Thomas Gingeras and colleagues study where and how functional information is stored in genomes. These efforts help explain the biological and clinical effects of disease-causing gene mutations in humans and other organisms. Gingeras is a leader of the ENCODE (ENCyclopedia of DNA Elements) and the mouseENCODE and modENCODE (model genome ENCODE) projects of the National Institutes of Health. His research has altered our understanding of the traditional boundaries of genes, revealing that almost the entire lengths of genomes in organisms ranging from bacteria to humans can be transcribed into RNA (pervasive transcription) and that most RNA products made by a cell are not destined to be translated into proteins (noncoding, or ncRNAs). In fact, ncRNAs are proving to be involved in a variety of other important biological functions. Some have been shown to be critical components in the pre- and postranscriptional and translational processes, as scaffolds upon which large protein complexes are assembled and as extracellular signals. The initial studies that led to these observations have been extended to cover the entire human genome.

There has been a growing appreciation in recent years that gene function is frequently context-dependent, with a large part of that context provided by the activities of other genes. But trying to understand how genes interact to produce function is a hugely complicated problem and one that appears likely to become more so as genomic information becomes more detailed. Jesse Gillis and colleagues are computational biologists who are presently challenging an oft-taken approach to the problem in which the functions of genes are interpreted in the context of networks derived from gene association data. Such networks consist of millions of interactions across thousands of genes, derived from protein-binding assays, RNA coexpression analysis, and other sources. Historically, many attempts to understand gene function through networks have leveraged a biological principle known as “guilt by association.” It suggests that genes with related functions tend to share properties (e.g., physical interactions). In the past decade, this approach has been scaled up for application to large gene networks, becoming a favored way to grapple with the complex interdependencies of gene functions in the face of floods of genomics and proteomics data. Gillis’ work centers on identifying the limits of the approach and making fundamental improvements to its operation, as well as applying those improvements to neuropsychiatric gene network data.

Gholson Lyon’s lab focuses on analyzing human genetic variation and its role in severe neuropsychiatric disorders and rare diseases, including Tourette syndrome, attention-deficit hyperactivity disorder (ADHD), obsessive compulsive disorder (OCD), intellectual disability, autism, and schizophrenia. By recruiting large groups of related individuals living in the same geographic location (e.g., Utah), Lyon’s lab can study the breadth and depth of genetic variants in a similar environmental background. Using the exome—the parts of the genome that code for protein—and whole-genome sequencing, his lab looks for mutations that segregate with syndromes in the various populations. A second focus of the Lyon lab is to study the mechanistic basis of a new rare disease that they described in 2011. This is the first human disease involving a defect in the amino-terminal acetylation of proteins, a common modification of eukaryotic proteins carried out by amino-terminal acetyltransferases (NATs). The team has been using several different cellular model systems to better understand the disease pathophysiology and the basic process of amino-terminal acetylation. This year, Lyon collaborated with a team of researchers from other universities and companies to use precision medicine to successfully treat a patient with severe OCD. His symptoms were treated with deep brain stimulation, and the team used whole-genome sequencing to try to understand the molecular basis of his disease. The patient experienced significant relief from his symptoms and his quality of life returned, suggesting that similar methods may hold tremendous promise in the future.
The insights of W. Richard McCombie and colleagues have led to the introduction and optimization of novel methods of high-throughput genome sequencing. His team has made it possible to catalog variation among individual organisms in a way that would have been unthinkable 5 years ago. They have brought online a new generation of Illumina sequencers and optimized their function to a level at which eight to 10 trillion DNA bases can be sequenced in a month. McCombie’s team has been involved in international efforts culminating in genome sequences for maize, rice, and this year for bread wheat—three of the world’s most important food crops. They have also had an important role in projects to sequence the flowering plant Arabidopsis thaliana, the fission yeast Schizosaccharomyces pombe, as well as Homo sapiens and other important genomes. McCombie’s group is currently involved in several important projects to resequence genes in patient samples that are of special interest to human health, including DISC1 (a strong candidate gene for schizophrenia), looking for genetic variants implicated in bipolar illness and major recurrent depression. With Memorial Sloan-Kettering Cancer Center, they are using a method called hybrid selection, developed with Greg Hannon, to look at mutations in samples collected from patients with prostate cancer.

