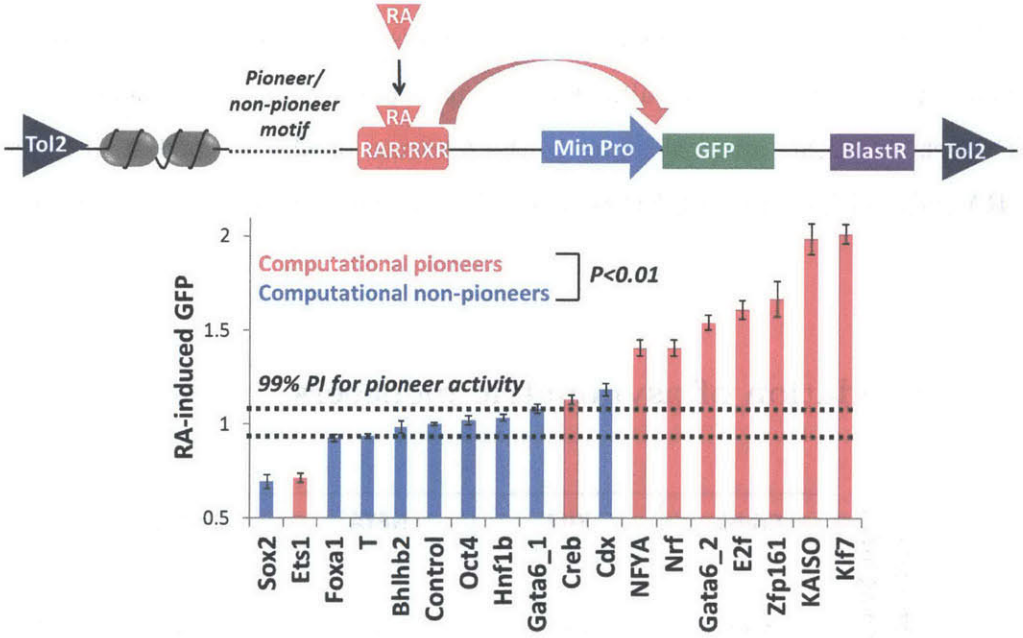

Figure 4-5: Design of the biological reporter assay (top panel) and results compared to computational prediction (bottom panel). All but one computationally predicted pioneer activates the reporter.

We experimentally tested the ability of a variety of predicted pioneer and control motifs to open up surrounding chromatin and allow other TFs to bind. To evaluate these criteria in a high-throughput, functional assay, we designed 18 versions of a reporter vector driven by a strong RXR:RAR motif directly adjacent to a pioneer or non-pioneer motif at a locus >1 kb from a minimal promoter and GFP reporter gene (Figure 4-5). We chose the RXR:RAR motif for three reasons. First, RXR:RAR binding shows no effect on surrounding chromatin in the computational analysis. Second, nuclear hormone receptors, which bind the RXR:RAR motif, respond primarily to surrounding chromatin state rather than specific cofactor interactions [7] (also see later text). Third, the RXR:RAR motif allows strong inducible expression of GFP upon addition of retinoic acid (RA), allowing a straightforward quantitative readout of cellular fluorescence intensity. We inserted this vector into the genome of mESC by means of Tol2 transposition35 followed by antibiotic selection, allowing for random genomic integration in a highly polyclonal fashion (>1,000 distinct clones per reporter line), thus controlling for site-specific effects. Consistent with this idea, biological replicates of several lines produced from distinct rounds of Tol2 transposition yielded highly reproducible results. We then used flow cytometry to measure cellular GFP levels in mESC after 24 hours in the presence or absence of RA, interpreting the RA-induced increase in GFP as a correlate of the accessibility of the RXR:RAR site.

4.2.4 Validation of asymmetric pioneeers

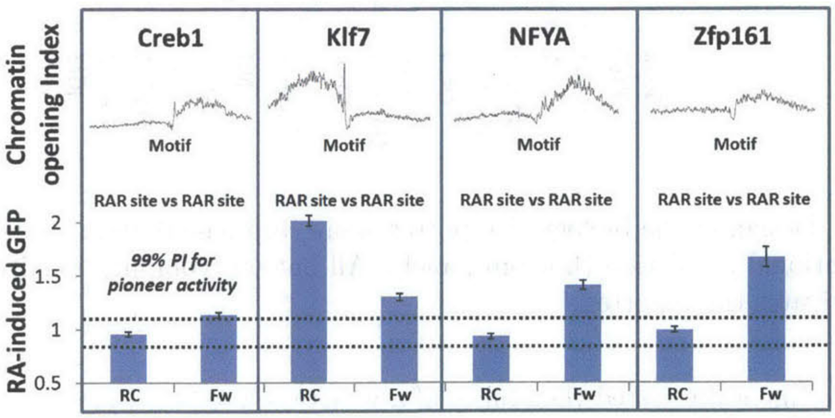

Figure 4-6: Using directional reporters show that computationally predicted directional pioneers result in nearly exactly the predicted activation patterns.

