Statistical Approach for Functionally Validating Transcription Factor Bindings Using Population SNP and Gene Expression Data

Jing Xiang

September 2017
CMU-ML-17-104

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Seyoung Kim, Chair
Geoff Gordon
Carl Kingsford
Steffi Oesterreich

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was sponsored by the National Science Foundation grant number MCB1149885, the National Institutes of Health grant number R01GM093156, the Pennsylvania Department of Health's Big Data for Better Health (BD4BH) initiative, the Pittsburgh Life Sciences Greenhouse's Facilities and Opportunity Fund, and a gift from Life Technologies.

Keywords: genomics, gene expression, microarray, transcriptional gene regulation, gene regulatory networks, perturbation studies, sparse regression, probabilistic graphical models, structure learning, estrogen receptor, breast cancer

If the map doesn't agree with the ground, the map is wrong. $\sim$ *Gordon Livingston*

Abstract

Understanding transcriptional gene regulation is an important step to understanding how essential mechanisms are controlled in biological systems. Functional assays such as ChIP-seq and DNase I have been used to obtain a binding map of transcription factor (TF) binding sites on DNA and to determine the transcriptional regulatory network of TFs and their target genes. However, binding alone may not result in a change in target gene expression. Experimental approaches to identifying functional binding events involve performing artificial TF knockdown experiments or genome editing [31, 45, 70] and then declaring the differentially expressed genes as functionally validated target genes. Instead of artificial perturbation, in order to functionally validate the TF binding map, we propose to leverage the naturally-occurring genetic variations as the source of perturbations that vary gene expressions and to analyze population single nucleotide polymorphism (SNP) and gene expression data. Experimental approaches typically target either a single TF or a family of TFs. In addition, in a single experiment, you must choose whether to perturb TF concentration through RNA interference or CRISPR interference, or TF binding affinity through genome editing. However, our approach is potentially more powerful because any aspects of the TF-target interaction, including TF concentration and TF binding affinity, can be perturbed by a large number of SNPs found across the genome simultaneously and the effects are learned in a single analysis.

In this thesis, we first introduce a statistical approach, based on conditional Gaussian Bayesian networks, that integrates population SNP and gene expression data with TF binding data to validate the TF binding map. We developed an efficient learning algorithm for learning the gene regulatory network by using TF binding data as prior knowledge, and selecting the TF-target interactions that are validated based on population SNP and gene-expression data. Given the estimated network, we perform inference on the estimated probabilistic graphical models to determine downstream genes that are differentially expressed due to the effect of the TF-target interactions.

We apply our method to learn transcriptional regulatory networks in lymphoblastoid cell lines (LCLs) and breast cancer tumours. First, we demonstrate our approach for validation of the TF binding map derived from ENCODE DNase I and ChIP-seq data from 71 TFs in LCLs, with SNP and gene expression data from the 1000 genomes and HapMap 3 projects respectively. We examined functional target genes that were validated under perturbation of TF concentration and TF binding affinity. Finally, we apply our method to perform TF binding map validation for ER and its coregulators which include 38 TFs obtained from Cistrome TF binding data, by using The Cancer Genome Atlas SNP and expression data from breast cancer tumors. We identified many previously known interactions between ER and its coregulators. We also found expression quantitative trait loci (eQTLs) in local binding regions of target genes that are potential super enhancers and eQTLs in coding regions that may affect the protein structure of important regulators.

Acknowledgments

I would like to thank my advisor Seyoung Kim for her guidance and patience. I am stronger academically and personally because of my experience working with you. To Geoff Gordon, your help with navigating this process was invaluable. Thanks for all your efforts towards understanding my research and giving helpful feedback. To Carl Kingsford, thanks for adopting me into your academic family when I needed it and for asking all the hard questions. Finally to Steffi Oesterreich, thanks for all your insights, and for expanding my knowledge of cancer biology.

To Ryan Tibshirani, Aarti Singh and Roy Maxion, thanks so much for your encouragement. It's been such a privilege to get to know you this past year.

To Diane Stidle, Jamie Rossi and Angie Lusk, you were so supportive of me and so many amazing things happened because of your presence and guidance.

To the eccentric friends of my PhD life, Joey Gonzalez, Sue Ann Hong, Alona Fyshe, Min Xu, Ankur Parikh, Alnur Ali, Hao Wang, Benjo Cowley, Manzil Zaheer, and the ladies and gents of the original CMU PhD support group, thanks for all the support and the good times.

Finally, to my family, thanks for always being my anchor, no matter how many boxes I live out of as I travel around on this crazy journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

One of the ultimate challenges in biology is to understand the mechanisms that underly cellular processes. A large part of this is comprehending gene regulation, which is basically how the cell controls the production of proteins that carries out these mechanisms. An important control point is transcription during which the binding of transcription factors (TFs) determines which genes are transcribed. The protein products of these regulated genes then go on and trigger other mechanisms and have downstream effects. One primary goal of studying gene regulation is to elucidate the TF to target gene relationships and the resulting downstream pathways.

Recently, many experimental procedures have been developed to construct a genome wide binding map of TF binding sites on DNA. These include approaches such as chromatin immunoprecipitation sequencing (ChIP-seq) [40, 108] and DNase I sequencing (DNase I) [40] . These techniques help identify the binding events that occur between TFs and target genes which construct a TF binding map. However, many of the binding events do not result in a change in target gene expression. Thus, functional validation is required to determine whether the TF binding affects target gene expression. One common approach for such functional validations is to perform artificial gene knockdown experiments. In this kind of experiment, a TF's expression is reduced, most commonly via RNA interference (RNAi) [94] and more recently via CRISPR interference (CRISPRi) [31]. Another recent method is to perform genome editing using the CRISPR-Cas9 system [7] where a particular locus can be modified, such as a regulatory region of a TF [31, 45, 70]. For both experimental approaches, the genes that are differentially expressed as a result of the perturbation are determined. The well-known limitation of these approaches is that the differential expression of a gene does not necessarily mean that the gene is being directly targeted by the TF, but can be an indirect downstream effect.

Instead of using experimental approaches to perturb gene expression, we propose to validate the TF binding map by using naturally occurring genetic variants as the source of perturbations that vary the gene expression levels. We accomplish this by analysis of population single nucleotide polymorphism (SNP) and gene expression data. Using SNP perturbations for functional validation has several advantages over the experimental methods. First, a large number of SNPs are found across the genome, and can perturb any aspect of the TF-target interaction simultaneously. For instance, in addition to capturing the effects of concentration perturbations, we can also capture SNP perturbations of binding affinity through changes in TF binding sites or TF binding domains. In contrast, different experiments must be performed for perturbing concentration such as RNAi or CRISPRi, and perturbing TF binding sites and binding domains such as genome editing. Furthermore, while our approach can perturb the TF-target interaction for many TFs simultaneously, experimental methods can only target a single TF or family of TFs at a time. Another advantage is since expression quantitative trait loci (eQTL) mapping [4, 44] with population gene expression and SNP data is widely used to study the genetic architecture of various diseases and tissues types, it is easy to leverage existing data for our approach. This also means that we can perform functional validation whenever SNP and gene expression data are available, and that we are not restricted to particular cell lines or model organisms as is the case for the experimental approaches.

The main computational challenge for using SNP perturbations to functionally validate TF bindings is to decouple the large number of SNP perturbations that are simultaneously affecting all genes. To address this challenge, we present a statistical approach based on conditional Gaussian Bayesian networks to validate the TF binding map, constructed from TF binding sites, by learning the functional TF-target interactions from SNP and expression data. The statistical model incorporates TF binding map as prior information, and then the learning algorithm selects the TF-target interactions that are validated under SNP perturbations. We develop a learning algorithm that can learn the gene regulatory networks under SNP perturbations efficiently and accurately [161]. The validated target genes that are directly regulated by the TF-target interaction are identified in the model as edges in the network, whereas we can perform inference on the model to infer the indirect downstream effects of TF-target interactions. Thus, we are able to overcome the well-known limitation of the standard experimental approach and distinguish between direct targets and indirect differentially expressed genes. We demonstrate our method on ChIP-seq and DNase I data from the Encyclopedia of DNA Elements (ENCODE) Project [22] and SNP and expression data from the 1000 Genomes [20] and HapMap 3 projects [97, 140], all collected for lymphoblastoid cells. In addition, we apply our approach to TF binding data of estrogen receptor and its coregulators, and population SNP and expression data collected for breast cancer cells from the Cancer Genome Atlas network (http://cancergenome.nih.gov/).

1.2 Thesis goals

We now define the subgoals to address the problem of developing a statistical model that can model gene regulation with population SNP and expression data.

1. **Use A* search for Gaussian Bayesian network learning.** We select Bayesian networks to model the directional relationships between genes in the regulatory network and propose a new learning algorithm called A* lasso. A* lasso learns a sparse Bayesian network structure of TFs, target genes and downstream genes. For a small number of nodes in the network, A* lasso recovers the optimal sparse Bayesian network structure by solving a single optimization problem with the A* search algorithm, using lasso in its scoring system. For larger networks, we suggest a heuristic scheme that dramatically reduces computational time without substantially compromising the quality of solutions.

2. **Use conditional Gaussian Bayesian network to determine functional TF bindings by integrating TF binding data with gene expression and SNP data.** We construct a method that validates TF to target binding events from ChIP-seq and DNase I data with population gene expression data under SNP perturbations. We extend A* lasso so that it can estimate a conditional Gaussian Bayesian network where the nodes are gene-expressions conditioned on SNPs and the TF binding map is integrated as prior knowledge.
3. **Seek biological evidence for TF to target interactions validated by perturbing concentration and by perturbing binding affinity in human lymphoblastoid cells.** We investigate how the TF to target relationships that are validated by perturbing TF concentration compare with knockdown experiments and analyze whether SNP perturbations indeed disrupt binding affinity of TFs resulting in differential gene expression of target genes.
4. **To investigate transcriptional regulation of estrogen receptor (ER), its coregulators, and targets in breast cancer tumor cells.** We apply our method of validating the TF binding map to data obtained from breast cancer cell lines. We build the TF binding map from ENCODE ChIP-seq and DNase I hypersensitivity data, and validate it with TCGA gene expression and SNP data. We perform detailed analysis of the gene regulatory network for ER and its coregulators. We also investigate the SNPs that affect target gene expression and look for potential enhancers. Finally, we identify SNPs in coding regions of TF and ER upstream regulators and examine their effect on protein structure.

1.3 High-throughput omics data

The subgoals presented above depend on several data sources. Instead of validating TF binding events with experimental approaches, we propose to use a statistical model based on population SNP and gene expression data. Population SNP and gene expression data is often collected for eQTL mapping, which is a popular approach for studying the effects of genetic variation on gene expressions. In our approach, we can harness the information that has already been collected to train our statistical model. A brief description of the data sources is provided below.

1.3.1 Microarray expression profiling

Microarrays are used to measure gene expression from a particular cell or tissue. For population expression data, we will be using microarray data from the HapMap 3 population [97, 140]. It is very useful for genome-wide studies because this technology allows the expression of tens of thousands of genes to be quantified simultaneously on one chip. A microarray experiment begins with the extraction of messenger RNA (mRNA) from cells or tissue. This mRNA is then amplified, converted to cDNA, labeled with fluorescent dyes and hybridized to the microarray chip. Each chip contains thousands of DNA probes which contain short sequences that will hybridize with the prepared sample. After hybridization, the arrays are then placed in a specialized scanner that will quantify the intensity of each probe. The scanned images are then analyzed using software that can subtract background noise, normalize the data to generate output expression levels for each gene. A detailed review of microarray technology is available from [126].

While microarray technology provides genome wide expression levels at a low cost, it also has certain limitations. There is a lot of noise in the measurement due to cross-hybridization [105, 119] and there is limited dynamic range of detection due to noise and saturation of signals [156]. In addition, detection of different isoforms is usually not possible. This has lead to the development of newer technologies such as RNA sequencing.

1.3.2 RNA sequencing

RNA sequencing (RNA-seq) is part of a suite of next-generation sequencing (NGS) tools that have recently become available, and is used to measure gene expression [89, 107, 156]. This technology is rapidly replacing microarrays and we will be using this type of data for analysis of gene regulation in cancer. In an RNA-seq experiment, the sample mRNA is converted into a library of cDNA fragments. High-throughout sequencing technology is then used to sequence each fragment. The resulting reads are then aligned to a reference genome. The reads are then quantified and statistical tools are used to normalize the data producing the expression levels for each gene.

Although RNA-seq has slightly greater experiment costs and computational costs due to requirement of sequencing and alignment, it has advantages over expression profiling by microarray. It has very low noise in the signal because it does not suffer from cross-hybridization in the experiment and it does not have an upper limit for quantification since it is counting sequences. Thus, a larger dynamic range of expression levels can be detected compared to microarrays. Another important advantage is that RNA-seq can detect transcripts that do not correspond to existing genomic sequence. Instead of mapping reads to a reference genome, they can be mapped to a de novo assembly of the transcriptome. This allows for the identification of novel transcripts and is useful for studies on non-model organisms [107].

1.3.3 Chromatin immunoprecipitation sequencing

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a technique used to identify the genomic locations where DNA-binding proteins such as TFs or histones bind [40, 108]. We use publicly available ChIP-seq data from the ENCODE project [22] to determine candidate TF to target gene relationships that we can then validate. These DNA-binding proteins bind preferentially to specific DNA sequences. However, these locations cannot be predicted using DNA sequence features alone. Thus, ChIP-seq is a way of finding these binding locations in cells. In a typical ChIP-seq experiment, a specific DNA-binding protein that associates with the DNA fragments are cross linked to the DNA *in vivo* by treatment with formaldehyde. The chromatin is then broken up by sonication resulting in fragments of DNA. Next, an antibody specific to the protein is used to immunoprecipitate the DNA-protein complex. Finally, the cross links are removed and the free DNA that was bound by the protein is sequenced. The resulting sequences are then aligned to the reference genome to find out where exactly in the genome the protein was found.

1.3.4 DNase I hypersensitive sites

DNase I digestion followed by sequencing [40] is a technique to detect open chromatin, these are regions where TFs can bind and are likely to have gene regulatory functions. The DNase I endonuclease preferentially digests unbound, open chromatin. The DNase I hypersensitive sites correspond to regions of unbound chromatin, free of nucleosomes. The DNase-seq experiments will determine the regions that are enriched for DNase I hypersensitive sites. We use this data from the ENCODE project [22] in addition to ChIP-seq data to determine the candidate TF binding map that is to be validated.

1.3.5 SNP data

Single nucleotide polymorphisms (SNPs) are variations in a single nucleotide that occurs at many positions in the genome. These genetic variations are identified for individuals in a population because they can potentially result in differences in both protein products and regulatory sequences. We use publicly available population SNP data from the 1000 Genomes Project [20]. In addition, for cancer analysis, we will use data from the TCGA Research Network: http://cancergenome.nih.gov/. SNP data can be generated by using SNP arrays or next generation sequencing [93] (NGS). SNP arrays can genotype thousands of previously known SNPs in one experiment [76]. In NGS methods, the genome is digested in small fragments that get sequenced and aligned to a reference genome. Once the fragments of individuals of population are aligned to a reference genome, the first step is to do variant calling. Next, genotype calling is performed which identifies the correct genotype for each individual at each site. Then SNP loci are determined by comparing the genome sequences across samples. The details of this process is reviewed in [104].

1.3.6 Knockdown experiments with RNA interference

One common method of determining whether TF binding is functional is by gene knockdown of the TF. A popular technique of achieving gene silencing is by RNA interference (RNAi) [94]. We use the RNAi experiments done by Cusanovich *et al.* [26] for a HapMap lymphoblastoid cell line and compare the results with our population-based approach. RNAi gene knockdown involves using siRNAs to target and silence TF expression. These experiments must be done for a single TF and a single cell line at a time. After TF knockdown, a profiling technique such as microarray or RNA-seq can be used to measure the gene expressions of the sample. The success of these experiments is dependent on siRNA targeting, that is the ability to silence the target TF and have relatively few off-target effects. Thus, when using this technique many knockdown experiments may be discarded due to low efficiency. In the study by Cusanovich *et al.* [26], out of 112, 53 TF knockdown experiments were discarded for low efficiency.

1.4 Background on computational methods

1.4.1 Bayesian networks

Bayesian networks are widely used for modeling complex distributions over a large number of variables, where the dependency between variables are represented as directed edges [75]. A Bayesian network provides a compact representation of a probability distribution. This distribution can be factorized by using its corresponding directed acyclic graph (DAG). Given a Bayesian network, inference can be performed to find probabilities of certain variables given observations of others. Bayesian networks are useful for modeling settings where the relationships between variables are directed and has often been used to model networks in biology [67].

1.4.2 Structure learning for Bayesian networks

Given the structure, estimating the parameters of a Bayesian network is straightforward, however learning the network structure is a difficult problem. In fact, learning the Bayesian network structure has been known to be an NP-hard problem [18] because of the constraint that the network structure has to be a directed acyclic graph (DAG). Many of the exact methods that have been developed for recovering the optimal structure are computationally expensive and require exponential computation time [71, 133]. Approximate methods based on heuristic search are more computationally efficient, but they recover a suboptimal structure.

Many of the existing algorithms are based on scoring each candidate graph and finding a graph with the best score, where the score decomposes for each variable given its parents in a DAG. Although methods may differ in the scoring method that they use (e.g., MDL [77], BIC [127], and BDe [51]), most of these algorithms, whether exact methods or heuristic search techniques, have a two-stage learning process. In Stage 1, candidate parent sets for each node are identified while ignoring the DAG constraint. Then, Stage 2 employs various algorithms to search for the best-scoring network structure that satisfies the DAG constraint by limiting the search space to the candidate parent sets from Stage 1. For Stage 1, methods such as sparse candidate [36], max-min parents children [151], and total conditioning [109] algorithms have been previously proposed. For Stage 2, exact methods based on dynamic programming [71, 133] and A* search algorithm [166] as well as inexact methods such as heuristic search technique [151] and linear programming formulation [63] have been developed. These approaches have been developed primarily for discrete variables, and regardless of whether exact or inexact methods are used in Stage 2, Stage 1 involves exponential computation time and space.

For continuous variables, L_1-regularized Markov blanket (L1MB) [124] was proposed as a two-stage method that uses lasso to select candidate parents for each variable in Stage 1 and performs heuristic search for DAG structure and variable ordering in Stage 2. Although a two-stage approach can reduce the search space by pruning candidate parent sets in Stage 1, Huang *et al.* [59] observed that applying lasso in Stage 1 as in L1MB is likely to miss the true parents in a high-dimensional setting, thereby limiting the quality of the solution in Stage 2. They proposed the sparse Bayesian network (SBN) algorithm that formulates the problem of Bayesian network structure learning as a single-stage optimization problem and transforms it into a lasso-type optimization to obtain an approximate solution. Then, they applied a heuristic search to refine the solution as a post-processing step. We develop a single-stage algorithm that recovers the optimal solution with less computation time than previous exact algorithms. We modified our algorithm using heuristic schemes to make it even more computationally efficient and practical for learning large networks without significantly reducing the quality of the solution.

1.4.3 A* Search

A* search is a search algorithm that has been widely used in path-finding. In A* search, the goal is to find the shortest path between the start node and the goal node, where the paths between nodes are weighted. To find the optimal path, A* search will construct a tree of paths from the start node, expanding nodes at every step which result in partial paths, until one of the paths reaches the goal node. At each iteration, A* will investigate each partial path and estimate the total cost to reach the goal. A* will then select the partial path that minimizes that cost. This cost is represented by:

$$f = g + h \tag{1.1}$$

where f is the estimate of the total cost, g is the exact cost of the path from the start node to the current node, and h is the heuristic estimate of the cost from the current node to the goal. The heuristic function is defined for each specific problem. To find the optimal path, the only property that it must satisfy is that it is admissible, which means that it never overestimates the actual cost to get to the goal node. If the heuristic also satisfies the property of consistency, then the optimal path to a node is always the first one followed. A heuristic is consistent if for two adjacent nodes x and y,

$$h(x) \leq c(x, y) + h(y). \tag{1.2}$$

This means that the estimated cost of reaching the goal from x is no greater than the step cost of getting to y plus the estimated cost of reaching the goal from y.

The most common way to implement A* search is with a priority queue. At each step, the node with the lowest cost f is selected popped off the queue, the f values of its neighbors are computed and added to the queue. This continues until a goal node has the minimum f value in the queue. The heuristic function h value is 0 at this point and this solution has the shortest path. A full description of A* search is provided in Russell and Norvig [120]. We incorporate A* search in our learning algorithm to significantly prune the search space while guaranteeing the optimality of the solution. We then further limit the search space of A* search with heuristic schemes to improve computation time of our algorithm for large networks.

1.4.4 Learning gene regulatory networks from expression data

Constructing gene regulatory networks is an important step in understanding the mechanisms behind biological processes. In the post genomic era, gene expression data is available for the whole genome which facilitates the understanding of organisms at the organism level, instead of pathways that involved a only few genes. However, this requires computational methods in order to model a network over a large number of genes. In this section, we review modern approaches to estimating gene regulatory networks from gene expression data. Further information is provided in review articles (e.g., [67], [154], [5]).

Logical networks

One of the earliest and simplest modeling strategies was to use logical models [154]. This was introduced by Kauffmann and Thomas in 1973 [46, 146]. Their application was to reconstruct the regulatory network that controlled the development of sea urchin embryos. Logical models have variables such as gene or proteins, and represent the state of the variables at a discrete level. The system is then assumed to update at synchronous time steps. This type of modeling is useful when the data is qualitative. There are also variations of logic models that include fuzzy logic [98].

Boolean networks

Another simple strategy for modeling gene regulatory networks are Boolean networks [154, 162]. In Boolean networks, each variable such as a gene or protein takes on two possible values, on or off. It is a directed graph of a set of binary variables called nodes and each node's value at a particular time is determined by its parents through a Boolean function. The states of all nodes are updated simultaneously (synchronously). Similar to logical networks, boolean networks can be useful when a small number of variables are available and the data is qualitative. The limitations of this method are that gene expression values must be discretized and that we must assume that states are changing synchronizing which is generally not true in biological systems. The structure of Boolean networks can be learned by observation of the gene expression measurements after perturbation experiments [61].

Bayesian Networks

As discussed previously, a common approach to building gene regulatory networks is using Bayesian networks. Bayesian networks are more flexible than logical networks and boolean networks allowing both discrete and continuous variables. They can be used to estimate static networks and dynamic bayesian networks (DBNs) can be used for temporal settings. As described previously, the challenge of using Bayesian networks is learning the structure. Nir Friedman demonstrated how to learn Bayesian networks from gene expression data and used it to model the gene regulatory networks of yeast [37]. A variation on Bayesian networks specifically to model regulatory relationships was proposed by Segal *et al.* called Module Networks [128, 129]. Module networks can be described as a Bayesian network in which the variables in the same module share parents and parameters. The authors propose a learning algorithm that finds sets of variables with similar behaviour that are then grouped as a module.

Correlation-based methods

Another category of methods to generate gene regulatory networks are correlation-based methods. These methods are based on defining a gene coexpression similarity matrix computed from the pairwise correlation coefficients between two genes and their expressions. Then, some thresholding technique is applyed to the similarity measures to determine whether the connections are meaningful or not [16, 78, 168]. While these methods are simple to implement are computationally efficiently, there is not a systematic method to select the threshold.

Partial-correlation based methods

In addition to correlation methods, there are partial-correlation based methods. These are based on gaussian graphical models where the conditional dependencies are inferred from the inverse of the covariance matrix, called the precision matrix. Two examples of these methods with different learning algorithms are SPACE [110], and GENENET [122]. Both correlation-based and partial-correlation based methods have the characteristic that they produce undirected networks which may not be appropriate for applications where the directed nature of gene regulation is preferred.

Information theoretic approaches

Finally, there are methods that use mutual information (MI) to determine the dependency amongst genes. ARACNE [88] is one of the most cited methods in this category. These methods also produce undirected networks but an advantage of information theory based methods is their ability to identify non-linear dependencies which will be missed by correlation-based methods.

Of the methods that build gene regulatory networks from continuous variables, we use Bayesian networks because it is most appropriate for our objectives. Because we are validating TF to target interactions, our application requires a directed network. Our model also facilitates the ability to incorporate the prior knowledge of TF binding data as candidate edges in the network. Our method is more flexible than module networks because we do not make assumptions about groups of genes. In addition, the sparsity of biological networks is incorporated directly in our model with the lasso penalty.

1.4.5 Constructing transcriptional regulatory networks from TF binding data

Transcription is an important control point of gene regulation. Thus, elucidating the transcriptional regulatory network has become a crucial step in understanding gene regulation. The transcriptional regulatory network consists of TFs which are proteins that bind to regulatory sequences in the DNA, which influence the transcription of target genes, affecting expression levels. As experimental techniques to profile TF binding such as chromatin immunoprecipitation (ChIP) techniques [157] such as ChIP-chip [14] and ChIP-seq [108] became available, it became possible to use this information to construct transcriptional networks more accurately.

Once TF binding was available, researchers began exploiting it to build transcriptional regulatory networks. Lee *et al.* used the genome-wide location analysis [117], which uses a modified ChIP-chip technique to identify where particular yeast TFs were bound. They use these TF-target relationships to build simple network motifs. These network motifs can then be used to form larger networks. They also integrate expression data to find groups of genes that are both bound and similarly expressed. Bar-Joseph and *et al.* [6] proposed a similar approach called Genetic Regulatory Modules (GRAM) which integrates TF binding data with gene expression data to construct gene modules of TFs and target genes. Their method was also applied to yeast data.

Recently, large amounts of TF binding data have been used to discover transcriptional relationships. Public consortiums such as ENCODE [22] have generated the functional data and projects like Cistrome [92] have provided at platform for organizing this data and making it easily accessible to researchers. For humans, the ENCODE project has facilitated the network analysis for 119 TFs solely based on TF binding data. Thorough analysis of the network found that the TFs coassociate in particular combinations near different targets. In addition, TFs coassociate in different patterns depending on whether the binding occurs close to the target gene or far away. TFs form a hierarchy where the middle level has the most information flow bottlenecks and the most regulatory collaboration between TFs. Detailed findings and conclusions are discussed in Gerstein *et al.*[42].

Perturbation experiments have been used in combination with TF binding data to confirm functional TF-target relationships. The methods discussed above primarily rely on the TF binding data to determine target genes. However, ChIP-seq data has false positives and thus does not guarantee functional binding. One way of addressing this issue is finding the overlap between bound genes and differentially expressed genes after TF knockdown. This is commonly done through RNA interference as demonstrated by Cusanovich *et al.* on lymphoblastoid cells [26]. There are several limitations of this approach. It is tedious because a perturbation experiment and subsequent microarray must be performed for each TF. In addition, the method is a single-gene perturbation and does not account for multiple perturbations. Furthermore, while a gene may be differentially expressed, this technique does not confirm that it was directly caused by perturbation of the TF, it could be a downstream effect.

In our approach, we replace the artificial perturbations by using the natural genetic variation as perturbations to the system. We then validate the TF-target interactions that occur under these SNP perturbations. By using a statistical approach to model all the SNP perturbations, we can handle multiple perturbations of each gene and we can distinguish between direct perturbations to the TF-target interaction and downstream effects.

1.4.6 Mapping expression quantitative trait loci (eQTLs) using variation and gene expression data

Driven by the completion of the human genome and subsequent cataloguing of all genetic variants in humans and other species, researchers invested significant efforts into understanding how these variants affect phenotypes. A specific group of variants, called expression quantitive trait loci (eQTLs) are those that influence the expression of genes. Doing genome-wide eQTL mapping was initially proposed in 2001 [65] and then carried out in yeast in 2002 [12] with the gene expression data from microarrays. Since then, there have been many studies on humans and other model organisms. Albert and Kryglyak's review article has summarized these early eQTL mapping studies [4]. Today, the availability of RNA-seq data and allele specific expression allows for isoform-specific eQTL mapping [141].

The problem of eQTL mapping has computational and statistical challenges because of the large number of genetic variants. SNP arrays can have up to a million SNPs and with the advances in modern sequencing techniques, there can be millions of SNPs produced by each study. Traditionally, eQTL analysis is carried out by performing an association test between the SNP and gene across individuals of a population. The association test method can be either a correlation or regression analysis. The association analysis is performed on each pair of SNP and gene separately.

Once the single SNP, single gene association tests are performed, a multiple testing correction must be performed. Many methods have been developed for multiple testing correction that vary in how stringent they are. For example, Bonferroni correction controls for Family Wise Error [56] but is too stringent in practice. More commonly, methods that control the False Discovery Rate (FDR) [139] and random permutations [11] are used.

In addition to association tests, there have been statistical learning approaches for eQTL mapping. One class of methods that was developed by the machine learning community is sparse linear regression methods.In this setting, the SNPs are the covariates and the gene expressions are the dependent variables. Because of the large number of SNPs, this problem generally suffers from the challenge of having a high-dimensional covariate matrix with the number of covariates being much larger than the number of samples. Sparse regression methods can be justified by assuming that only a few number of SNPs are associated for a particular gene expression. The lasso method is commonly used for eQTL analysis [148]. After running Lasso, only a few regression coefficients corresponding to the SNPs are non-zero. These are selected as eQTL associations.

Both association tests and performing Lasso assume that the SNPs and gene expressions are not correlated. There have been many methods that have been developed that are variations of Lasso that address the structure of the data. For example, elastic net regularization adds a ridge regression penalty which allows the method to select strongly correlated variables together [172]. Group lasso allows groups of covariates to be selected as a single unit [35]. Tree guided group lasso is used to solve the problem of multi-task regression that addresses the structure of the dependent variables [69]. For instance, the gene expressions can be represented as a tree structure. Fused lasso can be used in settings where the regression coefficents change in a smooth fashion. This makes sense when the same data is collected and varies temporally or spatially [149].

Mapping eQTLs is incorporated into our approach for ChIP-seq binding validation as we select the TF-target interactions that are validated under SNP perturbations. In order to do this, we implement a Gaussian Bayesian Network conditioned on SNPs. The SNPs that influence TF or target gene expressions are then selected by lasso as part of the optimization procedure.

1.4.7 Evaluating the effect of SNPs on gene expression regulation using CRISPR/Cas9 system

Recent work providing efficient site-specific genome editing with the CRISPR-Cas9 system has opened up a new avenue for studying gene regulation. While RNA interference can only modify TF expression, CRISPR-based screens can be used to study regulatory sequences, enhancer elements and non-coding sequences [7, 74, 116]. In addition, CRISPR-Cas9 engineered cells with different SNP genotypes can be used to study gene expression under a variety of conditions including drug treatment [114]. However, using this system for genome editing is expensive and time consuming since an experiment must be performed for each mutation and each cell line. It is much more efficient to identify possible regulatory SNPs computationally, refine the list of

SNPs and then using CRISPR-screening for final validation.

1.5 Thesis contributions

This thesis addresses the problem of validating the TF binding map constructed from experimental techniques. The existing experimental approaches use artificial perturbations via TF knockdowns or genome editing and then observe target gene expressions in order to find validated TF to target relationships. In our approach, we use genetic variation in populations as the source of perturbations of gene expressions levels. The main challenge of our method is how to model the large number of SNP perturbations that affect multiple genes simultaneously. In this thesis, we specify a statistical model, conditional Gaussian Bayesian networks and develop the learning algorithm, we demonstrate how to use the model to validate TF-to-target relationships on lymphoblastoid cell lines, and we then use the method to understand gene regulation of estrogen receptor and important cofactors in breast cancer.

Below, we describe the contributions of each thesis chapter.

Contributions of Chapter 2

- We propose a new single-stage algorithm called A* lasso that performs structure learning of a Gaussian Bayesian network (GBN) in a high dimensional setting. Our method finds the optimal network structure while significantly improving upon the computation time of the state-of-the-art optimal Bayesian network learning algorithm DP lasso, by using the A* search algorithm to prune the search space.
- We propose a heuristic scheme to further limit the search space for larger problems.
- In experiments, we demonstrate that A* lasso substantially improves computation time of previous optimal structure learning algorithms and that the heuristic scheme can substantially improve computation time without significantly compromising the quality of the solution.

Contributions of Chapter 3

- We extend the GBN of Chapter 2 to a conditional Gaussian Bayesian network (cGBN) so that we can model a network of gene expressions conditioned on SNPs, with candidate TF-to-target edges generated from the TF binding map being incorporated as prior knowledge.
- We demonstrate TF binding map validation on ENCODE ChIP-seq and DNase I data from lymphoblastoid cell lines by using SNP and gene expression data from the 1000 genomes and HapMap 3 projects.
- Unlike the experimental approaches to functional validation where the bound and differentially expressed genes are considered validated, we show that our method is able to distinguish between validated target genes and differentially expressed downstream genes.
- We show evidence that eQTLs detected near target genes disrupt binding affinity of TFs.

- We find a subset of eQTLs in coding regions of TFs which result in missense mutations that may affect TF protein structure.
- We compare the validated targets from our computational approach with those from knockdown studies. We show that epistatic interactions between TFs lead to different expression patterns of downstream genes between the two methods, which results in different sets of target genes.

Contributions of Chapter 4

- We demonstrate our computational approach on validating the TF binding map of ER and its coregulators in breast cancer tumours, derived from the Cistrome Project, by using SNP and expression data from the TCGA.
- From studying the gene regulatory network of ER and its coregulators, we find several interactions estimated from our model that were supported by experimental evidence in the literature.
- From analyzing the eQTLs in regulatory regions of target genes, we found eQTLs that are potentially super enhancers and act as important regulatory elements.
- By examining eQTLs in coding regions of protein kinases and TFs, we identified several missense SNPs that are near or contained within protein domains.

Chapter 2

A* Lasso: Using A* search for structure learning of Gaussian Bayesian networks

2.1 Motivation

In order to learn the directional relationships between genes in a regulatory network, we estimate a Gaussian Bayesian network structure of genes from gene expression data. We propose a new algorithm, called A* lasso, for learning a sparse Bayesian network structure with continuous variables in high-dimensional space. Our method is a single-stage algorithm that finds the optimal network structure with a sparse set of parents while ensuring the DAG constraint is satisfied. It significantly improves the computation time compared to the state-of-the-art optimal Bayesian network learning algorithm.

We first show that we can represent this structure learning problem as finding the shortest path from the start state to the goal state of a graph, and then show that this problem can be solved by familiar graph-search techniques. In particular, we show that this problem can be solved by dynamic programming (DP) where the scores are computed by incorporating a lasso-based scoring method (DP lasso). While previous approaches based on DP required identifying the exponential number of candidate parent sets and their scores for each variable in Stage 1 before applying DP in Stage 2 [71, 133], our approach effectively combines the score computation in Stage 1 within Stage 2 via lasso optimization. Because the number of states DP lasso must consider is exponential in the number of variables, it is not feasible for more than 20 nodes. Thus, we present A* lasso [161] which significantly prunes the search space of DP by incorporating the A* search algorithm [120], while guaranteeing the optimality of the solution. Since in practice, A* search can still be expensive compared to heuristic methods especially when trying to estimate large gene networks, we explore heuristic schemes that further limit the search space of A* lasso. We demonstrate in our experiments that this heuristic approach can substantially improve the computation time without significantly compromising the quality of the solution, especially on large Bayesian networks.

2.2 Problem definition: Bayesian network structure learning

A Bayesian network is a probabilistic graphical model defined over a DAG G with a set of $q = |V|$ nodes $V = \{v_1, \ldots, v_q\}$, where each node v_j is associated with a random variable Y_j [72]. For our purposes, the random variables Y_j represent genes. The probability model associated with G in a Bayesian network factorizes as $p(Y_1, \ldots, Y_p) = \prod_{j=1}^{p} p(Y_j|\text{Pa}(Y_j))$, where $p(Y_j|\text{Pa}(Y_j))$ is the conditional probability distribution for Y_j given its parents $\text{Pa}(Y_j)$ with directed edges from each node in $\text{Pa}(Y_j)$ to Y_j in G. We assume continuous random variables and use a linear regression model for the conditional probability distribution of each node $Y_j = \text{Pa}(Y_j)'\boldsymbol{\beta}_j + \epsilon$, where $\boldsymbol{\beta}_j = \{\beta_{jk}\text{'s for } Y_k \in \text{Pa}(Y_j)\}$ is the vector of unknown parameters to be estimated from data and ϵ is the noise distributed as $\sim N(0, 1)$.

Given a dataset $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_q]$, where $\mathbf{y}_j$ is a vector of n observations for random variable Y_j, our goal is to estimate the graph structure G and the parameters $\boldsymbol{\beta}_j$'s jointly. We formulate this problem as that of obtaining a sparse estimate of $\boldsymbol{\beta}_j$'s, under the constraint that the overall graph structure G should not contain directed cycles. Then, the nonzero elements of $\boldsymbol{\beta}_j$'s indicate the presence of edges in G. We obtain an estimate of Bayesian network structure and parameters by minimizing the negative log likelihood of data with sparsity enforcing L_1 penalty as follows:

$$\min_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p} \sum_{j=1}^{p} \parallel \mathbf{y}_j - \mathbf{Y}_{-j}'\boldsymbol{\beta}_j \parallel_2^2 + \lambda \sum_{j=1}^{p} \parallel \boldsymbol{\beta}_j \parallel_1 \quad \text{s.t. } G \in \text{DAG}, \tag{2.1}$$

where $\mathbf{Y}_{-j}$ represents all columns of $\mathbf{Y}$ excluding $\mathbf{y}_j$, assuming all other variables are candidate parents of node v_j. Given the estimate of $\boldsymbol{\beta}_j$'s, the set of parents for node v_j can be found as the support of $\boldsymbol{\beta}_j$, $S(\boldsymbol{\beta}_j) = \{v_i|\beta_{ji} \neq 0\}$. The λ is the regularization parameter that determines the amount of sparsity in $\boldsymbol{\beta}_j$'s and can be determined by cross-validation.

We notice that if the acyclicity constraint is ignored, Equation (3.2) decomposes into individual lasso estimations for each node:

$$\text{LassoScore}(v_j|V\backslash v_j) = \min_{\boldsymbol{\beta}_j} \parallel \mathbf{x}_j - \mathbf{x}_{-j}'\boldsymbol{\beta}_j \parallel_2^2 + \lambda \parallel \boldsymbol{\beta}_j \parallel_1,$$

where $V\backslash v_j$ represents the set of all nodes in V excluding v_j. The above lasso optimization problem can be solved efficiently with the shooting algorithm [39]. However, the main challenge in optimizing Equation (3.2) arises from ensuring that the $\boldsymbol{\beta}_j$'s satisfy the DAG constraint.

2.3 DP Lasso: Dynamic programming with Lasso as the score function

The problem of learning a Bayesian network structure that satisfies the constraint of no directed cycles can be cast as that of learning an optimal ordering of variables [72]. Once the optimal variable ordering is given, the constraint of no directed cycles can be trivially enforced by constraining the parents of each variable in the local conditional probability distribution to be a subset of the nodes that precede the given node in the ordering. We let $\Pi^V = [\pi_1^V, \ldots, \pi_{|V|}^V]$

denote an ordering of the nodes in V, where π_j^V indicates the node $v \in V$ in the jth position of the ordering, and $\Pi_{\prec v_j}^V$ denote the set of nodes in V that precede node v_j in ordering Π^V.