Using multidisciplinary approaches that combine computational analysis, modeling, and prediction with experimental verification, Doreen Ware’s lab seeks a deeper understanding of the evolution of genome sequences in plants and their implications for agricultural improvement. By looking comparatively across the genomes of plants in the same lineage, they seek answers to the following questions: How are genes conserved and lost over time? What are the fates of duplicated genes? What is the impact of structural variation on phenotypic variation? Ware’s team also studies gene regulation in plants, focusing on gene regulatory networks, targeting transcription factors and microRNA genes with the objective of understanding how these parts of the plant genome work together in determining spatial and temporal expression of genes. The lab had an important role in the project to produce a haplotype map reference genome of maize, spearheading the most comprehensive analysis of the crop yet. This has provided important information on the variation of the reference genome, as well as comparative data showing changes in the genome acquired through domestication and breeding. They have devoted special attention to examining diversity within maize, grape, and tomato, aiming to accelerate the development of strategies to introduce new germplasm that is needed to meet demands of increasing population and a changing environment. The lab also has brought fully sequenced genomes into an integrated data framework, to enhance the power of their comparative studies. This past year, Ware was named as its principal investigator for the National Science Foundation-funded Gramene project, a comparative genomics resource for agriculturally important crops and models to support sustainable food and fuel production. Ware, as principal investigator for plants, has also helped lead an effort funded by the Department of Energy to create—out of many separate streams of biological information—a single, integrated cyber-“knowledgebase” for plants and microbial life.
Our laboratory is currently involved in phase 3 of the Encyclopedia of DNA Elements (ENCODE3) project that began in July of 2012. This project endeavors to identify and characterize the functional elements present in the human genome sequence. Our contribution to this effort has been and continues to be the cataloging of the sites of long (>200 nucleotides) and short (<200 nucleotides) RNA transcription, the positions of 5′ and 3′ termini, as well as the splice sites present in human cells/tissues. In previous phases of the ENCODE project, the biological materials that served as a source for analyses were mostly cancer cell lines, whereas in this phase, the project has sought to analyze both human primary cell lines and tissues. This year, long RNAs were sequenced (RNA-Seq) from more than 30 primary cells, and a similar collection of human tissues was analyzed. A significant proportion of the year has been spent in attempting to centralize the analyses of the RNA-Seq data sets to the Data Coordination Center (DCC) established by the ENCODE project at Stanford University. To accomplish this, an evaluation of the available software tools that are needed in an analysis pipeline to map, quantify, and model data obtained from the RNA-Seq is currently being carried out by the DCC. Related to this evaluation effort was our participation this year in two independent evaluations carried out by the RGASP (RNA-Seq Genome Annotation Assessment Project) consortium of various RNA-Seq data-mapping software as well as transcript modeling/quantification tools. To assess the performance of current mapping software, we were invited as a developer of the STAR RNA-Seq mapping program to process four large human and mouse RNA-Seq data sets as part of the RGASP evaluation. A total of 26 mapping protocols based on 11 programs and pipelines were evaluated (Engström et al. 2013). STAR (spliced transcripts’ alignment to a reference; created in our laboratory by Dobin et al. 2013) was determined to be one of the best performing mapping tools tested.

