**PREVIOUS COMPUTATIONAL EFFORTS**

We have made extensive contributions in the analysis of genomics data, especially with regard to (A) constructing networks, (B) analyzing them, (C) building predictive models to relate gene expression with TF binding and histone modifications, and (D) identifying and annotate ncRNAs . Here we outline our main contributions in each of these aspects of genomics research.

1. **Contributions innetwork construction**

***General network construction.*** We have developed several methods to construct networks based upon genome features ([Jansen et al., 2002](#_ENREF_37)). Based on this work, we combined several heterogeneous biological datasets to increase the power of prediction ([Edwards et al., 2002](#_ENREF_20); [Lu et al., 2005](#_ENREF_56); [Xia et al., 2006](#_ENREF_84)) and we developed new machine-learning techniques to support these research goals ([Yip et al., 2009](#_ENREF_89)). In 2008, this work placed first in theDialogue for Reverse Engineering Assessments and Methods(DREAM, [www.the-dream-project.org](http://www.the-dream-project.org)) competition for the *in silico* network prediction challenge. In addition, we have participated in many experimental network determination projects ([Borneman et al., 2006](#_ENREF_6); [Li et al., 2004](#_ENREF_50)), in a continual effort to refine our methodologies and keep them at the cutting edge.

***Construction of proximal regulatory network.*** We have developed several computational approaches based on data from cutting-edge next-generation sequencing technologies (such as ChIP-seq) to help construct proximal regulatory networks and identify regulatory targets of TFs. To more effectively utilize the ChIP-seq data for network construction, we developed a method called PeakSeq ([Rozowsky et al., 2009](#_ENREF_74)) to define the binding peaks of TFs. PeakSeq constructs local thresholds using input signals from genomic DNA without an enrichment process to simulate the null process for the background. The variability in the background signal reflects the accessibility of the DNA due to the chromatin state of the genome. PeakSeq then identifies TF binding regions by identifying peaks that are significantly enriched relative to the background signal. PeakSeq is a widely used and highly versatile tool for identifying TF binding sites from ChIP-seq data.

In addition, we have also proposed a probabilistic model, referred to as target identification from profiles (TIP), that identifies a given TF’s target genes based on ChIP-seq data ([Cheng et al., 2011a](#_ENREF_12)). Given a ChIP-seq dataset for a particular TF, we initially identify all genes bound by the TF. Then, our method characterizes the binding profile of the TF by averaging its binding signal at each position around the transcription start site (TSS) for these genes. We can subsequently calculate the binding strength of a given TF for a particular gene by using the characteristic profile to obtain the weighted sum of the TF binding signal from nucleotides surrounding the TSS. Finally, we can estimate the significance levels for observed binding behavior between a TF and gene by assuming a normal distribution of regulatory scores. Compared to the peak-based method, our model identifies a more reliable set of TF targets. In particular, when TIP was used to identify target genes of ERα, the resultant set of genes was more responsive to estrogen treatment than genes identified using peak-based methods. Though not yet tested, we believe that the triangulation of TF binding of the three members of the ER enhancesome (ER, FOXA1, and GATA3) would provide an even greater estimate of the binding sites that serve a growth function in these cell lines.