Algorithms based on DP have been developed to learn the optimal variable ordering for Bayesian networks [144]. These approaches are based on the observation that the score of the optimal ordering of the full set of nodes V can be decomposed into (a) the optimal score for the first node in the ordering, given a choice of the first node and (b) the score of the optimal ordering of the nodes excluding the first node. The optimal variable ordering can be constructed by recursively applying this decomposition to select the first node in the ordering and to find the optimal ordering of the set of remaining nodes $U \in V$. This recursion is given as follows, with an initial call of the recursion with $U = V$:

$$\text{OptScore}(U) = \min_{v_j \in U} \text{OptScore}(U \backslash v_j) + \text{BestScore}(v_j | V \backslash U) \tag{2.2}$$

$$\pi_1^U = \operatorname*{argmin}_{v_j \in U} \text{OptScore}(U \backslash v_j) + \text{BestScore}(v_j | V \backslash U), \tag{2.3}$$

where BestScore$(v_j|V\backslash U)$ is the optimal score of v_j under the optimal choice of parents from $V\backslash U$.

In order to obtain BestScore$(v_j|V\backslash U)$ in Equations (2.2) and (2.3), for the case of discrete variables, many previous approaches enumerated all possible subsets of V as candidate sets of parents for node v_j to precompute BestScore$(v_j|V\backslash U)$ in Stage 1 before applying DP in Stage 2 [71, 133]. While this approach may perform well in a low-dimensional setting, in a high-dimensional setting, a two-stage method is likely to miss the true parent sets in Stage 1, which in turn affects the performance of Stage 2 [59]. In this chapter, we consider the high-dimensional setting and present a single-stage method that applies lasso to obtain BestScore$(v_j|V\backslash U)$ within DP as follows:

$$\begin{aligned} \text{BestScore}(v_j|V\backslash U) &= \text{LassoScore}(v_j|V\backslash U) \\ &= \min_{\boldsymbol{\beta}_j, S(\boldsymbol{\beta}_j) \subseteq V\backslash U} \parallel \mathbf{x}_j - \mathbf{x}_{-j}{}'\boldsymbol{\beta}_j \parallel_2^2 + \lambda \parallel \boldsymbol{\beta}_j \parallel_1 . \end{aligned}$$

The constraint $S(\boldsymbol{\beta}_j) \subseteq V\backslash U$ in the above lasso optimization can be trivially maintained by setting the β_{jk} for $v_k \in U$ to 0 and optimizing only for the other β_{jk}'s. When applying the recursion in Equations (2.2) and (2.3), DP takes advantage of the overlapping subproblems to prune the search space of orderings, since the problem of computing OptScore(U) for $U \subseteq V$ can appear as a subproblem of scoring orderings of any larger subsets of V that contain U.

The problem of finding the optimal variable ordering can be viewed as that of finding the shortest path from the start state to the goal state in a search space given as a subset lattice (Fig. 2.1). The search space consists of a set of states, each of which is associated with one of the $2^{|V|}$ possible subsets of nodes in V. The start state is the empty set $\{\}$ and the goal state is the set of all variables V. A valid move in this search space is defined from a state for subset Q_s to another state for subset $Q_{s'}$, only if $Q_{s'}$ contains one additional node to Q_s. Each move to the next state corresponds to adding a node at the end of the ordering of the nodes in the previous state. The cost of such a move is given by BestScore$(v|Q_s)$, where $v = Q_{s'}\backslash Q_s$. Each path from the start state to the goal state gives one possible ordering of nodes. Figure 2.1 illustrates the search space, where each state is associated with a Q_s. DP finds the shortest path from the

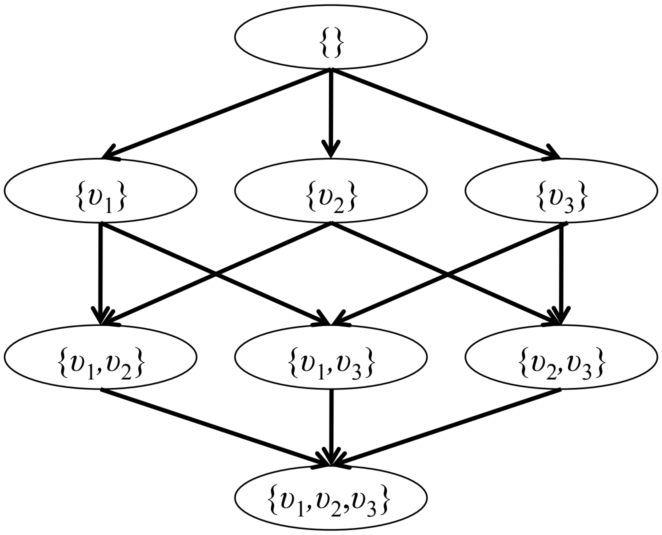

Figure 2.1: Search space of variable ordering for three variables $V = \{v_1, v_2, v_3\}$.

start state to the goal state that corresponds to the optimal variable ordering by considering all possible paths in this search space and visiting all $2^{|V|}$ states.

2.4 A* Lasso for Pruning Search Space

As discussed in the previous section, DP considers all $2^{|V|}$ states in the subset lattice to find the optimal variable ordering. Thus, it is not sufficiently efficient to be practical for problems with more than 20 nodes. On the other hand, a greedy algorithm is computationally efficient because it explores a single variable ordering by greedily selecting the most promising next state based on BestScore$(v|Q_s)$, but it returns a suboptimal solution. In this section, we propose A* lasso that incorporates the A* search algorithm [120] to construct the optimal variable ordering in the search space of the subset lattice. We show that this strategy can significantly prune the search space compared to DP, while maintaining the optimality of the solution.

When selecting the next move in the process of constructing a path in the search space, instead of greedily selecting the move, A* search also accounts for the estimate of the future cost given by a heuristic function $h(Q_s)$ that will be incurred to reach the goal state from the candidate next state. Although the exact future cost is not known until A* search constructs the full path by reaching the goal state, a reasonable estimate of the future cost can be obtained by ignoring the directed acyclicity constraint. It is well-known that A* search is guaranteed to find the shortest path if the heuristic function $h(Q_s)$ is *admissible* [120], meaning that $h(Q_s)$ is always an underestimate of the true cost of reaching the goal state. Below, we describe an admissible heuristic for A* lasso.

While exploring the search space, A* search algorithm assigns a score $f(Q_s)$ to each state s and its corresponding subset Q_s of variables for which the ordering has been determined. A* search algorithm computes this score $f(Q_s)$ as the sum of the cost $g(Q_s)$ that has been incurred so far to reach the current state from the start state and an estimate of the cost $h(Q_s)$ that will be incurred to reach the goal state from the current state:

$$f(Q_s) = g(Q_s) + h(Q_s). \tag{2.4}$$

More specifically, given the ordering Π^{Q_s} of variables in Q_s that has been constructed along the path from the start state to the state for Q_s, the cost that has been incurred so far is defined as

$$g(Q_s) = \sum_{v_j \in Q_s} \text{LassoScore}(v_j | \Pi^{Q_s}_{\prec v_j}) \tag{2.5}$$

and the heuristic function for the estimate of the future cost to reach the goal state is defined as:

$$h(Q_s) = \sum_{v_j \in V \backslash Q_s} \text{LassoScore}(v_j | V \backslash v_j) \tag{2.6}$$

Note that the heuristic function is admissible, or an underestimate of the true cost, since the constraint of no directed cycles is ignored and each variable in $V \backslash Q_s$ is free to choose any variables in V as its parents, which lowers the lasso objective value.

When the search space is a graph where multiple paths can reach the same state, we can further improve efficiency if the heuristic function has the property of *consistency* in addition to admissibility. A consistent heuristic always satisfies $h(Q_s) \leq h(Q_{s'}) + \text{LassoScore}(v_k | Q_s)$, where $\text{LassoScore}(v_k | Q_s)$ is the cost of moving from state Q_s to state $Q_{s'}$ with $\{v_k\} = Q_{s'} \backslash Q_s$. Consistency ensures that the first path found by A* search to reach the given state is always the shortest path to that state [120]. This allows us to prune the search when we reach the same state via a different path later in the search. The following proposition states that our heuristic function is consistent.

Proposition 1 *The heuristic in Equation (2.6) is consistent.*

Proof For any successor state $Q_{s'}$ of Q_s, let $v_k = Q_{s'} \backslash Q_s$.

$$\begin{aligned} h(Q_s) &= \sum_{v_j \in V \backslash Q_s} \text{LassoScore}(v_j | V \backslash v_j) \\ &= \sum_{v_j \in V \backslash Q_s, v_j \neq v_k} \text{LassoScore}(v_j | V \backslash v_j) + \text{LassoScore}(v_k | V \backslash v_k) \\ &\leq h(Q_{s'}) + \text{LassoScore}(v_k | Q_s), \end{aligned}$$

where $\text{LassoScore}(v_k | Q_s)$ is the true cost of moving from state Q_s to $Q_{s'}$. The inequality above holds because v_k has fewer parents to choose from in $\text{LassoScore}(v_k | Q_s)$ than in $\text{LassoScore}(v_k | V \backslash v_k)$. Thus, our heuristic in Equation (2.6) is consistent. ■

Given a consistent heuristic, many paths that go through the same state can be pruned by maintaining an *OPEN* list and a *CLOSED* list during A* search. In practice, the *OPEN* list can be implemented with a priority queue and the *CLOSED* list can be implemented with a hash table.

```
Input   : $\mathbf{X}$, $V$, $\lambda$
Output: Optimal variable ordering $\Pi^V$
Initialize OPEN to an empty queue;
Initialize CLOSED to an empty set;
Compute LassoScore($v_j|V\backslash v_j$) for all $v_j \in V$;
OPEN.insert(($Q_s = \{\}, f(Q_s) = h(\{\}), g(Q_s) = 0, \Pi^{Q_s} = [\,]$));
while true do
    ($Q_s, f(Q_s), g(Q_s), \Pi^{Q_s}$) $\leftarrow$ OPEN.pop();
    if h($Q_s$) = 0 then
        Return $\Pi^V \leftarrow \Pi^{Q_s}$;
    end
    CLOSED $\leftarrow$ CLOSED $\cup\{Q_s\}$;
    foreach $v \in V\backslash Q_s$ do
        $Q_{s'} \leftarrow Q_s \cup \{v\}$;
        if $Q_{s'} \notin$ CLOSED then
            Compute LassoScore($v|Q_s$) with lasso shooting algorithm;
            $g(Q_{s'}) \leftarrow g(Q_s) +$ LassoScore($v|Q_s$);
            $h(Q_{s'}) \leftarrow h(Q_s) -$ LassoScore($v|V\backslash v$);
            $f(Q_{s'}) \leftarrow g(Q_{s'}) + h(Q_{s'})$;
            $\Pi^{Q_{s'}} \leftarrow [\Pi^{Q_s}, v]$;
            OPEN.insert($L = (Q_{s'}, f(Q_{s'}), g(Q_{s'}), \Pi^{Q_{s'}})$);
        end
    end
end
```

Algorithm 1: A* lasso for learning Bayesian network structure

The *OPEN* list is a priority queue that maintains all the intermediate results $(Q_s, f(Q_s), g(Q_s), \Pi^{Q_s})$'s for a partial construction of the variable ordering up to Q_s at the frontier of the search, sorted according to the score $f(Q_s)$. During search, A* lasso pops from the *OPEN* list the partial construction of ordering with the lowest score $f(Q_s)$, visits the successor states by adding another node to the ordering Π^{Q_s}, and queues the results onto the *OPEN* list. Any state that has been popped by A* lasso is placed in the *CLOSED* list. The states that have been placed in the *CLOSED* list are not considered again, even if A* search reaches these states through different paths later in the search.

The full algorithm for A* lasso is given in Algorithm 1. As in DP with lasso, A* lasso is a single-stage algorithm that solves lasso within A* search. Every time A* lasso moves from state Q_s to the next state $Q_{s'}$ in the search space, LassoScore($v_j|\Pi^{Q_s}_{\prec v_j}$) for $\{v_j\} = Q_{s'}\backslash Q_s$ is computed with the shooting algorithm and added to $g(Q_s)$ to obtain $g(Q_{s'})$. The heuristic score $h(Q_{s'})$ can be precomputed as LassoScore($v_j|V\backslash v_j$) for all $v_j \in V$ for a simple look-up during A* search.

2.5 Simulation experiments

We perform simulation experiments to evaluate the accuracy the estimated networks and measure the computation time of our method. We created a group of small networks under 20 nodes and obtained the structure of several benchmarks networks between 20 and 60 nodes from the Bayesian Network Repository. In addition, we used the tiling technique [152] to generate two networks of approximately 300 nodes so that we could evaluate our method on larger graphs. Given the Bayesian network structures, we set the parameters $\boldsymbol{\beta}_j$ for each conditional probability distribution of node v_j such that $\beta_{jk} \sim \pm Uniform[l, u]$ for predetermined values for u and l if node v_k is a parent of node v_j and $\beta_{jk} = 0$ otherwise. We then generated data from each Bayesian network by forward sampling with noise $\epsilon \sim N(0, 1)$ in the regression model, given the true variable ordering. All data were mean centered.

We compare our method to several other methods including DP with lasso for an exact method, L1MB for heuristic search, and SBN for an optimization-based approximate method. We downloaded the software implementations of L1MB and SBN from the authors' website. For L1MB, we increased the authors' recommended number of evaluations 2500 to 10 000 in Stage 2 heuristic search for all networks except the two larger networks of around 300 nodes (Alarm 2 and Hailfinder 2), where we used two different settings of 50 000 and 100 000 evaluations. We also evaluated A* lasso with the heuristic scheme with the queue sizes of 5, 100, 200, and 1000.

DP, A* lasso, and A* lasso with a limited queue size require a selection of the regularization parameter λ with cross-validation. In order to determine the optimal value for λ, for different values of λ, we trained a model on a training set, performed an ordinary least squares re-estimation of the non-zero elements of $\boldsymbol{\beta}_j$ to remove the bias introduced by the L_1 penalty, and computed prediction errors on the validation set. Then, we selected the value of λ that gives the smallest prediction error as the optimal λ. We used a training set of 200 samples for relatively small networks with under 60 nodes and a training set of 500 samples for the two large networks with around 300 nodes. We used a validation set of 500 samples. For L1MB and SBN, we used a similar strategy to select the regularization parameters, while mainly following the strategy suggested by the authors and in their software implementation. All methods were implemented in Matlab and were run on computers with 2.4 GHz processors.

We first compare A* lasso to DP lasso which both yield optimal solutions. We record both the number of states considered and the computational time. We used a dataset generated from a true model with $\beta_{jk} \sim \pm Uniform[1.2, 1.5]$. It can be seen Figure 2.2A that DP considers all possible states $2^{|V|}$ in the search space that grows exponentially with the number of nodes. It is clear that A* lasso visits significantly fewer states than DP lasso, visiting less than 10% of the number of states for the network with 20 nodes. As a result, the computation time required for A* lasso is also dramatically reduced compared to DP lasso (Fig. 2.2B).

For networks with greater than 20 nodes, the exact methods become computationally expensive. For these networks, we limit the size of the queue in A* lasso, and use a heuristic scheme to prune the queue. We compare this approach, with L1MB for heuristic search, and SBN for an optimization-based approximate method. For A* lasso with the heuristic scheme, we use queue sizes of 5, 100, 200, and 1000. Compared with exact methods, A* lasso with heuristic scheme reduces both number of states visited and computation time. For instance, A* lasso with a queue

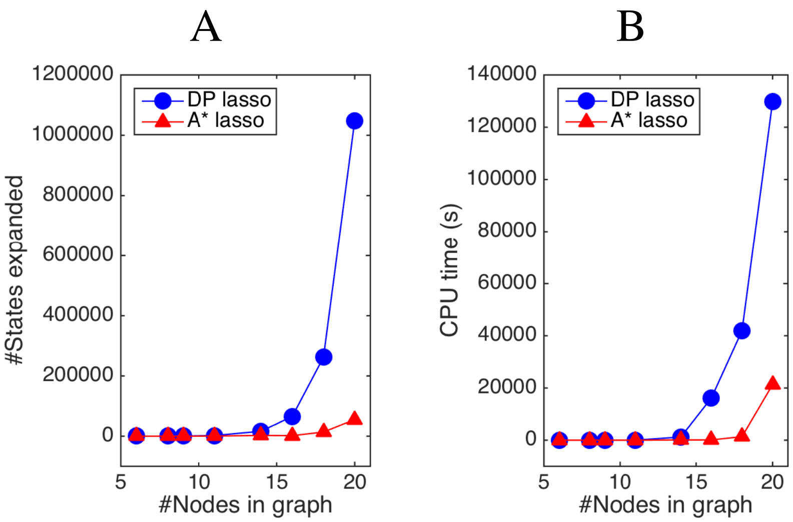

Figure 2.2: Comparison of DP lasso and optimal A* lasso on simulated data. A) The number of states expanded. B) Computational time.

limit of 1000 expands 0.2% of the states of DP lasso and takes 0.1% of the time for the network with 20 nodes. Figure 2.3 shows the number of states expanded for A* lasso of various limits and Figure 2.4 shows the computational time for A* lasso with various limits and the competing methods L1MB and SBN. We note that the computation time for A* lasso with a small queue of 5 or 100 is comparable to that of L1MB and SBN.

In general, we found that the extent of pruning of the search space by A* lasso compared to DP depends on the strengths of edges ($\boldsymbol{\beta}_j$ values) in the true model. We applied DP and A* lasso to datasets of 200 samples generated from each of the networks under each of the three settings for the true edge strengths, $\pm Uniform[1.2, 1.5]$, $\pm Uniform[1, 1.2]$, and $\pm Uniform[0.8, 1]$. As can be seen from the computation time and the number of states visited by DP and A* lasso in Figures 2.5 as the strengths of edges increase, the number of states visited by A* lasso and the computation time tend to decrease. The results in Figure 2.5 indicate that the efficiency of A* lasso is affected by the signal-to-noise ratio.

In order to evaluate the accuracy of the Bayesian network structures recovered by each method, we make use of the fact that two Bayesian network structures are indistinguishable if they belong to the same equivalence class, where an equivalence class is defined as the set of networks with the same skeleton and v-structures. The skeleton of a Bayesian network is defined as the edge connectivities ignoring edge directions and a v-structure is defined as the local graph structure over three variables, with two variables pointing to the other variables (i.e., $A \rightarrow B \leftarrow C$). We evaluate the performance of the different methods by comparing the estimated network structure with the true network structure in terms of skeleton and v-structures and computing the precision and recall.

The precision/recall curves for the skeleton and v-structures of the models estimated by the different methods are shown in Figure 2.6 and Figure 2.7, respectively. Each curve was obtained as an average over the results from 30 different datasets for the large graphs (Alarm 2 and Hailfinder 2) and from 50 different datasets for the smaller graphs (Barley and Hailfinder). The data

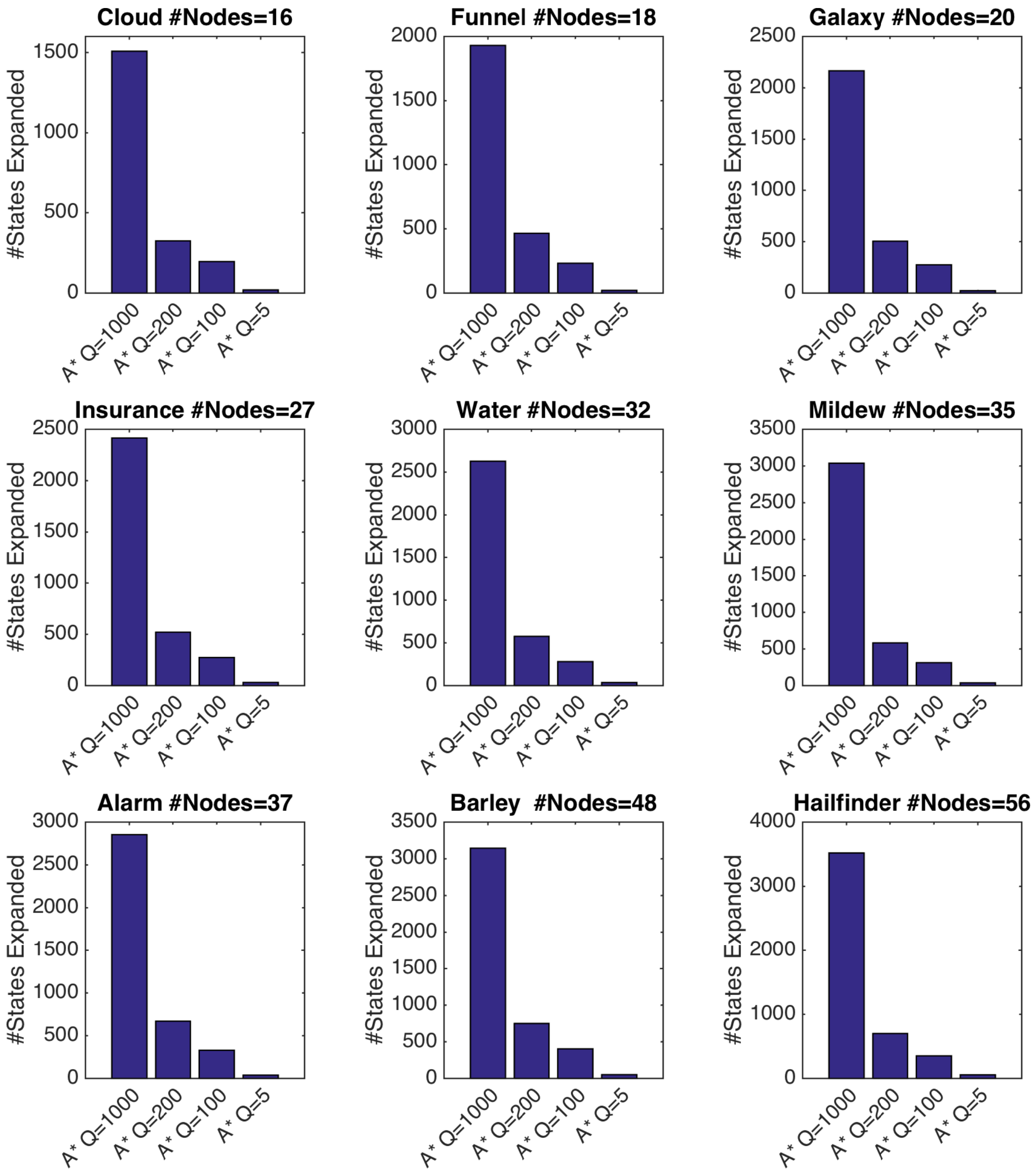

Figure 2.3: Comparison of the number of states required for Bayesian network structure learning algorithms on data simulated from benchmark Bayesian networks..

was simulated under the settings $\beta_{jk} \sim \pm Uniform[0.5, 1]$ and $\beta_{jk} \sim \pm Uniform[0.4, 0.7]$ respectively. For the benchmark Bayesian networks, we used A* lasso with different queue sizes, including 100, 200, and 1000, whereas for the two large networks (Alarm 2 and Hailfinder 2) that require more computation time, we used A* lasso with queue size of 5 and 100. For L1MB, 10000 evaluations were used for benchmark networks, and we used two different settings of 50

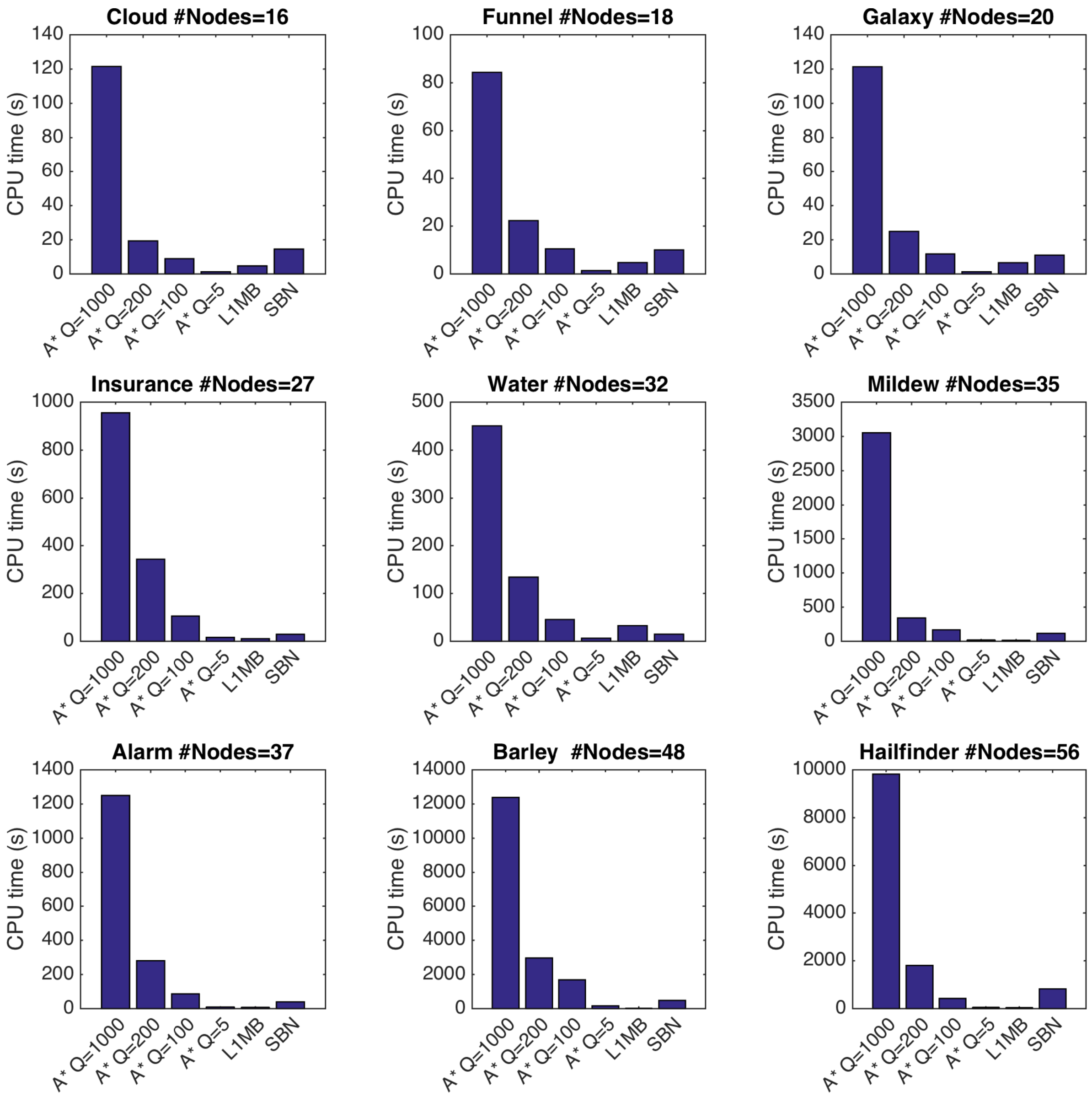

Figure 2.4: Comparison of the computational time required for Bayesian network structure learning algorithms on data simulated from benchmark Bayesian networks.

000 and 100 000 evaluations for larger networks. As can be seen in Figures 2.6 and 2.7, all methods perform relatively well on identifying the true skeletons, but find it significantly more challenging to recover the true v-structures. We find that although increasing the size of queues in A* lasso generally improves the performance, even with smaller queue sizes, A* lasso outperforms L1MB and SBN. While A* lasso with a limited queue size preforms consistently well on smaller networks, it significantly outperforms the other methods on the larger graphs such as Alarm 2 and Hailfinder 2, even with a queue size of 5 and even when the number of evaluations for L1MB has been increased to 50 000 and 100 000. This demonstrates that while limiting

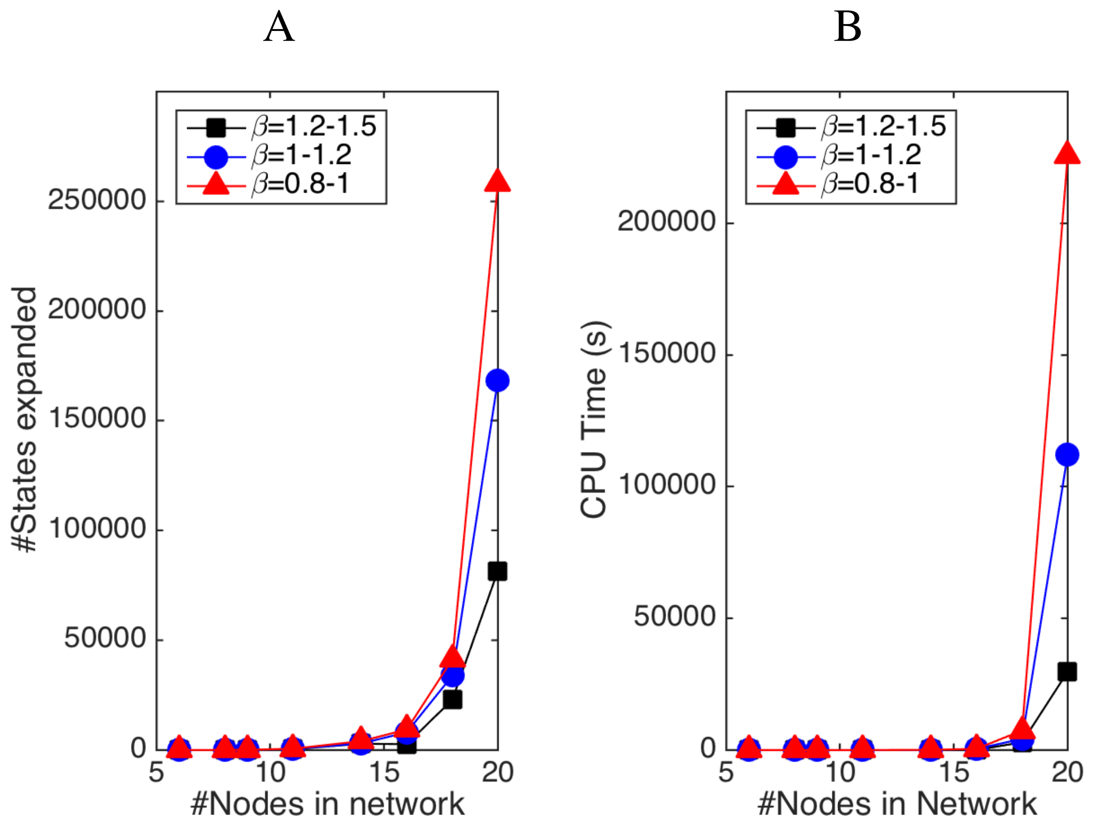

Figure 2.5: A* lasso computation time and states expanded under different edge strengths β_j's. A) The number of states expanded. B) Computational time.

the queue size in A* lasso will not guarantee the optimality of the solution, it still reduces the computation time of A* lasso dramatically without substantially compromising the quality of the solution. In addition, we compare the performance of the different methods in terms of prediction errors on independent test datasets in Figure 2.8. We find that the prediction errors of A* lasso are consistently lower even with a limited queue size.

2.5.1 Analysis of S&P stock data

We applied the methods on the daily stock price data of the S&P 500 companies to learn a Bayesian network that models the dependencies in prices among different stocks. We obtained the stock prices of 125 companies over 1500 time points between Jan 3, 2007 and Dec 17, 2012. We estimated a Bayesian network using the first 1000 time points with the different methods, and then computed prediction errors on the last 500 time points. For L1MB, we used two settings for the number of evaluations, 50 000 and 100 000. We applied A* lasso with different queue limits of 5, 100, and 200. The prediction accuracies for the various methods are shown in Figure 2.9. Our method obtains lower prediction errors than the other methods, even with the smaller queue sizes.

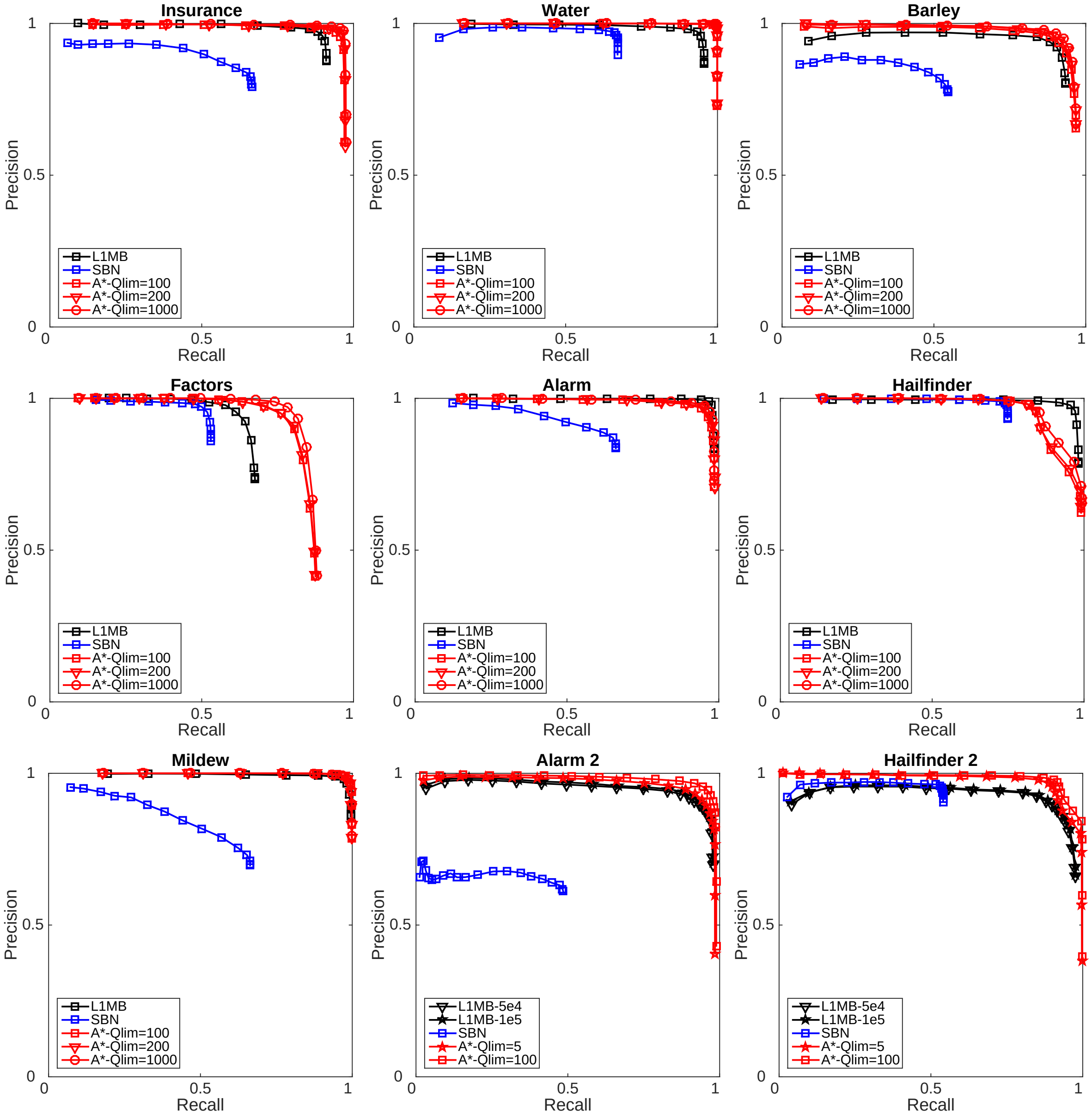

Figure 2.6: Precision/recall curves for the recovery of skeletons of the network structure of benchmark Bayesian networks.

2.6 Conclusions

In this chapter, we considered the problem of learning a Bayesian network structure and proposed A* lasso that guarantees the optimality of the solution while reducing the computational time of the well-known exact methods based on DP. We proposed a simple heuristic scheme that further improves the computation time but does not significantly reduce the quality of the solution.

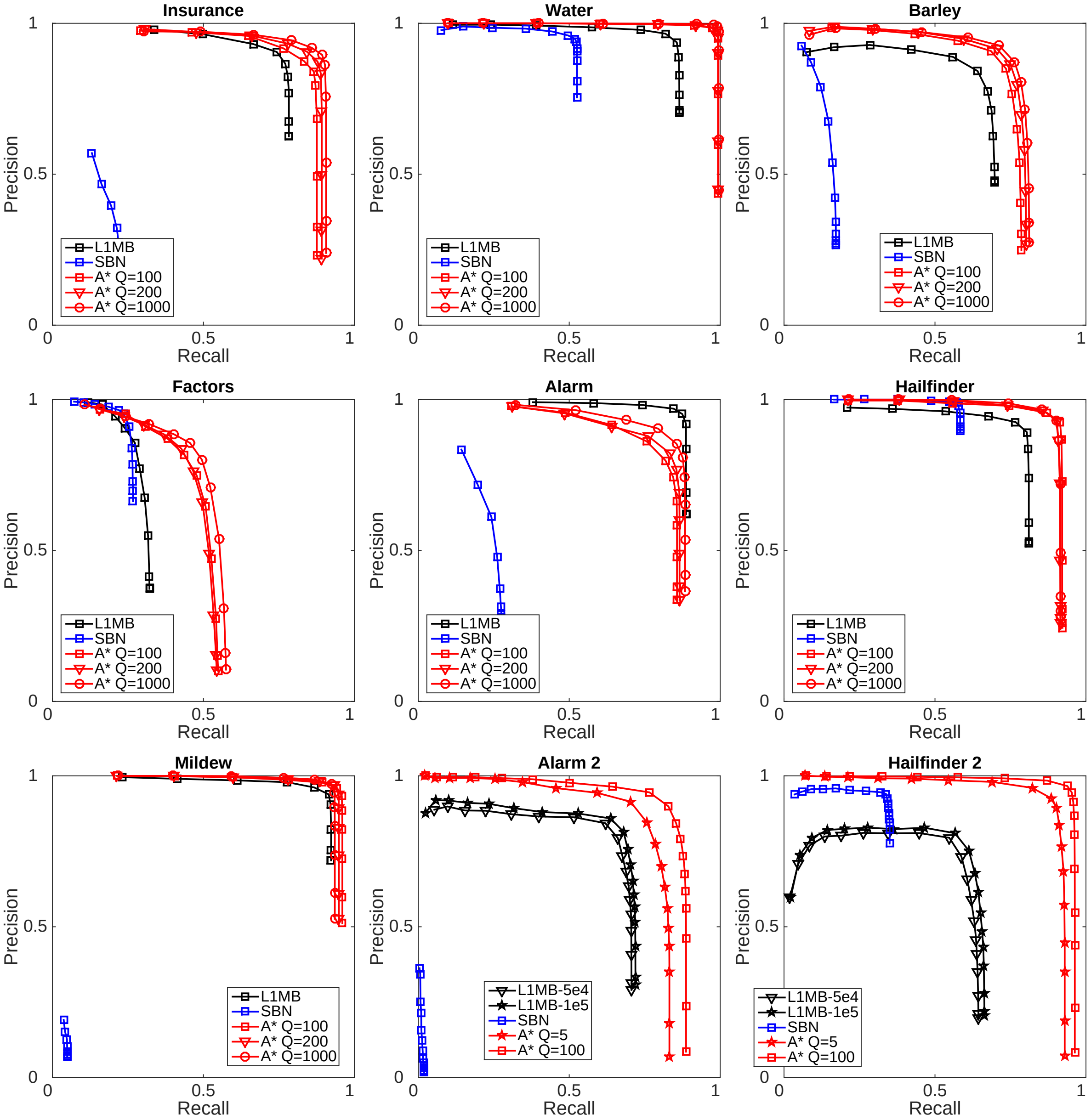

Figure 2.7: Precision/recall curves for the recovery of v-structures of the network structure of benchmark Bayesian networks.