Evidence exists that TFs deposit histone marks asymmetrically36. We identified a subset of pioneer TF families that open chromatin more significantly on one side of their motif than on the other (Figure 4-6). We call factors that possess this novel asymmetrical chromatin opening ability directional pioneers. To quantify directional pioneer activity, we measured the expected difference in chromatin opening on either side of each pioneer motif, identifying strong directional pioneer activity in the Klf/Sp, NFYA, Creb/ATF and Zfp161 pioneer TF families. As we cannot observe directional pioneer activity at palindromic motifs because PIQ cannot orient them, we note that the directional pioneer TF Creb/ATF has multiple PWMs, one of which is non-palindromic. Although directional motifs are known to be important at promoters [4], our analyses exclude TSS-adjacent regions and we do not find appreciable transcript production or promoter-characteristic histone marks at distal pioneer sites. Thus, the unidirectional opening of chromatin relative to pioneer TF motif appears to represent a property of certain TFs that to our knowledge has not been described. To experimentally assess directional pioneer activity, we performed reporter analysis on four motifs displaying strongly directional pioneer activity (Figure efdirpioneer), placing both motif orientations relative to the RXR:RAR site. In all four cases, RA-induced GFP was significantly stronger in the direction predicted to have higher pioneer activity (Figure efdirpioneer), and as predicted, NFYA, Creb and Zfp161 only open chromatin in a single direction from their motif. Directional pioneer activity does not occur during transient transfection, suggesting that this activity occurs through interaction with the local chromatin state.

4.2.5 Pioneers enable binding of 'settler' factors

Next we reasoned that classifying TFs by their interactions with chromatin might reveal distinctions in how TFs choose binding sites. As pioneers have been shown to scan nucleosomal DNA for their motifs[12], we reasoned that they may be more likely than other TFs to bind to their motif wherever it occurs. To assess this idea, we devised a metric to indicate the likelihood of a TF to bind to an instance of its motif, the correlation of PWM score and binding probability (referred to hereafter as motif

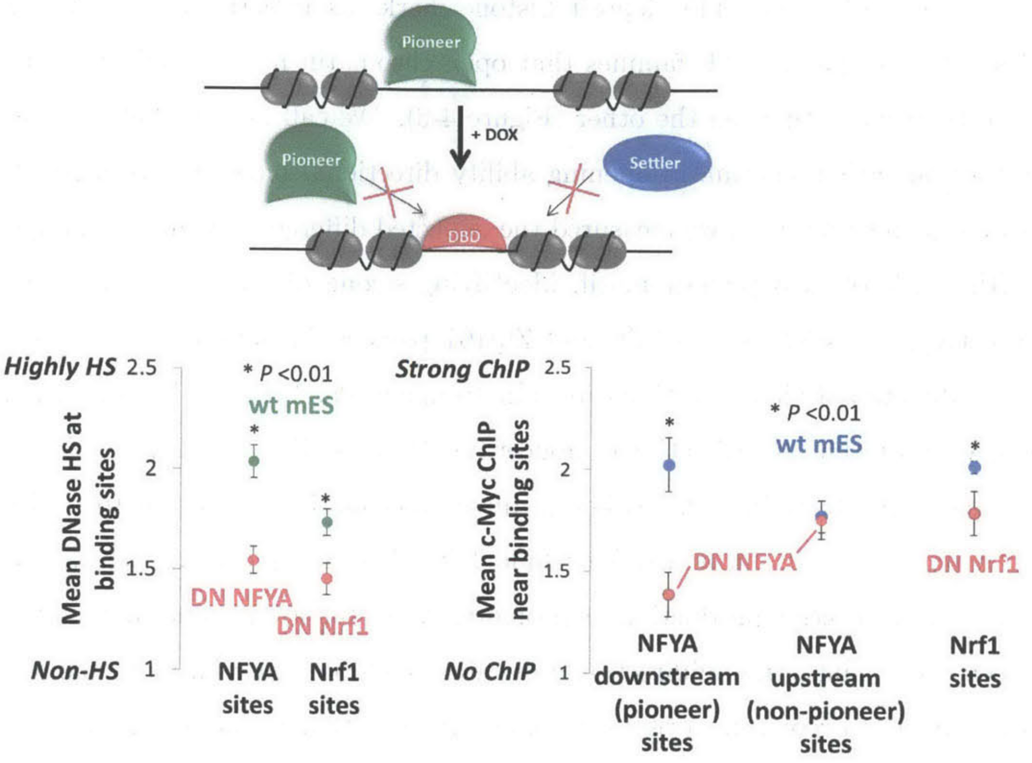

Figure 4-7: Design of a dominant negative knockout experiment to assess whether pioneers cause nearby transcription factor binding (top panel). Experimental evidence matches computational prediction of both magnitude and direction of pioneer activity in aiding nearby factor binding (bottom panel).

