One of the biggest moments in the history of science was when Johannes Kepler discovered the laws of planetary motion by mining the systematic astronomical catalogs he inherited from Tycho Brahe. In the post-genomics era, we have the privilege of categorizing the genomics, epigenomics, transcriptomics, and interactomics of hundreds of species. Following the footsteps of Kepler, I aim to model the quantitative relationships implicit in these big omics catalogs. Unlike planetary systems consisting of only a handful of bodies, biological systems are mediated by the interactions between thousands of molecules. To unlock this level of complexity, network-based statistical models are particular useful.

# Previous Accomplishments

*Statistical Models of Gene Expression.* Gene expression is output of a complex regulatory process with input signals from transcription factors (TFs) and histone modifications (HMs), two interrelated components that affect the DNA region upstream of a gene. To quantify the relationship between TFs/HMs and gene expression, my lab has constructed linear and non-linear models that utilize the signals of multiple TFs/HMs in the transcription start site (TSS) proximal to genes as the input to “predict” expression levels of protein coding genes models in multiple organisms ranging from yeast to human (Science '10 330:1775, Genome Res '12 22:1658, Genome Bio '12 13:R53, Nature '12 489:57, NAR '12 40:553).

*Building Networks.* The formalism of networks provides a common language for understanding a wide range of complex systems, so my lab has developed many network-science based approaches for building the wiring and connectivity relationships of biological networks. My lab developed methods for predicting networks from heterogeneous biological datasets including genome features (Genome Res '02 12:37, Science '03 302:449, Genome Res '05 15:945, JMB '06 357:339, Bioinformatics '09 25:243). In addition, we have participated in many experimental network determination projects, to refine and keep our methodologies at the cutting edge (Nature '06 440:637, Science '04 303:540, Genes Dev '06 20:435). Recently, we developed the machine-learning approaches for identifying individual proximal and distal edges together with miRNA target prediction algorithms, we have completed the ambitious goal of constructing draft regulatory networks for humans and model organisms based on the mod/ENCODE datasets (Science '10 330:1775, Nature '11 471:527, Nature '12 489:91).

*Analysis of Network Topology.* Biological networks are normally large in scale, but organized with topological structures in the form of interacting modules. Statistics such as ‘eccentricity' and ‘betweenness' are helpful to explain the connectivity and behavior of nodes in a network (Bioinformatics '07 23:2163, PLoS CB '07 3:e59). My lab 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 (Nature '04 431:308). We further developed a scalable approach to identify nearly complete, fully connected modules (defective cliques) present in network interactions (Bioinformatics '06 22:823). We also developed a method to extract variable metabolic modules from meta-genomic data, enabling us to identify pathways that are expressed under different environmental conditions (PNAS '09 106:1374), and a spectral analysis framework to identify connection patterns across three datasets, and applied it in a variety of genomic contexts including chemogenomics data (Genome Bio '11 12:R32). Moreover, we found that the hierarchy rather than the connectivity better reflects the importance of regulators and examined the degree of collaboration among different regulators (PNAS '06 103:14724, Sci Signal '10 3:ra79, PLoS CB '10 6:e1000755, Nature '12 489:91). We have found that in *E. coli*, yeast, and human the highest degree of collaboration is between regulators from the middle level, which is analogous to a corporate setting in which middle managers play an important organizational role (PNAS '10 107:6841). Besides developing above algorithms, my lab also constructed much software for network analysis including Topnet, tYNA and PubNet (NAR '04 32:328, Bioinformatics '06 22:2968, Genome Bio '05 6:R80).

*Mapping Variation onto Networks.* My lab uses networks as a framework for integrating a great variety of genomic variation/mutation data across individuals and organisms and studying their impact on biological systems. We have found that functionally significant and highly conserved genes tend to be more central in interaction and regulatory networks (i.e. more connectivity is associated with more constraint) -- but not in metabolic pathways, where the highly central genes have more duplicated copies and are more tolerant to loss-of-function mutations (Genome Bio '06 7:R55, PLoS CB '13 9:e1002886, PLoS CB '09 5:e1000413). Moreover, we examined the impact of adaptive evolution to protein interaction networks, and found that proteins under positive selections tend to locate at network periphery (PNAS '07 104:20274). We have contrasted this more-connectivity-more-constraint pattern in biological networks to what is found in the designed network of a computer operating system (the Linux call graph), which has the opposite pattern (PNAS '10 107:9186). We also developed a framework to quantify the differences between networks in a unified fashion via looking at the degree of rewiring between different networks (PLoS CB '11 7:e1001050). Finally, we have demonstrated that networks can be used practically to prioritize the most deleterious variants in cancers (Science '13 342:6154).

# Future Research Direction

My lab aims to develop methods by integrating statistical models and network analysis to predict biological mechanisms underlying network wiring, especially those driven by genomic variation. This work could have a big *payoff* in discovering novel biological perturbations for increasing bio-fuel productions in plants and microbes.

*(1) Combinatorics of Regulation & Enhancers.* Because current statistical models of gene expression do not really consider the intricate combinatorics of the many different combinations of regulatory factors (RFs) in a network way, we want to develop scalable predictive models to identify genome-wide regulatory networks and predict regulatory mechanisms underlying network connectivity. For example, because the target genes can be regulated by the TFs/HMs that bind to proximal and distal DNA regions, we will use machine-learning methods to model gene expression changes based on the TFs/HMs signals close to transcription start sites of targets, and/or enhancers having distal regulatory interactions to targets, which can be identified by ChIA-PET. The enhancer (or other RFs such as miRNAs) models for predicting gene expressions will be built from our methods.

*(2) Circuit Logic Formalism.* Different RFs work together to control gene expression, in a fashion like a logic circuit in many cases. Thus, my lab is also going to study the cooperativity between RFs using the formalism of circuit logic; i.e., to see whether gene expression has an activating or a repressive influence by combinatorial effects of RFs. We will integrate logic circuits from VLSI designs used in electronic chips into our analyses. Boolean models are also very efficient and scalable for analyzing large-scale genomic data. We plan to build a multiplex regulatory network composed of various RFs such as TFs, enhancers, miRNAs and lncRNAs along with their target genes, which captures not only the regulatory models identified above for each RF but also the cooperative connections between RFs.

*(3) DBNs.* To make our methods scalable to increasing amounts of new data and more genomic features, my lab plans to extend our predictive models based on Dynamic Bayesian network (DBN), which is a probabilistic graphical model that describes dependencies and conditional independencies for a set of random variables. It provides a coherent framework for predicting random-variable behavior through the integration of prior knowledge and new data. We want to incorporate the prior systematic regulatory networks from above, and to simultaneously identify regulatory pathways with local behavior (e.g., direct regulation) and global behavior (e.g., indirect regulation and/or conditional independence).

*(4) Impact of Variation.* Finally, my lab will use predicted regulatory networks to study how genomic variations affect regulatory mechanisms. For example, we will search the regulatory pathways associated with the genes for bio-fuel productions such as cell wall genes in plants, and predict their expression changes when their RFs are associated with variations/mutations. A list of top candidate regulatory pathways to increase bio-fuel productions can be provided for further experimental validations.