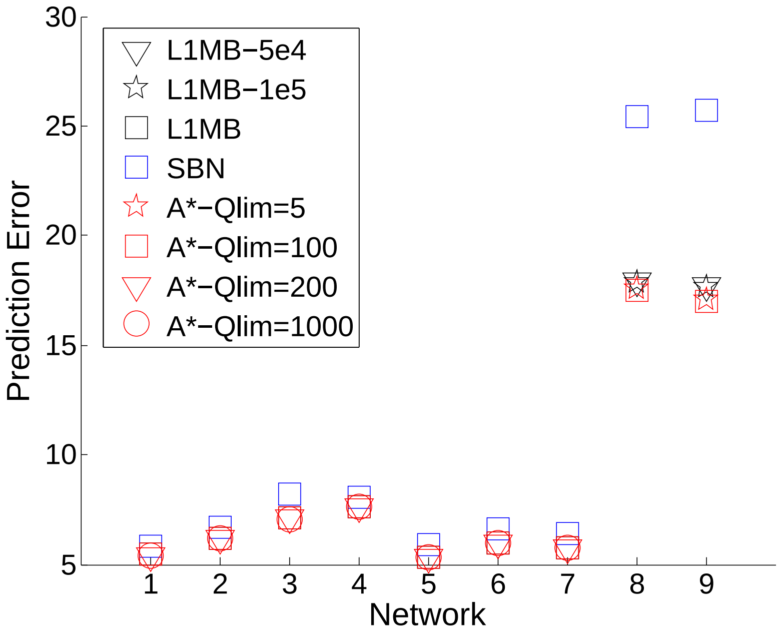

Figure 2.8: Prediction errors for benchmark Bayesian networks. The x-axis labels indicate different benchmark Bayesian networks for 1: Factors, 2: Alarm, 3: Barley, 4: Hailfinder, 5: Insurance, 6: Mildew, 7: Water, 8: Alarm 2, and 9: Hailfinder 2.

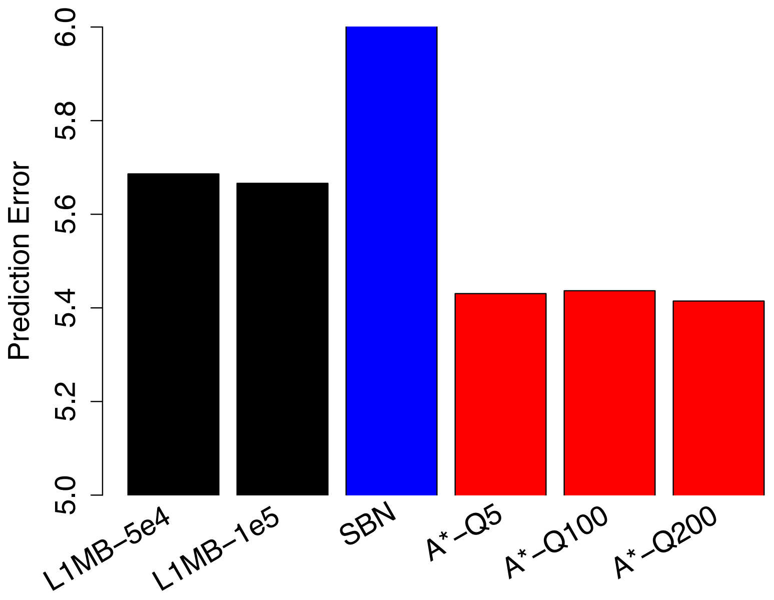

Figure 2.9: Prediction errors for S&P stock price data.

Chapter 3

TF binding map validation with conditional Gaussian Bayesian networks

3.1 Motivation

We propose to validate TF bindings by using population expression and SNP data, and we propose using conditional Gaussian Bayesian networks (cGBNs) as the statistical model to achieve this objective. These cGBNs are an extension of GBNs discussed in the previous section where some variables represent genes, and are conditioned on other variables representing SNPs. While A* lasso was originally developed to learn GBNs, we extend it to learn a cGBN. This extension is straightforward, as it involves augmenting the variable ordering over genes, with the conditioning variables which are SNPs at the beginning of the ordering. In addition to modeling direct influence of TFs or SNPs on target genes, the downstream effects of such interactions can also be determined by performing inference on the model.

3.2 Methods

3.2.1 Learning Gene Regulatory Networks Under SNP Perturbations

In order to determine whether TF bindings on DNA have functional consequences on gene expressions, we propose to leverage an eQTL dataset that captures how naturally occurring genetic variants perturb a transcriptional regulatory system. We introduce a computational methodology, based on conditional Gaussian Bayesian networks (cGBNs), for integrating TF binding data with an eQTL dataset, to identify functional TF-target interactions, along with the overall gene regulatory network and eQTLs that perturb this network. In this section, we first describe our approach for learning a gene regulatory network under SNP perturbations from population gene expression and SNP data. Then, we show how ChIP-seq data can be integrated into our model and learning algorithm as prior knowledge to select the TF-target interactions that are validated under SNP perturbations of gene expressions.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$ denote the expression levels of q genes and $\boldsymbol{X} = (X_1, \ldots, X_p)$ the SNP genotypes of p SNPs for the same individual, where $X_j \in \{0, 1, 2\}$ for the minor allele frequency at SNP j. We model the gene regulatory network as a directed graph over q genes, and the SNP perturbations of the gene expressions as edges from p SNPs to q genes. Then, each gene expression Y_j can be influenced by the expression levels of other gene-expression regulators or by genetic variants that have edges pointing to Y_j. We define a conditional Gaussian Bayesian network as a probability density over this graph that factorizes as follows:

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{j=1}^{q} p(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}_{\mathrm{pa}(j)}), \tag{3.1}$$

where $\boldsymbol{Y}_{\mathrm{pa}(j)}$ is the set of gene expressions regulating the expression Y_j and $\boldsymbol{X}_{\mathrm{pa}(j)}$ is the set of SNPs perturbing Y_j. We model each probability factor using a linear regression model:

$$p(Y_j|\boldsymbol{Y}_{\mathrm{pa}(j)}, \boldsymbol{X}_{\mathrm{pa}(j)}) = \mathcal{N}(\mathbf{Y}_{\mathrm{pa}(j)}\boldsymbol{\beta}_j + \boldsymbol{X}_{\mathrm{pa}(j)}\boldsymbol{\alpha}_j, \sigma_j^2),$$

where $\boldsymbol{\beta}_j = \{\beta_{jk}|Y_k \in \boldsymbol{Y}_{\mathrm{pa}(j)}\}$ and $\boldsymbol{\alpha}_j = \{\alpha_{jk}|X_k \in \boldsymbol{X}_{\mathrm{pa}(j)}\}$ are the regression parameters associated with edges in the graph, modeling the strengths of expression regulations by $\boldsymbol{Y}_{\mathrm{pa}(j)}$ and SNP perturbations by $\boldsymbol{X}_{\mathrm{pa}(j)}$, respectively, and σ_j^2 models the noise.

In order to simultaneously estimate the graph structure and regression parameters from data, we extend A* lasso from our previous work for learning Gaussian Bayesian networks to the case of conditional Gaussian Bayesian networks. Given gene expression data $\boldsymbol{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_q]$ and SNP data $\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$, where $\mathbf{y}_j$ and $\mathbf{x}_k$ are vectors of length n for observations from n individuals for gene Y_j and SNP X_k, A* lasso jointly learns the network structure and edge weights by maximizing the following L_1 regularized log-likelihood of data, under the constraint that the graph over $\mathbf{Y}$ is a directed acyclic graph:

$$\min_{\boldsymbol{\beta}_j, \boldsymbol{\alpha}_j, \sigma_j^2} \sum_{j=1}^{q} \left(\frac{n}{2}\log(\sigma_j^2) + \frac{1}{2\sigma_j^2} \parallel \mathbf{y}_j - \boldsymbol{Y}_{-j}\boldsymbol{\beta}_j - \boldsymbol{X}\boldsymbol{\alpha}_j \parallel_2^2 + \frac{\lambda}{\sigma_j} \parallel \boldsymbol{\beta}_j \parallel_1 + \frac{\gamma}{\sigma_j} \parallel \boldsymbol{\alpha}_j \parallel_1 \right) \tag{3.2}$$

where $\boldsymbol{Y}_{-j}$ is the gene expression data for all genes except for gene j. The L_1 regularization $||\mathbf{c}||_1 = \sum_{k=1}^{K} |c_k|$ for vector $\mathbf{c} = [c_1, \ldots, c_K]$ plays the role of setting a small number of elements in $\mathbf{c}$ to non-zero values to determine the network structure. The non-zero elements in $\boldsymbol{\beta}_j$'s correspond to the presence of edges in the gene network, and the non-zero elements in $\boldsymbol{\alpha}_j$'s correspond to the presence of SNPs that perturb the expression levels. λ and γ are the regularization parameters that control the amount of sparsity in $\boldsymbol{\beta}_j$'s and $\boldsymbol{\alpha}_j$'s and are determined by cross-validation.

To solve Eq. (3.2) for learning the conditional model $p(\boldsymbol{Y}|\boldsymbol{X})$ in Eq. (3.1), we extend A* lasso that we have previously developed which significantly improves the computation time for learning an optimal Gaussian Bayesian network $p(\boldsymbol{Y})$. Our learning algorithm simultaneously solves two problems, one for learning the network structure and the other for learning the parameters associated with the network edges. The structure learning problem is cast as that of finding a topological ordering of the variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ such that edges go only from left to right in the ordering, which is then solved with dynamic programming, combined with A* algorithm to reduce the search space of variable orderings. We learn the model parameters jointly with the network structure by embedding lasso as a scoring system within the dynamic programming. We modify the original A* lasso to learn a conditional model, by augmenting the variable ordering over $\boldsymbol{Y}$ with the conditioning variables $\boldsymbol{X}$ at the beginning of the ordering.

3.2.2 Validating TF Bindings in ChIP-seq with SNP and Gene-expression Data

Our computational approach leverages perturbations of multiple mechanisms behind TF gene regulation by a large number of SNPs across genomes in order to perform functional validation. In order to determine which TF bindings in ChIP-seq data are functional, we integrate the ChIP-seq TF binding information into the A* lasso learning procedure described above. The candidate TF-target interactions identified by ChIP-seq provide prior knowledge on the transcription network structure, which is then updated by A* lasso to include only the validated TF-target interactions after seeing the gene expression and genetic variation data. In particular, we consider three different ways that SNPs affect TF-target interactions to modify the expressions of the target genes and further downstream genes which are described below:

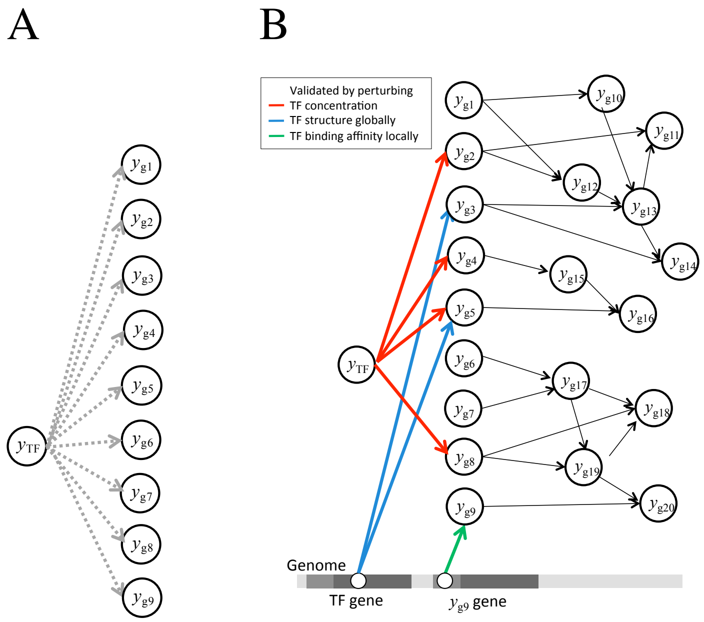

Figure 3.1: Illustration of our approach for validating TF binding events. A) Candidate edges between TFs and target genes are collected from ChIP-seq and DNase I data. B) The candidate edges are validated in three different ways. The edges that are detected in the statistical model are shown. Target genes validated by concentration appear as an edge from the TF to target gene (red). Target genes validated by local perturbations of TF binding affinity appear as an edge from a SNP in the regulatory region of the target gene to the target gene (green). Target genes validated by global SNP perturbation appear as an edge from a SNP in the TF to the target gene (blue). C) Only a subset of the candidate edges represent functional binding events. The resulting validated edges are shown. The colors of the edges correspond with the type of statistical edge used for validation.

- **Perturbation of TF concentration:** The first type of SNP perturbation we consider is the change in TF concentration, due to either SNPs or the expression variation of upstream genes, that induces changes in target gene expressions. Functional target genes validated under this scenario are represented in our cGBN as edges from the TF expression to the target gene expressions (red edges in Figure 3.1B).

- **Local perturbation of TF binding affinity:** Another type of SNP perturbation we consider is SNPs in the regulatory regions of the TF bound genes that modify the short sequence bound by a TF and thus the binding affinity for the TF to influence the expressions of the bound genes. Such cis eQTLs of functional target genes are represented in our cGBN as edges from SNPs in the regulatory region of the target genes to the target gene expressions (green edges in Figure 3.1B).
- **Global SNP perturbation of TF structure:** The final type of SNP perturbation we consider is SNPs located in the TF coding region that can influence the expressions of many target genes globally, by modifying TF amino acid sequence in the case of non-synonymous mutations or by modifying TF translation, folding, or splicing in the case of synonymous mutations. Such trans eQTLs are represented in our model as edges from SNPs in the coding regions of TFs to the target gene expressions (blue edges in Figure 3.1B).

Each candidate target gene derived from the TF binding map as a gene near each binding site (Figure 3.1A) provides three types of candidate edges corresponding to the above three scenarios. Once the statistical model is estimated (Figure 3.1B), we examine this model to determine which of the candidate target genes are functional targets of the TF (Figure 3.1C). A candidate target gene identified by ChIP-seq may be validated under one or more scenarios above. In addition, once population genome and gene expression data are available, this dataset can be used to validate TF binding events in ChIP-seq data for multiple different TFs simultaneously.

In order to reduce the computation time for learning our network model over a large number of gene expressions and SNPs, we further reduce the A* lasso search space over network structures by making additional assumptions on the network edge connectivities. First, we focus on learning the regulatory network over the TFs with ChIP-seq data and their downstream genes, and assume that those non-TFs without ChIP-seq data are candidate downstream genes of the TFs. This assumption is easily implemented within A* lasso by constraining all TFs to be in front of the rest of the genes in the variable ordering. In addition, we assume that the downstream genes of TFs form regulatory modules, where edge connections exist from TFs to genes in each module and among genes within each module, but not between modules. Under this assumption, we first applied hierarchical clustering to all candidate downstream genes to find 40 gene clusters of sizes 100-400 genes, and then applied A* lasso on TFs and each module separately.

3.2.3 Inferring Downstream Effects of Functional TF Bindings

Given the estimated model, we determine the effects of perturbations of TF-target interactions on the expressions of downstream genes via probabilistic inference. Such inferred downstream perturbation effects are analogous to differential gene expressions after experimental TF knockdown. However, unlike in experimental TF knockdown, among the differentially expressed genes, our approach can further distinguish between the genes directly regulated by the TF and the downstream genes indirectly affected by the TF regulation. As a result, while the experimental approach declares all bound and differentially expressed genes as functionally validated

(Fig. 3.2A), our approach finds only a subset of the bound and differentially expressed genes as directed regulated and functionally validated target genes (Fig. 3.2B).

To learn the downstream effects of TF concentration perturbation, we infer from $p(\mathbf{Y}|\mathbf{X})$ the conditional density $p(\mathbf{Y}_{\mathrm{TF_d}}|Y_{\mathrm{TF}}, \mathbf{Y}_{-\mathrm{TF_d}}, \mathbf{X})$, which represents the expressions of TF downstream genes $\mathbf{Y}_{\mathrm{TF_d}}$ given the perturbed TF concentration Y_{TF}, assuming the expressions of all the other genes $\mathbf{Y}_{-\mathrm{TF_d}}$ and genotypes $\mathbf{X}$ are fixed. This conditional probability density is Gaussian with mean $E[\mathbf{Y}_{\mathrm{TF_d}}] = [Y_{\mathrm{TF}}, \boldsymbol{Y}_{-\mathrm{TF_d}}^{\top}, \boldsymbol{X}^{\top}]\boldsymbol{\Theta}$ in a linear regression form, where covariates are the conditioning variables and regression coefficients are given in $\boldsymbol{\Theta}$, a $(q - |\mathbf{Y}_{\mathrm{TF_d}}| + p) \times |\mathbf{Y}_{\mathrm{TF_d}}|$ matrix. The row of $\boldsymbol{\Theta}$ for Y_{TF} quantifies the downstream effect sizes of TF concentration perturbation on each of the downstream genes.

Similarly, to determine the downstream effects of perturbations by either cis or trans eQTLs of target genes, we infer from $p(\mathbf{Y}|\mathbf{X})$ the conditional density $p(Y_A, \mathbf{Y}_{A_\mathrm{d}}|\mathbf{Y}_{-A_\mathrm{d}}, \mathbf{X})$ for target gene Y_A with an eQTL in $\mathbf{X}$. This conditional probability is again Gaussian distributed with mean $E[Y_A, \mathbf{Y}_{A_\mathrm{d}}] = [\boldsymbol{Y}_{-A_\mathrm{d}}^{\top}, \boldsymbol{X}^{\top}]\mathbf{K}$, where $\mathbf{K}$ is a $(|\boldsymbol{Y}_{-A_\mathrm{d}}| + p) \times |Y_A, \mathbf{Y}_{A_\mathrm{d}}|$ matrix of regression coefficients. The row of $\mathbf{K}$ for SNPs represent the strengths of SNP effects on the downstream genes.

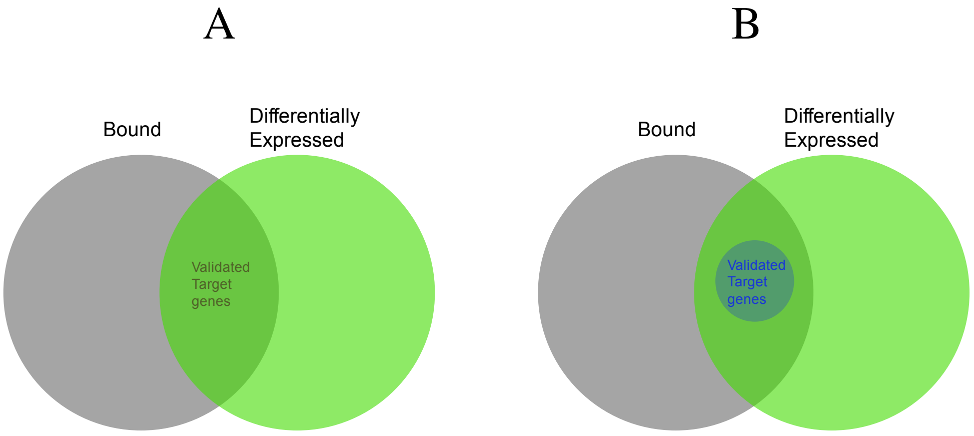

Figure 3.2: Comparison between our approach and TF knockdown experiments for functional validation of the TF binding map. A) With TF knockdown experiments, the validated targets are those that are bound and differentially expressed. B) In our approach, the validated targets are bound and directly targeted by the TF, a subset of the genes that are bound and differentially expressed.

While in general inference tasks in probabilistic graphical models is computationally expensive, efficient inference algorithms can be obtained when the local conditional probability densities $p(Y_k|\boldsymbol{Y}_k^{\mathrm{pa}}, \boldsymbol{X})$'s are Gaussian [72]. In our inference tasks, the desired conditional densities for TFs' downstream effects are in the form of $p(\boldsymbol{Y}_A|\boldsymbol{Y}_B, \boldsymbol{X})$, where $\boldsymbol{Y}_A$ and $\boldsymbol{Y}_B$ are two disjoint sets of gene expression variables. In order to derived this conditional density, we first re-write it as $p(\boldsymbol{Y}_A|\boldsymbol{Y}_B, \boldsymbol{X}) = p(\boldsymbol{Y}_A, \boldsymbol{Y}_B|\boldsymbol{X})/p(\boldsymbol{Y}_B|\boldsymbol{X})$. Then, we find the numerator $p(\boldsymbol{Y}_A, \boldsymbol{Y}_B|\boldsymbol{X})$ as a multivariate Gaussian density that is equivalent to our original conditional Gaussian Bayesian network in Eq. (3.1), and find the denominator $p(\boldsymbol{Y}_B|\boldsymbol{X})$ from $p(\boldsymbol{Y}_A, \boldsymbol{Y}_B|\boldsymbol{X})$ via marginalization, which is a straightforward operation in Gaussian densities. Then, our desired conditional density $p(\boldsymbol{Y}_A|\boldsymbol{Y}_B, \boldsymbol{X})$ can be obtained from $p(\boldsymbol{Y}_A, \boldsymbol{Y}_B|\boldsymbol{X})$ and $p(\boldsymbol{Y}_B|\boldsymbol{X})$ with the standard result in multivariate Gaussians: $p(\boldsymbol{Y}_B|\boldsymbol{Y}_A, \boldsymbol{X}) = \mathcal{N}([\boldsymbol{\mu}_B - \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\mu}_A] + \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{Y}_A,\ \ \boldsymbol{\Sigma}_{BB} - \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{AB})$, given $p(\boldsymbol{Y}_A, \boldsymbol{Y}_B|\boldsymbol{X}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB}\end{bmatrix}\right)$.

In order to compute the multivariate Gaussian form $p(\boldsymbol{Y}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ from the factorized model in Eq. (3.1), we recursively construct this density, visiting each Y_j in the toplogical ordering of the genes found by A* lasso. Let $\Pi = [\pi_1, \ldots, \pi_q]$ represent the topological ordering of genes in $\boldsymbol{Y}$ found by A* lasso, where edges are allowed only from left to right. Assume ($\boldsymbol{\alpha}_k$, $\boldsymbol{\beta}_k$)'s, $k = 1, \ldots, q$, are ordered according to Π. Also, assume the elements of $\boldsymbol{\beta}_j$ and $\boldsymbol{Y}$ are ordered according to Π. Given the factorized density

$$p(\boldsymbol{Y}|\boldsymbol{X}, \Pi) = p(Y_1|\boldsymbol{X}) \prod_{k=2}^{q} p(Y_k|\boldsymbol{Y}_{1:(k-1)}, \boldsymbol{X}),$$

we initialize with

$$p(Y_1|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{\alpha}_1^\top \boldsymbol{X}, \sigma_1^2) = \mathcal{N}(\mu_1, \boldsymbol{\Sigma}_{1,1}),$$

and compute the partial joint distribution iteratively for $k = 2, \ldots, q$

$$\begin{aligned}
p(\boldsymbol{Y}_{1:k}|\boldsymbol{X}) &= p(\boldsymbol{Y}_k|\boldsymbol{Y}_{1:(k-1)}, \boldsymbol{X}) p(\boldsymbol{Y}_{1:(k-1)}|\boldsymbol{X}) \\
&= \mathcal{N}(\boldsymbol{\beta}_k^\top \boldsymbol{Y}_{1:(k-1)} + \boldsymbol{\alpha}_k^\top \boldsymbol{X}, \sigma_k^2) \mathcal{N}(\boldsymbol{\mu}_{1:(k-1)}, \boldsymbol{\Sigma}_{1:(k-1),1:(k-1)}) \\
&= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_{1:(k-1)} \\ \boldsymbol{\beta}_k^\top \boldsymbol{\mu} + \boldsymbol{\alpha}_k^\top \boldsymbol{X}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{1:(k-1),1:(k-1)} & \boldsymbol{\Sigma}_{1:(k-1),1:(k-1)}\boldsymbol{\beta}_k \\ \boldsymbol{\beta}_k^\top \boldsymbol{\Sigma}_{1:(k-1),1:(k-1)} & \sigma_k^2 + \boldsymbol{\beta}_k^\top \boldsymbol{\Sigma}_{1:(k-1),1:(k-1)}\boldsymbol{\beta}_{,k}\end{bmatrix}\right) \\
&= \mathcal{N}\left(\boldsymbol{\mu}_{1:k}, \boldsymbol{\Sigma}_{1:k,1:k}\right).
\end{aligned}$$

```
Input  : $\mathbf{x}^i$, $\mathbf{y}^i$, $(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \sigma_j^2)$ for $j = 1, \ldots, q$, $\Pi$
Output: $\boldsymbol{\mu}$, $\Sigma$
Initialize $\boldsymbol{\mu} = \mathbf{0}_q$, $\boldsymbol{\Sigma} = \mathbf{0}_{q \times q}$;
for $k \leftarrow 1$ to $q$ do
    $\mu_k \leftarrow \boldsymbol{\beta}_k^\top \boldsymbol{\mu} + \boldsymbol{\alpha}_k^\top \mathbf{x}^i$;
    $\Sigma_{kk} \leftarrow \sigma_k^2 + \boldsymbol{\beta}_k^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_k$;
    $\Sigma_{1:(k-1),k} \leftarrow \Sigma_{1:(k-1),k} + \Sigma_{1:(k-1),1:(k-1)} \boldsymbol{\beta}_k$;
    $\Sigma_{k,1:(k-1)} \leftarrow \Sigma_{1:(k-1),k}$;
end
Set $p(\mathbf{Y}|\mathbf{X}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$;
```

Algorithm 2: Procedure to compute $P(\mathbf{y}^i|\mathbf{x}^i)$.

3.3 Datasets

We applied our computational approach to determine whether the TF bindings in the ENCODE ChIP-seq and DNase-seq data [22] are functional, using the SNP and gene expression data from the HapMap 3 and 1000 Genomes Project population [20, 97, 140].

We downloaded the ENCODE ChIP-seq data for 71 TFs and the DNase I hypersensitivity sites both for the GM12878 LCL, processed with the ENCODE uniform peak calling pipeline [22]. For the five TFs whose ChIP-seq data are available from multiple experiments, we took the union of the binding sites from all experiments. We constructed the TF binding map from ChIP-seq binding sites for each TF that also overlapped with DNase I hypersensitivity regions. Given the TF binding map, we determined the set of genes potentially targeted by the TF as those genes that contained binding regions within 10kb from the transcription start and end sites of the gene.

In order to functionally validate the candidate target genes suggested by ChIP-seq data, we identified 520 individuals whose LCL gene expression levels were profiled in a previous study of HapMap 3 population [97, 140] and whose genome sequences were available from the 1000 Genomes Project Phase 3 [20]. We downloaded the expression data for 21,800 probes that was collected with the Illumina Human-6 v2 Expression BeadChip platform and analyzed in [97, 140]. We first filtered out the probes with standard deviation less than 0.2, if the probes correspond to non-TF genes without ChIP-seq results, and discarded the redundant probes that recognize the same transcripts. The remaining 9,940 probes were included in our network learning. Of the total 9,940 probes, 8,505 probes represented gene-level expressions with one-to-one correspondence to genes, as they recognized either all transcripts of the gene or the single transcript that exists for that gene. Our probe data also included 1,176 probes corresponding to transcript-specific expressions for 1,048 genes that have multiple transcripts. For a small subset of these genes (259 genes), probes for gene-level expressions were available in addition to probes for transcript-specific expressions and we included them in our analysis.

For the same 520 individuals above with expression data, we obtained the genome sequence data from the 1000 Genome Project Phase 3 [20]. After filtering out SNPs with minor allele frequency less than 0.05, we included in our analysis the 87,267 biallelic SNPs in the promoter and exon regions of each gene with expression data, where the promoter region was defined as 2000bp from the transcription start site.

3.4 Results

3.4.1 Functional validation of TF bindings in lymphoblastoid cell lines

We applied our approach for TF binding map validation to determine if the TF binding events from the ENCODE TF binding data in LCL lead to changes in target gene expressions, using the SNP and gene expression data from the HapMap 3 and 1000 Genome Project population [20, 97, 140]. We used the TF binding map from ENCODE as a structural prior for our cGBN and then we estimated a cGBN by combining this prior knowledge with the expression and SNP genotype data.

We first examined our estimated model to determine which of the bound genes in the TF binding map are functionally validated. For each TF, we extracted from the estimated model the validated target genes under each of the three perturbation scenarios, as genes with edges from TF expression, genes with cis eQTLs located within the TF bound region, and genes with trans eQTLs located in TF coding region. As can be seen in Figure 3.3A, the fraction of functionally validated target genes to candidate target genes in the binding map varied from 3% to 69%

depending on the TF. We found that a large fraction of the functional target genes were those validated under the perturbations of TF concentrations (Figure 3.3B), and that relatively fewer genes were validated under the perturbations of regulatory sequences or TF coding sequences (Figures 3.3C-D). In addition, we found that for each TF, more functional target genes were perturbed by their cis eQTLs than by trans eQTLs located in the TF coding region. For example, 71 out of 83 TFs had more than 500 target genes validated under perturbations by cis eQTLs (Figure 3.3C), whereas none of the TFs had more than 500 target genes validated under perturbations by trans eQTLs in the TF coding region (Figure 3.3D). This is consistent with the observation from previous studies that trans regulatory elements tend to be evolutionarily more conserved than cis regulatory elements, because of the global impact of the potentially damaging changes in trans regulatory elements [15, 138, 160]. Finally, we found that the TF expression variances are highly correlated with the number of target genes validated under the perturbation of TF concentration ($R = 0.80$, Figure 3.4). This shows that our approach of using SNP perturbations to functionally validate TF bindings is most effective when there exists sufficient variations in TF expression levels in the population.

Next, we examined the differential gene expression induced by each individual perturbation as described above, via probabilistic inference on our cGBN. While an experimental knockdown approach considers all bound and differentially expressed genes as functional targets of the TF, our approach finds only a subset of those genes as functional targets, as it further distinguishes direct targets of the TF from indirectly affected downstream genes with non-functional binding. Given the estimated cGBN, we compared the differential gene expressions of the functionally validated target genes with those of further downstream genes with non-functional binding. We first performed probabilistic inference to obtain downstream effect sizes on each downstream gene of the TF, for each individual perturbation of TF concentrations, regulatory regions, and coding sequences. Then, at different levels of downstream effect sizes, we compared the set of differentially-expressed and bound genes (green bars in Figure 3.5) with a subset of those genes, which consists of functionally validated target genes (red curves in Figure 3.5). As can be seen in Figure 3.5, our approach found that for each perturbation type and for each TF, 0.02 to 100 % of the bound and differentially expressed genes were validated as functional targets of the given TF. This shows that our computational model and learning algorithm can determine the direct and indirect targets by statistically assessing the direct and indirect probabilistic dependencies among the gene expression variables. This allows our approach to potentially identify functional targets of each TF with higher accuracy, compared to exprimental method. In addition, we found that the downstream effect sizes vary across TFs and that those TFs with stronger downstream perturbation effects tend to have a greater number of validated genes. Among the three types of perturbations we consider, the downstream effect sizes were the strongest for the perturbations of TF concentrations and the weakest for trans eQTLs in TFs coding region for target gene expression.

To characterize the biological functions that the TFs are regulating, we looked for the Gene Ontology (GO) terms that are enriched in the set of validated target genes for each TF, using the GO annotations. For those TFs with more than 20 target genes validated under each of the three scenarios, we show in Figure 3.6 the significantly enriched GO slim terms (FDR of 5%) and the corresponding p-values, grouping together the related GO slim terms. We show the GO slim terms enriched in at least four TFs for the TF concentration perturbation (Figure 3.6A), and in one or more TFs for the local perturbation of target gene regulatory sequences and global perturbations in TF coding sequences (Figures 3.6A and 3.6B). As can be seen in Figure 3.6A, for the TF concentration perturbation, the GO category of immune system processing is enriched for many of the TFs. This is consistent with the fact that an LCL is a human B cell immortalized after Epstein-Barr virus infection and thus has the phenotypes of highly activated B cells [38]. In addition, since activated immune cells undergo cell proliferation and potential changes in metabolic processes [62], GO terms related to cell growth (e.g., cell cycle, cell proliferation) and metabolic processes (e.g., biosynthetic and catabolic processes) were found enriched for many of the same TFs whose target genes were enriched for immune system processing. For the cases of local and global SNP perturbations of TF binding affinities (Figures 3.6A and 3.6B), the enrichment was overall weaker than the case of TF concentration perturbation, mainly because there were far fewer validated target genes. The enrichment results under the perturbations of TF concentrations and under the perturbations of cis eQTLs of target genes were similar in that the GO terms related to metabolic processes were enriched for many of the TFs in both cases.

Finally, to see if the enrichments of the immune system processing GO category in Figure 3.6A indicate a B cell immune response, we performed a similar GO enrichment analysis with the more fine-grained GO categories under the immune system processing GO category in the hierarchy. As can be seen in Figure 3.7, many TFs had target gene sets that are enriched in GO terms related to B cell activities. Overall, we found the biological functions that these TFs regulate are consistent with what is known about the B cell immune response in LCL.

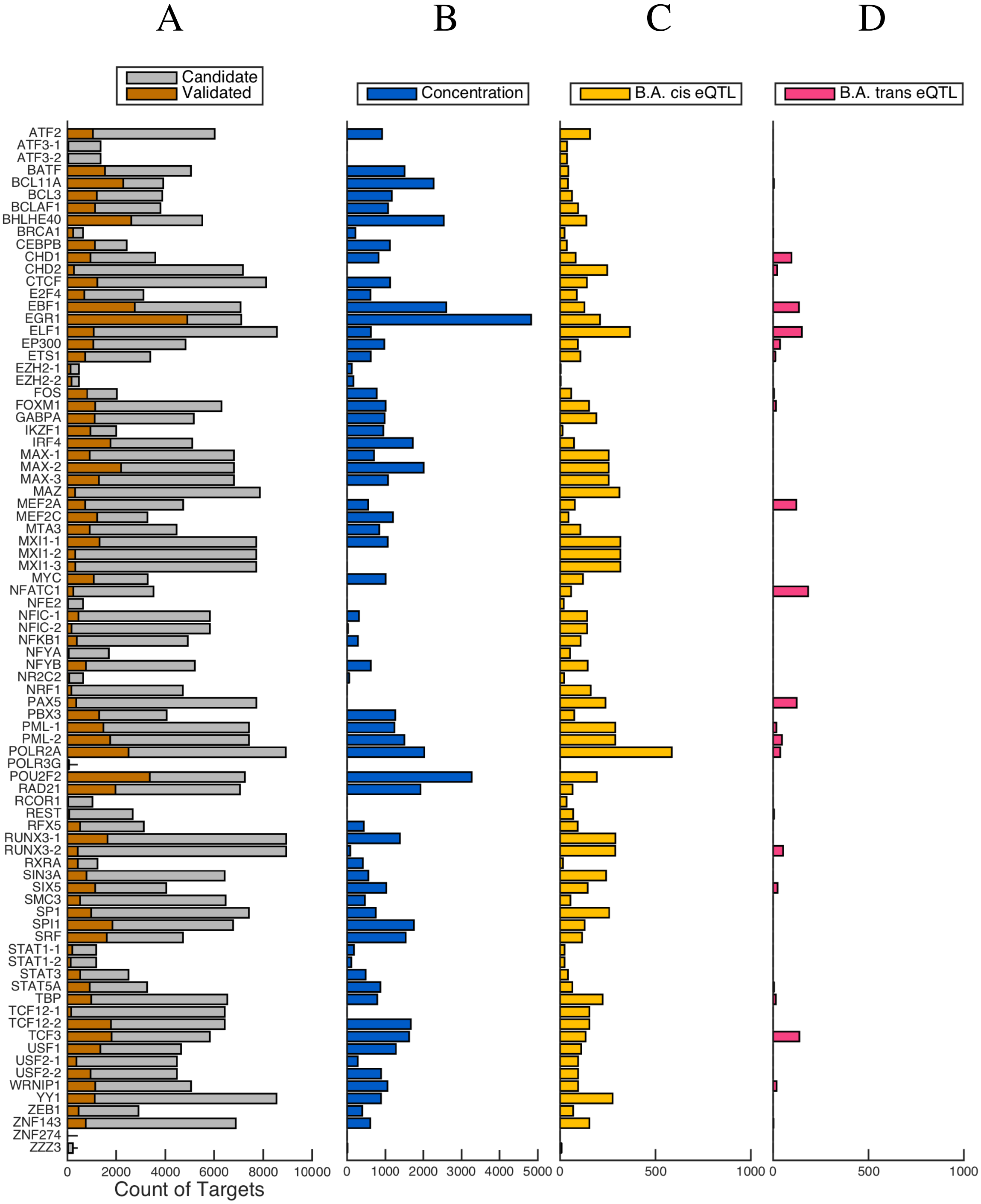

Figure 3.3: The number of validated target genes in TF binding map for LCLs. For TFs that had multiple transcripts in the study, each transcript is numbered. A) The gray bars represent the number of target genes in the TF binding map for each TF. The brown bars represent the number of target genes that were validated using our approach. B), C), and D) show the total number of target genes validated by perturbing TF concentration, perturbing TF binding affinities locally, and perturbing the TF structure globally.