Model Genome (mod)ENCODE: Human-Fly-Worm Comparative Analyses of Transcriptomes
C. Davis, A. Dobin, C. Zaleski

Both the ENCODE and modENCODE consortia have generated a community resource containing large amounts of RNA-Seq data from a wide variety of samples, resulting in a comprehensive annotation for the human, worm, and fly genomes. Extensive data integration reveals fundamental principles of transcription, conserved across highly divergent animals. In particular, by clustering expression profiles, we discovered conserved coexpression modules shared between the organisms, many of which are enriched in developmental genes. We used these to align the stages in worm and fly development, finding the expected embryo-to-embryo and larvae-to-larvae pairings in addition to a novel pairing between worm embryo and fly pupae. Furthermore, we found that the extent of per-base-pair noncanonical noncoding transcription is similar between the organisms. Finally, we found that the gene expression levels in the organisms, both coding and noncoding, can be consistently predicted by their upstream histone marks using a common “universal model.”

Mouse-ENCODE Project: Human–Mouse Comparative Analyses of Transcriptomes
C. Davis, A. Dobin, C. Xue, C. Zaleski

Mammalian genomes are pervasively transcribed, with much of the transcribed regions appearing to
lack the associated signatures of selective constraint often considered indicative of biological function. However, by analyzing the transcriptomes of a large and diverse panel of human cell lines and mouse tissues, we have observed that many of the transcriptional features associated with pervasive transcription, specifically intergenic, antisense, and chimeric transcription, are conserved between human and mouse even in genomic regions that lack signatures of evolutionary constraint. We also observed that the degree of conservation of the levels of expression and splicing is independent from the conservation of promoter and splicing sequences. These results reveal an evolutionary tendency to maintain transcriptional levels and organization that are not easily traceable to the conservation of primary sequence.

Database and Software Development
A. Dobin, C. Xue, C. Zaleski

Database Development
The infrastructure of our laboratory’s analysis programs is used for the continued development and maintenance of the large databases collected as part of the human–mouse and model genome (mod)ENCODE projects. The notable achievements in this area include the following:

- Continued management and organization of sequencing data consisting of fastq files (raw data) and genome-mapped data files (bam, bigwig, contig). During this year, we have added ~30 TB (terabytes) of sequencing data.
- Continued improvement and development of the RNA sequencing production pipeline. This consists of sequence mapping into multiple formats, generation of “contigs” and “splice junctions,” several new multiple QC steps, and file transfer of data files to outside locations (DCC and our Barcelona collaborators).
- Implementation of an ENCODE data submission client to interface with the primary repository at the University of California, Santa Cruz. A set of web-based forms connects to a server at CSHL, which translates the information to schema-compliant structure. The data are then submitted to a REST (representational state transfer) service and are viewable via the DCC’s public interface.

- Integrated our lab experimental processes with a new CSHL sequencing LIMS (laboratory information management system). Data tracking, processing of unique IDs, and management of completed sequencing runs were all migrated to the new system.
- Continued coordination with the CSHL IT department to manage our primary storage hardware migration. More than 40 TB of data will be migrated to a new hardware system and connected to the shared computing cluster as well as our own lab’s servers. Code migration, testing, and QC are ongoing processes.

Software Development
- STAR development and user support: STAR is the RNA-Seq aligner developed and published in 2013 by Dobin et al. This paper has been cited 40 times (Google Scholar) in the less than 1 year and the software has been downloaded more than 3000 times in this period. Efforts were concentrated on improving the algorithm, introducing new features, and fixing bugs, with 26 patches, and the paper was publicly released. Extensive user support is provided by e-mail and through the STAR user group.

- Circular RNA (circRNA): Interest in circRNAs has been re-ignited by recent publications demonstrating that circRNAs are present in various species and cell types, as well as their possible role as microRNA sponges. A. Dobin, a member of our lab, has developed an accurate and efficient method of circRNA discovery in high-throughput RNA-Seq data, based on the chimeric detection algorithm of our RNA-Seq aligner STAR. The method does not rely on annotations or intronic motifs and thus allows for unbiased “de novo” detection of circRNAs. Using this algorithm, we analyzed the circRNA repertoire in hundreds of human cell lines of the ENCODE transcriptome data set. Thousands of circRNAs are reproducibly detected in each of the samples. We demonstrate that circRNAs are significantly enriched in non-polyadenylated RNA populations compared to the polyadenylated RNAs, corroborating the circular nature of the molecules. Most of the detected circRNAs possess canonical intron motifs and connect protein-coding exons, implicating the standard splicing machinery in the biogenesis of the
circRNAs. Some but not all of the circular molecules appear to correspond to alternatively spliced exons and can be attributed to the "lariat model." By comparing the circRNA and linear RNA abundances across a variety of cell types, we show that many circular RNAs are strongly regulated.