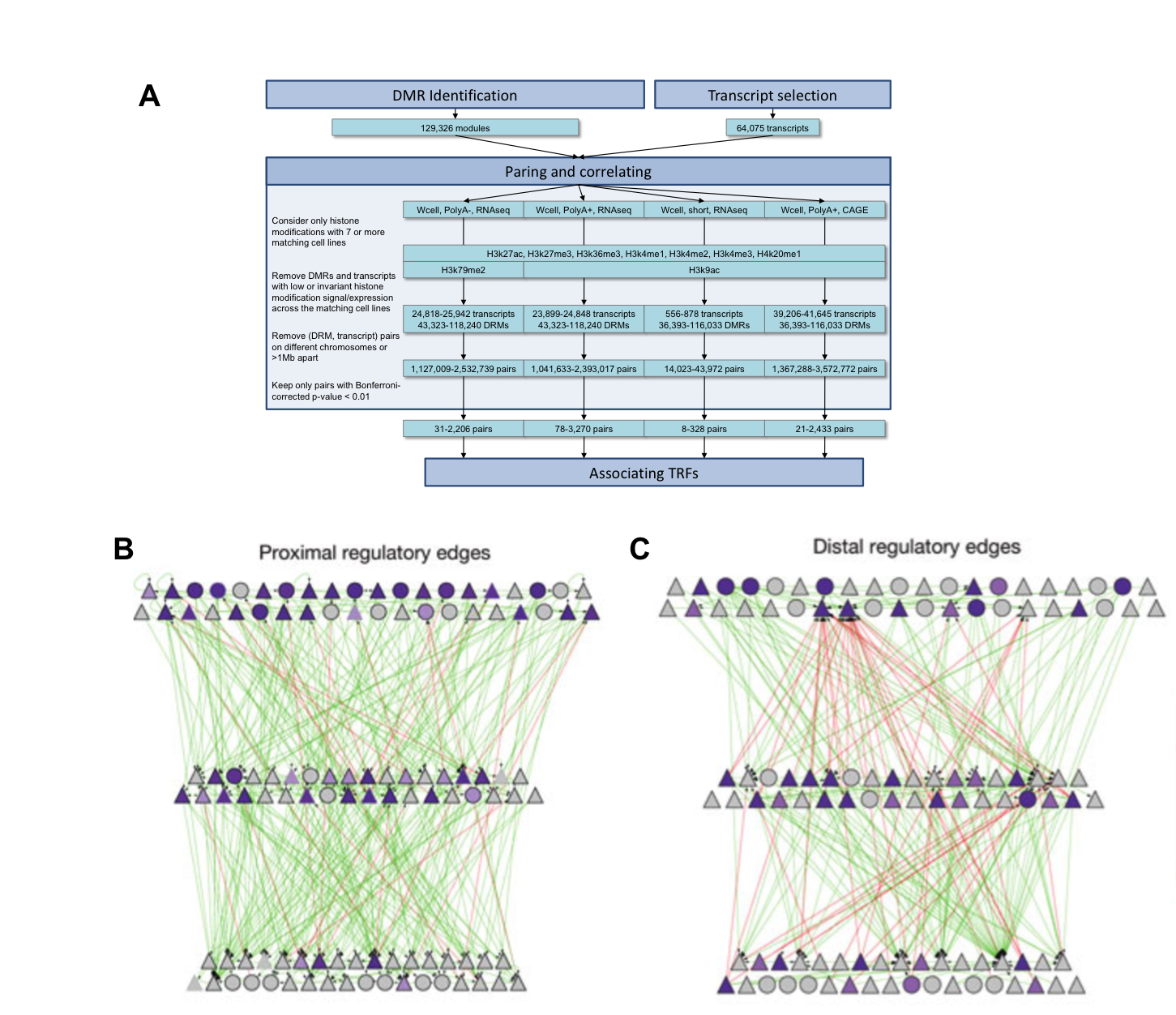
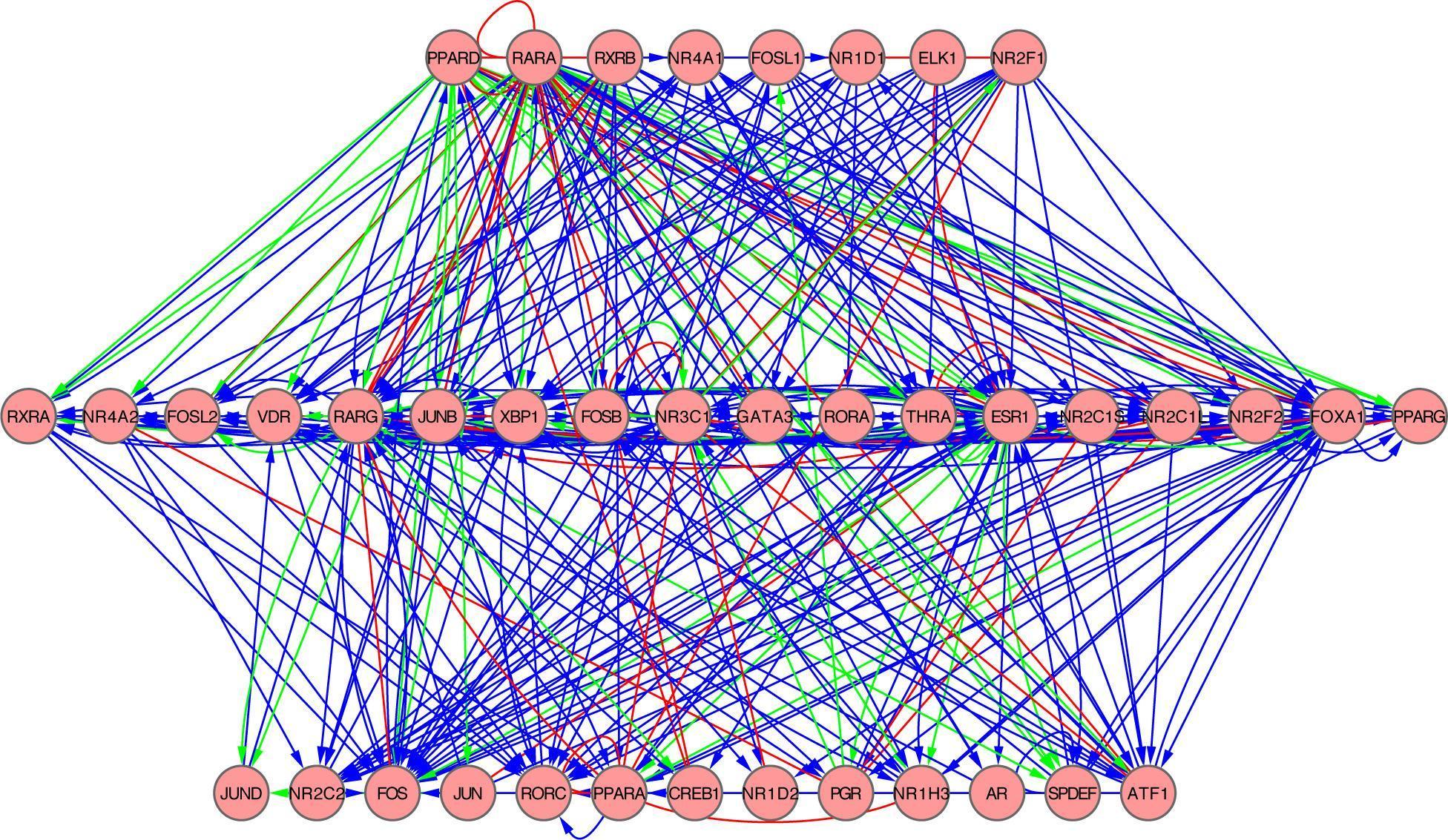
***Constructing distal networks and identifying enhancers.*** We have developed machine-learning methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal regulatory regions (DRM) ([Yip et al., 2010](#_ENREF_87)). By correlating the binding signals around DRMs with respect to expression of transcripts, we developed a computational pipeline to identify potential enhancers and the transcripts associated with them (Figure 1A). As published in ([Yip et al., 2012](#_ENREF_88)), we have validated some of the results from our pipeline by experiments, which show a fairly high predictive accuracy. The enhancers and their targets form a distal regulatory network (Figure 1C), and when analyzed along with the corresponding proximal regulatory network (Figure 1B), provide a more comprehensive and complete view that incorporates multiple dimensions of transcriptional regulation into the network.