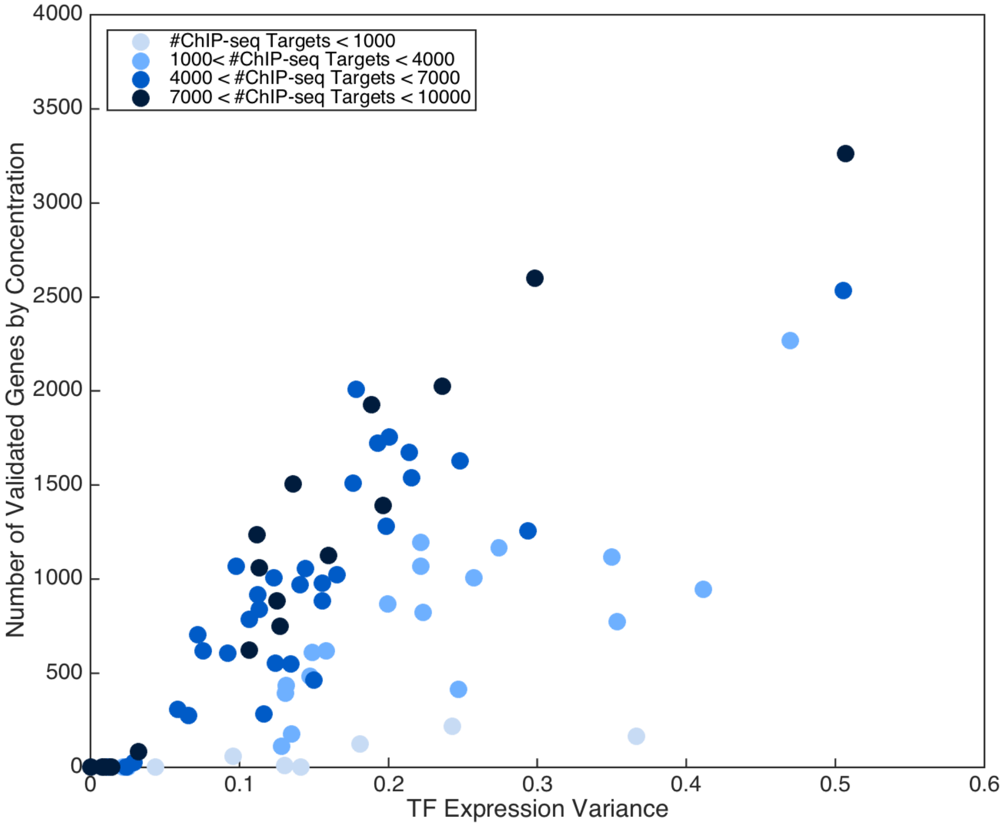

Figure 3.4: The effects of TF expression variance on functional validation of TF bindings. Each data point represents a different TF. The number of validated target genes is correlated with the TF expression variance (R=0.80).

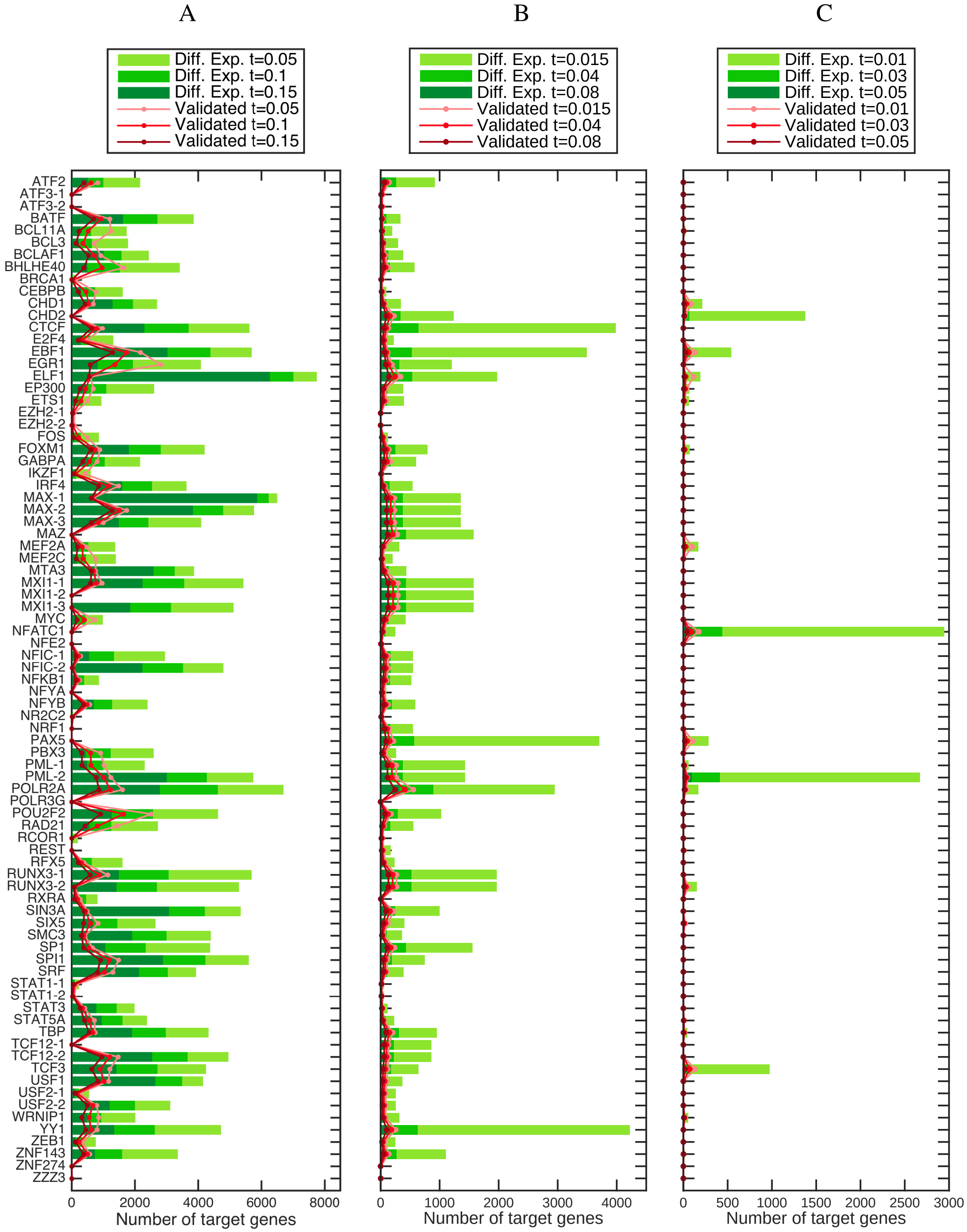

Figure 3.5: Validated target genes and differentially expressed and bound genes for each TF determined by our approach. Each panel represents genes that are either validated (red curves) or differentially expressed and bound (green bars) by perturbing A) TF concentration, B) TF binding affinities locally, and C) TF structure globally.

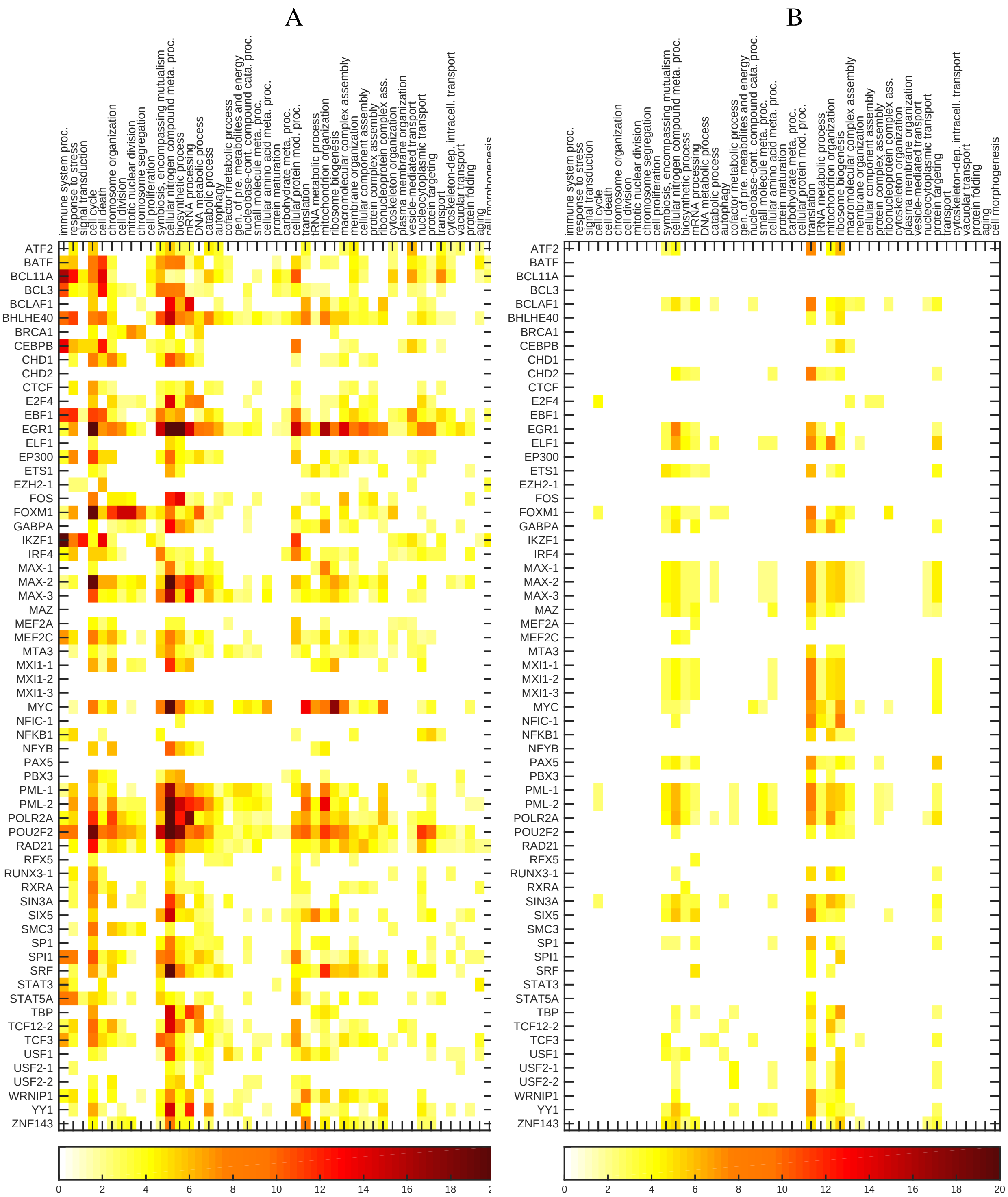

Figure 3.6: GO enrichment analysis of validated target genes. The x-axis are GO ontology terms and y-axis list TFs. A) GO analysis of target genes validated by concentration. B) GO analysis of target genes validated by local perturbations from cis eQTLs.

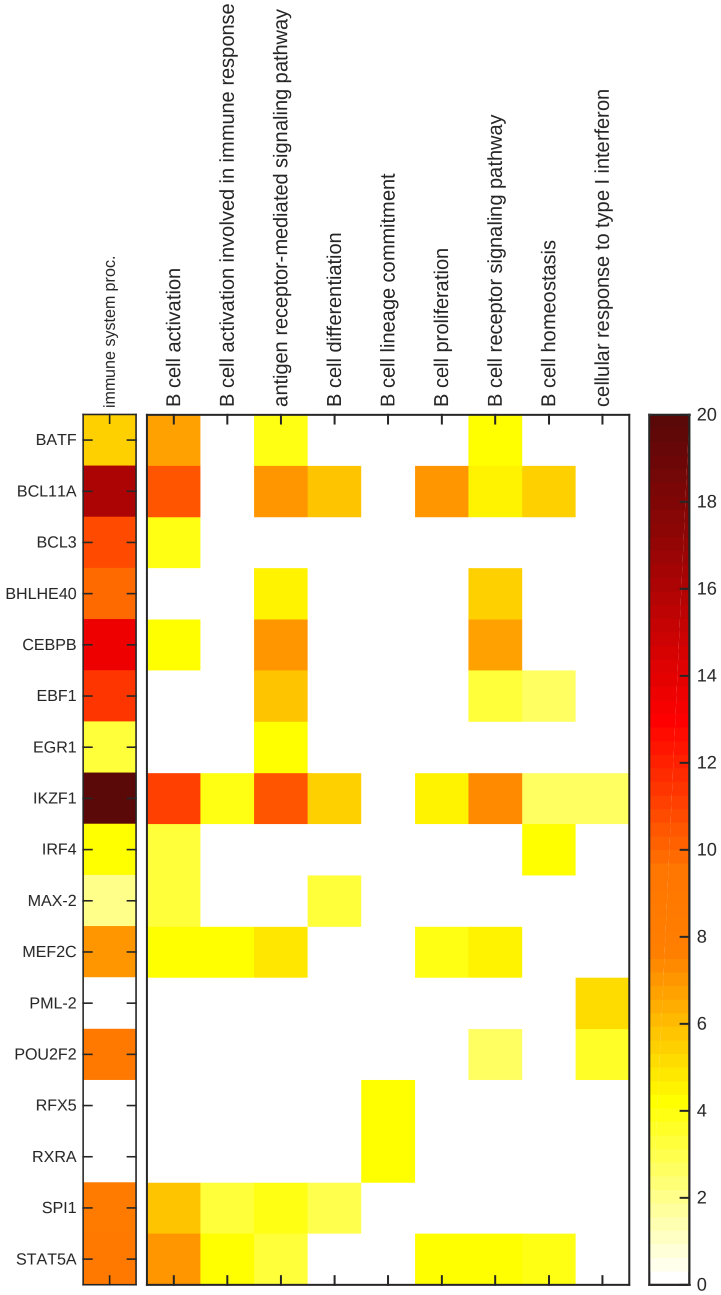

Figure 3.7: GO enrichment analysis of selected TFs that show enrichment for more specific GO categories under the broad immune system processing GO category.

3.4.2 TF-target interactions validated by perturbing TF concentration

First, we examine the bound genes whose expressions are affected under the perturbations of TF concentrations, as identified by our analysis. In particular, to assess the effectiveness of our approach, we compare our results with those obtained from TF RNAi experiments in the LCL from a previous study [26]. Since both RNAi and SNP perturbations of TF concentrations vary TF expression levels, downstream effects of such perturbations may be similar in both cases. On the other hand, RNAi perturbs a single gene at a time, whereas in an eQTL data, a large number of genes are perturbed simultaneously by a large number of SNPs. Thus, the downstream effects may be different between the two approaches, just as there is a significant difference in gene expression patterns between single knockdown and double knockdown. Below, we explore such similarities and differences in the perturbation effects and their impact on functional validation of binding maps between experimental and our computational approaches.

We obtained the functionally validated target genes from the previous RNAi experiments [26] as those genes that are bound and differentially expressed after TF knockdown. Then, we compared these genes with an equivalent set of genes in our analysis, the set of bound and differentially expressed genes under TF concentration perturbation, a subset of which is the target genes functionally validated by our computational approach. For 14 out of 72 TFs included in our study, microarray gene expression data were available for a HapMap LCL (GM19238) before and after RNAi with knockdown efficiency above 50% measured by qPCR [26]. From the 4,661 probe measurements that matched the probes used in our analysis, we determined the genes differentially expressed after RNAi, using the same procedure in [26] and applying the likelihood ratio test followed by multiple testing correction (FDR < 0.05). Among the TF bound genes in our binding map (Figure 3.8A), we examined the overlap between the differentially expressed genes after TF RNAi and those genes in our analysis with downstream effect sizes greater than .05 under TF concentration perturbation (Figure 3.8B). We also examined the amount of TF concentration perturbation in our population data as the sample variance of TF expression, since this can directly affect the amount of differential expression of downstream genes (the red line in Figure 3.8B).

For TFs such as PAX5 and TCF12-1 whose expressions are not naturally perturbed and thus, have little variability across samples, our computational approach found no or few differentially expressed downstream genes (Figure 3.8B). However, for the same TFs, the TF knockdown experiments revealed many differentially expressed genes. This suggests that in general, if there is no or little naturally-occurring variation in TF expression, it is not feasible to leverage the TF concentration perturbation to functionally validate target genes of the TF or to reveal downstream genes regulated by the TF. Then, an artificial perturbation via experimental knockdown is necessary to determine whether the TF bindings near the candidate target genes are functional.

On the other hand, for EZH2-1 and EZH2-2, even though the TF expressions had substantial variability across individuals in our data, neither experimental nor naturally-occurring perturbation of the TF expression led to significant differential expression of any downstream genes (Figure 3.8B). Thus, for those TFs, we conclude that the results are in agreement between our computational method and the experimental method.

For the other remaining 10 TFs in Figure 3.8B, the TF expressions were perturbed both in TF knockdown and in the population in nature, but the sets of bound and differentially expressed genes after perturbations only partly agreed between the two approaches. We hypothesize an epistatic interaction between a TF and its co-regulators as the primary underlying cause of this discrepancy in perturbation effects. Without epistasis, the regulatory effect of each TF on its target gene would be independent of other regulators of the target gene. Thus, naturally-occurring perturbations and experimental knockdown studies would reveal a nearly identical set of differentially expressed downstream genes. Even though unlike in TF knockdown many gene expressions are simultaneously perturbed in SNP perturbation, the effect of perturbation of each individual TF concentration, as is teased out by our computational approach, would be analogous to the TF knockdown effect. The two types of perturbations may differ in the magnitudes of TF expression perturbations, but otherwise, would induce the same downstream effects.

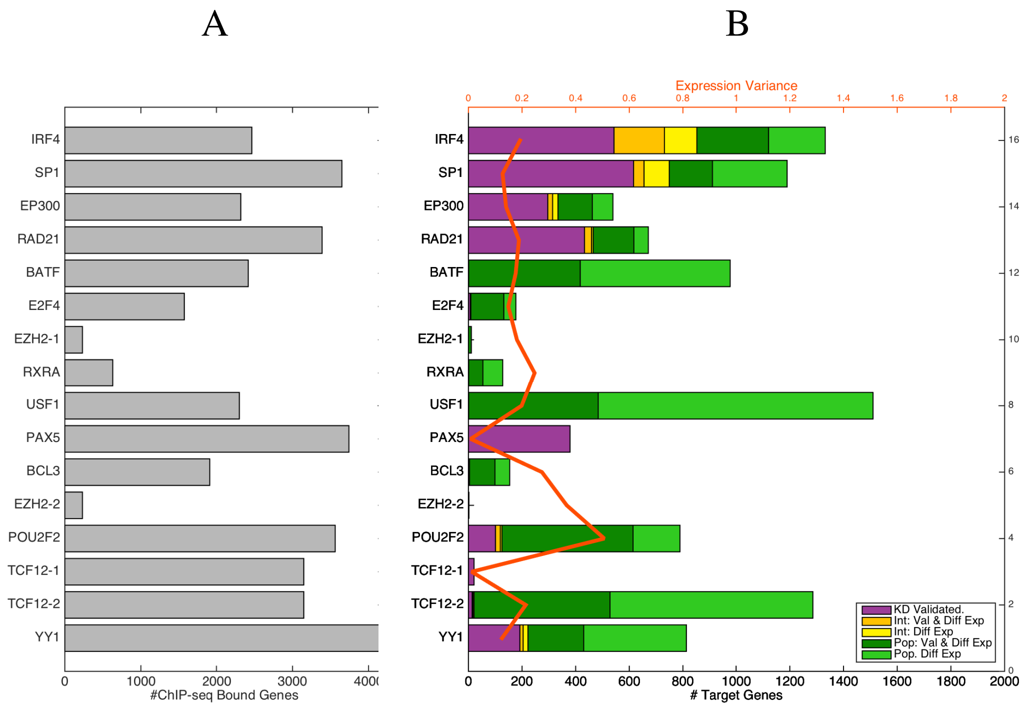

Figure 3.8: Comparison of validated target genes between TF knockdown experiments and our approach. A) Target genes from the TF binding map. B) The bar graph indicates number of target genes validated by each method. The line graph shows the expression variance of each TF in the population.

However, when a TF epistatically interacts with other co-regulators to regulate target genes, TF knockdown and SNP perturbations of the TF and its co-regulators would lead to different downstream expression patterns. Under epistasis, the regulatory effect of the TF on target genes is dependent on the states of co-regulators in the cell environment. Then, the key difference between the two perturbation approaches is that a TF knockdown experiment is typically performed on cells with a single genetic background representing a single state of the co-regulators, whereas an eQTL dataset contains many samples with different genetic backgrounds and thus with different configurations of TF and co-regulator states. This difference in turn leads to differences in the sets of downstream genes that can be revealed by either type of TF perturbations. A TF knockdown can reveal downstream genes only if the states of co-regulators in the cell under study do not mask the effects of TF knockdown. Otherwise, SNP perturbations, where potentially many more configurations are represented across samples, can be potentially more powerful, when the data are analyzed with an appropriate computational method that models epistatic interactions. On the other hand, SNP perturbations can reveal downstream genes, only if there exist samples in the eQTL dataset in which the states of co-regulators do not mask the effects of TF perturbations. Otherwise, an artificial perturbation is necessary to reveal the downstream genes. Overall, there are downstream genes that can be revealed only by either of the perturbation methods, while some downstream genes can be revealed by both methods if the co-regulator states do not mask the TF perturbation effect in the TF knockdown and in some of the samples in the eQTL dataset.

The power of SNP perturbations in eQTL data to identify epistatically regulated downstream genes can be fully realized if the data are analyzed with an approach that fully models the epistatic interactions of regulators. Our cGBN has limited power as it assumes independent influence of each regulator on its target. Instead of fully modeling epistasis, which is computationally expensive due to a large number of possible interactions that need to be considered, we use a relatively simple statistical method to assess the presence of epistasis in the regulatory relationships that have been identified by our cGBN.

For each TF, we fit a linear model that models two-way epistatic interactions, each of which involves the given TF and one other co-regulator and can potentially influence the bound and differentially-expressed genes identified by either the TF knockdown or our computational approach. Recall the original conditional probability density is written as $p(\boldsymbol{Y}_{\mathrm{TF_d}}|Y_{\mathrm{TF}}, \boldsymbol{Y}_{-\mathrm{TF_d}}, \boldsymbol{X})$ for downstream genes $\boldsymbol{Y}_{\mathrm{TF_d}}$, conditional on TF itself Y_{TF}, the rest of the genes $\boldsymbol{Y}_{-\mathrm{TF_d}}$, and SNPs $\boldsymbol{X}$, $E[\boldsymbol{Y}_{\mathrm{TF_d}}] = [Y_{\mathrm{TF}}, \boldsymbol{Y}^{\top}_{-\mathrm{TF_d}}, \boldsymbol{X}^{\top}]\boldsymbol{\Theta}$. For the linear model augmented with two-way interaction terms, $E[\boldsymbol{Y}_{\mathrm{TF_d}}] = [Y_{\mathrm{TF}}, \boldsymbol{Y}^{\top}_{-\mathrm{TF_d}}, \boldsymbol{X}^{\top}, \boldsymbol{\Phi}(Y_{\mathrm{TF)}}]\boldsymbol{\Psi}$ where the set of non-downstream genes $\boldsymbol{Y}_{-\mathrm{TF_d}}$ is limited to those genes with the corresponding entries in $\boldsymbol{\Theta} > 0.15$ and $\boldsymbol{\Phi}(Y_{\mathrm{TF}}) = Y_{\mathrm{TF}} \times Y_{1:k}$, where $Y_{1:k} = [Y_1, \ldots, Y_k]$ consists of TFs in $\boldsymbol{Y}_{-\mathrm{TF_d}}$ that potentially interact with Y_{TF}. To fit the model, we optimize the L_1 regularized negative log-likelihood using a different regularization parameter for the linear terms, interaction terms, and SNPs. The best regularization parameters were selected by cross validation.

Next, we examined for each TF how many of the bound and differentially expressed genes in either type of perturbations were found to be regulated by epistatically interacting TF and its co-regulators (Figure 3.9). Across TFs, most of the bound genes that were differentially expressed only in TF knockdown were found to be epistatically regulated, while the remaining genes are likely to be the ones that are not perturbed in nature due to the masking effects of co-regulators of the TF. In addition, some of the bound genes that were differentially expressed only in our analysis of eQTL data are likely to be genes with strong epistatic effects that could be captured by our cGBN but were found to be unaffected in the TF knockdown due to the masking effects of the interaction partners. As can be seen in Figure 3.9, many of these genes were found to be epistatically regulated by the TF and another co-regulator, while other genes may be under two-way epistasis that were not modeled in our analysis or under higher-order epistasis.

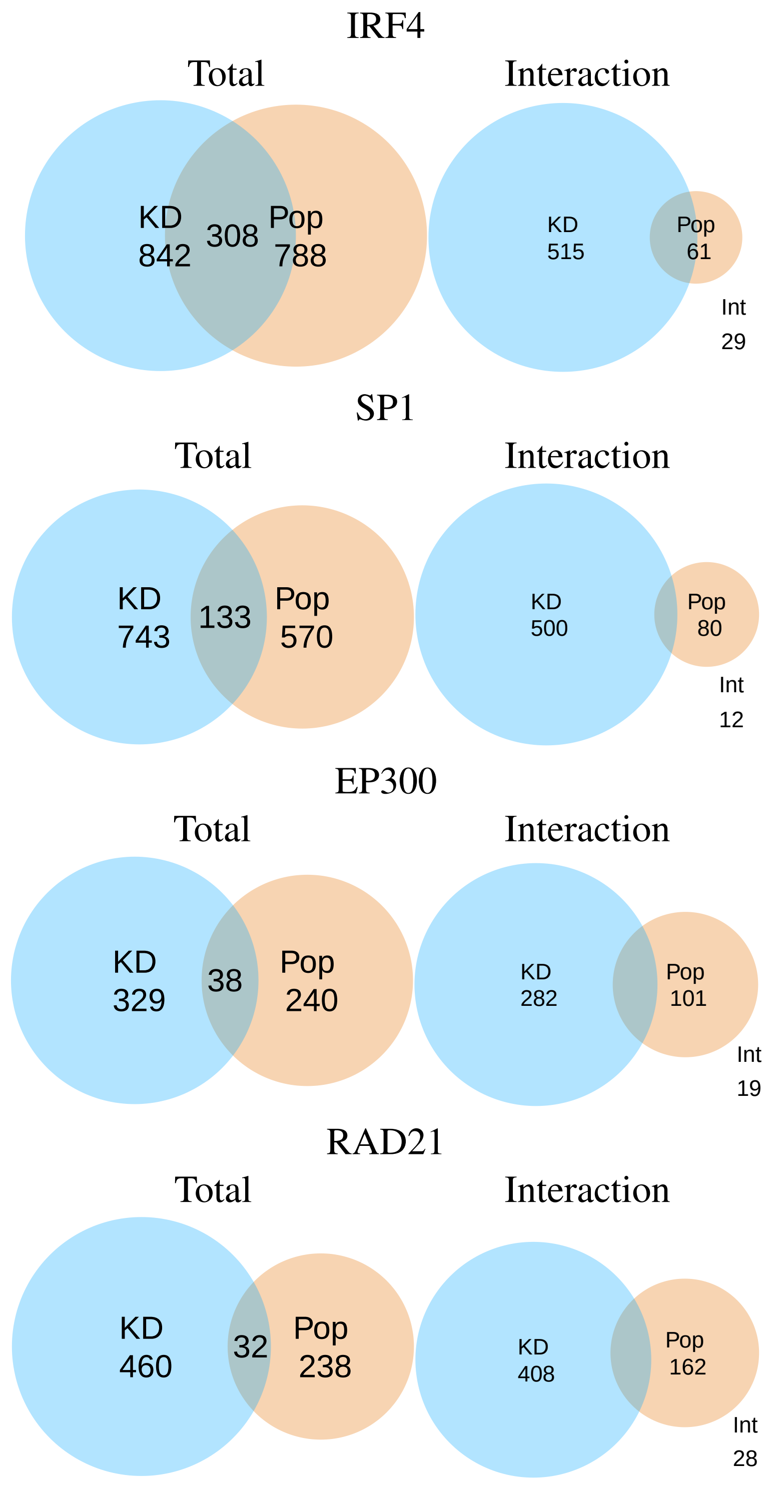

Figure 3.9: Two-way interaction effects of bound and differentially expressed genes. Left column: Total number of bound and differentially expressed genes that were identified by TF knockdown or our computational approach using population data. Right column: The number of bound and differentially expressed genes with two-way interaction effects involving the TF and co-regulators.

3.4.3 Perturbing TF binding affinities

Now, we turn to the other perturbation scenarios and examine the TF bound genes whose expressions are modified by the genetic variants in the target gene regulatory sequences or the TF coding sequences. In particular, we investigate whether any of the eQTLs of the validated target genes affect TF binding affinities. For each validated target gene, we examine whether its cis eQTLs change DNA motif sequences recognized by TF and whether trans eQTLs in the TF coding region change the TF structure.

Perturbing TF binding affinities locally

In this section, we examine whether the cis eQTLs in the TF bound regions disrupt binding affinities of the DNA motif sequences recognized by TF. ChIP-seq and DNase-seq technology provides only the information on broad DNA regions bound by TF, but not the precise location on DNA where TF bindings occur. We use the previously known TFBS motif models to pinpoint TF binding sites (TFBSs) on the binding map and then, assess the impact of the cis eQTLs on TF binding affinities of motif matching sequences.

In order to narrow down TFBSs within the bound region, where genetic variants can alter binding affinities, we first scanned the genome of the same cell line (GM12878) whose TF binding data were available from the ENCODE project, with motif position weight matrices (PWMs) from TRANSFAC and JASPAR databases [90, 91]. We focused on the 58 TFs whose PWMs were available from the databases and identified TFBS motif matches located within the promoter regions of the validated target genes, defined as 2000bp from the transcription start site. For many of the TFs, multiple PWMs were available, each derived from different data sources (e.g., SELEX, ChIP-seq, DNA binding arrays, protein binding arrays, and 3D-structure-based energy calculations) or compiled from the literature and individual genomic sites. For PWMs derived from ChIP-seq, we used the ones from LCL ChIP-seq data processed with MDSCAN [84], but if a PWM from LCL is not available, we used PWMs from another normal cell line. For PWMs obtained from SELEX, we selected for our analysis the PWM from the most recent SELEX experiment, including both homodimer and heterodimer cases. For PWMs compiled from many genomic sites or from the literature, we first favored a factor-specific PWM over a family PWM, a PWM derived from human data over non-human data, and PWMs constructed from a greater number of genomic sites. We added pseudocounts to entries in PWMs and re-normalized the PWMs. The pseudocounts were set to 0.004 for ‘C’ and ‘G’ and 0.006 for ‘A’ and ‘T’, if PWMs were available as a normalized probability matrix. For PWMs with unnormalized counts, the pseudocounts were set to 0.04 for ‘C’ and ‘G’ and 0.06 for ‘A’ and ‘T’, if the PWM was derived from a large number of binding sites (e.g., PWMs obtained through SELEX, DNA binding array data, or 3D structure based energy calculations), and 0.4 for ‘C’ and ‘G’ and 0.6 for ‘A’ and ‘T’, if the PWM was derived from a small number of binding sites (e.g., PWMs compiled from individual genomic binding sites or from the literature). A position-specific scoring matrix was then constructed from the resulting PWMs using background nucleotide frequencies with 40% GC content. We assessed the significance of motif matches at $\alpha = 0.001$, based on the null distribution obtained by enumerating and scoring all possible sub-sequences of motif length via dynamic programming. For motifs with length 12 or greater, we used an approximate null distribution by binning the scores. All computations were made using the Biopython package [19].

Next, we identified those motif matches that contain cis eQTLs found by our learning algorithm. For 50 TFs out of the 58 TFs with known PWMs, there was at least one motif match in bound genes overlapping with cis eQTLs. Many of the cis eQTLs located in the bound promoter regions of the validated target genes overlapped with motif matches (Figure A.1). Across the 50 TFs, the fraction of cis eQTLs that coincide with motif matches for each TF ranged from 2.0% to 48.8%. These cis eQTLs of the validated target genes that lie on the TF motif matches could potentially change the binding affinity of the short DNA sequences recognized by TF to influence the expression of genes near the eQTLs.

To see if the eQTLs lying on TF motif matches indeed disrupt TF binding affinity, we compared the effects of eQTLs on motif match scores with those of other SNPs in the bound promoter regions that are not eQTLs. To measure SNP effects on binding affinity, we defined a score delta as a difference in motif match scores of two short sequences that are identical except for different alleles at SNP locus. For the 50 TFs with at least one motif match overlapping with cis eQTLs, we computed score deltas for all motif matches with cis eQTLs (5,165 motif matches across all TFs) and also for all motif matches with the other SNPs that are not cis eQTLs (722 motif matches across all TFs), and then compared the two score delta distributions. As can be seen in Figure 3.10A, overall the cis eQTLs resulted in higher score deltas on motif matching sequences than the other SNPs (rank sum test p-value = 0.0286). We also examined average score deltas for eQTLs and non eQTLs within each TF, after averaging over all motif matches for the TF. We found that among the 50 TFs, 29 TFs had higher average score deltas for cis eQTLs than other SNPs that are not cis eQTLs (Figure 3.10A). This provides evidence that when bound genes are validated under the perturbation by cis eQTLs in our approach, those cis eQTLs are likely to change the binding affinities of regulatory sequences recognized by TFs to lead to expression changes of the bound genes.

When an eQTL overlaps with motif matching sequences for multiple TFs, we do not have knowledge of which TF's binding site is affected by the eQTL. In such cases, so far, we assumed the eQTL influences the binding affinities of all TFs with overlapping motif matches. Instead, we now hypothesize that an eQTL influences the binding of TF with the largest score delta, to see if such an assignment of eQTL to a TF better differentiates between the two score delta distributions for eQTLs and the other SNPs. We first computed the max score delta for each eQTL, defined as the maximum score delta over overlapping TF motif matches. Then, we compared the max score delta distribution across the 302 eQTLs with the score delta distribution that we obtained above for the other SNPs. As can be seen in Figure 3.11A, the max score deltas for eQTLs are significantly higher than the score deltas for the other SNPs (rank sum test, p-value=1.6×10^{-16}). We also examined the eQTL max score deltas for each TF, and compared the max score deltas averaged over motif matches within each TF with the average score deltas for the other SNPs for the same TF. Out of the 45 TFs with more than one motif matches assigned with the max score delta, for 35 TFs, max score deltas for eQTLs were larger than score deltas for the other SNPs (Figure 3.11B). Our results show score delta distributions between SNPs and cis eQTLs were significantly different when cis eQTLs were assigned to TFs with the strongest evidence for a change in binding affinity.

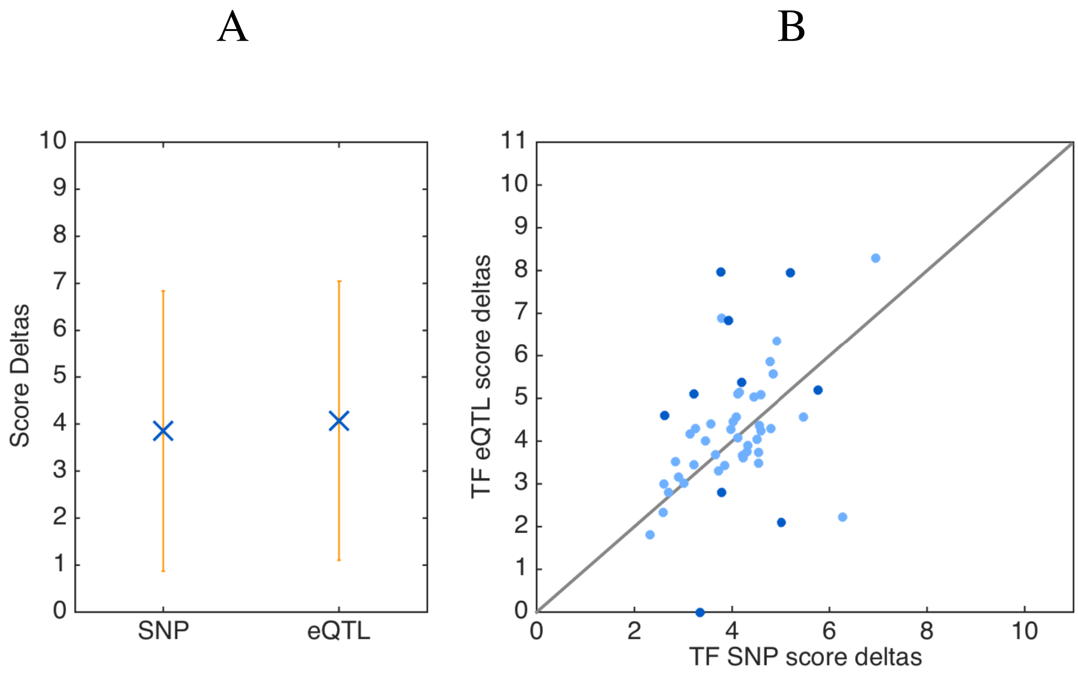

Figure 3.10: Comparing the effects of eQTLs and non eQTLs on TF binding affinities. A) The distributions of score deltas for eQTLs and non eQTLs. The mean and standard deviation are shown as blue dots and orange lines. B) Average score deltas for eQTLs and non eQTLs computed for each TF. Dark blue points represent TFs with significant difference ($p < 0.1$) in the distribution of score deltas.

Perturbing TF structure globally

In order to determine whether trans eQTLs that reside in the TF gene regions and affect target gene expression, by modifying the structure of the TF, we investigated the location of these changes in the protein sequence. To alter protein structure, the trans eQTL must result in a missense mutation that generates a different amino acid sequence. We first found the trans eQTLs located in coding regions of TFs. We then annotated these SNPs using the UCSC variant annotator to find the missense SNPs [118]. To see if these missense mutations affect TF protein structure, we also collected SIFT scores and protein binding domain information from the UCSC variant annotator [118], the Uniprot database [23], and ScanProsite [29]. Out of 301 trans eQTLs, 48 eQTLs were located in coding regions of TFs. Out of these 48 variants, 16 of them were annotated as missense variants. We found that 7/16 of the missense SNPs were associated with binding domains, 6 of them had nearby binding domains that are between 30-100 amino acids away, and 1 of them overlapped with the binding domain. In addition, 3/18 were considered damaging with a SIFT score<0.05, two of these were associated with a binding domain, and one of them was not. In total, we found evidence for structural changes in 8/16 of these trans eQTLs covering 8 different TFs out of 11 TFs. These eQTLs are described in detail in Table 3.4.3.

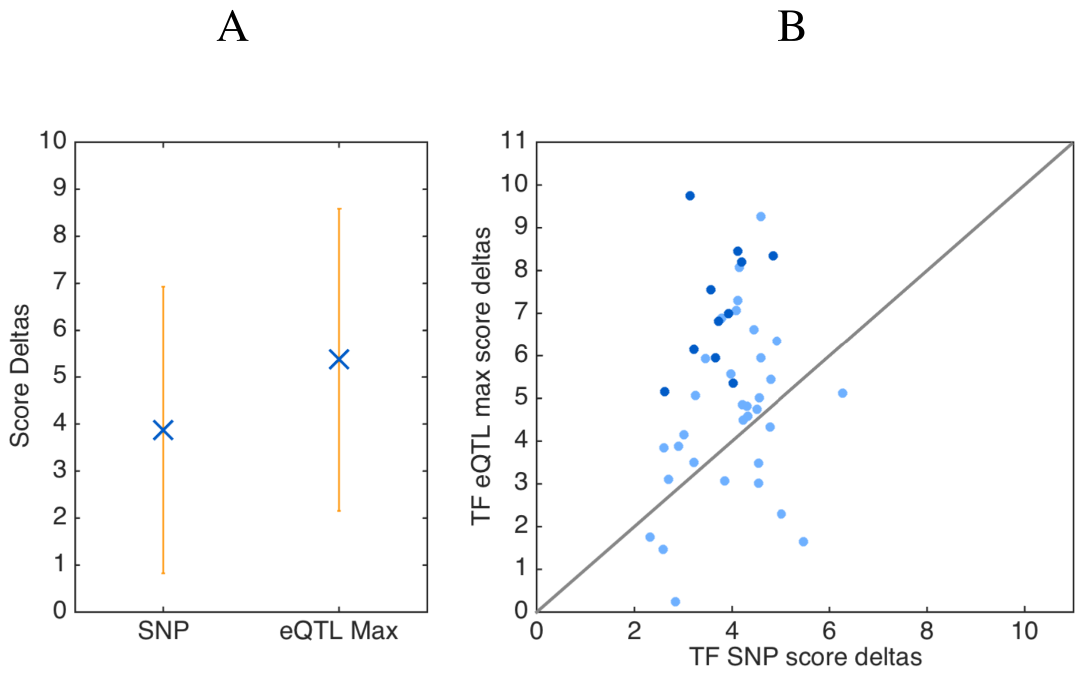

Figure 3.11: Comparing the effects of eQTLs and non eQTLs on TF binding affinities where eQTLs are assigned to the TF with the largest score delta. A) The distributions of max score deltas for eQTLs and score deltas for non eQTLs. The mean and standard deviation are shown as blue dots and orange lines. B) Averaged max score deltas for eQTLs and average score deltas for non eQTLs computed for each TF. Dark blue points represent TFs with significant difference ($p < 0.1$) in the distribution of score deltas.