De Novo DNA Demethylation and Noncoding Transcription Define Active Intergenic Regulatory Elements
F. Schlesinger

Deep sequencing of mammalian DNA methylomes has uncovered a previously unpredicted number of discrete hypomethylated regions in intergenic space (iHMRs). In collaboration with the Hannon laboratory at CSHL, we combined whole-genome bisulfite sequencing data with extensive gene expression and chromatin-state data to define functional classes of iHMRs and to reconstruct the dynamics of their establishment in a developmental setting. Comparing HMR profiles in embryonic stem and primary blood cells, we show that iHMRs mark an exclusive subset of active DNase-hypersensitive sites (DHS) and that both developmentally constitutive and cell-type-specific iHMRs display chromatin states typical of distinct regulatory elements. We also observe that iHMR changes are more predictive of nearby gene activity than the promoter HMR itself and that expression of noncoding RNAs within the iHMR accompanies full activation and complete demethylation of mature B-cell enhancers. Conserved sequence features corresponding to iHMR transcript start sites, including a discernible TATA motif, suggest a conserved, functional role for transcription in these regions. Similarly, we explored both primate-specific and human population variation at iHMRs, finding that although enhancer iHMRs are more variable in sequence and methylation status than any other functional class, conservation of the TATA box is highly predictive of iHMR transcript start sites, including a discernible TATA motif, suggesting a conserved, functional role for transcription in these regions. Overall, our analysis allowed us to construct a three-step timeline in which (1) intergenic DHS are preestablished in the stem cell, (2) partial demethylation of blood-specific intergenic DHS occurs in blood progenitors, and (3) complete iHMR formation and transcription coincide with enhancer activation in lymphoid-specified cells (Schlesinger et al. 2013).

High-Fidelity Promoter Profiling Reveals Widespread Alternative Promoter Usage and Transposon-Driven Developmental Gene Expression
P. Batut, A. Dobin

Many eukaryotic genes possess multiple alternative promoters with distinct expression specificities. Therefore, comprehensively annotating promoters and deciphering their individual regulatory dynamics are critical for gene expression profiling applications and for our understanding of regulatory complexity. We introduce RAMPAGE (RNA annotation and mapping of promoters for the analysis of gene expression), a novel promoter activity profiling approach that combines extremely specific 5′-complete cDNA sequencing with an integrated data analysis workflow, to address the limitations of current techniques. RAMPAGE features a streamlined protocol for fast and easy generation of highly multiplexed sequencing libraries, offers very high transcription start site specificity, generates accurate and reproducible promoter expression measurements, and yields extensive transcript connectivity information through paired-end cDNA sequencing. We used RAMPAGE in a genome-wide study of promoter activity throughout 36 stages of the life cycle of Drosophila melanogaster and describe here a comprehensive data set that represents the first available developmental time course of promoter usage. We found that >40% of developmentally expressed genes have at least two promoters and that alternative promoters generally implement distinct regulatory programs. Transposable elements, long proposed to have a central role in the evolution of their host genomes through their ability to regulate gene expression, contribute at least 1300 promoters shaping the developmental transcriptome of D. melanogaster. Hundreds of these promoters drive the expression of annotated genes, and transposons often impart their own expression specificity upon the genes they regulate. These observations provide support for the theory that transposons may drive regulatory innovation through the distribution of stereotyped cis-regulatory modules throughout their host genomes (Batut et al. 2013).
Currently, two distinct populations of extracellular vesicles (EVs)—exosomes (<100 nm) and microvesicles (100–1000 nm) have been classified. These two types of EVs are believed to possess key differences in their mode of release from host cells. However, the qualitative and quantitative differences in the cargo RNAs between these two subpopulations of EVs have been relatively unexplored.