Figure 1. A. Associating transcription factors to target genes ([Yip et al., 2012](#_ENREF_88)). B. Proximal regulatory network. C) Distal regulatory network ([Gerstein et al., 2012](#_ENREF_26))

***Network constructions using ENCODE, modENCODE and other system-wide data.*** Using the machine-learning approaches we developed for identifying individual proximal and distal edges together with miRNA target prediction (and other) algorithms, we have completed the highly ambitious goal of constructing highly integrated regulatory networks for humans and model organisms based on the ENCODE ([Gerstein et al., 2012](#_ENREF_26)) and modENCODE datasets ([Negre et al., 2011](#_ENREF_67)). These integrated networks consist of three major types of regulation: TF-gene, TF-miRNA and miRNA-gene, showing rich statistical patterns. For instance, the human regulatory network uniquely displays distinct preferences for binding at proximal and distal regions. The proximal-distal binding preference is a property of the intergenic space in the human genome, which is much larger relative to the genomes of other model organisms. This difference leads to a larger amount of distal binding. Furthermore, in the human regulatory network, the more highly connected TFs are more likely to exhibit allele-specific binding and gene expression. More recently, we built a regulatory map for 24 nuclear receptors and 14 breast-cancer-associated TFs that are expressed in the breast cancer cell line MCF-7 (Figure 2). The resulting network reveals a highly interconnected regulatory matrix with extensive “crosstalk” between NRs and other breast-cancer-associated TFs. We show that large numbers of factors bind in a coordinated fashion to target regions throughout the genome. The highly occupied targets are associated with active chromatin state and hormone-responsive gene expression.