3.5 Conclusions

In this chapter, we presented a computational approach for validating the TF binding map by using SNP perturbations of gene expressions. We discussed how to learn a cGBN that models the gene regulatory network and outlined an efficient learning and inference procedure. ChIP-seq and DNase I data was used as prior knowledge and our method selected the functional TF-target interactions that were validated under SNP perturbations. We compared the target genes validated by concentration with knockdown experiments and found that epistasis amongst the TFs exists. By analyzing the cis eQTLs that affect target gene expression, we found evidence that cis eQTLs alter binding affinity by modifying TFBSs in the promoter regions of target genes. In addition, we found several trans eQTLs that change the protein sequence of the TF which might alter TF protein structure.

Table 3.1: Regulatory SNPs

Gene	Refseq ID	rSNP	TFBSs	Mean Score Delta	Function
CCL5	NM_002985	rs2107538	EP300, MAZ, SP1	6.5	This rSNP has been linked to atopic dermatitis, atherosclerosis, and prostate cancer. It is also produced by B-cells and attracts monocytes and other leukocytes [53, 68, 99, 158].
FECH	NM_000140	rs17063905	MAZ[1]	5.1	The rSNP was originally studied in the context of erythropoeitic protoporphyria (EPP), a condition resulting from FECH enzyme deficiency [47].
LCT	NM_002299	rs56064699			This rSNP is in the lactase gene [41].

[1] The TFBSs of this TF are also within the ChIP-seq binding site.

Table 3.2: Missense SNPs

TF (RefSeq ID)	SNP ID	SIFT[1]	Amino Acid position[2]	Affected Domain	Domain Type
BRCA1 (NM_007299)	rs1799966	T(0.05)	509	538-632	BRCT
ELF1 (NM_172373)	rs1056820	T(0.84)	343	208-290	ETS
ELF1 (NM_172373)	rs7799	T(0.23)	58		
EP300 (NM_001429)	rs20551	T(0.32)	997	1067-1139	Bromo
FOXM1 (NM_202003)	rs3742076	T(0.91)	628		
NFATC1 (NM_172387)	rs754093	D(0.00)	738	690-930	Trans-activation Domain
PML (NM_033238)	rs5742915	T(1.00)	645		
PML (NM_033239)	rs743580	T(1.00)	772		
PML (NM_033239)	rs743581	T(0.25)	780		
PML (NM_033239)	rs743582	T(0.06)	802		
REST (NM_005612)	rs3796529	T(0.14)	797		
RUNX3 (NM_001031680)	rs6672420	D(0.01)	18	68-196	Runt Domain
SIX5 (NM_175875)	rs2341097	D(0.04)	693		
SIX5 (NM_175875)	rs2014576	T(0.23)	635		
TCF3 (NM_003200)	rs2074888	T(0.38)	492	549-602	bHLH
ZNF143 (NM_003442)	rs10743108	T(0.52)	561	417-440	Zinc Finger

[1] SIFT is a tool based on sequence homology that is used to predict whether amino acid substitutions are likely to affect protein function [102]. Through the UCSC Variant Annotation Integrator [118], we access the SIFT scores. A SIFT score is a probability of observing the substituted amino acid at that position. The scores range from 0 to 1 with a value between 0 and 0.05 indicating that the substitution is predicted to affect protein function [132].

[2] Given the particular transcript of the TF and the variant, the UCSC Variant Annotation Integrator [118] will locate the codon that is affected and the position of the amino acid in the amino acid sequence that is affected.

Chapter 4

Estrogen receptor and cofactor regulation in breast cancer cells

4.1 Motivation

We apply our approach of validating the TF binding map on breast cancer cells in order to study estrogen receptor (ER) regulation. ER plays an important role in breast cancer progression. When activated by the binding of estrogen, ER translocates to the nucleus and binds to target genes which stimulates transcription [121]. Clinicians have found that disrupting ER function is an effective therapeutic strategy [48]. However, researchers are still trying to understand the precise mechanisms of ER action in breast cancer cells. Many coregulatory TFs have been found to be crucial for transcriptional activity of ER [52, 86]. In addition, growth factor receptor tyrosine kinase including those in the ErbB family and the protein kinase families of several signal transduction pathways such as PIK3, AKT and MAPK act upstream of the transcription factors, including ER, to enhance mechanisms leading to cellular proliferation, growth and survival [167]. Thus, these growth factor receptors and protein kinase pathways have been investigated as potential targets for drug therapy. In this study, we use TF binding map validation to elucidate the transcriptional network of the TFs. We also learn the interactions of the protein kinases in order to study the mechanisms of upstream regulation.

While breast cancer is heterogenous with five major subtypes: Luminal A, Luminal B, Basal, Claudin-low and HER2 [57], we focus on luminal A breast cancer. In order to determine the candidate TF-to-target relationships, we obtain ChIP-seq and DNase-seq data from the Cistrome Project database [92] for MCF-7 and T47D cell lines, which are used as example luminal A cell lines [57, 79]. To validate these relationships, we train our cGBN using SNP and expression data from the TCGA Research Network (http://cancergenome.nih.gov/), selecting only patients that have subtype luminal A breast cancers. We then analyze the TF-target relationships that are learned and discuss the effects of local perturbations of TF binding affinity and global perturbations of TF structure.

4.2 Methods

4.2.1 Modeling the gene regulatory network

To validate the TF binding map using SNP perturbations of gene expressions for luminal A breast cancer population, we estimate a cGBN. We follow the same approach that was performed for lymphoblastoid cells as described in Chapter 3 Section 3.2.1. The TF binding map constructed from ChIP-seq and DNase I data was incorporated into our model and learning algorithm as prior knowledge. Then, TF-target interactions that are validated under SNP perturbations of gene expressions are selected by the learning algorithm. Figure 4.1 provides an overview of the TF-target interactions that are validated by each mechanism. The TF-target interactions are validated by concentration, if the TF's expression influences the target gene's expression. The TF-target interaction is validated by local perturbations of binding affinity, if there exists SNP in the regulatory region of the target gene that is associated with the target gene. These SNPs are then referred to as cis eQTLs. In addition, the TF-target interaction is validated by global perturbations, if there exists an SNP in the coding region of the TF that is associated with the target gene. These SNPS are trans eQTLs.

We also validate interactions between protein kinases and TFs (Fig. 4.1). A interaction between a protein kinase and a TF is validated by concentration, if the perturbation in the expression of the protein kinase affects the expression of the TF. In addition, we can validate the protein kinase-TF interaction by binding affinity, if there exists a SNP in the coding region of the protein kinase that influences the expression of the TF. This indicates that a change in the structure of the protein kinase is affecting the expression of the TF. We refer to these SNPs are trans eQTLs.

We learn the network structure such that it reflects the sequence of interactions in the ER biological pathways. Protein kinases are recruited to the ER complex and act upstream of ER and other TFs. The genes pertaining to these protein kinases and their respective families are listed in Table 4.1. We constrain our network structure such that the protein kinases influence the TFs, and the TFs then regulate target genes. We enforce this constraint during learning by partially fixing the variable ordering a priori such that the kinase nodes appear before the TF nodes in the ordering.

4.2.2 Selection of genomic variants

SNPs that were selected for the model were those that were useful in identifying local perturbations of binding affinity or structural changes of TFs or upstream regulators. SNPs within binding sites near target genes were included in order to detect local perturbations of TF binding affinity. In addition, SNPs from the exon regions of TFs were included to identify changes in the structure of the TF. Similarly, SNPs from exon regions of upstream regulators were included to capture SNP perturbations of protein structure.

4.2.3 Experimental design and parameter selection

To learn the gene regulatory network structure, we perform network estimation on the gene expressions conditional on SNPs. Because protein kinases are recruited to the ER complex and are

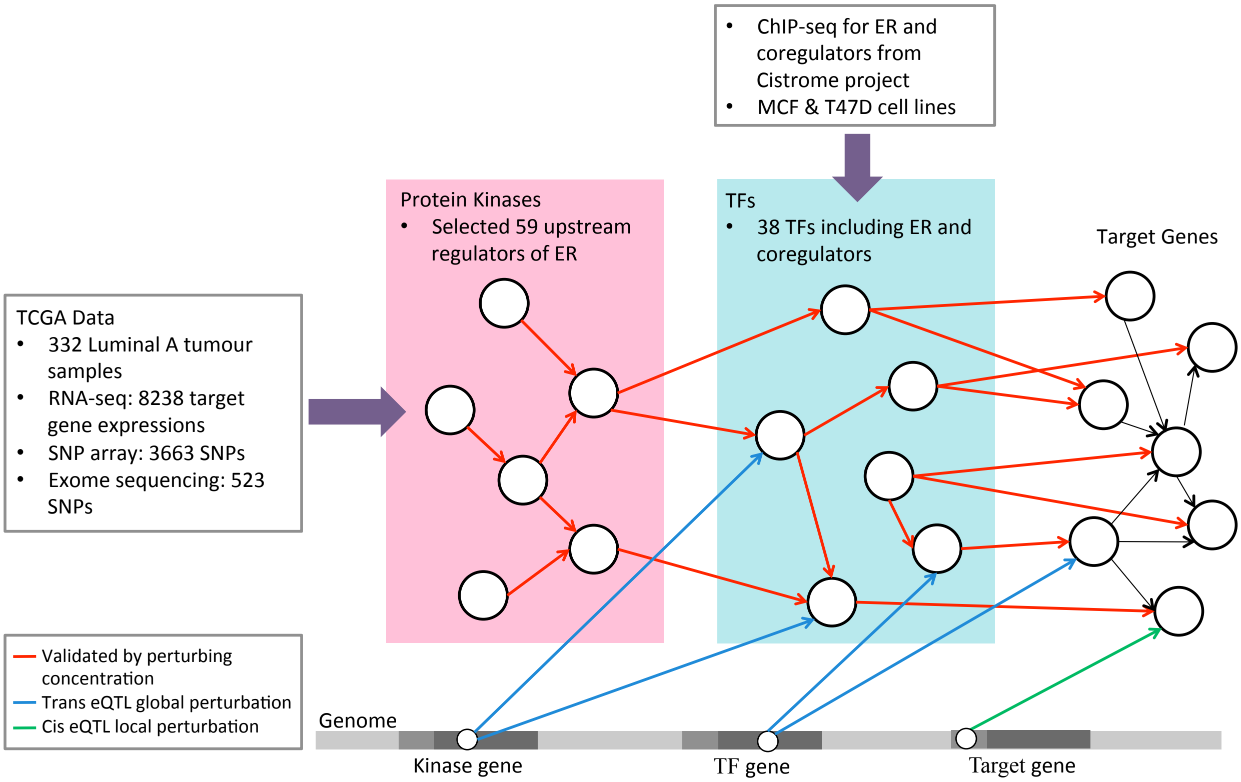
ChIP-seq for ER and coregulators from Cistrome project
MCF & T47D cell lines
Protein Kinases
Selected 59 upstream regulators of ER
TFs
38 TFs including ER and coregulators
Target Genes
TCGA Data
332 Luminal A tumour samples
RNA-seq: 8238 target gene expressions
SNP array: 3663 SNPs
Exome sequencing: 523 SNPs
Validated by perturbing concentration
Trans eQTL global perturbation
Cis eQTL local perturbation
Genome
Kinase gene
TF gene
Target gene

Figure 4.1: Overview of gene regulatory network learned for Luminal A breast cancer cells from SNP and expression data. Red edges indicate those edges that are validated by perturbation of concentration. Blue edges between SNPs and gene expressions mark those edges validated by global perturbations. The associated SNPs are referred to as trans eQTLs. Green edges between SNPs and target gene expressions indicate local perturbations. The associate SNP is referred to as a cis eQTL.

upstream of the TFs, we learn their structure first, and they can have directed edges to TFs. The TFs can have directed edges to other TFs or candidate target genes. The rest of the genes (candidate target genes and other genes) can interact with each other freely. To reduce computational time, we approximate the structure of the entire network as follows. We divide the non-TF genes into clusters that were selected based on hierarchical clustering of the gene expression data. The cluster sizes vary between 200-600 genes. After determining the network structure over the protein kinases and TFs, we then add a cluster of non-TF genes and continue learning the structure. The regularization parameters were selected by cross validation. We explored a range of λ's between 40 and 200, and a range of γ's between 20 and 200. We cross-validated with a 80%20% split of the data. There were 332 luminal A breast cancer samples in total, which means that 265 samples were using for training and 67 samples were used as test data.

Family	Gene Names
Akt	AKT1, AKT2, AKT3
GSK	GSK3A, GSK3A
mTOR	MTOR
ERK	MAPK1, MAPK3
JAK	JAK1, JAK2, JAK3, TK2
PI3K	PIK3CA, PIK3CB, PIK3CG, PIK3CD, PIK3R1, PIK3R2, PIK3R3, PIK3R4, PIK3R5, PIK3C2A, PIK3C2B, PIK3C2G, PIK3C3
SRC	SRC, FYN, YES1, LCK, LYN, HCK, BLK
MAPK	MAPK8, MAPK9, MAPK10, MAPK11, MAPK12, MAPK13, MAPK14, MAPK7, MAPK6, MAPK4, MAPK15
MEK	MAP2K1, MAP2K2
RAF	RAF1, BRAF
S6K	RPS6KA1, RPS6KB1
ErbB	EGFR, ERBB2, ERBB3, ERBB4
IGFR	IGF1R
InsR	INSR
FGFR	FGFR1, FGFR2, FGFR3, FGFR4

Table 4.1: List of upstream regulators of ER grouped by families.

4.3 Datasets

4.3.1 ChIP-seq and DNase I data for the TF binding map

In order to determine the TF binding map in luminal A breast cancer cells, we obtain ChIP-seq and DNase I hypersensitivity data from previous studies performed on luminal A cell lines. We downloaded all ChIP-seq and DNase I studies for MCF-7 and T47D cell lines from the Cistrome Project database [92] that were either done in full media or were estrogen treated. The list of TFs with ChIP-seq data, which cell lines data was available for, and the corresponding studies that generated it is listed in Table 4.2. DNase I data was available for both MCF-7 and T47D cell lines [50, 147] and was obtained from the Cistrome Project database [92]

For each TF and each cell line, we generate a consensus set of binding sites. If a TF had multiple samples within the same study, we determined a consensus set of binding sites for that study by taking the majority vote across all samples using the DiffBind R package [136]. Similarly, for TFs that had multiple ChIP-seq studies available, we found a consensus set of binding sites by taking the majority vote across all studies.

To filter the resulting binding sites for the regions that exist in open chromatin, we use the information from DNase I hypersensitivity data. For each cell line, we find the regions of the TF's ChIP-seq binding sites that are contained in DNase I hypersensitivity sites. The resulting set of binding sites make up the TF binding map for each TF and each cell line. For TFs, that had data available for both MCF-7 and T47D cell lines, we took the union of the binding sites to create a single set of binding sites.

To generate candidate TF to target gene relationships, we must associate target genes with the TF binding map of a particular TF. We consider a gene a candidate target of a particular TF if there are binding sites for the TF within 100kb upstream of its transcription start site.

4.3.2 Genome and Transcriptome data from breast cancer cohorts

SNP Data

To validate TF-target gene relationships in MCF-7 and T47D cell lines under SNP perturbations, we identified 332 individuals of luminal A subtype with both SNP and expression data available from the TCGA project. SNP data in upstream regulatory regions of genes were obtained from SNP arrays. SNPs in coding regions of TFs and upstream regulators were primarily taken from exome sequencing data.

In order to process SNP array data, we filter SNPs based on confidence and minor allele frequency (MAF). To do this, we first downloaded TCGA SNP array data. The original set of 906,600 SNPs was first filtered based on confidence. SNPs that were included had at least 80% of the samples called with confidence$<$0.1. Next, SNPs were filtered based on minor allele frequency with those with MAF$>$0.01 being included in the analysis. We threshold the SNPs by MAF because we can only detect SNP perturbations of gene expression where there exists variation in the population.

To study SNP perturbation of binding sites, we select SNPs for our analysis that were located in the TF binding map associated with each gene. A total of 3646 SNPs were found in the TF binding map near TFs and target genes. Table 4.3.2 specifies the number of genes that had SNPs in the TF binding map and whether they had somatic or germline mutations. If at least one individual differed at the SNP between its normal and tumor cells, we listed this SNP as a somatic mutation. We also show the number of genes with SNPs after thresholding at MAF$>$0.01. In our analysis, 3210 genes had SNPs located within the TF binding map, while 5066 genes (including TFs) did not.

To find SNP perturbations of protein structure for upstream regulators and TFs, we include those SNPs that were in exon regions which potentially affect the protein sequence. The SNPs were obtained from exome sequencing data and SNP arrays from the TCGA. The pool of variants considered from exome sequencing data were those that were present both in the population of individuals sequenced at UCSC and in the population of individuals sequenced at WUSC. These SNPs were then filtered such that those with MAF> 0.01 were included in our analysis. A small number of SNPs located in exon regions from SNP arrays: three SNPs found in TFs and 14 SNPs found in of upstream regulators were also included. The total number of somatic and germline mutations collected for TFs and upstream regulators in exon regions are listed in Figures 4.3 and 4.4.

RNA-seq data

To validate the candidate target genes proposed by ChIP-seq, we identified 332 individuals with luminal A breast cancer cells and had their gene expression levels available. We obtained RNA-seq Version 2 gene-level expression data from the TCGA. We obtained processed data where the gene-level expression measurement was already provided (Level 3 data). There was originally a

TFs	MCF-7	T47D	Reference
BRD4	✓		[101]
CTBP1	✓		[30]
CTCF	✓		[60]
DNase I	✓	✓	[50, 147]
EHMT2	✓		[131]
EP300	✓	✓	[73, 96, 145]
ESR1	✓	✓	[34, 60, 66, 81, 96, 101, 145]
FOS	✓		[66]
FOXA1	✓	✓	[34, 60, 66, 85, 143]
GATA3	✓	✓	[1, 73, 85, 145]
GREB1	✓		[96]
HECTD1	✓		[81]
HIF1A		✓	[169]
JUN	✓		[66]
KLF4	✓		[95]
LMTK3	✓		[163]
MBD3	✓		[130]
MED1	✓		[85]
MTA3	✓		[131]
MYC	✓		[22]
NCAPG	✓		[81]
NCAPG2	✓		[81]
NR2F2	✓		[95]
NR5A2	✓		[9]
NRF1		✓	[169]
POLR2A	✓		[66]
PR	✓	✓	[96]
RAD21	✓		[123]
RELA	✓		[34]
RXRA	✓		[95]
SPDEF	✓		[33]
STAG1	✓		[123]
TDRD3	✓		[164]
TFAP2A	✓		[143]
TFAP2C	✓		[85]
TLE3	✓		[95]
TOP2B	✓		[87]
TP53	✓		[171]
TRIM24	✓		[150]

Table 4.2: ChIP-seq data from the Cistrome Project that was included in our study.

total of 17023 genes. We included expression data from 59 upstream regulators and TFs. For the rest of the genes, after log-transforming the data and filtering for standard deviation >0.8, we included the 8238 target genes that passed the threshold.

	Number of genes	Number of genes (SNP MAF >0.01)
Genes with SNPs	3347	3210
Genes with somatic mutations	3274	3173
Genes with germline mutations	103	77

Table 4.3: Somatic and germline mutations in nearby upstream regions of genes.

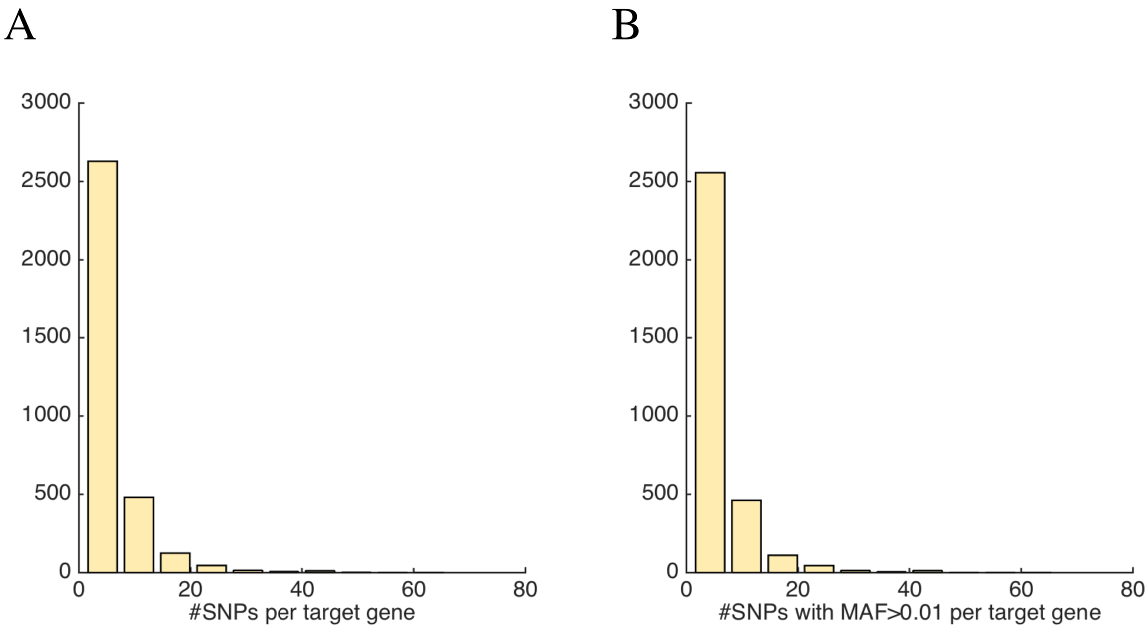

Figure 4.2: SNPs within 100K upstream of the transcription start site of target genes. A) Histogram of number of SNPs associated with target genes. B) Histogram of number of SNPs associated with target genes with MAF>0.01.

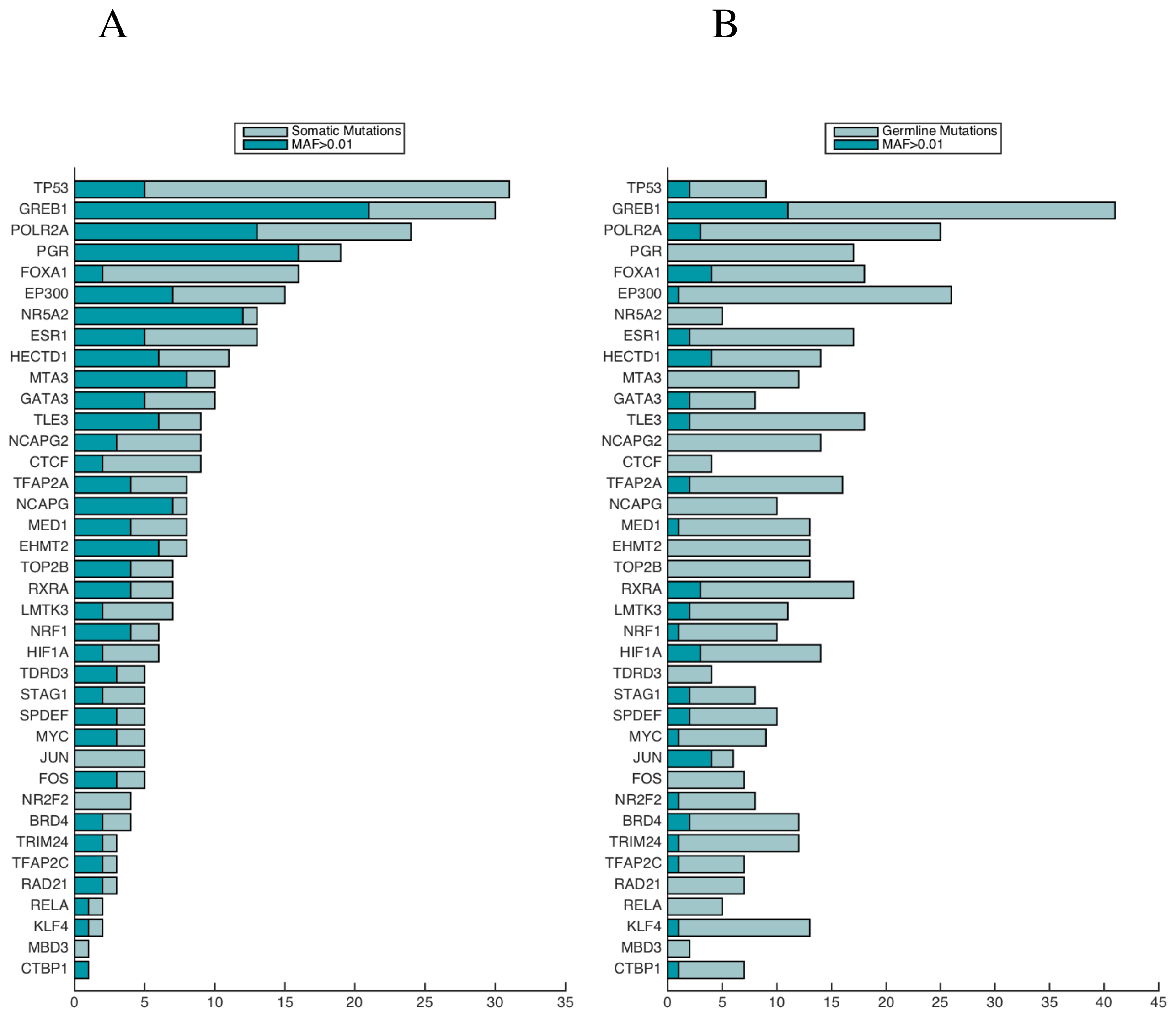

Figure 4.3: Somatic and germline mutations collected in exon regions of TFs. A) Number of somatic mutations for in luminal A samples. B) Number of germline mutations in luminal A samples.

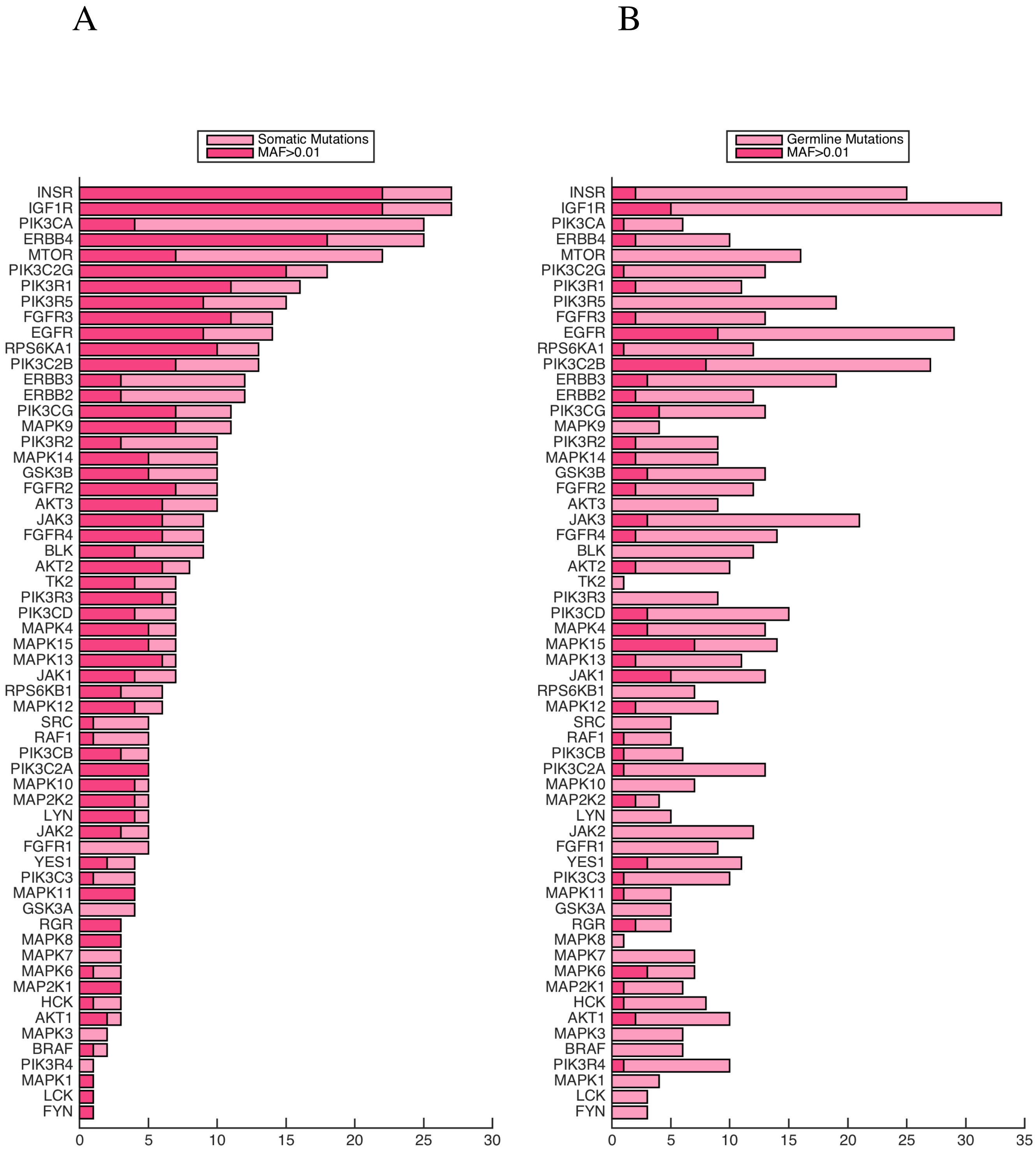

Figure 4.4: Somatic and germline mutations collected in exon regions of ER upstream regulators. A) Number of somatic mutations in luminal A samples. B) Number of germline mutations in luminal A samples.

4.4 Results

4.4.1 Overview of upstream regulation and validated TF binding events

We validate the TF binding map of ER and other pertinent TFs and investigate the influence of upstream regulators. To do this, we validate the TF-binding map determined from ChIP-seq and DNase I data from the Cistrome project using SNP and expression data from the TCGA. Overall, we validated TF bindings for 38 TFs and looked at interactions of 59 upstream regulators. We functionally validated these interactions by using expression and SNP data from 8238 genes and 4,302 SNPs.

The estimated network structure of TFs involved in ER regulation contains previously known interactions. We plotted the network structure of ER and other important coregulating TFs that was generated by our method. In Figure 4.5A we show the correlation between the gene expressions of the TFs. In Figure 4.5B, we show the adjacency matrix that corresponds to the network structure that was estimated. The strength of the interaction is indicated by color. We found that several edges between TFs listed in Table 4.4 are known interactions that have been previously validated by experimental approaches in the literature. A diagram of the estimated network of TFs is shown in Figure 4.6 with the known interactions highlighted. From examining this estimated network, we can conclude that FOXA1 is an important coregulator of ER that shares direct interactions towards GATA3, PGR. In addition, ER expression influences GREB1 and TLE3.

From	To	Reference
ESR1	GATA3	[153, 159]
ESR1	PGR	[96]
ESR1	GREB1	[95, 115]
ESR1	TLE3	[64, 100]
FOXA1	ESR1	[60]
FOXA1	MYC	[103]
FOXA1	PGR	[3]
FOS	GREB1	
PGR	GREB1	[170]

Table 4.4: The regulatory relationships from TFs to targets identified by our algorithm and supported by evidence in the literature.

Examining the network structure of the upstream regulators which are protein kinases that are recruited to the ER complex show that families of protein kinases interact with each other. We plotted the correlation between gene expressions of upstream regulators in Figure 4.7A, and the adjacency matrix corresponding to the network structure of these genes in Figure 4.7B. It can be seen that several members of the Src family of kinases, LYK, FYN, BLK, LYN, and HCK, are connected. In addition, many of the MAPK and PIK3 families are connected in the network (Fig. 4.7). MTOR is also connected to these families which is an interesting observation since PI3Ks play a role in the MTOR signaling pathway [112].

Studying the network interactions of upstream regulators show that two particular groups of protein kinases regulate two distinct groups of TFs, one involved in transcription regulation,

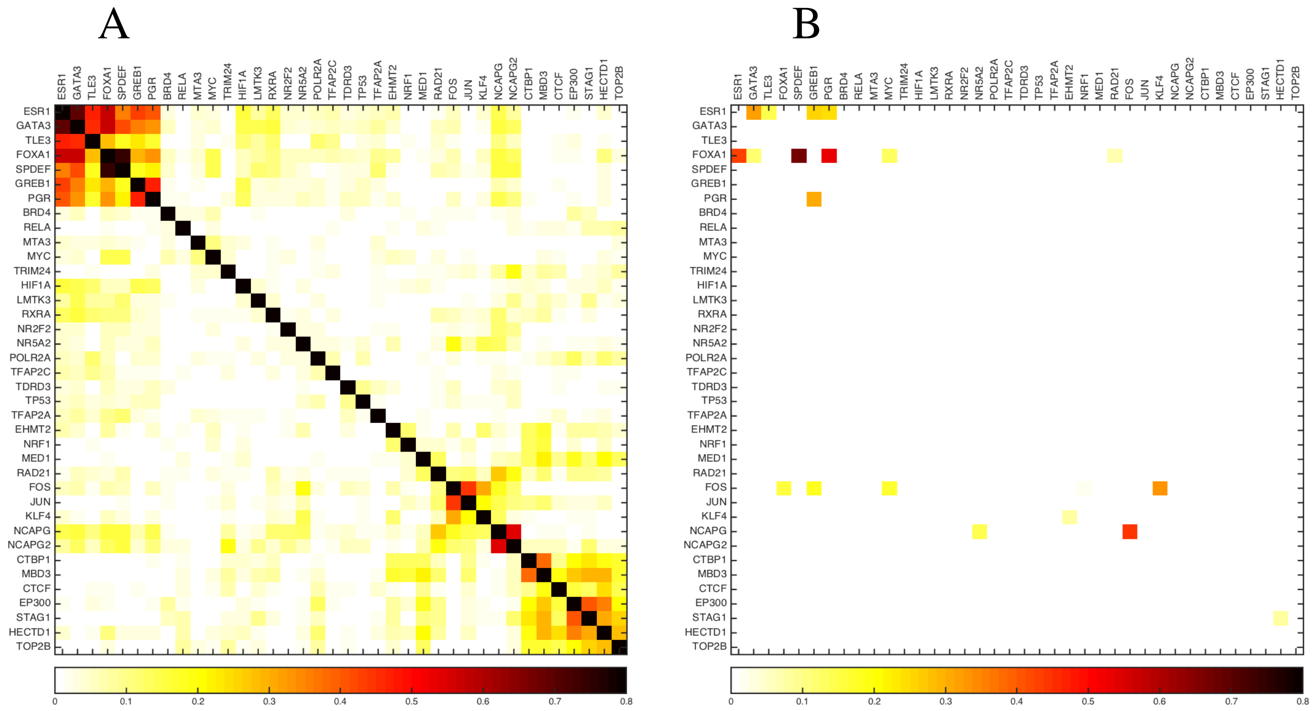

Figure 4.5: Correlation and estimated network structure of TFs. A) The correlation R^2 of gene expressions of TFs. B) The edges of the estimated network between TFs.

and other involved in chromatin regulation. We illustrate the correlation pattern between upstream regulators and TFs in Figure 4.8A and the adjacency matrix showing the network estimated by our method in Figure 4.8B. The network recovered shows that a group of protein kinases MAPK4, EGFR, ERBB3, ERBB4 and IGF1R regulate a group of TFs ESR1, GATA3 and FOXA1. The activity of these TFs have previously been shown to be important for luminal tumours. Specifically, ESR1 drives tumour growth and both FOXA1 and GATA3 have been shown to influence transcription of ESR1 [145]. In addition, a different group of protein kinases MAPK2, PIK3CA, MAPK8, PIK3C2A, and MAP2K2 have direct interactions with MBD3, CTCF, EP300 which are involved in chromatin remodeling and STAG1 and TOP2B which participate in DNA replication.

We quantify the validated interactions between upstream regulators and TFs. For each upstream regulator, we show the overall number of TFs that it regulates in Figure 4.9A. In addition, we also show those edges validated by concentration (Fig.4.9B), and those edges validated by global SNP perturbations in exon regions of upstream regulators (Fig.4.9C). Figure 4.10A illustrates the total number of SNPs that were located in exon regions of upstream regulators and how many of them were selected as trans eQTLs. Figure 4.10B shows the number of TFs regulated per trans eQTL. This indicates that most trans eQTL affects a small number of TFs.

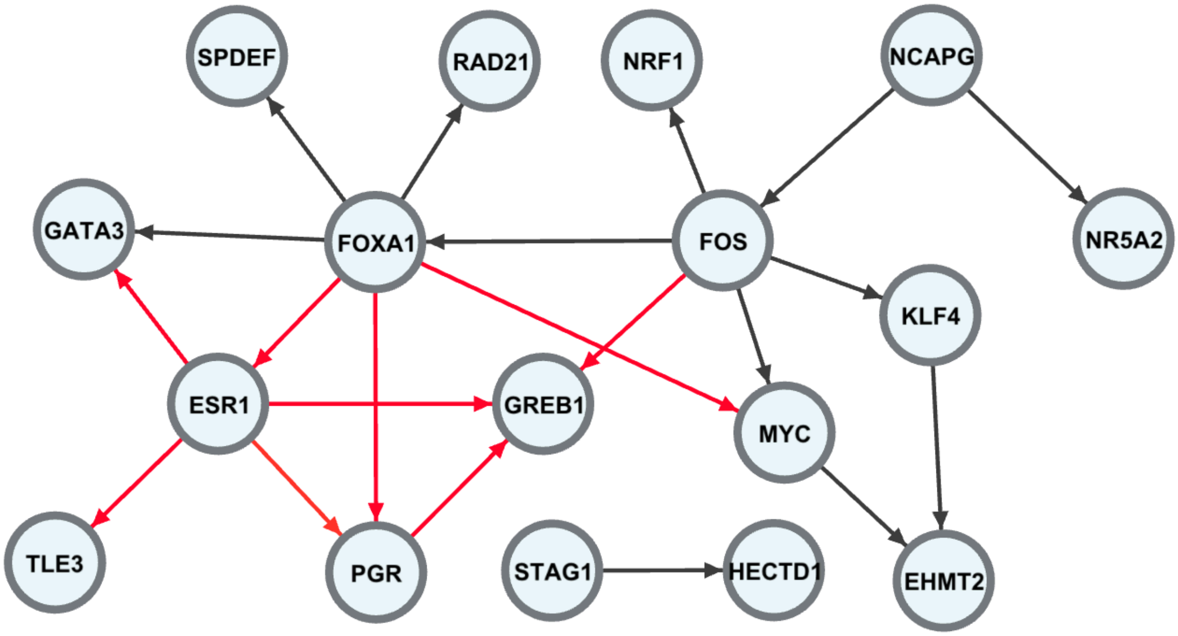

Figure 4.6: Estimated network structure of ER and TF coregulators. Black indicates edges recovered by our model, red indicates edges supported by evidence in the literature.