**EV RNAs as Possible Cancer Biomarkers**

The repertoire of EVs released by four cancer cell lines (K562, HELA, MCF7, and U2OS) and two primary cell lines (BJ and HUVEC) were studied. These primary cell lines are frequently the cell types most likely to be present in the tumor microenvironment. Using previously published protocols, we isolated separately the two families of EVs. Using nanoparticle tracking analysis, we quantified and analyzed the size distribution of the two families; we found that the exosomes had a size distribution ranging from 10 nm to 300 nm, with a median of 135 nm, and were released at an average across cell lines from 10 to 100 exosomes per cell. We found that microvesicles had a size distribution ranging from 50 nm to 400 nm with a median of 225 nm, but that they were less numerous than exosomes. A very large degree of overlap in size between these two subgroups, as well as all present techniques being unable to clearly separate out the two populations solely defined by their difference in size, led us to ask the question of whether the two subgroups were just an artificial construct or whether they were really two physiologically and biochemically distinct vesicles being released by cells. To answer this question, we used RNA-Seq to investigate the transcriptomes of both the EV subgroups. Unexpectedly, there was a very low correlation between the transcript types composing the contents of the two EV subgroups. This result is consistent with the likelihood that there are two distinct populations of EVs that are not merely different in their size distribution but are also distinct in their RNA content. A comparison of the RNA transcripts of exosomes to source cells in both cancer and primary-cell-derived exosomes showed that there were mainly two families of RNAs that were enriched in the exosomes when compared to source cell, microRNA (miRNA), and Y-RNA. Of the Y-RNA family, RNY5 was found to be the predominant molecule. Other interesting molecules that were seen to be enriched in exosomes were miRNA-17-92a, miRNA-93, and miRNA-103b. Comparison of exosomal and microvesicle RNA contents derived from different cell types revealed that the contents were cell-type specific, increasing the likelihood of identifying specific disease cell-derived exosomes within human body fluids.

**Intercellular Signaling**

The intercellular transfer of exosomes and their RNA contents have been investigated by light and electron microscopy, and their contents were studied using RNA-Seq. To demonstrate the release of exosomes in the extracellular environment, K562 cells were labeled with membrane lipid dye PKH26. Live cell imaging was performed and demonstrated the release of labeled vesicles from the cells to the extracellular environment followed by the uptake of the exosomes by different recipient cells using a transwell cell-culture system. The labeled exosomes have also been directly incubated with multiple human and mouse recipient cell types (K562, BJ, 3T3) that were labeled with lipid dye PKH67 and Hoechst 33342. In both scenarios, live imaging on the recipient cells clearly revealed the uptake of labeled exosomes by the recipient cells. To investigate the subcellular localization of the exosomes upon uptake in the recipient cells, multiple subcellular organelles in the recipient cells were labeled before being incubated with prelabeled exosomes, and live imaging was performed. Live imaging revealed that (1) the exosomes are primarily localized in the cytoplasm in the recipient cells upon uptake and do not colocalize with any particular organelle, (2) the exosomes do not cross the nuclear membrane in the recipient cells, but occupy a characteristic perinuclear localization, and (3) the exosomes, upon uptake, are not fused with lysosomes or P bodies. To demonstrate the uptake of exosomal RNA cargo in recipient cells, metabolic labeling of RNA with ethynyl-uridine (EU) was performed. Briefly, donor K562 cells were incubated with EU for 24 h, and exosomes were isolated from the conditioned medium. The isolated exosomes were then added to recipient cells (NIH-3T3) and incubated for 2 h. The exosomal RNA was then detected using click chemistry in the recipient cells and
imaged by fluorescent microscopy. Actinomycin D was used to block the endogenous transcription from any EU carry over during the exosome preparation.