1. **Contributions in network analysis**

Over the last few years, we have developed a number of tools to analyze the organization and structure of biological networks including identifying the importance of a node in a single network and identifying the modular structure inherent within several biological networks.

***Centrality measures and network modules.*** We have developed a number of different methods to analyze networks that allow us to identify modules and hubs. Starting with local properties of nodes, we have completed numerous studies correlating a node’s tendency to act as a hub with various forms of “essentiality” (i.e., the degree to which a given node is essential for various functions in a network) ([Yu et al., 2004a](#_ENREF_92)). In addition to hubs, we found that bottlenecks in the network are also important ([Yu et al., 2007](#_ENREF_93)) .

Nodes in networks tend to work together as small structures called network motifs. We found that in many of the regulatory networks we constructed in human, worm and fly, the small modular motifs have been evolutionarily reused to create complex transcriptional regulatory networks. The feed-forward loop is over represented in these networks and is used to filter the input stimuli regulating the transcriptional machinery across different hierarchical levels to modulate the expression level of different genes.

**Figure 2.** **Hierarchical regulatory network in MCF7 cells**. The network was inferred for three classes of edges representing the following types of functional interaction between NRs: type I (binding of one TF within 50 kb range of the promoter of the other TF, depicted in blue); type II (expression level of one factor affects expression of the target gene of another factor, depicted in red); type III (activation of one NR affects the expression of another factor, depicted in green).

Cellular networks are also organized in the form of interacting modules, whereby nodes in a module tend to have a larger density of edges connecting them. Biologically, the genes within a module of a genetic regulatory network are co-regulated. We developed various methods to identify the functional modules of various networks. For example, by mapping gene-expression data onto the regulatory network of yeast, we identified different sub-networks that are active in different conditions ([Luscombe et al., 2004](#_ENREF_58)). We developed a method to extract metabolic modules from metagenomic data, enabling us to identify pathways that are expressed under different environmental conditions (Figure 3B) ([Gianoulis et al., 2009](#_ENREF_29)). We have also developed a way to identify nearly complete, fully connected modules (cliques) present in network interactions ([Yu et al., 2006](#_ENREF_95)) and we have been using networks to map various kinds of functional genomics data ([Gerstein et al., 2012](#_ENREF_26)).

***Building hierarchies.*** In addition to analyzing the topology of gene-regulatory networks, we have also probed the direction of information flow in these networks. We found that gene-regulatory networks are composed of hierarchical structures dominated by downward information flow and that some TFs act as top master regulators to govern the transcription of downstream TFs. We developed methods to determine the hierarchical organization of regulatory networks and applied them to analyze the regulatory networks of a variety of species from yeast to human, including networks constructed from ENCODE, modENCODE and MCF7 data ([Cheng et al., 2011b](#_ENREF_13); [Gerstein et al., 2012](#_ENREF_26); [Gerstein et al., 2010](#_ENREF_27); [Yan et al., 2010](#_ENREF_85)). In these hierarchical networks, TFs are organized into three levels, whereby TFs at the top tend to act as regulators while TFs at the bottom tend to be targets of regulation.

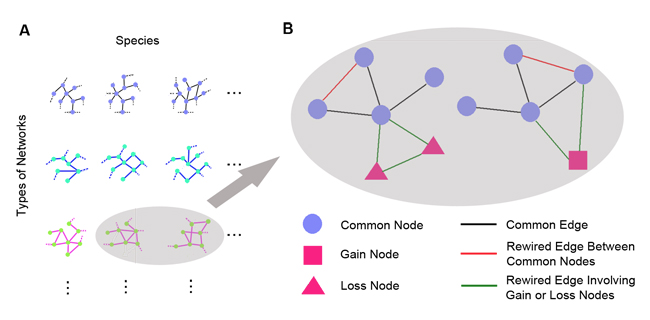
**Figure 3**. **A.** Comparison of management (left) and transcriptional regulatory (right) hierarchies ([Yu and Gerstein, 2006](#_ENREF_91)). **B.** Metabolic network modules active in a specific environment ([Gianoulis et al., 2009](#_ENREF_29)). **C**. Positive selection occurs on the periphery of the human protein interaction network (Kim et al. 2007).