We summarize the overall gene regulatory network by quantifying the TF-target interactions. From the estimated GBN, we determined that the fraction of validated target genes to candidate target genes according to the TF binding map was between 1.4% to 46.5% depending on the TF (Fig 4.11A). For each TF, we then further categorize the validated genes by the mechanism by which they were validated. Figure 4.11B shows the target genes that were validated by concentration. Figure 4.11C shows that target genes that were validated by local SNP perturbations in regulatory regions of target genes. Figures 4.11D shows the target genes validated by global SNP perturbations from exon regions of TFs. Figure 4.12A illustrates the total number of SNPs that were located in exon regions of TFs and how many of them were selected as trans eQTLs. Figure 4.12B shows the number of target genes per trans eQTL. It can be observed that that number of most of the target genes were validated by perturbation of concentration. ESR1, FOS, FOXA1, GATA3 and GREB1 are highly active as regulators as they have a large number of target genes. In addition, there are a number of trans eQTLs that have greater than five target genes illustrating that these global SNP perturbations can exhibit an effect across multiple target genes.

After examining the results from perturbing concentration, we find that the number of candidate targets and the expression variance of the TF in the population affects the number of target genes that can be validated by our approach. There are a couple factors that affect that number of targets recovered for each TF. First, the number of candidates targets affect how many targets we are able to validate (Fig. 4.13). In addition, we find that the expression variance of the TF in the population affects the number of target genes that can be validated by our approach (Fig. 4.13). A TF that has a high expression variance also has a large number of target genes assuming that the TF binding map suggests many candidate targets for this TF.

The highly active TF regulators also share many targets. We plotted the adjacency matrix where each position shows the number of common targets validated by perturbation of concentrated that were shared by the pair of TFs. ESR1, GATA3, FOXA1 and GREB1 all share many targets (Fig. 4.14) indicating that together they coordinate target genes that lead to luminal A breast cancer in the cell.

4.4.2 Overview of differentially expressed genes

In order to identify genes that are differentially expressed as a result of TF-target interactions, we performed inference on the cGBN. We identified those genes that were differentially expressed as a result of perturbation of TF concentration, local perturbations of target genes, and global perturbations of TF gene regions (Fig. 4.15). By examining the genes that were differentially expressed due to concentration, it can be seen that despite having few validated targets, MYC, NRF2, RAD21, STAG1, TFAP2C have many differentially expressed genes indicating that the downstream effects of these TFs may be important in breast cancer progression. The TFs MYC, RAD21, STAG1 are all involved in modifying the chromatin during cell cycle progression and TFAP2C has been found to regulate ER expression in breast cancer [111]. Many of the same TFs

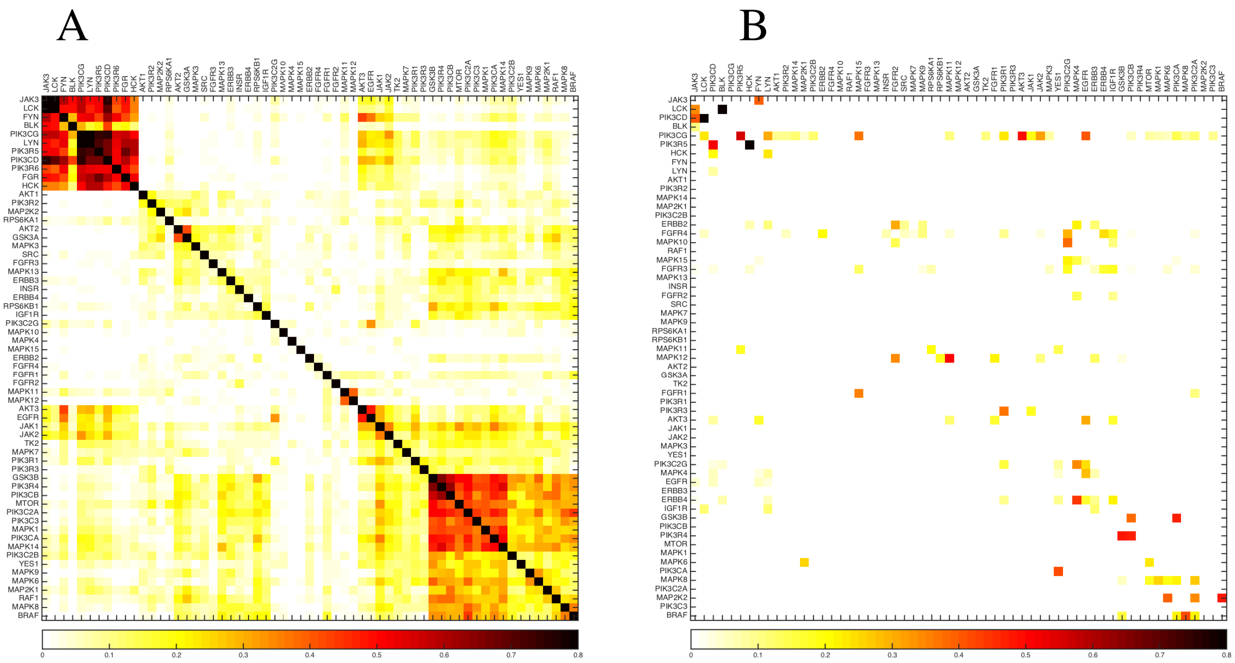

Figure 4.7: Correlation and estimated network structure of upstream regulators. A) The correlation R^2 of gene expressions of ER upstream regulators . B) The edges of the estimated network among ER upstream regulators.

RAD21, STAG1, TFAP2C also have a relatively large number of differentially expressed genes as a result of SNP perturbations locally and globally. In addition, local SNP perturbations of CTCF and CTBP1 also have a larger number of differentially expressed genes. These TFs were discussed previously as being chromatin modifiers.

Next, we identify differentially expressed genes as a result of the upstream regulators to TF interactions by performing inference. We determined those genes that were differentially expressed as a result of perturbation of the concentration of the upstream regulators and perturbations of their gene regions (Fig. 4.16). From examining the downstream genes of protein kinases, we observe that EGFR, ERBB2-4, IGF1R, FGGR3-4 have a large number of differentially expressed genes. This is expected since it is known that the EGF and ErbB family receptors are over expressed in breast cancer [137, 165].

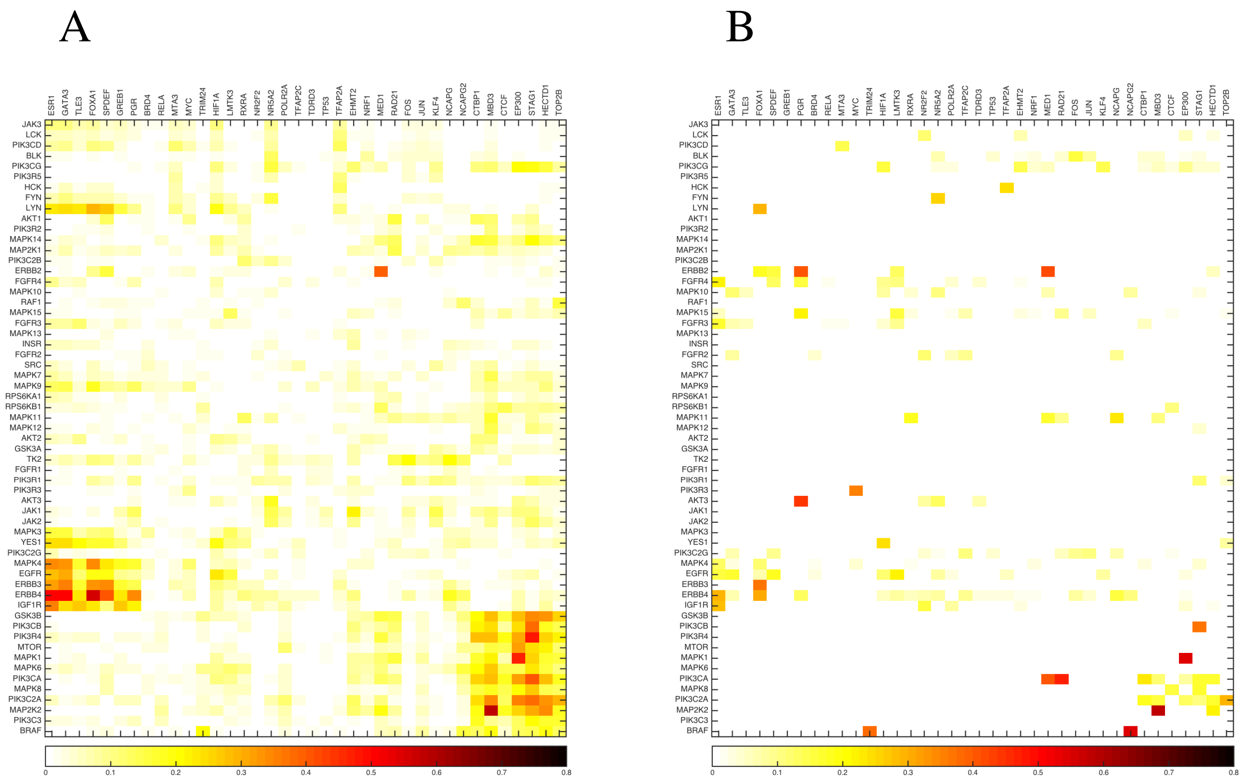

Figure 4.8: Correlation and network structure of upstream regulators to TFs. A) The correlation R^2 of gene expressions of upstream regulators to TFs. B) The edges of the estimated network between upstream regulators and TFs.

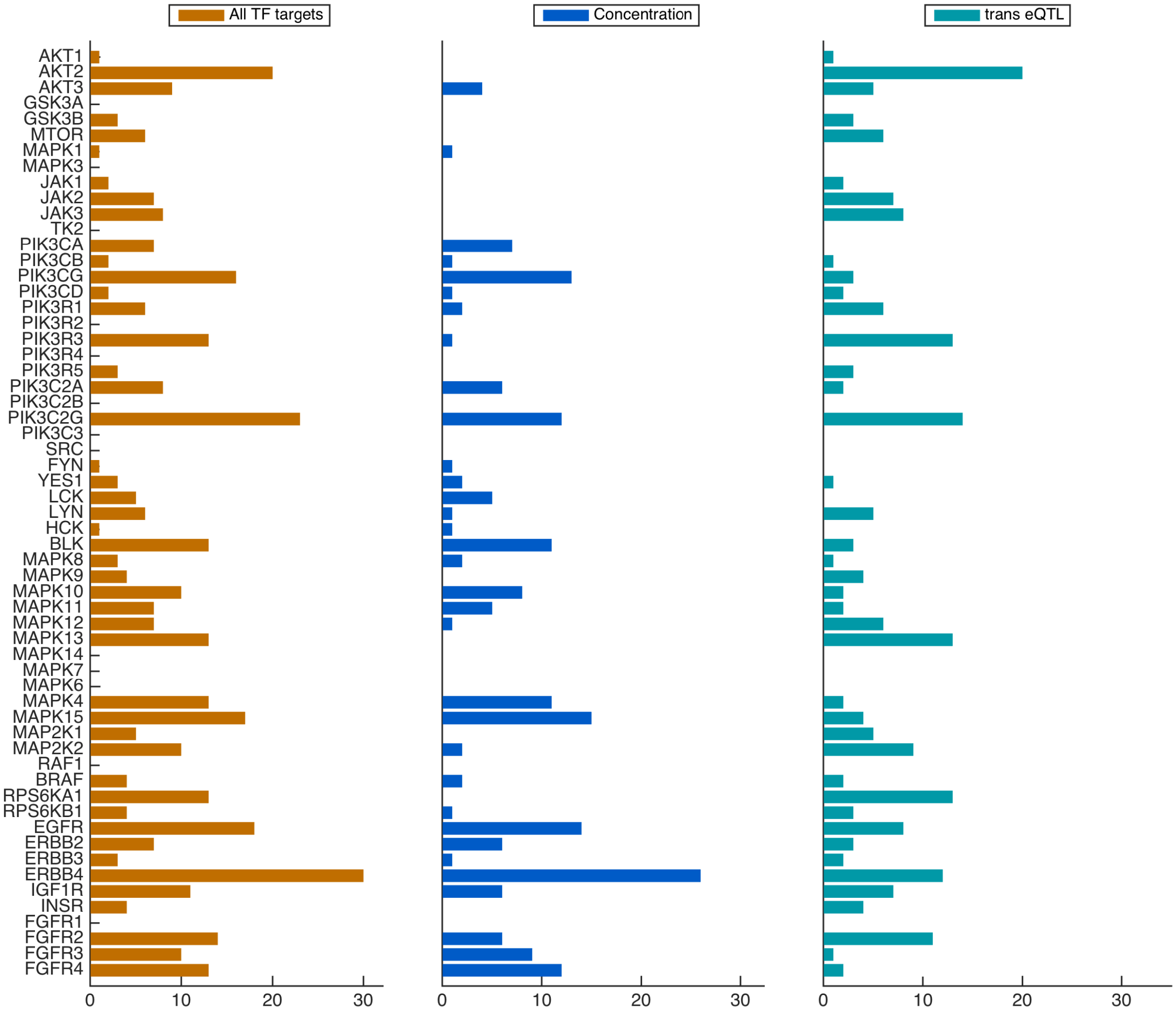

Figure 4.9: ER upstream regulators and their validated target genes. The set of potential target genes are ER and its TF coregulators. A) All target genes. B) All target genes validated by concentration. C) All target genes validated by global SNP perturbations.

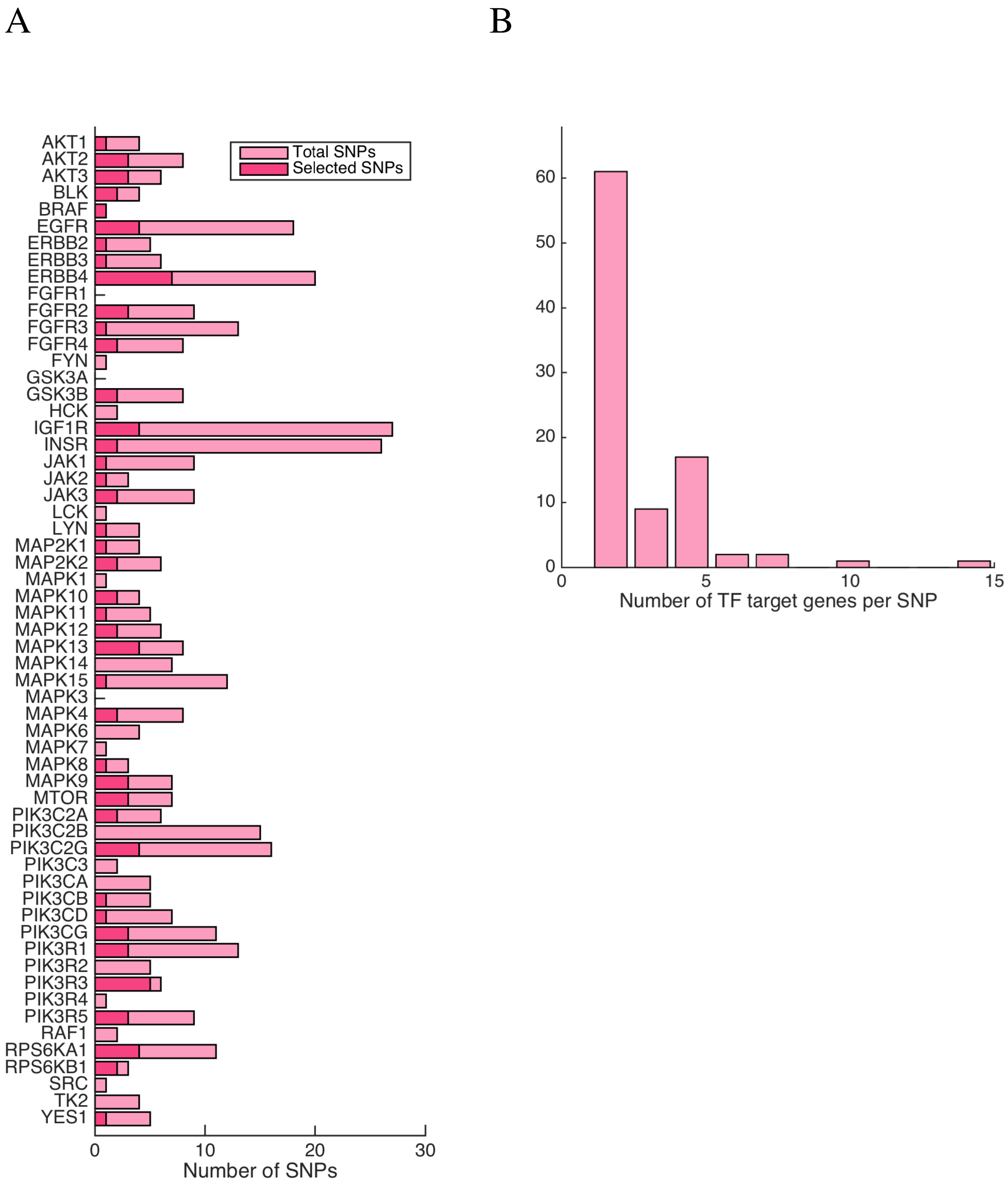

Figure 4.10: Global perturbations of TF expressions by SNPs located in exon regions of ER upstream regulators. A) Total number of SNPs in exon regions and those selected as eQTLs. B) Histogram of the number of targets genes per trans eQTL.

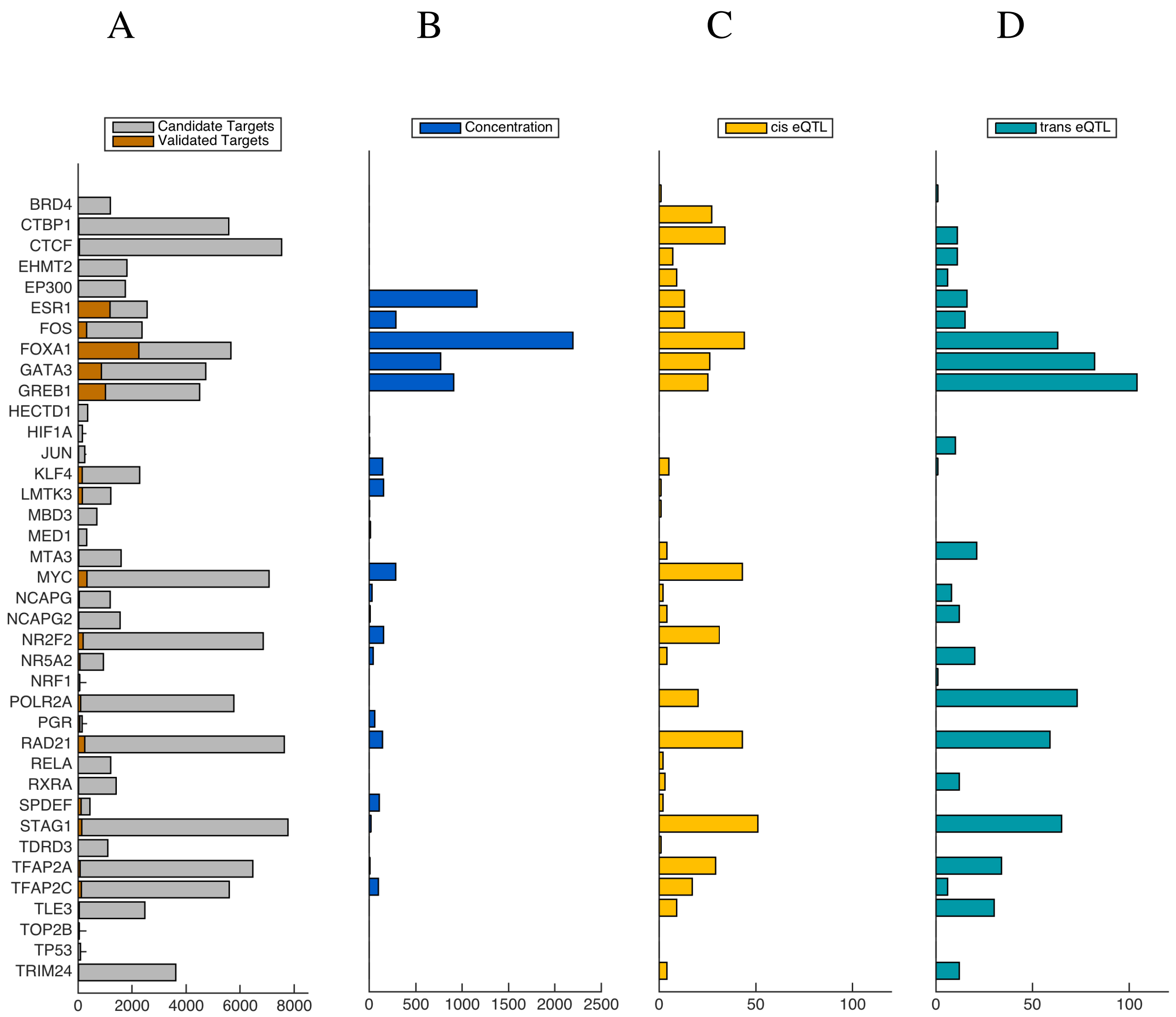

Figure 4.11: Summary of validated TF to target relationships. A) The candidate target genes are shown in gray and the validated target genes across all mechanisms are shown in brown. B) Target genes validated by concentration. C) Target genes validated by local SNP perturbations of binding affinity D) Target genes validation by global SNP perturbations.

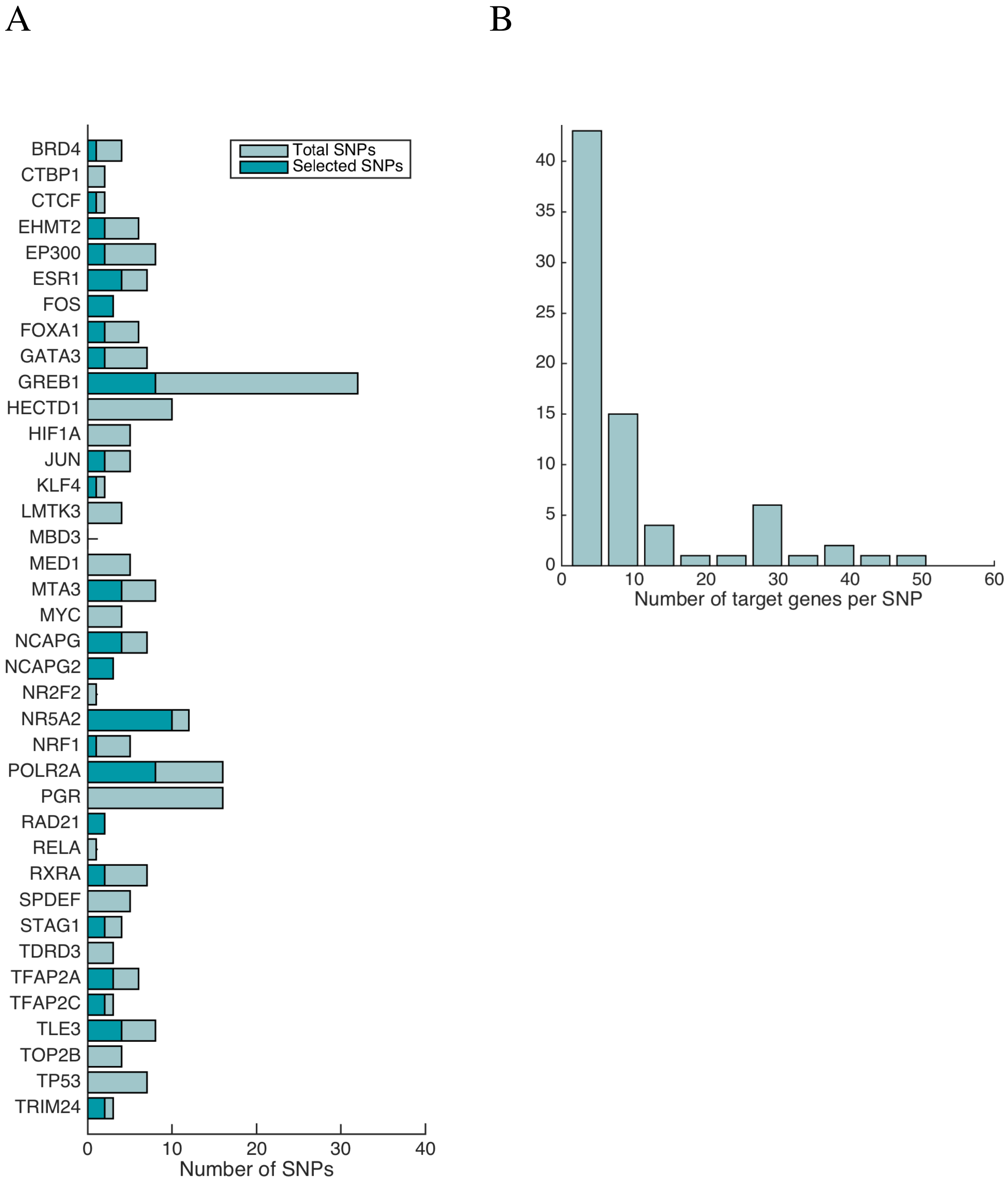

Figure 4.12: Global perturbations of SNPs in exon regions of TFs. A) Total number of SNPs in exon regions and those selected as eQTLs. B) Histogram of the number of targets genes per trans eQTL.

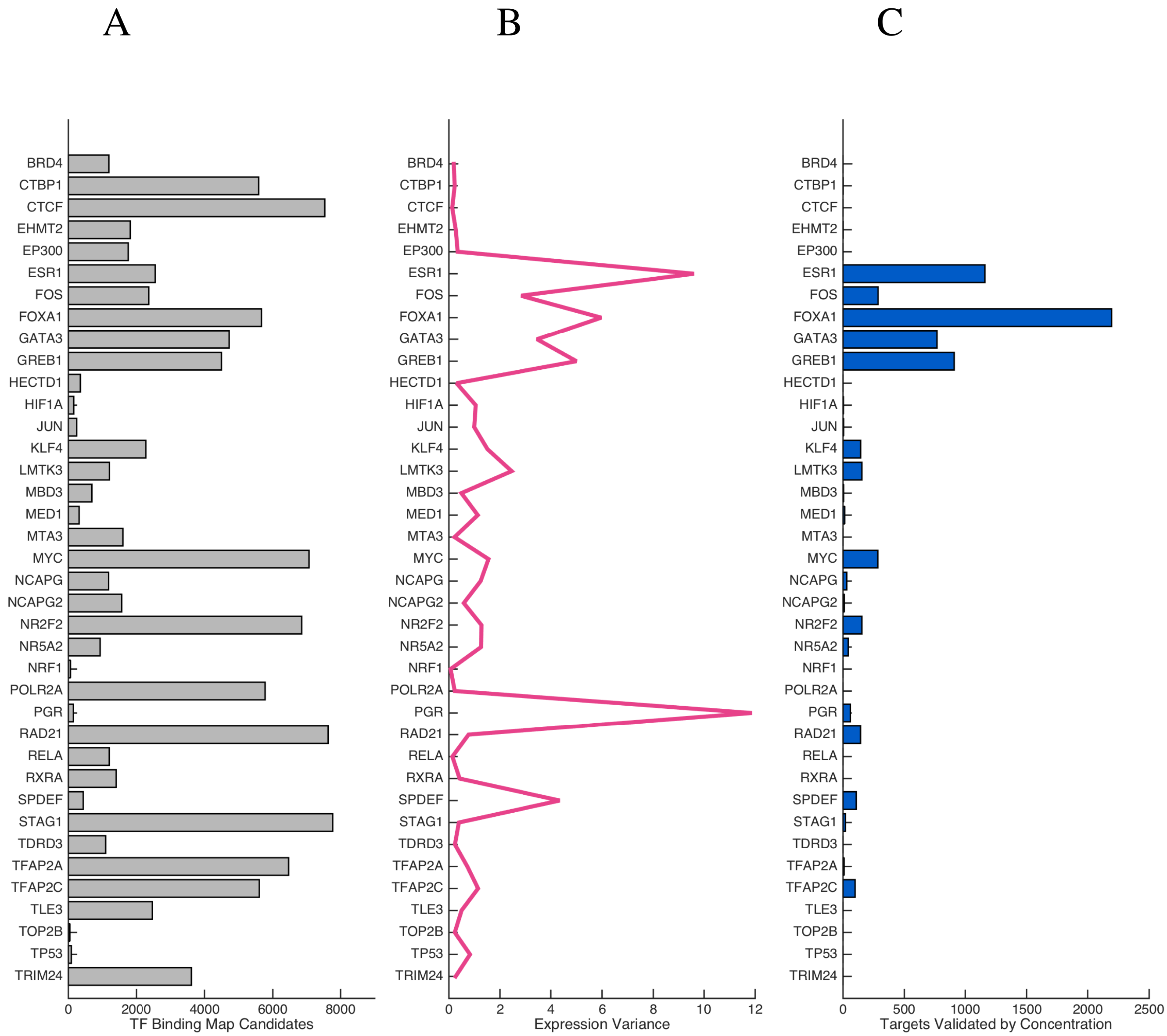

Figure 4.13: TF binding events validated by concentration. A) Candidate edges between TFs and target genes from the TF binding map. B) Expression variance of TF gene expression data. C) Number of target genes validated by concentration.

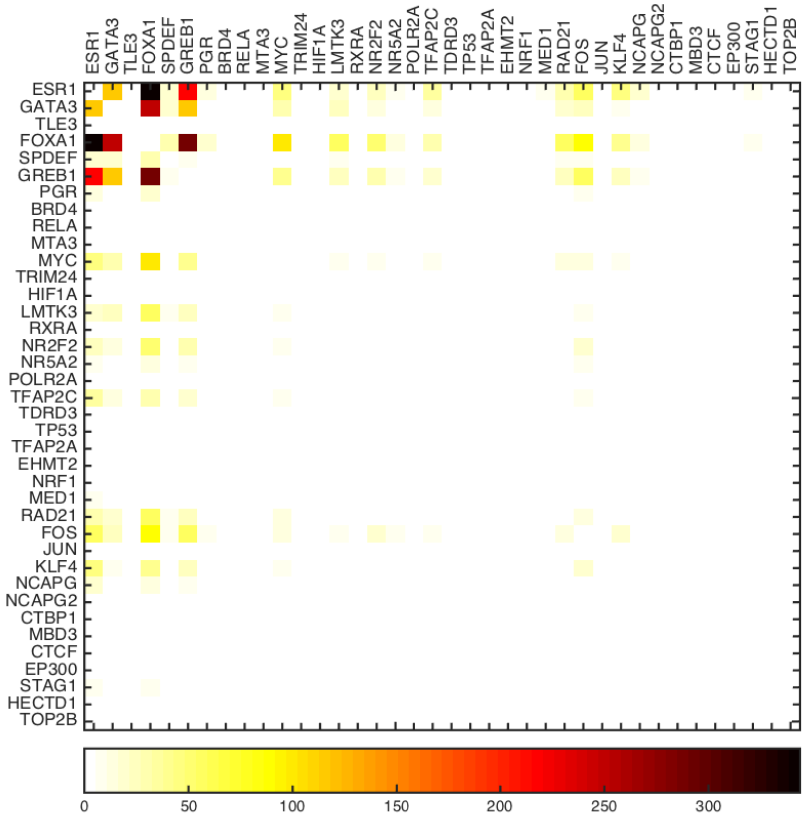

Figure 4.14: Number of common target genes of TFs validated by concentration.

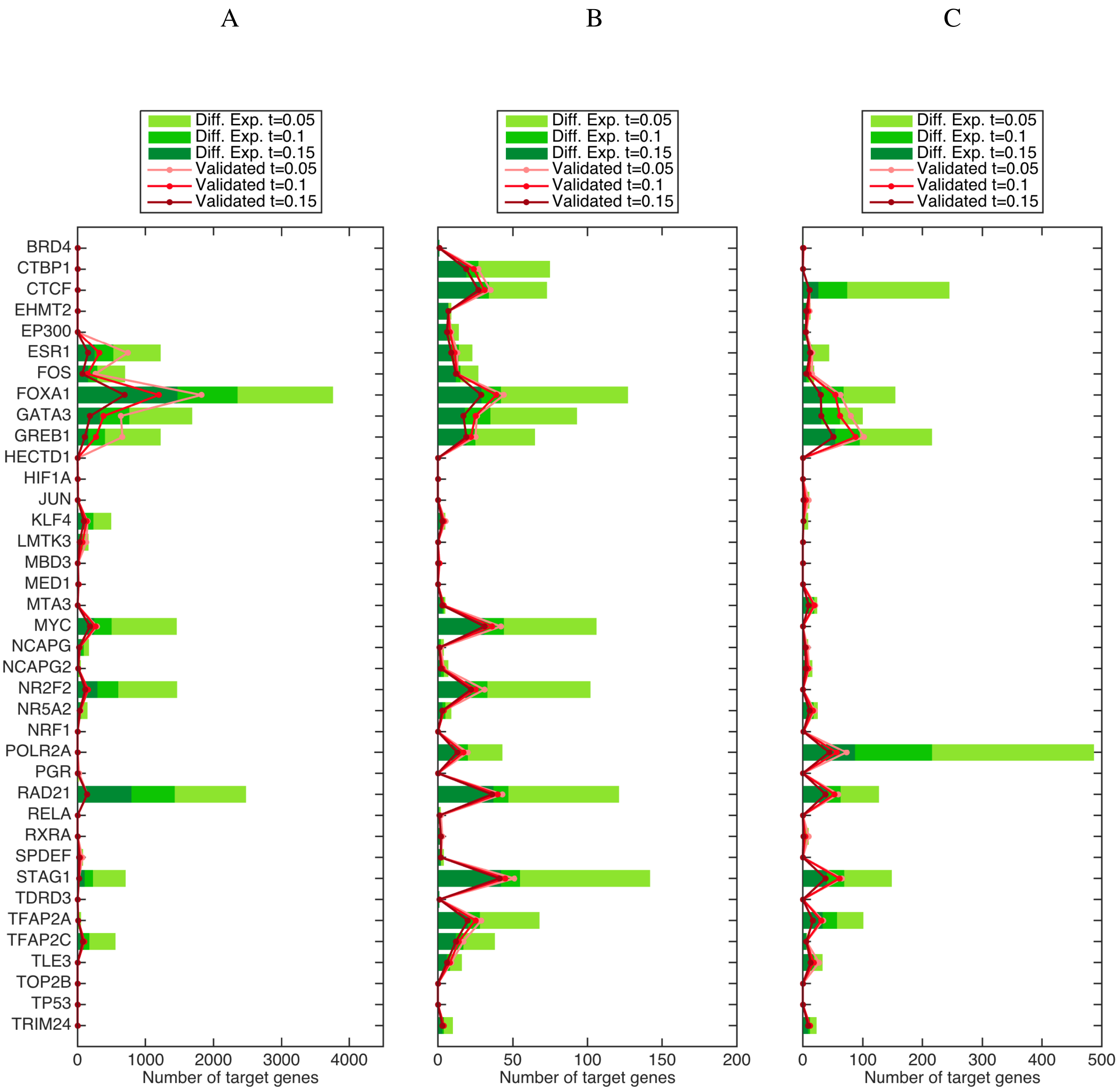

Figure 4.15: Differentially expressed genes for each TF. Each panel represents genes that are differentially expressed by A) TF concentration, B) TF binding affinities locally, and C) TF binding affinities globally.

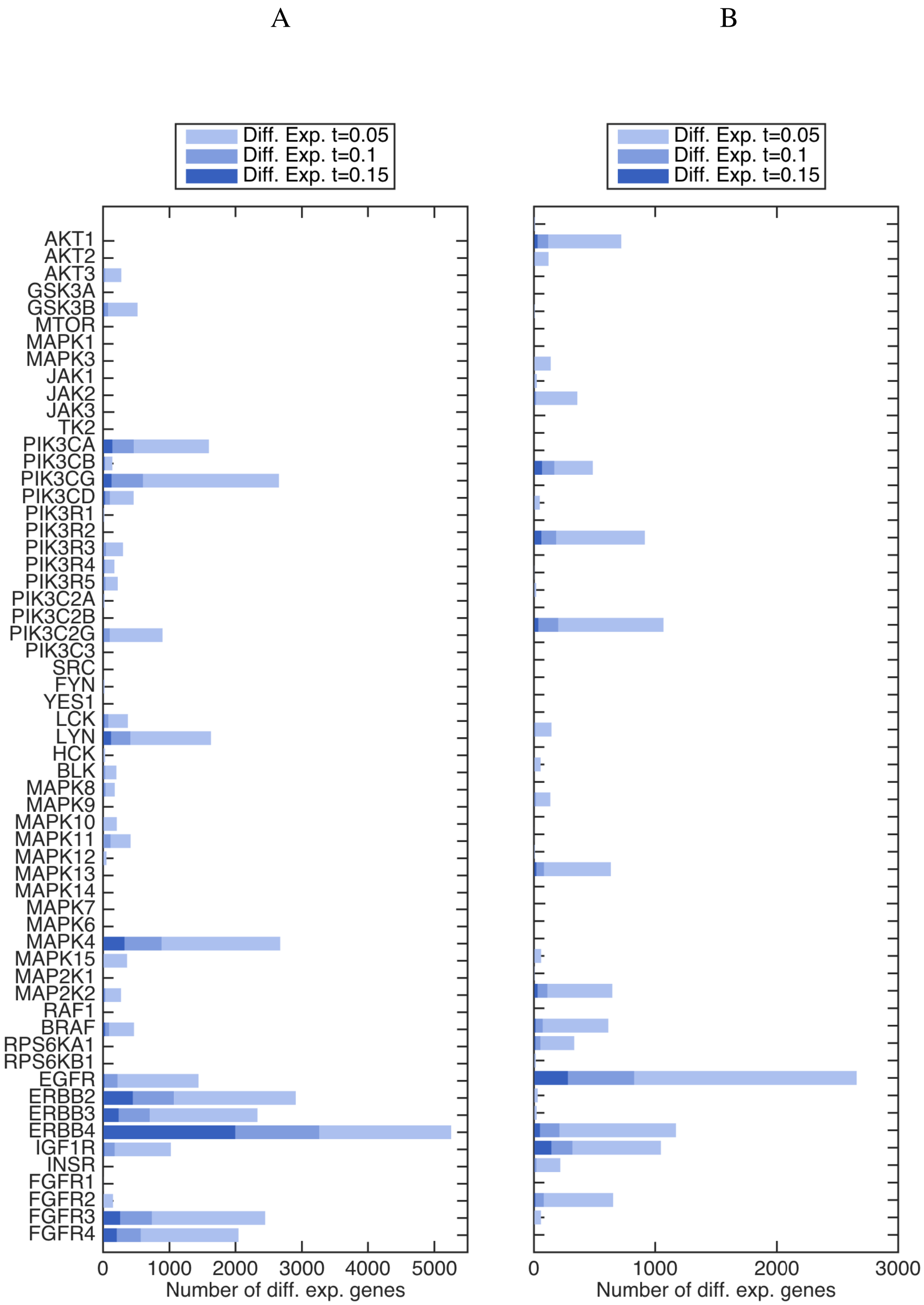

Figure 4.16: Differentially expressed genes for each upstream regulator. Each panel represents genes that are differentially expressed by A) concentration, B) global binding affinities.