The transfer of RNA through exosomes among cells was also investigated by another orthogonal technique, namely, RNA-Seq. One member of the Y-RNA small noncoding gene RNY5 is expressed abundantly in humans as well as in many lower evolutionary organisms, but it is genetically missing in rodents. Human RNY5 RNA transcripts are particularly abundant in exosomes secreted by K562 cells and thus can serve as excellent genetic markers for intercellular (and interspecies) RNA transfer through exosomes. K562 exosomes were isolated and incubated with mouse HB-4 cells for various durations (0, 6, 12, 24, and 48 h). Small RNA was isolated from recipient mouse cells, and the presence of RNY5 transcripts in mouse HB-4 cells was detected using RNA-Seq. Interestingly, the time-course experiment revealed the stability of RNY5 in mouse cells. Maximum expression of RNY5 was detected at a 12-h incubation period in the recipient cells.

One of the most well-known and studied functions of tumor-derived exosomes is its role in inducing apoptosis upon transfer in the recipient immune cells. The ability of tumor exosomes to induce apoptosis in immune cells is generally regarded as a mechanism of immune evasion for the tumor cells. The mechanism through which apoptosis is induced is thought to be protein mediated (Fas-Ligand pathway). In this project, we decided to investigate the role of RNA cargo of tumor exosomes upon transfer in two other important primary cells of the tumor microenvironment, namely, primary fibroblast (BJ) and endothelial (HUVEC) cells.

K562 and BJ cells were incubated with K562 exosomes for 24 h, and live-dead cell counting was performed using Hoechst 33342 (live) and Yo-pro1 (dead) dye. Surprisingly, we discovered K562 exosomes also induce apoptosis in primary fibroblast cells (BJ) but not in the tumor cells (K562) themselves. Thus, the ability to induce apoptosis by tumor exosomes is not restricted to immune cell lineage alone. To investigate whether the observed phenotype is caused by exosomal RNA or exosomal proteins/lipids, RNA was isolated from K562 exosomes and lipofected into BJ and K562 cells using Lipofectamine 2000. Consistent with previous experiment, apoptosis was detected again in BJ cells alone but not in K562 cells, thus suggesting that apoptosis is caused by exosomal RNA and not exosomal proteins/lipids. Scrambled RNA and mock lipofectamine treatment were performed as negative controls—neither of which caused apoptosis in any of the cells.

One of the most abundant transcripts present in K562 exosomes is a 32-mer 5′ fragment of a small noncoding RNA, RNY5. RNY5 is a Polymerase-III-transcribed gene of the Y-RNA family (Y1, Y3, Y4, and Y5) and is thought to have a role in DNA replication and RNA degradation (when bound with protein Ro). To investigate whether RNY5 transcripts in exosomal cargo are responsible for causing apoptosis, K562 and BJ cells were transfected with a synthetic RNY5 fragment (32-mer) with Lipofectamine 2000. Consistent with previous experiments with K562 exosomes and exosomal RNA, transfection of the synthetic RNY5 fragment alone resulted in apoptosis in BJ cells but not in K562 cells. Mock lipofectamine treatment and nonspecific scrambled sequences (Allstate negative control RNA) were used as negative controls; neither of which caused cell death in any cells.

Full-length RNY5 has a characteristic secondary structure of two double-stranded RNA stem structures called lower and upper stem, separated by an inner loop that is conserved among all Y-RNA genes. Exosomes carry a 5′ 32-nucleotide fragment of RNY5 only, and thus, only the 5′ side of the lower and upper stem and inner loop is present in exosomes. The double-stranded upper stem in full-length RNY5 is considered to be the motif responsible for its role in DNA replication. Because exosomes contain only the 5′ side of the stem structures, the 5′ side of the upper stem (8-mer sequence GUUGUGGG) from nucleotides 14–21 was investigated for its role in apoptosis. K562 and BJ cells were transfected with a normal synthetic y5 fragment (32-mer, synthetic Y5 fragment with 8-mer deleted; Hy5-δ-us) and 8-mer-scrambled (Hy5-us-scram). Neither hy5-δ-us nor hy5-us-scram was able to induce cell death in either of the cells any more, thus suggesting that the 8-mer sequence is the motif in RNY5 responsible for causing apoptosis in primary cells.