Through cross-species analysis, we also found that regulatory factors are hierarchically organized in all organisms. In a hierarchical organization, the factors at the top of the hierarchy are most influential, as reflected by their highly correlative binding and gene-expression profiles. The factors at the top level are under stronger selective pressure and are more conserved evolutionarily. In comparison, the middle level contains many elements characterized by bottlenecks in information flow and are highly connected by miRNA and distal regulatory elements. These cross-species analyses thus highlight the general features of regulatory networks conserved in evolution.

***Relating network and genomic variation.*** Recently, we used networks to improve our understanding of genomic variants ([Khurana et al., 2013b](#_ENREF_42)). In ([Khurana et al., 2013a](#_ENREF_41)), we built a multi-layered network that incorporated information from heterogeneous data sources such as protein-protein interactions and metabolic, phosphorylation, signaling, genetic, and regulatory networks. In general, population variants are more likely to be deleterious when they occur in genes or in regulatory elements associated with hubs in the multi-layered networks, indicating that a gene’s interactions likely influence the selective pressures on acting on it. Connectivity is also related to selective pressure in noncoding regions, as transcription binding motifs with greater connectivity tend to be under stronger evolutionary pressure ([Khurana et al., 2013b](#_ENREF_42)). We built a workflow model to prioritize noncoding mutations in disease variants based on these patterns of negative selection in functional variants. In addition, we showed that proteins under positive selection are found on the network and on the cellular periphery, an indication of how human variation is arranged with respect to the interactome ([Kim et al., 2007](#_ENREF_43)). Finally, we have also developed methods to describe allelic binding with TFs in a network frame ([Rozowsky et al., 2011](#_ENREF_73)).

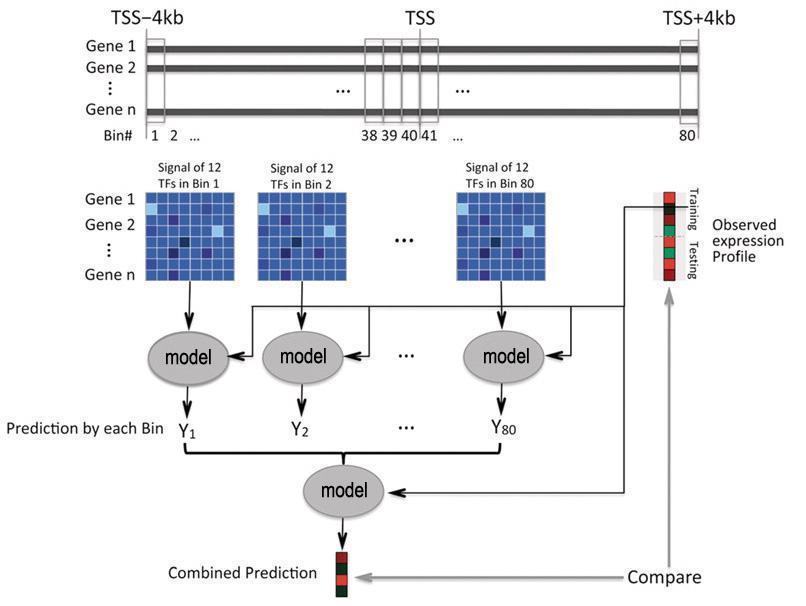
***Integrating networks with other biological data.*** To further illustrate the value of the network concept, we have also combined network analyses with many other types of biological data. For example, to decipher the complexity of protein-interaction networks, we integrated structural protein information, i.e., types of protein interfaces ([Bhardwaj et al., 2011](#_ENREF_4)). This work showed that much of the debate on the degree to which hubs are conserved could be resolved by focusing on the number of structural interfaces of a protein rather than its number of partners. It also suggested different models for network evolution through gene duplication, depending on whether or not a newly created protein connects to a pre-existing structural interface.