4.4.3 Perturbing binding affinities of transcription factors

To observe whether cis eQTLs of validated target genes influence gene expression through modifying transcription factor binding sites (TFBSs), we find the cis eQTLs that overlap with TFBSs in the upstream region of the transcription start site (TSS). The TFBSs were collected by scanning sequences of the TCGA luminal A breast cancer population with motif position weight matrices (PWMs) from the the TRANSFAC and JASPAR databases [90, 91]. We were able to obtain PWMs for 22 out of the 38 TFs in our analysis. We only scanned regions that were in the TF binding map, that is regions that were were both ChIP-seq bound and in open chromatin, as determined by ChIP-seq and DNase I hypersensitivity data from the Cistrome Project [92]. For each target gene, we scanned a region that is 100kb upstream of the TSS. Given the PWM, we derived the corresponding log-odds scores and used the resulting position specific scoring matrix (PSSM) to score TFBSs. We thresholded TFBSs such that they passed a PWM-specific score threshold (p <0.001). We collected all the TFBSs, those binding sites that were polymorphic in the population. There were a total of 12793 polymorphic loci in the TF binding map and 472 of them overlapped with eQTLs. Of those, 1240 polymorphic loci intersected with TFBSs and 61 of those overlapped with eQTLs. In total, there were 3646 SNPs and we determined that 150 of them were eQTLs. Of these, there were 949 SNPs that intersected with TFBSs and 40 eQTLs that intersected with TFBSs.

To see if the cis eQTLs in TFBSs do change TF binding affinity, we compared the differences between the effects of SNPs and those of cis eQTLs. For each polymorphic site in the population, we measured the effect of a SNP or eQTL on binding affinity by finding the difference in PWM scores between the TFBSs that contain alternate alleles. We refer to these differences as score deltas. Detailed discussion of how these score deltas are obtained is provided in Section 3.4.3. As can be observed from Figure 4.17A, we found that the score delta distributions of cis eQTLs compared to SNPs were not significantly different according to the rank sum test (p=0.59).

Next, we investigated whether assigning the TFBSs to cis eQTLs that exhibited more evidence of influencing the binding site would change who the score dealt distributions compare. We make the same assumption described in Section 3.4.3 that the TF with the largest score delta is most likely affected by the cis eQTL. Once we compare max TF score deltas to the mean score deltas of other SNPs, we see that the distributions are significantly different (p=0.02, Fig. 4.17B, eQTL max TF score deltas are denoted eQTL Max).

We then compare the average of score deltas across all SNPs for each TF against all cis eQTLs for each TF (Fig. 4.18A). In addition, we compare the average of score deltas across all SNPs for each TF against all cis eQTLs for each max TF (Fig. 4.18B). Both comparisons for performed for the same 10 TFs. For 6/10 TFs, the average of cis eQTLs in Figure 4.18A was larger than that of SNPs, one TF had a significant difference in score delta distributions (p<0.1). After cis eQTLs were assigned to the max TF, for 7/10 TFs, the average of cis eQTLs in Figure 4.18B was larger than that of SNPs, again one TF had a significant different in score delta distributions (p<0.1). To summarize, the score deltas between SNPs and cis eQTLs are similar when averaged over all PWM models, but the distributions are different when cis eQTLs were assigned to TFs with the strongest evidence. However, comparing these score delta distributions specifically for each TF does not yield differences in score deltas between SNPs and cis eQTLs. For this study, because we were relying on SNP arrays in non-coding regions, there were fewer SNPs. Now whole genome sequencing is available and we are able to analyze more SNPs in this region, these results might change.

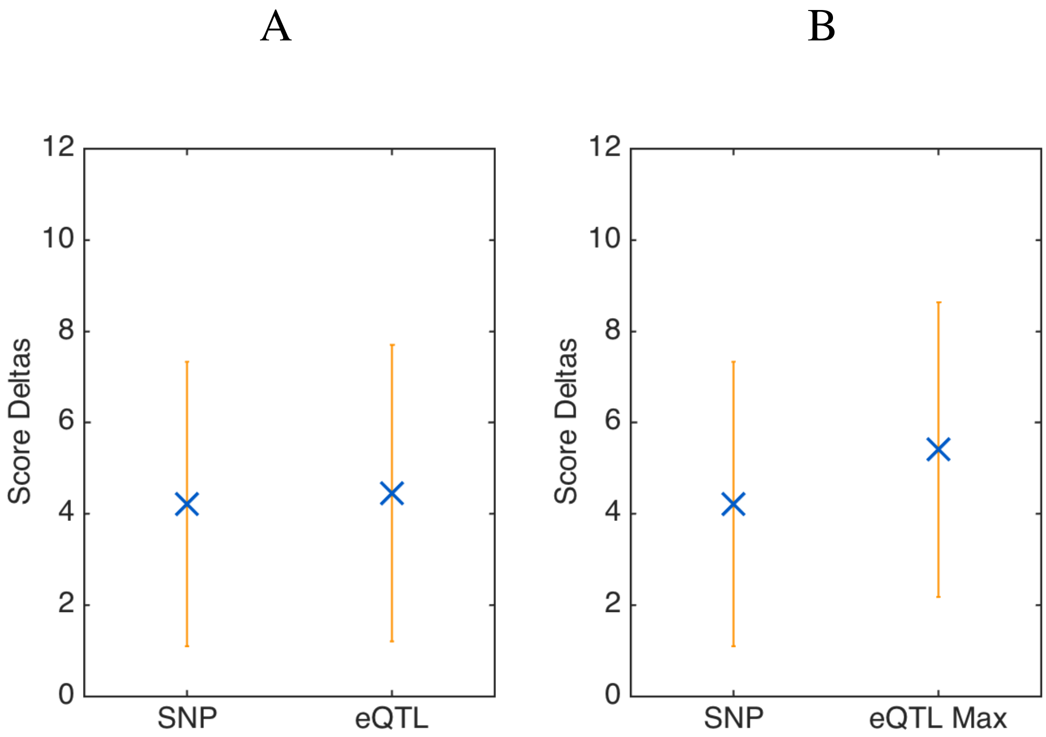

Figure 4.17: The effects of eQTLs on TF binding affinities. The distributions of score deltas for SNPs, eQTLs and of maximum score deltas over TFs for eQTLs. The mean and standard deviation are shown as blue dots and orange lines.

4.4.4 Identification of super enhancers

By investigating the cis eQTLs of target genes and the TF motif matches that overlap, we identify several sites that may be super enhancers. Super enhancers are sites where many TFs bind and drive gene expression [55, 113]. From examining the list of cis eQTLs, we find many where multiple TFs have binding sites that overlap with that eQTLs (Tab. 4.5). Because, histone H3K27ac distinguishes actively used enhancers [24], we identify those eQTLs in regions with H3K27ac enrichment (Fig4.19). In order identify active enhancer regions we used histone H3K27ac enrichment data from the Cistrome Project database [92]. We downloaded studies that were performed on MCF-7 cells with estrogen treatment [13, 49, 80, 145] and T47D cells in full media [155]. We aggregated all the studies by taking the union of all sites. We list the eQTLs in Table 4.20 that had at least two TFs bind and were also in H3K27ac enriched regions (highlighted pink in Figure 4.5).

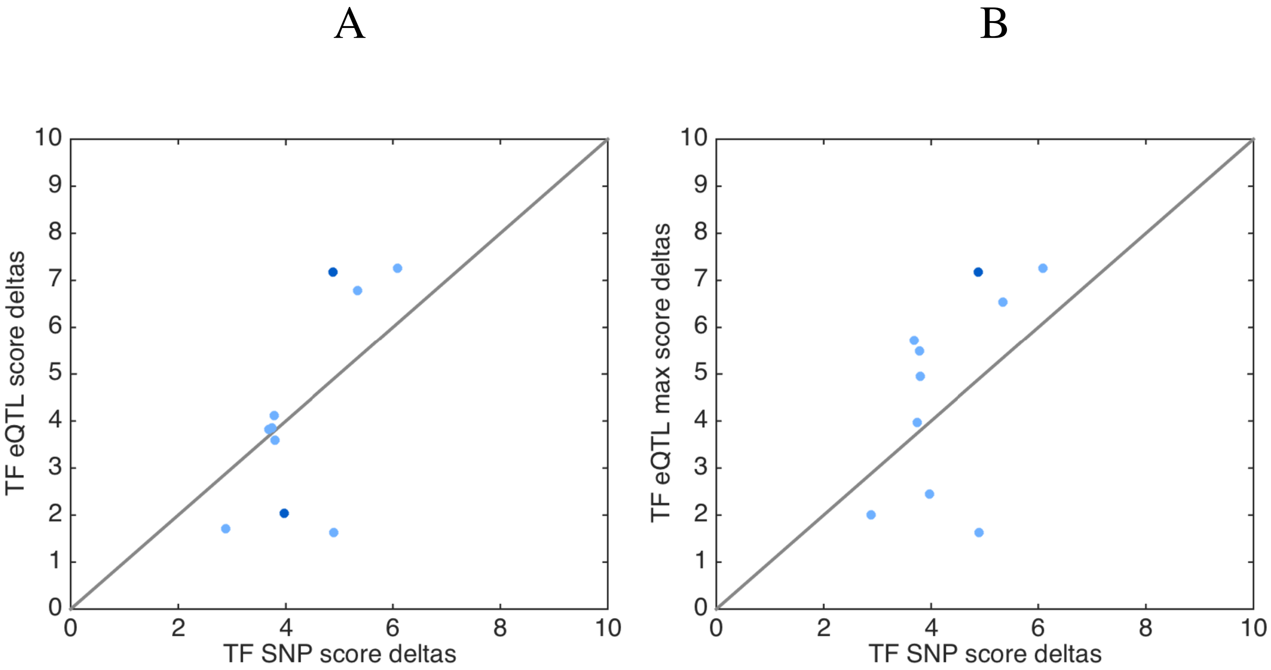

Figure 4.18: The effects of eQTLs on TF binding affinities. TF averages for SNPs and eQTLs computed for each TF. Dark blue points represent TFs with significant difference in the distribution of score deltas for SNPs and eQTLs.

4.4.5 The effect of global perturbations of SNPs from upstream regulators and transcription factors

To see if SNP perturbations of TFs have global effects on target genes, we study the SNPs in exon regions of TFs that influence target gene expression . There were 233 SNPs in exon regions of TFs and 75 of them were selected as eQTLs. We annotated the 75 trans eQTLs using the UCSC Variant Annotator (cite) and 30 of them were annotated as either synonymous or missense mutations. These 30 trans eQTLs, the corresponding annotation of the variant and the number of targets that they influence are illustrated in Table 4.6. We also indicate whether the mutation was a somatic mutation in at least one patient.

Next, we study SNPs located within upstream regulators to see whether they influence the gene expression of TFs. There are 402 SNPs located in exon regions of upstream regulators, and 96 of them were selected as eQTLs. Similar to the procedure for trans eQTLs of TFs, we annotated these trans eQTLs using the UCSC Variant Annotator (cite) and 43 of them were annotated as either synonymous or missense mutations. We list these 43 trans eQTLs, the annotation of the variant and the TF genes that they influence in Table 4.6. In addition, we indicate whether the mutation was a somatic mutation in at least one patient.

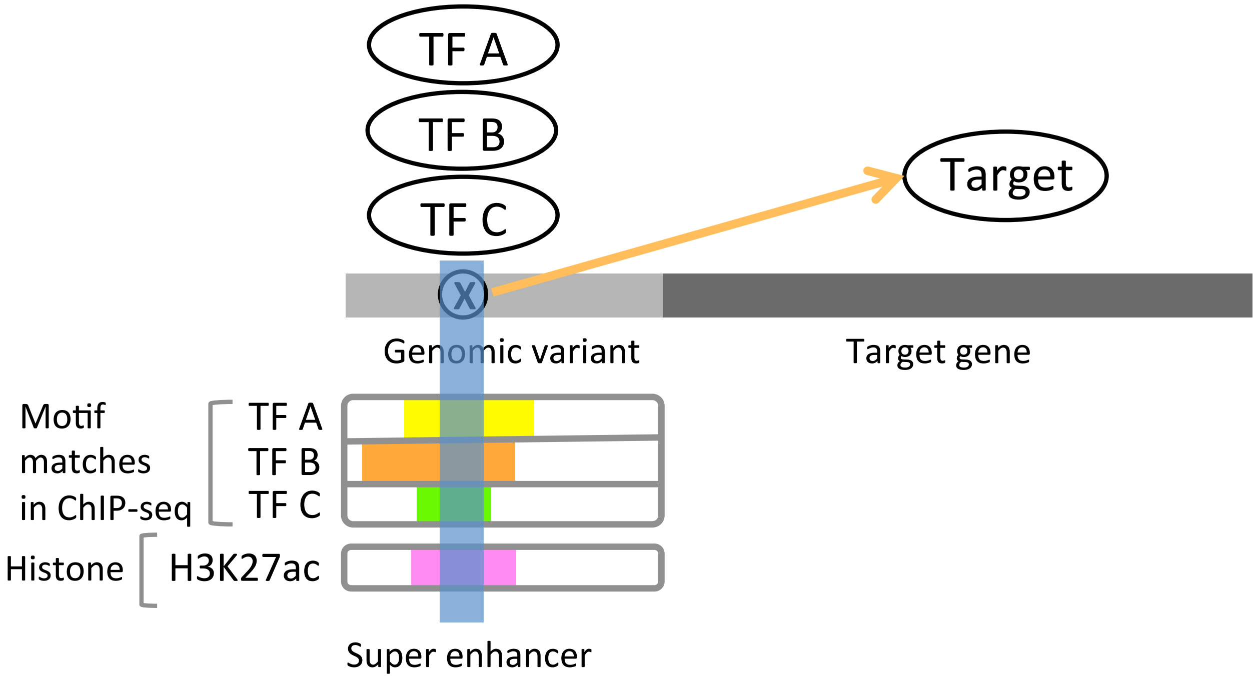

Figure 4.19: Identification of super enhancers. Of the cis eQTLs that were found to associate with target genes. We identify those that are contained in multiple TFBSs and also overlap with H3K27ac regions.

eQTL	Target	TFs with binding sites
rs584438	GJD3	NR2F2, TFAP2A, TFAP2C
rs2249851	STXBP1	CTBP1, TFAP2A
rs1078272	TFF2	CTCF, MYC
rs6130959	WFDC3	FOXA1, RAD21
rs7185427	GCSH	FOXA1,MYC
rs2289226	SLC40A1	NR2F2, TFAP2A, TFAP2C

Figure 4.20: Identified eQTLs that overlap with super enhancers.

4.5 Conclusions

In this chapter, we applied our computational approach of validating the TF binding map to studying ER regulation in Luminal A breast cancer cells. We used ChIP-seq and DNase I data from the Cistrome project as prior knowledge for our cGBN, and trained the network using SNP and expression data from the TCGA Research network. We were able to identify several interactions between ER and its coregulators that were supported by evidence from the literature. We identified both the validated target genes and differentially expressed genes for both TF coregulators and upstream regulators of ER. By analyzing local SNP perturbations of target gene expression, we found cis eQTLs that are possible super enhancers. In addition, we found trans

eQTL	TF	Distance	AATK	AQP3	BCL2L15	BMP6	C21orf128	C5AR2	CCDC62	EFHD1	EMP2	GCSH	GDF10	GJD3	GSTM5	HIST1H1D	HOXC10	IL20	KRT10	KRT12	KRT9	MEIS1	NAALADL1	PDZK1	PHACTR3	RTDR1	SFT2D2	SHISA9	SIGLEC5	SLC1A5	SLC40A1	STXBP1	SUSD3	TCTE3	TFF2	TMEM215	TPM1	TTC39B	USH1G	USP18	WFDC3	ZIC2	ZNF678
rs2249851	CTBP1	50331																													🟪											
rs1078272	CTCF	25705																																🟪								
rs12103867	CTCF	44906	🟦																																							
rs1215114	CTCF	57265																																			🟦					
rs1329004	CTCF	79709																																	🟦							
rs1838173	CTCF	36123														🟦																									
rs3850616	CTCF	71842												🟦																											
rs900347	EP300	938																					🟦																		
rs6546227	ESR1	45467																			🟦																				
rs8105903	ESR1	16																											🟦												
rs900347	ESR1	938																					🟦																		
rs757274	FOS	42406																🟦	🟦																						
rs757274	FOS	87005																🟦	🟦																						
rs10500391	FOXA1	34356																									🟦														
rs10853784	FOXA1	47076					🟦																																		
rs220149	FOXA1	11816				🟪																																			
rs458480	FOXA1	93387																																					🟦		
rs4845139	FOXA1	68951															🟦																								
rs6130959	FOXA1	80911																																						🟪	
rs6546227	FOXA1	45467																			🟦																				
rs7185427	FOXA1	40411									🟪																														
rs7986370	FOXA1	55653																																							🟦	
rs9325891	FOXA1	58330										🟦																													
rs1044228	GATA3	97263																																				🟦			
rs11059356	GATA3	95252						🟦																																	
rs3782094	GATA3	66461																				🟦																			
rs458480	GATA3	93387																																					🟦		
rs5751603	GATA3	40841																							🟦																
rs1078272	MYC	25705																																🟪								
rs10813990	MYC	59039		🟦																																						
rs10853784	MYC	47076					🟦																																		
rs2393592	MYC	50344													🟪																										
rs4656572	MYC	327																								🟪															
rs6546227	MYC	45467																			🟦																				
rs7185427	MYC	40411									🟪																														
rs2289226	NR2F2	80668																												🟪											
rs3800550	NR2F2	39021																															🟪								
rs4075583	NR2F2	408																																		🟦					
rs584438	NR2F2	78227											🟪																												
rs6479445	NR2F2	54498																														🟪									
rs900347	NR2F2	938																					🟦																		
rs971173	NR2F2	17745			🟦																																					
rs1329004	RAD21	79709																																	🟦							
rs1840549	RAD21	7773																		🟦																					
rs2814086	RAD21	81826																																								🟦
rs3850616	RAD21	71842												🟦																											
rs6130959	RAD21	80911																																						🟪	
rs7260152	RAD21	26578																										🟦													
rs1729252	TFAP2A	42184							🟦																																
rs2249851	TFAP2A	50331																													🟪											
rs2289226	TFAP2A	80668																												🟪											
rs2426800	TFAP2A	5098																						🟦																	
rs3850616	TFAP2A	71842												🟦																											
rs584438	TFAP2A	78227											🟪																												
rs900347	TFAP2A	938																					🟦																		
rs1840549	TFAP2C	7773																		🟦																					
rs1990716	TFAP2C	21332								🟪																															
rs2289226	TFAP2C	80668																												🟪											
rs3850616	TFAP2C	71842												🟦																											
rs584438	TFAP2C	78227											🟪																												
rs6910546	TFAP2C	53471				🟦																																				
rs900347	TFAP2C	938																					🟦																		

Table 4.5: Motif matches that overlap with eQTLs in nearby regions of target genes.

eQTLs that result in missense mutations affecting important regions in the protein sequence such as binding domains for TFs and kinase domains for protein kinases.

eQTL	TF	Somatic	Mutation
rs535586	EHMT2	✓	syn
rs7887	EHMT2	✓	mis
rs20551	EP300	✓	mis
rs20552	EP300	✓	syn
rs1801132	ESR1	✓	syn
rs2077647	ESR1	✓	syn
rs2228480	ESR1		syn
rs1046117	FOS	✓	syn
rs7144658	FOXA1	✓	mis
rs10929757	GREB1	✓	mis
rs2304402	GREB1	✓	mis
rs36030386	GREB1	✓	mis
rs4669751	GREB1	✓	mis
rs2302212	NCAPG	✓	syn
rs1060060	NR5A2	✓	syn
rs2821368	NR5A2	✓	syn
rs1882094	NRF1	✓	syn
rs2071504	POLR2A	✓	syn
rs2228128	POLR2A	✓	syn
rs2228129	POLR2A	✓	syn
rs2228132	POLR2A	✓	syn
rs2301609	POLR2A	✓	syn
rs1050838	RAD21	✓	syn
rs9860801	STAG1	✓	syn
rs3734391	TFAP2A	✓	syn
rs35023929	TFAP2C		syn
rs1057865	TLE3	✓	syn
rs2133977	TLE3	✓	syn
rs2228178	TLE3	✓	syn
rs33935215	TRIM24	✓	syn

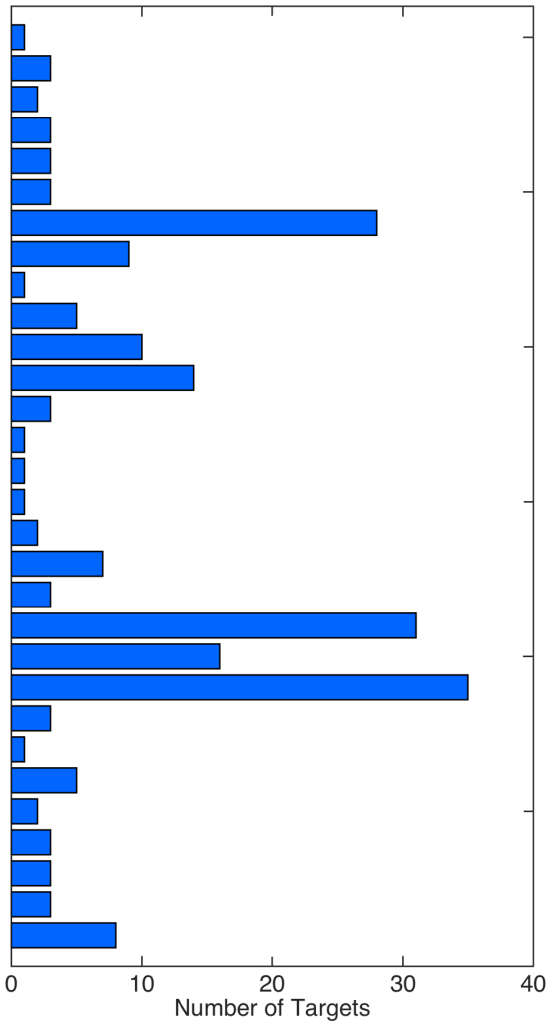

Table 4.6: Trans eQTLs in the coding regions of TFs.

eQTL	TF	Somatic	Mutation	BRD4	CTBP1	CTCF	EHMT2	EP300	ESR1	FOS	FOXA1	GATA3	GREB1	HECTD1	HIF1A	JUN	KLF4	LMTK3	MBD3	MED1	MTA3	MYC	NCAPG	NCAPG2	NR2F2	NR5A2	NRF1	POLR2A	PGR	RAD21	RELA	SPDEF	STAG1	TDRD3	TFAP2A	TFAP2C	TLE3	TOP2B	TP53	TRIM24
rs1130233	AKT1	✓	syn																														■								
rs2306234	BLK	✓	syn																■																						
rs3816668	BLK	✓	syn					■																■																	
rs9648696	BRAF	✓	syn											■				■																							
rs1050171	EGFR		syn								■		■																												
rs2072454	EGFR	✓	syn																	■	■							■													
rs2227983	EGFR	✓	mis									■																													
rs2227984	EGFR	✓	syn													■																	■								
rs1058808	ERBB2	✓	mis												■															■							■				
rs2271189	ERBB3	✓	syn					■																■																	
rs3748962	ERBB4	✓	syn																																				■		
rs1047057	FGFR2	✓	syn																															■				■			■
rs1047100	FGFR2	✓	syn				■																			■		■	■												
rs376618	FGFR4	✓	mis																																	■					
rs452885	FGFR4	✓	syn					■																												■					
rs2229765	IGF1R	✓	syn	■																								■			■				■	■				
rs2059806	INSR	✓	syn																■																						
rs2230588	JAK1	✓	syn																					■			■														
rs2230724	JAK2	✓	syn													■			■		■			■	■														■	■	
rs10250	MAP2K2	✓	syn			■		■						■						■											■	■	■							■	
rs2076139	MAPK11	✓	syn															■									■														
rs1129880	MAPK12	✓	syn				■																					■	■				■					■			
rs2272857	MAPK12	✓	syn				■																																■		
rs1059227	MAPK13	✓	syn					■							■							■							■											■	
rs12678428	MAPK15	✓	syn														■	■			■													■							
rs3752087	MAPK4	✓	mis												■																										
rs1064261	MTOR	✓	syn						■						■																										
rs11121691	MTOR	✓	syn																															■							
rs1135172	MTOR	✓	syn													■													■											■	
rs214936	PIK3C2A	✓	mis																						■																
rs11044004	PIK3C2G	✓	mis																			■			■																
rs12312266	PIK3C2G	✓	mis	■																																	■			
rs762537354	PIK3C2G	✓	mis				■			■	■	■			■								■	■	■											■			■		
rs1129293	PIK3CG	✓	syn																														■								
rs17847825	PIK3CG		mis																																	■					
rs706713	PIK3R1	✓	syn																																			■	■	■	
rs785467	PIK3R3	✓	mis																																			■			
rs785468	PIK3R3	✓	syn	■																																				
rs11650737	PIK3R5	✓	syn																																				■		
rs394811	PIK3R5	✓	syn					■								■																										
rs9915880	PIK3R5	✓	syn													■																						■		
rs1064196	RPS6KA1	✓	syn				■																		■										■	■				■	
rs11800553	RPS6KA1		syn											■								■		■																	

Table 4.7: Trans eQTLs in coding regions of ER upstream regulators.

eQTL	TF	Mutation	AA Pos.	SIFT	Domain Loc.	Domain
rs7144658	FOXA1	Mis	83	T(0.540000)	170 - 264	Fork head domain
rs20552	EP300	Syn	1061		1067-1139	Bromodomain
rs2228480	ESR1	Syn	594		262 - 595	Interaction with AKAP131
					264 - 595	Self-association
					311- 595	Transactivation AF-2

Table 4.8: Identified eQTLs that affect target genes and are near or overlap with protein domains in the protein sequence of TFs.

eQTL	Kinase	Mutation	AA Pos.	SIFT	Domain Loc/.	Domain
rs1058808	ERBB2	Mis	1170	D(0.030000)	720-987	Protein Kinase
					1011 - 1023	EF-hand calcium-binding domain
					1195 -1197	Interaction with PIK3C2
rs3752087	MAPK4	Mis	38	T(0.080000)	18 - 310	Protein kinase domain
rs214936	PIK3C2A	Mis	90	D(0.030000)	2-142	Interaction with clathrin
rs12312266	PIK3C2G	Mis	911	D(0.040000)	916 -1180	PI3 and PI4 kinase family profile
rs1050171	EGFR	Syn	742		667 - 934	Protein kinase domain

Table 4.9: Identified eQTLs that affect TFs and are near or overlap with protein domains in the protein sequence of protein kinases.

Chapter 5

Conclusions

5.1 Summary of the thesis

The main goal of this thesis was to validate the TF binding map that is constructed from functional data such as ChIP-seq and DNase I data. In order to achieve this goal, we proposed to use natural variation in the genome sequence as perturbations of gene expression. However, the challenge of doing this is how to decouple the effects of a large number of SNPs on gene expressions to determine which TF-to-target gene relationships are being affected. To address this challenge, we developed a statistical approach that incorporates TF binding map as prior information and then selects the TF-target interactions validated under SNP perturbations. We demonstrated this approach with data from lymphoblastoid cell lines. We also applied our approach to study ER regulation in breast cancer cells.

In Chapter 2, we first discussed our method for structure learning of a Gaussian Bayesian network. We developed a learning algorithm called A* lasso that estimates the structure of Bayesian network for continuous variables in a high dimensional setting. For small networks, this method used the A* search algorithm to find the optimal solution, while being more computationally efficient than dynamic programming. For larger networks, we developed a heuristic scheme that further prunes the search space of A* search such that we can learn the networks efficiently and without substantial compromise to the quality of the solutions.

In Chapter 3, we extended the statistical model to a conditional Gaussian Bayesian network of gene expressions conditioned on SNPs and demonstrated how to perform ChIP-seq validation on data from lymphoblastoid cells. The TF binding map was encoded as prior information and the TF-target interactions validated under SNP perturbations were selected. We used a single analysis that learned the validated TF-target interactions from SNP and expression data. We found that a relatively larger number of targets were validated by perturbations in concentration compared to the number of targets validated by local perturbations of binding affinity or global perturbations of TF structure. By analyzing the score differences of TF binding sites of SNPs compared with eQTLs found in the promoter regions of target genes, we found that eQTLs did affect binding affinity of TFs. By studying the locations of trans eQTLs in coding regions of TF, we found evidence that a subset of them may affect TF protein structure. Finally, we found that the target genes determined through our approach differ from the set of target genes determined by artificial perturbation studies. We found that there exists interaction effects amongst the TFs that may contribute to these differences.

In Chapter 4, we applied our method to study estrogen receptor (ER) regulation and its cofactors in breast cancer. For this analysis, we focused on luminal A breast cancers. After analyzing the connectivity amongst the TFs, we found several edges detected by our algorithm that are supported by the literature. By analyzing the eQTLs in the regulatory region of target genes, we found that several eQTLs are potentially super enhancers and thus may be important regulatory elements. In addition, by examining the eQTLs that are in coding regions of protein kinases and TFs, we found missense SNPs that may be important to protein structure or binding domains.

In summary, we proposed a computational approach for exploiting natural variation in populations to validate TF binding data. We presented a statistical model and developed the learning algorithm to estimate the model efficiently and accurately. We then demonstrated how to perform ChIP-seq validation on lymphoblastoid cell lines and showed how we can use this method to understand gene regulation in breast cancer biology.

5.2 Future Work

We conclude by discussing the limitations of our approach and directions for future work that extend the contributions of this thesis. We discuss the computational limitations of our approach and suggest technical extensions of our algorithm in order to improve computational efficiency. We discuss future work in defining guidelines and requirements that would allow other researchers to evaluate whether our method is appropriate for their data. We also make suggestions that would improve the power of our computational approach for validating TF-target regulatory relationships for both LCL and breast cancer regulatory networks.

5.2.1 Computational limitations

Limitations of modeling the transcriptional network with Gaussian Bayesian networks

There are inherent limitations of modeling the transcriptional network using conditional Gaussian Bayesian networks. First, Bayesian networks are directed acyclic graphs and they do not model cycles. But, it is well known that feedback loops occur in biological mechanisms. However, despite this limitation, we can still learn useful TF-to-target interactions and SNP perturbations of TF and targets using conditional Gaussian Bayesian networks and use this information to validate ChIP-seq candidate edges. In addition, we assume that gene expressions vary linearly with the expression of their parent genes and SNPs. This assumption allows us to exploit the computational advantages of Gaussian graphical models. However, the limitation of the Gaussian assumption is that it does not account for the situation where the expression of a particular gene varies non-linearly with respect to a particular parent gene or SNP. Further work can explore the possibility of using non-linear models.

Limitations of structure learning

In addition to constraints of the model, there are limitations of the structure learning procedure. For instance, we constrain the structure such that TFs are upstream of the target genes, and we group the target genes into clusters and learn the structure of each group independently. Similarly, we limit the priority queue in A* lasso and use a heuristic scheme to obtain a high-quality solution, but not the optimal solution. For both these issues, we trade off computational efficiency and estimation of the best possible structure. Furthermore, the structure that we estimate is dependent on the measurements of variables that we have. In directed graphical models, if a variable is missing, then it is equivalent to being marginalized out. As a result, two variables may be correlated even though there does not exist a causal relationship between them. When we estimate the network with missing variables, we may learn edges between variables that are correlated instead of those with true causal relationships. Specifically, when estimating networks on data from LCLs, we are limited by the measurements that are collected. In this case, we are limited by the ChIP-seq data which was collected for 71 TFs. Increased number of TF ChIP-seq experiments becoming available will allow us to improve the accuracy of the estimated transcriptional network.

Improving computational efficiency of A* lasso

While A* lasso is computationally efficient for structure learning of GBNs, for networks with thousands of nodes, the algorithm is still not fast enough. Because the bottleneck is doing the lasso optimization step, the algorithm can potentially be made faster with shotgun for lasso [10]. One potential approach to improving computational efficiency is modification of the current heuristic scheme. Using theoretical results to justify why the heuristic scheme is able to improve computational speed without substantially decreasing the quality of solutions may bring about insights that can inform extensions of this algorithm. There are two dimensions upon which the heuristic scheme may change. First, the actual heuristic function can be modified. For example instead of using the LassoScore without the DAG constraint, an alternative heuristic could be finding the LassoScore of a subset of nodes. Second, how the search is conducted can be altered. Currently, we limit the queue size and prune the solutions intelligently based on the depth of the states. However, there are alternative methods of speeding up search originally developed for planning applications. For example, Multi-Heuristic A* uses multiple heuristics at different parts of the search space in order to achieve solutions faster and overcome local minima [2]. There are also alternatives that incorporate randomness in the searches. For example, R* performs short-range weighted A* searches towards goal states selected at random and then reconstructs a solution from the paths produced by these small searches [82]. These modifications on the heuristic or the search strategy may be investigated in order to improve computational efficiency of A* lasso.

5.2.2 Recommendations for biological data

Defining sample size requirements

In order to make our approach accessible to other researchers in the biological domain, it would be useful to specify the number of samples required for successful learning. One strategy to provide these recommendations is to study sample complexity for conditional Gaussian Bayesian networks. It may be possible to borrow ideas from previous analyses on Gaussian Bayesian networks [43] and discrete Bayesian networks [28]. A more practical approach is to perform Bootstrap [32] simulations. In particular, we would generate multiple sample data sets of different sample sizes in order to evaluate the recovery of the network structure, given data sampled from known Bayesian network structures.

Defining appropriate TFs for TF binding map validation

It would be useful to evaluate whether TF-target regulatory relationships in the TF binding map can be validated successfully using our population-based approach. One important feature is variation in expression measurements across the population. If a TF is tightly regulated or if genetic variation does not vary the TF expression, then using an artificial perturbation may be more successful. In addition, if a TF is highly redundant, then even if the TF expression varies, we may not find the correct set of validated target genes because other TFs will take over its function. For these TFs, higher order knockdown experiments may be necessary for functional validation.

5.2.3 Expanding scope of the current study

Further investigation of epistatic interactions between TFs

We currently do not model epistatic interactions between TFs in our cGBN. A useful extension would be to model epistasis efficiently. Furthermore, comparing the results from our approach using population data with higher-order artificial perturbation studies would improve our understanding of cooperative TF regulation. For example, when double knockdown data are available, comparing the targets of those artificial perturbation studies with the targets determined from population analysis would provide more insight into how two-way interaction effects affect downstream target genes. In performing comparisons, it would be useful to be able to predict off-target effects of RNA interference and remove those genes from the target set [27].

Validation of TFs in ER signaling in breast cancer

After estimating the transcriptional network on luminal A breast cancer data, we find interactions of TFs and protein kinases with ER and known coregulators of ER that are potential discoveries. NCAPG and RAD21 have been previously characterized in cancer, but not much is known about these TFs in association with breast cancer [54, 83]. SPDEF and KLF4 have been previously reported to be involved in breast cancer but not specific to ER signaling [134, 135]. In addition, we find that a group of protein kinases MAPK4, EGFR, ERBB3, ERBB4, and IGF1R regulate

EP300, CTCF, MBD3, STAG1, and TOP2B in a tight cluster. While EP300 and CTCF have been tied to ER independently, there is not a lot known about their cooperation with each other, their interaction with the other TFs: MBD3, STAG1, TOP2B, or their upstream regulation. One strategy to confirm that these interactions are important for ER signaling is to replicate these interactions using another dataset. One future direction is to refit the model with data from the METABRIC project [25] and observe whether the same interactions are in the estimated network. Another direction that can be pursued is to use survival data from the breast cancer patients and determine whether these genes are associated with breast cancer survival.

Validation of eQTLs in breast cancer

We found both cis eQTLs and trans eQTLs that potentially affect binding sites or protein domains. We find six cis eQTLs that might be super enhancers. One of these eQTLs rs1078272 was previously reported to affect binding of both ER and RAD21 [58]. The rest of the five cis eQTLs require further investigation to determine how they might affect ER signaling. We found seven missense eQTLs among TF coding regions. Out of these seven eQTLs, one of them was found within the binding domain (rs7144658), and one near a binding domain (rs20551). We found 10 missense eQTLs among protein kinase coding regions. Out of these 10 eQTLs, 4 were within a protein and 3 were near a protein domain. The three eQTLs near protein domains were all previously reported as mutations in other cancers [17, 21]. Further efforts are required to determine whether these eQTLs are potential driver mutations in breast cancer. Future efforts should be directed towards determining whether these eQTLs exist as both germline and somatic mutations for different patients and to study their frequencies in normal and tumor populations. If we observe that particular eQTLs are being selected for or against in breast cancer tumors, this might indicate that these eQTLs are important in tumor progression.

Modeling gene regulation for subtypes of breast cancer

For studying gene regulation in breast cancers, studying all subtypes instead of simply focusing on luminal A breast cancers is the next step. This can be done by learning a cGBN for each of the five subtypes of breast cancer (luminal A, luminal B, basal, claudin-low and HER2). Alternatively, a single model can be estimated where certain edges of the network are common, and others deviate according to subtype. The latter approach may be advantageous because of the limited data available for each subtype. Previous methods proposed such as fused lasso may be useful for network estimation [106, 149].

Including rare variants in the analysis

Among luminal A tumours in TCGA data, we did not include variants with a minor allele frequency of less than 0.01. This eliminated any mutations that occurred in only one or two patients because our sample size is small. Collapsing multiple rare variants is a common approach to address the limited statistical power of individual rare variances because of their low frequency. There are two potential strategies to collapse variants. The first is based on proximity of variants in genome. For example the variants from a particular gene or segment of the genome may be combined [142]. In addition, we can also collapse variants based on their proximity in the 3-dimensional structure of the chromatin, which can now be obtained through chromatin conformation capture methods [8, 125].