**PUBLICATIONS**


usage and transposon-driven developmental gene expression. 


GENE NETWORK ANALYSIS

J. Gillis  S. Ballouz  W. Verleyen

Research in my lab centers on using computational methods to understand gene function. Computational biology has taken up the challenge of determining gene function mainly by attempting to interpret the activities of genes in the context of networks derived from gene association data. As data sets characterizing genes grow in size and complexity, it seems self-evident that computation can assist in inference as to gene function. Gene network analysis intended to provide insight into complex disorders is a dominant interest in the field. Gene associations (of various sorts) are believed to encode functional interaction, and this interaction is frequently shown to be able to substantially predict gene function across all functions. This approach, commonly called “guilt by association” (GBA), is embedded in everything from prioritization of de novo variants to uncovering novel molecular phenotypes or mechanisms of disease. Our research focuses on identifying limitations in the GBA approach and making fundamental improvements to its operation for the interpretation of neuropsychiatric genomics data. Postdocs Sara Ballouz and Wim Verleyen joined my lab in early 2013.

Analysis of RNA-Seq Coexpression Networks

RNA-Seq offers profound biological and technical advantages over microarray technologies, such as detecting the whole transcriptome and an improved dynamic range. However, as was once the case for microarrays, RNA-Seq’s utility is hobbled by our inability to determine sound consensus standards. One major reason is that RNA-Seq data have statistical properties quite different from those of microarray data, and thus the same principles do not apply for differential expression. Important questions about data quality, such as the role of biological noise in experimental design, are not readily addressed. A broader characterization of RNA-Seq data through meta-analysis offers one avenue to investigate problems of this nature.

Although differential expression analysis is a more common means for interpreting transcriptomic data, coexpression analysis is far more routine in the context of meta-analysis, with thousands of expression profiles aggregated to generate robust signatures using repurposed data. One advantage of coexpression methods is that they already tend to require meta-analysis of disparate data sets with quite different properties, so approaches for aggregating across data are already common. We have developed a series of methods appropriate for RNA-Seq meta-analysis and have produced reference networks for general use. RNA-Seq coexpression poses novel statistical and bioinformatics challenges. We have identified major confounds and developed appropriate control experiments necessary for network construction, laying the groundwork for functional analyses into RNA-Seq coexpression; we are particularly focusing on functional inference in noncoding RNA.

Bias Tradeoffs in the Creation and Analysis of Protein–Protein Interaction Networks

Networks constructed from aggregated protein–protein interaction data are commonplace in biology. But the studies these data are derived from were conducted with their own hypotheses and foci. Focusing on data from budding yeast present in BioGRID, we determine that many of the downstream signals present in network data are significantly impacted by biases in the original data. We determine the degree to which selection bias in favor of biologically interesting bait proteins goes down with study size, and we also find that promiscuity in prey contributes more substantially in larger studies. We analyze interaction studies over time with respect to data in gene ontology and find that reproducibly observed interactions are less likely to favor multifunctional proteins. We find that strong alignment between coexpression and protein–protein interaction data occurs only for extreme coexpression values and use these data to suggest candidates for targets likely to reveal novel biology in follow-up studies.
Assessing Identity, Redundancy, and Confounds in Gene Ontology Annotations Over Time

Gene ontology (GO) is a key means by which systems biologists operationalize gene function, making it a heavily relied upon tool in innumerable analyses and data interpretation exercises. GO annotations are often used as a gold standard, but it has widely appreciated imperfections. Ironically, it is very difficult to assess the properties of GO itself because there is no other comprehensive gold standard against which to hold it. Broadly speaking, assessment of GO has focused on three distinct attributes: the accuracy of annotations assigned to GO, GO’s structure independent of annotation, and the utility of GO and its annotations for the interpretation of data. Despite misgivings about the incompleteness of GO annotations, the use of GO “sets” as representing “functions” is now endemic. This is put to use in numerous applications such as “gene group enrichment,” gene network analysis, and gene function prediction. It is essential to understand the extent to which such applications are valid.