***Comparison of networks in terms of evolution and rewiring.*** We have also explored the evolution of networks and studied the conservation and variability of different parts of the network. We defined "interologs" and showed how to compare interaction networks between organisms. We also defined "regulogs" for transferring regulatory relationships between organisms ([Yu et al., 2004b](#_ENREF_94)). In addition, we developed a method to study network rewiring on all currently available biological networks (Figure 4). We noted that biological networks show a decreased rate of change over large time intervals. However, different types of biological networks consistently rewire at different rates ([Shou et al., 2011](#_ENREF_78)).



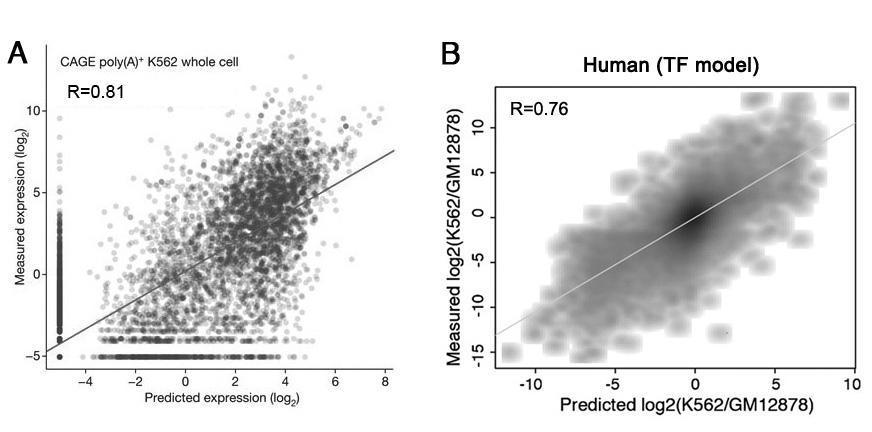
**Figure 4**. **Measuring network rewiring by comparing networks of species pairs**. **A.** Types of biological networks with currently available data for different species are collected. Selected types of commonplace networks with multiple time-point data are also collected. **B**. For each network type, we perform edge rewiring analysis for pairs of species. Three types of nodes are first identified as CNs, GNs and LNs. Four types of rewired edges are then identified and counted including gain/loss edges between CNs (red) and those involving GNs or LNs (green). Rewiring rate from comparing the networks is calculated**.**

***Web tools for network analysis.*** We have developed many network analysis web tools such as TopNet ([Yu et al., 2004c](#_ENREF_96)), [tYNA](http://tyna.gersteinlab.org/tyna/) ([Yip et al., 2006](#_ENREF_90)), and [PubNet](http://pubnet.papers.gersteinlab.org/) ([Douglas et al., 2005](#_ENREF_19)). These tools have been widely used by the research community to analyze network topology—i.e., to calculate hubs, “between-ness”, shortness of paths and degree of modularity.

**Figure 5.** **A two-layer supervised model for predicting gene expression levels based on TF binding signals**. DNA regions surrounding the TSS are divided into 80 bins, each of 100bp in size. In the first layer, for each bin a model is constructed to predict the expression values of genes using binding signals of multiple TFs as predictors. The predicted values from all bins are then combined by the second layer model to make the final prediction of gene expression. TSS: transcription start site.

***Predictive models of gene expression.*** Transcription factors and histone modifications are two interrelated components that regulate the transcriptional output of a gene. To quantify the relationship between TF binding and gene expression, we have constructed linear and non-linear models (Figure 5) that utilize the binding signals of multiple TFs in the transcription start site (TSS) proximal to genes as the input to “predict” expression levels of protein coding genes ([Cheng and Gerstein, 2012](#_ENREF_11)). Similarly, we have also constructed models to predict gene-expression levels based on histone modification signals at different positions proximal to the TSS of different genes ([Dong et al., 2012](#_ENREF_18); [Gerstein et al., 2010](#_ENREF_27)).

We applied these models in multiple organisms ranging from yeast to human. For example, a model based on binding signals of 40 TFs achieves high predictive accuracy in K562 (Figure 6A).

****

**Figure 6.** **TF binding predicts expression levels and differential expression of genes**. **A**. Expression of genes in K562 can be accurately predicted by TF binding signals in the promoter regions. **B**. Differential gene expression between K562 and GM12878 is highly predictive of TF binding differences in the two cell lines.