Appendix A

Supplementary Materials

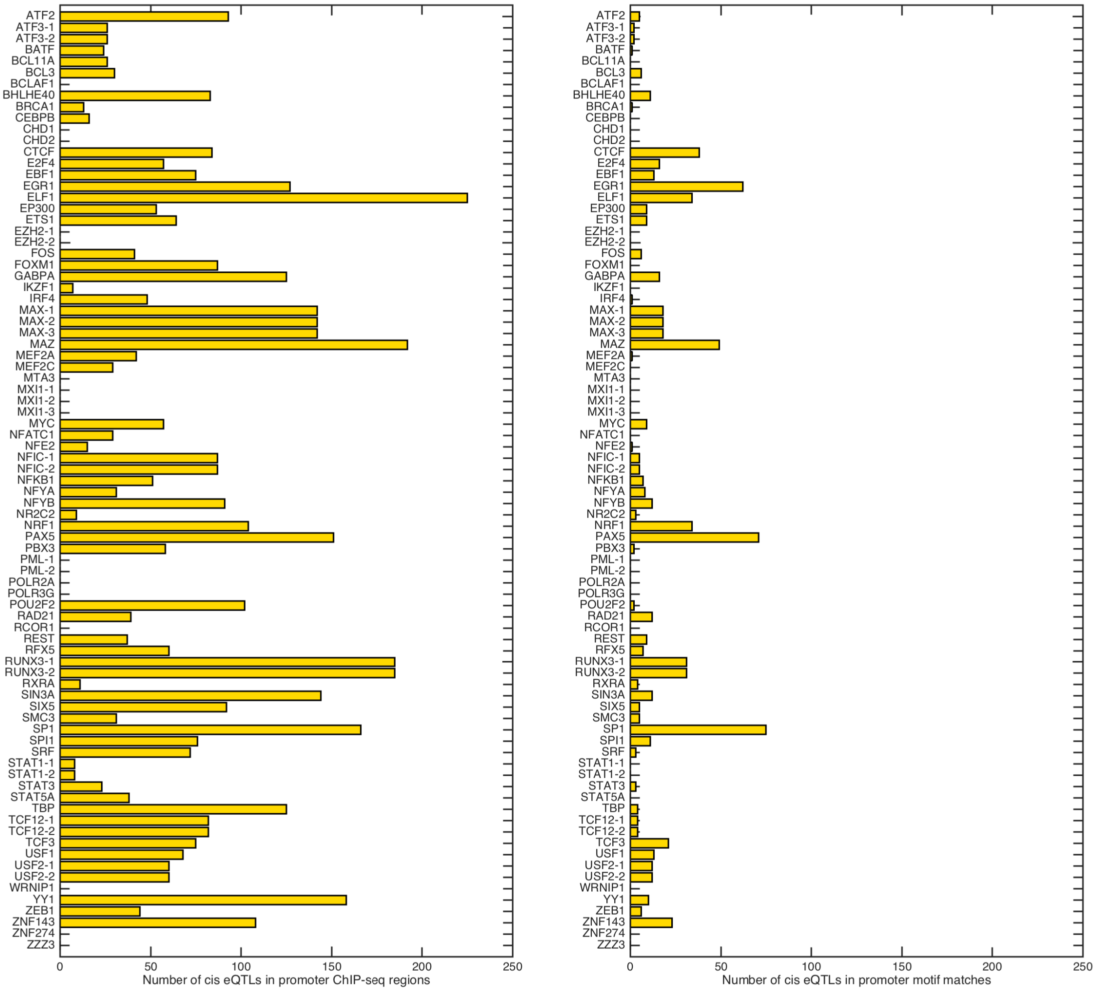

Figure A.1: Number of cis eQTLs in TF binding regions. A) Number of cis eQTLs located in promoter ChIP-seq regions. B) Number of cis eQTLs in promoter motif matches.

Bibliography

[1] Aleksandra B Adomas, Sara A Grimm, Christine Malone, Motoki Takaku, Jennifer K Sims, and Paul A Wade. Breast tumor specific mutation in gata3 affects physiological mechanisms regulating transcription factor turnover. *BMC cancer*, 14(1):278, 2014. 4.3.1

[2] Sandip Aine, Siddharth Swaminathan, Venkatraman Narayanan, Victor Hwang, and Maxim Likhachev. Multi-heuristic a. *International Journal of Robotics Research*, 35 (1-3):224–243, 2016. 5.2.1

[3] André Albergaria, Joana Paredes, Bárbara Sousa, Fernanda Milanezi, Vítor Carneiro, Joana Bastos, Sandra Costa, Daniella Vieira, Nair Lopes, Eric W Lam, et al. Expression of foxa1 and gata-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours. *Breast cancer research*, 11(3):R40, 2009. 4.4.1

[4] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015. 1.1, 1.4.6

[5] Jeffrey D Allen, Yang Xie, Min Chen, Luc Girard, and Guanghua Xiao. Comparing statistical methods for constructing large scale gene networks. *PloS one*, 7(1):e29348, 2012. 1.4.4

[6] Ziv Bar-Joseph, Georg K Gerber, Tong Ihn Lee, Nicola J Rinaldi, Jane Y Yoo, D Benjamin Gordon, Ernest Fraenkel, Tommi S Jaakkola, Richard A Young, David K Gifford, et al. Computational discovery of gene modules and regulatory networks. *Nature biotechnology*, 21(11):1337, 2003. 1.4.5

[7] Rodolphe Barrangou and Jennifer A Doudna. Applications of crispr technologies in research and beyond. *Nature biotechnology*, 34(9):933–941, 2016. 1.1, 1.4.7

[8] Jon-Matthew Belton, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. Hi–c: a comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276, 2012. 5.2.3

[9] Stéphanie Bianco, Mylène Brunelle, Maïka Jangal, Luca Magnani, and Nicolas Gévry. Lrh-1 governs vital transcriptional programs in endocrine-sensitive and-resistant breast cancer cells. *Cancer research*, 74(7):2015–2025, 2014. 4.3.1

[10] Joseph K Bradley, Aapo Kyrola, Danny Bickson, and Carlos Guestrin. Parallel coordinate descent for l 1-regularized loss minimization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 321–328. Omnipress, 2011. 5.2.1

[11] Rachel B Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102(5):1572–1577, 2005. 1.4.6

[12] Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, 2002. 1.4.6

[13] Mylène Brunelle, Alexei Nordell Markovits, Sébastien Rodrigue, Mathieu Lupien, Pierre-Étienne Jacques, and Nicolas Gévry. The histone variant h2a. z is an important regulator of enhancer activity. *Nucleic acids research*, 43(20):9742–9756, 2015. 4.4.4

[14] Michael J Buck and Jason D Lieb. Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83 (3):349–360, 2004. 1.4.5

[15] Sean B Carroll. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):25–36, 2008. 3.4.1

[16] Scott L Carter, Christian M Brechbühler, Michael Griffin, and Andrew T Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, 2004. 1.4.4

[17] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, 2012. 5.2.3

[18] David Maxwell Chickering. Learning Bayesian networks is NP-complete. In *Learning from data*, pages 121–130. Springer, 1996. 1.4.2

[19] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. 3.4.3

[20] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012. 1.1, 1.3.5, 3.3, 3.4.1

[21] AACR Project GENIE Consortium et al. Aacr project genie: Powering precision medicine through an international consortium. *Cancer Discovery*, 2017. 5.2.3

[22] ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. 1.1, 1.3.3, 1.3.4, 1.4.5, 3.3, 4.3.1

[23] UniProt Consortium et al. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014. 3.4.3

[24] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010. 4.4.4

[25] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda,

Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012. 5.2.3

[26] Darren A Cusanovich, Bryan Pavlovic, Jonathan K Pritchard, and Yoav Gilad. The functional consequences of variation in transcription factor binding. *PLoS Genetics*, 10(3): e1004226, 2014. 1.3.6, 1.4.5, 3.4.2

[27] Shaoli Das, Suman Ghosal, Jayprokas Chakrabarti, and Karol Kozak. Seedseq: off-target transcriptome database. *BioMed research international*, 2013, 2013. 5.2.3

[28] Sanjoy Dasgupta. The sample complexity of learning fixed-structure bayesian networks. *Machine Learning*, 29(2-3):165–180, 1997. 5.2.2

[29] Edouard De Castro, Christian JA Sigrist, Alexandre Gattiker, Virginie Bulliard, Petra S Langendijk-Genevaux, Elisabeth Gasteiger, Amos Bairoch, and Nicolas Hulo. Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic acids research*, 34(suppl 2):W362–W365, 2006. 3.4.3

[30] Li-Jun Di, Jung S Byun, Madeline M Wong, Clay Wakano, Tara Taylor, Sven Bilke, Songjoon Baek, Kent Hunter, Howard Yang, Maxwell Lee, et al. Genome-wide profiles of ctbp link metabolism with genome stability and epithelial reprogramming in breast cancer. *Nature communications*, 4:1449, 2013. 4.3.1

[31] Antonia A Dominguez, Wendell A Lim, and Lei S Qi. Beyond editing: repurposing crispr–cas9 for precision genome regulation and interrogation. *Nature reviews. Molecular cell biology*, 17(1):5, 2016. (document), 1.1

[32] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986. 5.2.2

[33] Michael NC Fletcher, Mauro AA Castro, Xin Wang, Ines De Santiago, Martin O'Reilly, Suet-Feung Chin, Oscar M Rueda, Carlos Caldas, Bruce AJ Ponder, Florian Markowetz, et al. Master regulators of fgfr2 signalling and breast cancer risk. *Nature communications*, 4, 2013. 4.3.1

[34] Hector L Franco, Anusha Nagari, and W Lee Kraus. Tnfα signaling exposes latent estrogen receptor binding sites to alter the breast cancer cell transcriptome. *Molecular cell*, 58 (1):21–34, 2015. 4.3.1

[35] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010. 1.4.6

[36] Nir Friedman, Iftach Nachman, and Dana Peér. Learning Bayesian network structure from massive datasets: the "Sparse Candidate" algorithm. In *Proceedings of the Fifteenth conference on Uncertainty in Artificial Intelligence*, pages 206–215. Morgan Kaufmann Publishers Inc., 1999. 1.4.2

[37] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000. 1.4.4

[38] Teresan Frisan, Victor Levitsky, and Maria Masucci. Generation of lymphoblastoid cell

lines (lcls). *Epstein-Barr Virus Protocols*, pages 125–127, 2001. 3.4.1

[39] Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998. 2.2

[40] Terrence S Furey. Chip–seq and beyond: new and improved methodologies to detect and characterize protein–dna interactions. *Nature Reviews Genetics*, 13(12):840–852, 2012. 1.1, 1.3.3, 1.3.4

[41] Pascale Gerbault. The onset of lactase persistence in europe. *Human heredity*, 76(3-4): 154–161, 2013. 3.1

[42] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, et al. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91, 2012. 1.4.5

[43] Asish Ghoshal and Jean Honorio. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. *arXiv preprint arXiv:1703.01196*, 2017. 5.2.2

[44] Greg Gibson, Joseph E Powell, and Urko M Marigorta. Expression quantitative trait locus analysis for translational medicine. *Genome medicine*, 7(1):60, 2015. 1.1

[45] Luke A Gilbert, Matthew H Larson, Leonardo Morsut, Zairan Liu, Gloria A Brar, Sandra E Torres, Noam Stern-Ginossar, Onn Brandman, Evan H Whitehead, Jennifer A Doudna, et al. Crispr-mediated modular rna-guided regulation of transcription in eukaryotes. *cell*, 154(2):442–451, 2013. (document), 1.1

[46] Leon Glass and Stuart A Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *Journal of theoretical Biology*, 39(1):103–129, 1973. 1.4.4

[47] Laurent Gouya, Herve Puy, Anne-Marie Robreau, Monique Bourgeois, Jerôme Lamoril, Vasco Da Silva, Bernard Grandchamp, and Jean-Charles Deybach. The penetrance of dominant erythropoietic protoporphyria is modulated by expression of wildtype fech. *Nature genetics*, 30(1):27–28, 2002. 3.1

[48] Jennifer M Gross and Douglas Yee. How does the estrogen receptor work? *Breast Cancer Research*, 4(2):62, 2002. 4.1

[49] Michael J Guertin, Xuesen Zhang, Lynne Anguish, Sohyoung Kim, Lyuba Varticovski, John T Lis, Gordon L Hager, and Scott A Coonrod. Targeted h3r26 deimination specifically facilitates estrogen receptor binding by modifying nucleosome structure. *PLoS genetics*, 10(9):e1004613, 2014. 4.4.4

[50] Housheng Hansen He, Clifford A Meyer, Mei Wei Chen, V Craig Jordan, Myles Brown, and X Shirley Liu. Differential dnase i hypersensitivity reveals factor-dependent chromatin dynamics. *Genome research*, 22(6):1015–1025, 2012. 4.3.1

[51] David Heckerman, Dan Geiger, and David M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995. 1.4.2

[52] Nina Heldring, Ashley Pike, Sandra Andersson, Jason Matthews, Guojun Cheng, Johan Hartman, Michel Tujague, Anders Ström, Eckardt Treuter, Margaret Warner, et al. Estro-

gen receptors: how do they signal and what are their targets. *Physiological reviews*, 87 (3):905–931, 2007. 4.1

[53] Christian Herder, Wouter Peeters, Thomas Illig, Jens Baumert, DP De Kleijn, Frans L Moll, Ulrike Poschen, Norman Klopp, M Muller-Nurasyid, Michael Roden, et al. Rantes/ccl5 and risk for coronary events: results from the monica/kora augsburg case-cohort, athero-express and cardiogram studies. *PloS one*, 6(12):e25734, 2011. 3.1

[54] Victoria K Hill, Jung-Sik Kim, and Todd Waldman. Cohesin mutations in human cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1866(1):1–11, 2016. 5.2.3

[55] Denes Hnisz, Brian J Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A Sigova, Heather A Hoke, and Richard A Young. Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–947, 2013. 4.4.4

[56] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988. 1.4.6

[57] Deborah L Holliday and Valerie Speirs. Choosing the right cell line for breast cancer research. *Breast Cancer Research*, 13(4):1, 2011. 4.1

[58] Chia-Ni Hsiung, Hou-Wei Chu, Yuan-Ling Huang, Wen-Cheng Chou, Ling-Yueh Hu, Huan-Ming Hsu, Pei-Ei Wu, Ming-Feng Hou, Jyh-Cherng Yu, and Chen-Yang Shen. Functional variants at the 21q22. 3 locus involved in breast cancer progression identified by screening of genome-wide estrogen response elements. *Breast Cancer Research*, 16(5):455, 2014. 5.2.3

[59] Shuai Huang, Jing Li, Jieping Ye, Adam Fleisher, Kewei Chen, Teresa Wu, Eric Reiman, Alzheimer's Disease Neuroimaging Initiative, et al. A sparse structure learning algorithm for gaussian bayesian network identification from high-dimensional data. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1328–1342, 2013. 1.4.2, 2.3

[60] Antoni Hurtado, Kelly A Holmes, Caryn S Ross-Innes, Dominic Schmidt, and Jason S Carroll. Foxa1 is a key determinant of estrogen receptor function and endocrine response. *Nature genetics*, 43(1):27–33, 2011. 4.3.1, 4.4.1

[61] Trey E Ideker, Vesteinn Thorsson, and Richard M Karp. Discovery of regulatory interactions through perturbation: inference and experimental design. In *Pacific symposium on biocomputing*, volume 5, pages 302–313, 2000. 1.4.4

[62] Sanford-Burnham Prebys Medical Discovery Institute. How b cell metabolism is controlled: Gsk3 acts as a metabolic checkpoint regulator in b cells., January 2017. URL www.sciencedaily.com/releases/2017/01/170123115239.htm. Retrieved August 26, 2017. 3.4.1

[63] Tommi S Jaakkola, David Sontag, Amir Globerson, Marina Meila, et al. Learning bayesian network structure using lp relaxations. In *AISTATS*, pages 358–365, 2010. 1.4.2

[64] Maïka Jangal, Jean-Philippe Couture, Stéphanie Bianco, Luca Magnani, Hisham Mohammed, and Nicolas Gévry. The transcriptional co-repressor tle3 suppresses basal signaling on a subset of estrogen receptor α target genes. *Nucleic acids research*, 42(18): 11339–11348, 2014. 4.4.1

[65] Ritsert C Jansen and Jan-Peter Nap. Genetical genomics: the added value from segregation. *TRENDS in Genetics*, 17(7):388–391, 2001. 1.4.6

[66] Roy Joseph, Yuriy L Orlov, Mikael Huss, Wenjie Sun, Say Li Kong, Leena Ukil, You Fu Pan, Guoliang Li, Michael Lim, Jane S Thomsen, et al. Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. *Molecular systems biology*, 6(1):456, 2010. 4.3.1

[67] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008. 1.4.1, 1.4.4

[68] LR Kidd, Dominique Z Jones, Erica N Rogers, Nayla C Kidd, Sydney Beache, James E Rudd, Camille Ragin, Maria Jackson, Norma McFarlane-Anderson, Marshall Tulloch-Reid, et al. Chemokine ligand 5 (ccl5) and chemokine receptor (ccr5) genetic variants and prostate cancer risk among men of african descent: a case-control study. *Hered Cancer Clin Pract*, 10:16–27, 2012. 3.1

[69] Seyoung Kim, Eric P Xing, et al. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *The Annals of Applied Statistics*, 6(3):1095–1117, 2012. 1.4.6

[70] Yea Woon Kim and AeRi Kim. Deletion of transcription factor binding motifs using the crispr/spcas9 system in the β-globin lcr. *Bioscience Reports*, 37(4):BSR20170976, 2017. (document), 1.1

[71] Mikko Koivisto and Kismat Sood. Exact bayesian structure discovery in bayesian networks. *Journal of Machine Learning Research*, 5(May):549–573, 2004. 1.4.2, 2.1, 2.3

[72] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 2.2, 2.3, 3.2.3

[73] Say Li Kong, Guoliang Li, Siang Lin Loh, Wing-Kin Sung, and Edison T Liu. Cellular reprogramming by the conjoint action of erα, foxa1, and gata3 to a ligand-inducible growth state. *Molecular systems biology*, 7(1):526, 2011. 4.3.1

[74] Gozde Korkmaz, Rui Lopes, Alejandro P Ugalde, Ekaterina Nevedomskaya, Ruiqi Han, Ksenia Myacheva, Wilbert Zwart, Ran Elkon, and Reuven Agami. Functional genetic screens for enhancer elements in the human genome using crispr-cas9. *Nature biotechnology*, 34(2):192–198, 2016. 1.4.7

[75] Timo JT Koski and John M Noble. A review of bayesian networks and structure learning. *Mathematica Applicanda*, 40(1):53–103, 2012. 1.4.1

[76] Thomas LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, page gkp552, 2009. 1.3.5

[77] Wai Lam and Fahiem Bacchus. Learning Bayesian belief networks: An approach based on the MDL principle. *Computational intelligence*, 10(3):269–293, 1994. 1.4.2

[78] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008. 1.4.4

[79] Adrian V Lee, Steffi Oesterreich, and Nancy E Davidson. Mcf-7 cells—changing the

course of breast cancer research and care for 45 years. *Journal of the National Cancer Institute*, 107(7):djv073, 2015. 4.1

[80] Wenbo Li, Dimple Notani, Qi Ma, Bogdan Tanasa, Esperanza Nunez, Aaron Yun Chen, Daria Merkurjev, Jie Zhang, Kenneth Ohgi, Xiaoyuan Song, et al. Functional roles of enhancer rnas for oestrogen-dependent transcriptional activation. *Nature*, 498(7455):516–520, 2013. 4.4.4

[81] Wenbo Li, Yiren Hu, Soohwan Oh, Qi Ma, Daria Merkurjev, Xiaoyuan Song, Xiang Zhou, Zhijie Liu, Bogdan Tanasa, Xin He, et al. Condensin i and ii complexes license full estrogen receptor α-dependent enhancer activation. *Molecular cell*, 59(2):188–202, 2015. 4.3.1

[82] Maxim Likhachev and Anthony Stentz. R* search. In *In Proceedings of the National Conference on Artificial Intelligence (AAAI*. Citeseer, 2008. 5.2.1

[83] Wanwei Liu, Bo Liang, Hongliang Liu, Yong Huang, Xiangbao Yin, Fan Zhou, Xin Yu, Qian Feng, Enliang Li, Zhenhong Zou, et al. Overexpression of non-smc condensin i complex subunit g serves as a promising prognostic marker and therapeutic target for hepatocellular carcinoma. *International Journal of Molecular Medicine*, 40(3):731–738, 2017. 5.2.3

[84] X Shirley Liu, Douglas L Brutlag, and Jun S Liu. An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature biotechnology*, 20(8):835, 2002. 3.4.3

[85] Zhijie Liu, Daria Merkurjev, Feng Yang, Wenbo Li, Soohwan Oh, Meyer J Friedman, Xiaoyuan Song, Feng Zhang, Qi Ma, Kenneth A Ohgi, et al. Enhancer activation requires trans-recruitment of a mega transcription factor complex. *Cell*, 159(2):358–373, 2014. 4.3.1

[86] Bramanandam Manavathi, Venkata SK Samanthapudi, and Vijay Narasimha Reddy Gajulapalli. Estrogen receptor coregulators and pioneer factors: the orchestrators of mammary gland cell fate and development. *Frontiers in cell and developmental biology*, 2:34, 2014. 4.1

[87] Catriona M Manville, Kayleigh Smith, Zbyslaw Sondka, Holly Rance, Simon Cockell, Ian G Cowell, Ka Cheong Lee, Nicholas J Morris, Kay Padget, Graham H Jackson, et al. Genome-wide chip-seq analysis of human top2b occupancy in mcf7 breast cancer epithelial cells. *Biology open*, 4(11):1436–1447, 2015. 4.3.1

[88] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(1):S7, 2006. 1.4.4

[89] Jeffrey A Martin and Zhong Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10):671–682, 2011. 1.3.2

[90] Anthony Mathelier, Xiaobei Zhao, Allen W Zhang, François Parcy, Rebecca Worsley-Hunt, David J Arenillas, Sorana Buchman, Chih-yu Chen, Alice Chou, Hans Ienasescu,

et al. Jaspar 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, 42(D1):D142–D147, 2014. 3.4.3, 4.4.3

[91] Volker Matys, Olga V Kel-Margoulis, Ellen Fricke, Ines Liebich, Sigrid Land, A Barre-Dirrie, Ingmar Reuter, D Chekmenev, Mathias Krull, Klaus Hornischer, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(suppl 1):D108–D110, 2006. 3.4.3, 4.4.3

[92] Shenglin Mei, Qian Qin, Qiu Wu, Hanfei Sun, Rongbin Zheng, Chongzhi Zang, Muyuan Zhu, Jiaxin Wu, Xiaohui Shi, Len Taing, et al. Cistrome data browser: a data portal for chip-seq and chromatin accessibility data in human and mouse. *Nucleic Acids Research*, page gkw983, 2016. 1.4.5, 4.1, 4.3.1, 4.4.3, 4.4.4

[93] Michael L Metzker. Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1):31–46, 2010. 1.3.5

[94] Simone Mocellin and Maurizio Provenzano. Rna interference: learning gene knock-down from cell physiology. *Journal of Translational Medicine*, 2(1):1, 2004. 1.1, 1.3.6

[95] Hisham Mohammed, Clive D'Santos, Aurelien A Serandour, H Raza Ali, Gordon D Brown, Alan Atkins, Oscar M Rueda, Kelly A Holmes, Vasiliki Theodorou, Jessica LL Robinson, et al. Endogenous purification reveals greb1 as a key estrogen receptor regulatory factor. *Cell reports*, 3(2):342–349, 2013. 4.3.1, 4.4.1

[96] Hisham Mohammed, I Alasdair Russell, Rory Stark, Oscar M Rueda, Theresa E Hickey, Gerard A Tarulli, Aurelien A Serandour, Stephen N Birrell, Alejandra Bruna, Amel Saadi, et al. Progesterone receptor modulates erα action in breast cancer. *Nature*, 523(7560): 313–317, 2015. 4.3.1, 4.4.1

[97] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464 (7289):773–777, 2010. 1.1, 1.3.1, 3.3, 3.4.1

[98] Melody K Morris, Julio Saez-Rodriguez, Peter K Sorger, and Douglas A Lauffenburger. Logic-based models for the analysis of cell signaling networks. *Biochemistry*, 49(15): 3216–3224, 2010. 1.4.4

[99] Chris G Mueller, Charlotte Boix, Wing-Hong Kwan, Cécile Daussy, Emilie Fournier, Wolf H Fridman, and Thierry J Molina. Critical role of monocytes to support normal b cell and diffuse large b cell lymphoma survival and proliferation. *Journal of leukocyte biology*, 82(3):567–575, 2007. 3.1

[100] Ali Naderi, Michelle Meyer, and Dennis H Dowhan. Cross-regulation between foxa1 and erbb2 signaling in estrogen receptor-negative breast cancer. *Neoplasia*, 14(4):283IN3–296, 2012. 4.4.1

[101] Sankari Nagarajan, Tareq Hossan, Malik Alawi, Zeynab Najafova, Daniela Indenbirken, Upasana Bedi, Hanna Taipaleenmäki, Isabel Ben-Batalla, Marina Scheller, Sonja Loges, et al. Bromodomain protein brd4 is required for estrogen receptor-dependent enhancer

activation and gene transcription. *Cell reports*, 8(2):460–469, 2014. 4.3.1

[102] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003. 3.2

[103] Min Ni, Yiwen Chen, Teng Fei, Dan Li, Elgene Lim, X Shirley Liu, and Myles Brown. Amplitude modulation of androgen signaling by c-myc. *Genes & development*, 27(7): 734–748, 2013. 4.4.1

[104] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011. 1.3.5

[105] Michał J Okoniewski and Crispin J Miller. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7(1):276, 2006. 1.3.1

[106] Nooshin Omranian, Jeanne MO Eloundou-Mbebi, Bernd Mueller-Roeber, and Zoran Nikoloski. Gene regulatory network inference using fused lasso on multiple data sets. *Scientific Reports*, 6, 2016. 5.2.3

[107] Alicia Oshlack, Mark D Robinson, and Matthew D Young. From rna-seq reads to differential expression results. *Genome biology*, 11(12):1, 2010. 1.3.2

[108] Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009. 1.1, 1.3.3, 1.4.5

[109] Jean-Philippe Pellet and André Elisseeff. Using Markov blankets for causal structure learning. *The Journal of Machine Learning Research*, 9:1295–1342, 2008. 1.4.2

[110] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009. 1.4.4

[111] Susan M Perkins, Casey Bales, Tudor Vladislav, Sandra Althouse, Kathy D Miller, George Sandusky, Sunil Badve, and Harikrishna Nakshatri. Tfap2c expression in breast cancer: correlation with overall survival beyond 10 years of initial diagnosis. *Breast cancer research and treatment*, 152(3):519–531, 2015. 4.4.2

[112] Camillo Porta, Chiara Paglino, and Alessandra Mosca. Targeting pi3k/akt/mtor signaling in cancer. *Frontiers in oncology*, 4, 2014. 4.4.1

[113] Sebastian Pott and Jason D Lieb. What are super-enhancers? *Nature genetics*, 47(1):8–12, 2015. 4.4.4

[114] Sisi Qin, James N Ingle, Mohan Liu, Jia Yu, D Lawrence Wickerham, Michiaki Kubo, Richard M Weinshilboum, and Liewei Wang. Calmodulin-like protein 3 is an estrogen receptor alpha coregulator for gene expression and drug response in a snp, estrogen, and serm-dependent fashion. *Breast Cancer Research*, 19(1):95, 2017. 1.4.7

[115] James M Rae, Michael D Johnson, Joshua O Scheys, Kevin E Cordero, José M Larios, and Marc E Lippman. Greb1 is a critical regulator of hormone dependent breast cancer growth. *Breast cancer research and treatment*, 92(2):141–149, 2005. 4.4.1

[116] Nisha Rajagopal, Sharanya Srinivasan, Kameron Kooshesh, Yuchun Guo, Matthew D Edwards, Budhaditya Banerjee, Tahin Syed, Bart JM Emons, David K Gifford, and Richard I Sherwood. High-throughput mapping of regulatory dna. *Nature biotechnology*, 34(2):167, 2016. 1.4.7

[117] Bing Ren, François Robert, John J Wyrick, Oscar Aparicio, Ezra G Jennings, Itamar Simon, Julia Zeitlinger, Jörg Schreiber, Nancy Hannett, Elenita Kanin, et al. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–2309, 2000. 1.4.5

[118] Kate R Rosenbloom, Joel Armstrong, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The UCSC genome browser database: 2015 update. *Nucleic acids research*, 43(D1):D670–D681, 2015. 3.4.3, 3.2

[119] Thomas E Royce, Joel S Rozowsky, and Mark B Gerstein. Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Research*, 35(15):e99, 2007. 1.3.1

[120] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. *Artificial intelligence: a modern approach*, volume 2. Prentice hall Upper Saddle River, 2003. 1.4.3, 2.1, 2.4, 2.4

[121] Sudipa Saha Roy and Ratna K Vadlamudi. Role of estrogen receptor signaling in breast cancer metastasis. *International journal of breast cancer*, 2012, 2011. 4.1

[122] Juliane Schäfer and Korbinian Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2004. 1.4.4

[123] Dominic Schmidt, Petra C Schwalie, Caryn S Ross-Innes, Antoni Hurtado, Gordon D Brown, Jason S Carroll, Paul Flicek, and Duncan T Odom. A ctcf-independent role for cohesin in tissue-specific transcription. *Genome research*, 20(5):578–588, 2010. 4.3.1

[124] Mark Schmidt, Alexandru Niculescu-Mizil, and Kevin Murphy. Learning graphical model structure using L1-regularization paths. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, pages 1278–1283, 2007. 1.4.2

[125] Anthony D Schmitt, Ming Hu, and Bing Ren. Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*, 17(12):743–755, 2016. 5.2.3

[126] Almut Schulze and Julian Downward. Navigating gene expression using microarrays — a technology review. *Nature Cell Biology*, 3(8):E190–E195, 2001. 1.3.1

[127] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978. 1.4.2

[128] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2):166, 2003. 1.4.4

[129] Eran Segal, Dana Pe'er, Aviv Regev, Daphne Koller, and Nir Friedman. Learning module networks. *Journal of Machine Learning Research*, 6(Apr):557–588, 2005. 1.4.4

[130] Takashi Shimbo, Ying Du, Sara A Grimm, Archana Dhasarathy, Deepak Mav, Ruchir R

Shah, Huidong Shi, and Paul A Wade. Mbd3 localizes at promoters, gene bodies and enhancers of active genes. *PLoS Genet*, 9(12):e1004028, 2013. 4.3.1

[131] Wenzhe Si, Wei Huang, Yu Zheng, Yang Yang, Xujun Liu, Lin Shan, Xing Zhou, Yue Wang, Dongxue Su, Jie Gao, et al. Dysfunction of the reciprocal feedback loop between gata3-and zeb2-nucleated repression programs contributes to breast cancer metastasis. *Cancer Cell*, 27(6):822–836, 2015. 4.3.1

[132] Ngak-Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C Ng. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*, 40(W1):W452–W457, 2012. 3.2

[133] Ajit Singh and Andrew Moore. Finding optimal Bayesian networks by dynamic programming. Technical Report 05-106, School of Computer Science, Carnegie Mellon University, 2005. 1.4.2, 2.1, 2.3

[134] Xiang Song, Yue-Ming Xing, Wei Wu, Guo-Hua Cheng, Feng Xiao, Gang Jin, Ying Liu, and Xin Zhao. Expression of krüppel-like factor 4 in breast cancer tissues and its effects on the proliferation of breast cancer mda-mb-231 cells. *Experimental and Therapeutic Medicine*, 13(5):2463–2467, 2017. 5.2.3

[135] Ashwani K Sood, Joseph Geradts, and Jessica Young. Prostate-derived ets factor, an oncogenic driver in breast cancer. *Tumor Biology*, 39(5):1010428317691688, 2017. 5.2.3

[136] Rory Stark and Gordon Brown. Diffbind: differential binding analysis of chip-seq peak data. *R package version*, 100, 2011. 4.3.1

[137] David F Stern. Tyrosine kinase signalling in breast cancer: Erbb family receptor tyrosine kinases. *Breast Cancer Research*, 2(3):176, 2000. 4.4.2

[138] David L Stern and Virginie Orgogozo. The loci of evolution: how predictable is genetic evolution? *Evolution*, 62(9):2155–2177, 2008. 3.4.1

[139] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003. 1.4.6

[140] Barbara E Stranger, Stephen B Montgomery, Antigone S Dimas, Leopold Parts, Oliver Stegle, Catherine E Ingle, Magda Sekowska, George Davey Smith, David Evans, Maria Gutierrez-Arcelus, et al. Patterns of cis regulatory variation in diverse human populations. *PLoS Genetics*, 8(4):e1002639, 2012. 1.1, 1.3.1, 3.3, 3.4.1

[141] Wei Sun and Yijuan Hu. eqtl mapping using rna-seq data. *Statistics in biosciences*, 5(1): 198–219, 2013. 1.4.6

[142] Yun Ju Sung, Keegan D Korthauer, Michael D Swartz, and Corinne D Engelman. Methods for collapsing multiple rare variants in whole-genome sequence data. *Genetic epidemiology*, 38(S1), 2014. 5.2.3

[143] Si Kee Tan, Zhen Hua Lin, Cheng Wei Chang, Vipin Varang, Kern Rei Chng, You Fu Pan, Eu Leong Yong, Wing Kin Sung, and Edwin Cheung. Ap-2γ regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription. *The EMBO journal*, 30(13):2569–2581, 2011. 4.3.1

[144] Marc Teyssier and Daphne Koller. Ordering-based search: A simple and effective al-

gorithm for learning Bayesian networks. In *Proceedings of the Twentieth conference on Uncertainty in Artificial Intelligence*, pages 584–590, 2005. 2.3

[145] Vasiliki Theodorou, Rory Stark, Suraj Menon, and Jason S Carroll. Gata3 acts upstream of foxa1 in mediating esr1 binding by shaping enhancer accessibility. *Genome research*, 23(1):12–22, 2013. 4.3.1, 4.4.1, 4.4.4

[146] René Thomas. Boolean formalization of genetic control circuits. *Journal of theoretical biology*, 42(3):563–585, 1973. 1.4.4

[147] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414): 75–82, 2012. 4.3.1

[148] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 1.4.6

[149] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. 1.4.6, 5.2.3

[150] Wen-Wei Tsai, Zhanxin Wang, Teresa T Yiu, Kadir C Akdemir, Weiya Xia, Stefan Winter, Cheng-Yu Tsai, Xiaobing Shi, Dirk Schwarzer, William Plunkett, et al. Trim24 links a non-canonical histone signature to breast cancer. *Nature*, 468(7326):927–932, 2010. 4.3.1

[151] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006. 1.4.2

[152] Ioannis Tsamardinos, Alexander Statnikov, Laura E Brown, and Constantin F Aliferis. Generating realistic large Bayesian networks by tiling. In *the Nineteenth International FLAIRS conference*, pages 592–597, 2006. 2.5

[153] Jerry Usary, Victor Llaca, Gamze Karaca, Shafaq Presswala, Mehmet Karaca, Xiaping He, Anita Langerød, Rolf Kåresen, Daniel S Oh, Lynn G Dressler, et al. Mutation of gata3 in human breast tumors. *Oncogene*, 23(46):7669, 2004. 4.4.1

[154] Nedumparambathmarath Vijesh, Swarup Kumar Chakrabarti, and Janardanan Sreekumar. Modeling of gene regulatory networks: A review. *Journal of Biomedical Science and Engineering*, 6(02):223, 2013. 1.4.4, 1.4.4, 1.4.4

[155] Yubao Wang, Tinghu Zhang, Nicholas Kwiatkowski, Brian J Abraham, Tong Ihn Lee, Shaozhen Xie, Haluk Yuzugullu, Thanh Von, Heyuan Li, Ziao Lin, et al. Cdk7-dependent transcriptional addiction in triple-negative breast cancer. *Cell*, 163(1):174–186, 2015. 4.4.4

[156] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009. 1.3.1, 1.3.2

[157] Julie Wells and Peggy J Farnham. Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation. *Methods*, 26(1):48–56, 2002. 1.4.5

[158] Thomas Werfel, Jonathan M Spergel, and Wieland Kiess. Atopic dermatitis in childhood and adolescence. *Dermatology*, 225(2):97–192, 2012. 3.1

[159] Brian J Wilson and Vincent Giguère. Meta-analysis of human cancer microarrays reveals gata3 is integral to the estrogen receptor alpha pathway. *Molecular cancer*, 7(1):49, 2008. 4.4.1

[160] Patricia J Wittkopp and Gizem Kalay. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature reviews. Genetics*, 13(1):59, 2012. 3.4.1

[161] Jing Xiang and Seyoung Kim. A* lasso for learning a sparse bayesian network structure for continuous variables. In *Advances in Neural Information Processing Systems*, pages 2418–2426, 2013. 1.1, 2.1

[162] Yufei Xiao. A tutorial on analysis and simulation of boolean gene regulatory network models. *Current genomics*, 10(7):511–525, 2009. 1.4.4

[163] Yichen Xu, Hua Zhang, Nicos Angelopoulos, Joao Nunes, Alistair Reid, Laki Buluwela, Luca Magnani, Justin Stebbing, Georgios Giamas, et al. Lmtk3 represses tumor suppressor-like genes through chromatin remodeling in breast cancer. *Cell reports*, 12(5): 837–849, 2015. 4.3.1

[164] Yanzhong Yang, Yue Lu, Alexsandra Espejo, Jiacai Wu, Wei Xu, Shoudan Liang, and Mark T Bedford. Tdrd3 is an effector molecule for arginine-methylated histone marks. *Molecular cell*, 40(6):1016–1023, 2010. 4.3.1

[165] Dihua Yu and Mien-Chie Hung. Overexpression of erbb2 in cancer and erbb2-targeting strategies. *Oncogene*, 19(53):6115, 2000. 4.4.2

[166] Changhe Yuan, Brandon Malone, and Xiaojian Wu. Learning optimal bayesian networks using a* search. In *IJCAI proceedings-international joint conference on artificial intelligence*, volume 22, pages 2186–2191. Citeseer, 2011. 1.4.2

[167] Daniel J Zabransky and Ben Ho Park. Estrogen receptor and receptor tyrosine kinase signaling: Use of combinatorial hormone and epidermal growth factor receptor/human epidermal growth factor receptor 2–targeted therapies for breast cancer. *Journal of Clinical Oncology*, 32(10):1084–1086, 2014. 4.1

[168] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005. 1.4.4

[169] Jing Zhang, Chengyang Wang, Xi Chen, Mamoru Takada, Cheng Fan, Xingnan Zheng, Haitao Wen, Yong Liu, Chenguang Wang, Richard G Pestell, et al. Egln2 associates with the nrf1-pgc1α complex and controls mitochondrial function in breast cancer. *The EMBO journal*, 34(23):2953–2970, 2015. 4.3.1

[170] Ze-Yi Zheng, Boon-Huat Bay, Swee-Eng Aw, and Valerie CL Lin. A novel antiestrogenic mechanism in progesterone receptor-transfected breast cancer cells. *Journal of Biological Chemistry*, 280(17):17480–17487, 2005. 4.4.1

[171] Jiajun Zhu, Morgan A Sammons, Greg Donahue, Zhixun Dou, Masoud Vedadi, Matthäus

Getlik, Dalia Barsyte-Lovejoy, Rima Al-Awar, Bryson W Katona, Ali Shilatifard, et al. Gain-of-function p53 mutants co-opt chromatin pathways to drive cancer growth. *Nature*, 525(7568):206–211, 2015. 4.3.1

[172] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. 1.4.6