dependence). Plotting motif dependence against the chromatin opening index, we find a statistically significant (P<0.01 in t-test) but imperfect positive correlation between motif dependence and chromatin opening (Figure 4-7), suggesting that pioneer TFs generally do not bind to a high fraction of their genomic motif candidates. Several non-pioneer TFs, including REST, also display strong motif dependence (Figure efsettler). Motif dependence is uncorrelated with motif information content, suggesting that it is not an artifact of database PWM quality. Thus, although pioneers TFs are more likely to bind their motifs than are non-pioneers, they still rely on facets other than their motif in a majority of their binding decisions. Among non-pioneer TFs, we reasoned that some TFs might be disproportionately dependent on the preexisting chromatin state as established by pioneer TFs. We explored this possibility computationally by measuring the correlation between DNase accessibility surrounding high-confidence TF motifs and binding probability. Plotting this metric against the chromatin opening index, which controls for TF-intrinsic chromatin opening, we found that TFs vary substantially in their dependence on chromatin openness in order to bind genomic DNA (Figure 4-7). A subset of TFs were highly likely to bind wherever their motif occurs in an open chromatin landscape but do not open chromatin themselves. We coin the term settler TFs to define the set of TFs whose binding is predominantly dependent on the openness of chromatin at their motifs. Chromatin dependence of TFs is graded, but a stringent cutoff gives an estimate that 131 of the 733 motifs (18%) act as settler TFs. The majority of non-pioneer TFs, which we term migrant TFs, bind only sporadically even when chromatin at their motifs is open and are presumably more heavily dependent on specific cofactor interactions. Accurate a priori prediction (AUC>0.9) of ChIP-seq genomic binding of settler TFs, such as members of the Myc/MAX, nuclear hormone receptor (i.e. RXR:RAR), Ap-2 and NF-yB families, can be obtained simply by measuring DNase accessibility surrounding their motifs, so settler TF binding can be accurately determined solely based on chromatin accessibility in the absence of ChIP or DNase profile information. Pioneer TF binding can also be predicted a priori by local DNase accessibility (Figure 4-7), presumably a result of pioneer-induced chromatin opening at binding sites either in the profiled developmental stage or at a prior timepoint. Thus, we have identified a class of settler TFs that to our knowledge has not been described that obey one simple rule, binding DNA when chromatin is open, establishing settler TFs as a class whose binding is directly dependent on the chromatin opening ability of pioneer TFs.

Chapter 5

Conclusions

We have proposed a new framework for analyzing DNase-seq data for transcription factor binding sites and determining their statistical significance. The methodology extends existing approaches to DNase-seq analysis by using machine learning techniques to control for noisy data while using the concept of background regions to determine the binding classifier as well as asses statistical significance. This differs substantially from the mostly un-supervised approaches in the literature which uses only the PWM match regions to determine binding. We believe that the use of background regions, as well as the covariance correction substantially increases the reliability of our calls.

Biologically, this improvement has led to the ability to detect motifs that open chromatin whenever they appear, which we classify as pioneer transcription factors, as well as a complementary set of proteins which we term settlers which bind at regions of chromatin that are already open. Various biological assays including reporter screens for DNase and ChIP qPCR match computational predictions and show that our transcription factor classes behave as expected.

5.1 Contributions

We have three primary contributions. The first of our contributions is pradigmatic; we focus upon assessing the statistical significance of binding calls using background tracks. This turns what was originally an unsupervised clustering problem into a semisupervised one: we can consider the motif-match set as a mixture of both positive and negative examples while the background set is purely composed of negative examples.

This supervised framework allows us to determine the binding threshold, as well as the possibility that a binding event occurs due to sequence bias rather than due to a transcription factor binding event.

Our second contribution is methodological: we extend current research in the fitting of Gaussian process time series to deal with a single extremely long stationary sequence. Exploiting both the block-toeplitz structure and the sparsity of the observed Poisson process allows us to obtain runtimes that are thousands of times faster than previously reported in the literature.