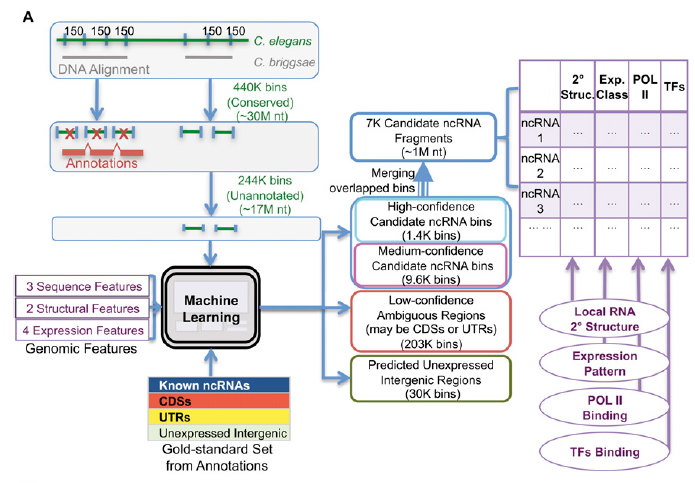
The predictive models revealed several important trends: (1) TF binding and histone modification profiles achieve comparable predictive accuracy of gene expression, each accounting for >60% of the variation in gene-expression levels. (2) TF binding and histone modification signals are highly redundant in predicting gene expression, such that inclusion of both variables as predictors does not improve prediction accuracy beyond models built using just one of these variables. (3) TF binding signals achieve the highest prediction accuracy at the TSS of genes and the accuracy decays rapidly when models consider TF binding regions close to the TSS, whereas histone modifications result in high accuracy in a relatively wide number of DNA regions (e.g., from upstream locations to transcribed regions of genes). (4) Gene-expression levels can be accurately predicted using a few TFs due to correlated binding signals among TFs.

We constructed TF and histone models for predicting expression levels of protein-coding and non-coding genes ([Cheng and Gerstein, 2012](#_ENREF_11)). Strikingly, the models trained solely on protein-coding genes also predicted the expression levels of non-coding genes, suggesting a common regulatory mechanism for ncRNA and protein-coding genes. In addition, our models indicate that, in different species, the functions of histone modifications are conserved. Furthermore, the models predict changes in gene expression in different cell lines or under different conditions. Specifically, more than 50% of the variation in differential gene expression between human K562 and GM12878 cell lines can be explained by differential TF binding ([Cheng et al., 2012](#_ENREF_10)) (Figure 6B).

***Non-coding RNA annotation and functional characterization.*** Long and short non-coding RNAs (ncRNA) are key post-transcriptional regulators of gene expression ([Mercer and Mattick, 2013](#_ENREF_64)). Although there are many methods to incorporate ncRNAs into regulatory networks, these are mostly concentrated on the short ncRNAs (miRNAs and other short interfering RNAs). Regarding long ncRNAs, only the regulatory effects of antisense transcripts have been studied since their targets are most easily identified.

Based upon data analysis using RNA-seq, we have developed many tools for identifying, quantifying, and characterizing pervasive and non-coding RNA transcription, including RSeqTools ([Habegger et al., 2011](#_ENREF_32)) (which are a suite of programs for processing RNA-seq data), lncRNA ([Gerstein et al., 2010](#_ENREF_27)) (which combines sequence conservation, secondary structure, protein-coding potential, and expression patterns to train a classifier for detection of ncRNAs), and DART ([Rozowsky et al., 2007](#_ENREF_75)) (which characterizes the novel transcribed regions from expression patterns in multiple samples). In addition, we have developed pipelines for the identification of pervasive transcription ([Lian et al., 2003](#_ENREF_52)) using RNA-seq and expression microarray data (Figure 7).

Taken together, our previous and current work demonstrates our expertise not only in the biology of estrogen receptors, but also in developing innovative genome technologies, in computational genomics and in the construction of regulatory networks that explain signaling behavior.

****

**Figure 7**. **lncRNA framework**. The expressed elements are divided into three classes; CDS, UTR, and ncRNA. For each element, we compute several features representing the structure, sequence, conservation, and expression levels. The features are input the machine learning algorithm to build a classifier that is then used to classify the transcribed regions that are not annotated.