Finally, our two computational contributions result in a biological finding regarding the binding of transcription factors. At least for a subset of factors we can model the binding of transcription factors using two rules: pioneers bind to their motif regardless of accessibility and open local chromatin while settlers recognize motif sites that are accessible and bind.

5.2 Future directions

Future extension of PIQ is in two directions: first we can extend PIQ to deal with more structured data such as differential or time-series DNase-seq data. This requires more rigorous statistical analysis of change in binding.

Another direction is to consider PIQ as a first order attempt to understand phenotype (chromatin accessibility) using sequence (motif matches). This framework of sequence to phenotype may allow us to model the semantic structure of the genome in terms of measurable high-throughtput sequencing data.

Bibliography

[1] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–22, 2008. 1097-4172 (Electronic) 0092-8674 (Linking) Journal Article Research Support, N.I.H., Extramural Research Support, N.I.H., Intramural Research Support, U.S. Gov't, Non-P.H.S.

[2] A. P. Boyle, L. Song, B. K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford, and T. S. Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res*, 21(3):456–64, 2011. 1549-5469 (Electronic) 1088-9051 (Linking) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't.

[3] E. H. Davidson. Emerging properties of animal gene regulatory networks. *Nature*, 468(7326):911–20, 2010.

[4] J. Eddy, A. C. Vallur, S. Varma, H. Liu, W. C. Reinhold, Y. Pommier, and N. Maizels. G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. *Nucleic Acids Res*, 39(12):4975–83, 2011. 1362-4962 (Electronic) 0305-1048 (Linking) Journal Article Research Support, N.I.H., Extramural Research Support, N.I.H., Intramural Research Support, Non-U.S. Gov't.

[5] Y. Guo, S. Mahony, and D. K. Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol*, 8(8):e1002638, 2012. 1553-7358 (Electronic) 1553-734X (Linking) Journal Article Research Support, N.I.H., Extramural.

[6] Y. Guo, G. Papachristoudis, R. C. Altshuler, G. K. Gerber, T. S. Jaakkola, D. K. Gifford, and S. Mahony. Discovering homotypic binding events at high spatial resolution. *Bioinformatics*, 26(24):3028–34, 2010. 1367-4811 (Electronic) 1367-4803 (Linking) Journal Article Research Support, N.I.H., Extramural.

[7] S. John, P. J. Sabo, R. E. Thurman, M. H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager, and J. A. Stamatoyannopoulos. Chromatin accessibility predetermines glucocorticoid receptor binding patterns. *Nat Genet*, 43(3):264–8, 2011.

[8] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–502, 2007. 1095-9203 (Electronic) 0036-8075 (Linking) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't.

[9] T. Minka. Propagation for approximate bayesian inference. *UAI'01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

[10] S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J. MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul, and J. A. Stamatoyannopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012. 1476-4687 (Electronic) 0028-0836 (Linking) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.

[11] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome Res*, 21(3):447–55, 2011. 1549-5469 (Electronic) 1088-9051 (Linking) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't.

[12] T. Sekiya, U. M. Muthurajan, K. Luger, A. V. Tulin, and K. S. Zaret. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor foxa. *Genes Dev*, 23(7):804–9, 2009.

[13] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012. 1476-4687 (Electronic) 0028-0836 (Linking) Journal Article Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S.

[14] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K.

Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012. 1476-4687 (Electronic) 0028-0836 (Linking) Journal Article Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S.

[15] E. Trompouki, T. V. Bowman, L. N. Lawton, Z. P. Fan, D. C. Wu, A. DiBiase, C. S. Martin, J. N. Cech, A. K. Sessa, J. L. Leblanc, P. Li, E. M. Durand, C. Mosimann, G. C. Heffner, G. Q. Daley, R. F. Paulson, R. A. Young, and L. I. Zon. Lineage regulators direct bmp and wnt pathways to cell-specific programs during differentiation and regeneration. *Cell*, 147(3):577–89, 2011. 1097-4172 (Electronic) 0092-8674 (Linking) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't.

[16] H. Weintraub and M. Groudine. Chromosomal subunits in active genes have an altered conformation. *Science*, 193(4256):848–56, 1976. 0036-8075 (Print) 0036-8075 (Linking) Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

[17] C. Wu. The 5' ends of drosophila heat shock genes in chromatin are hypersensitive to dnase i. *Nature*, 286(5776):854–60, 1980. 0028-0836 (Print) 0028-0836 (Linking) Journal Article Research Support, U.S. Gov't, P.H.S.

[18] R. A. Young. Control of the embryonic stem cell state. *Cell*, 144(6):940–54, 2011. 1097-4172 (Electronic) 0092-8674 (Linking) Journal Article Review.