The statement “the differentially expressed genes were enriched for genes with functions in cell growth” does not necessarily mean the same thing today as it did 5 years ago, because the definition of “cell growth genes” has changed in GO. Valid experimental results often become obsolete over time, but the reported facts of the experiment should not. But that is what happens when the gene ontology changes. This is of course to be expected, and the problem can be ameliorated by reporting which version of the gene annotations was used. We find that genes can alter their “functional identity” over time, with 20% of genes not matching to themselves (by semantic similarity) after 2 years. We consider ensuring independence of GO from the data sets to which it is being applied as an absolute minimum standard, and our results show that at least some protein interaction data do not meet this standard. We discovered that many entries in protein interaction databases are due to the same published reports that are used for GO annotations, with 66% of assessed GO groups exhibiting this confound. In our experience, among systems biologists, there seems to be a broadly appreciated disjunction between the true utility of GO and how often it is used, even if this is rarely acknowledged in the peer-reviewed literature. The use of GO annotations is often regarded as a minimally interesting validation of results, but not safe to use for discovery purposes. We believe the problems we have identified are among the underlying sources of these mixed feelings about GO. If it is “too easy” to obtain interesting results using GO, and those results do not consistently hold up, then GO’s utility for such purposes is limited.

Meta-Analysis across Computational Approaches Predicting Gene Function

The use of our burgeoning genomics data to characterize gene function has been one of the central research objectives of computational biology in the postgenomics era. Despite this intense focus, the abundance of public data, and methods development in machine learning and “big data” analysis, progress has been surprisingly uncertain. Attaching novel functions to genes based on prior data remains difficult, and a large number of genes still have comparatively little information attached to them.

To better understand the forces driving algorithmic and data performance, developers have focused on comparative assessment. Despite some important progress enabled by these assessments, the underlying mechanisms driving methods performance are poorly understood. Each developer may use separate data resources, different algorithms, and quite specific means for parsing the data in combination, making it quite hard to explore factors affecting performance. Field-wide progress is hampered by our inability to know why a specific algorithm worked on a particular data set for a potentially narrow task.

To overcome this lack of detail in assessment, we implemented representative samples of cutting-edge “default” machine learning algorithms and obtained performances across our set comparable to those of previous critical assessments at identical tasks with well-characterized data and fully motivated parameter choices. Because we now have in-house versions of multiple algorithms, we can explore variance and data dependencies in a way not typically possible.

We focused on data based on networks derived from protein–protein interaction, sequence similarity, aggregated coexpression, and semantic similarity to study underlying patterns of performance. Our infrastructure allows us to characterize in detail why aggregation improves performance, where results are robust
and reproducible, and what artifacts are potentially problematic in data interpretation. We plan to make this replication across methods available as a public resource, allowing geneticists using function prediction methods to better assess reproducibility and data dependencies in target genes of interest.

Assessing the First Computational Gene Function Prediction Assessment

The computational assignment of gene function remains a difficult but important task in bioinformatics. The establishment of the first critical assessment of functional annotation (CAFA) was aimed at increasing progress in the field. We demonstrated an independent assessment of the results of CAFA, aimed at identifying challenges in assessment and at understanding trends in prediction performance. We found that well-accepted methods based on sequence similarity (i.e., BLAST) have a dominant effect. Many of the most informative predictions turned out to be either recovering existing knowledge about sequence similarity or were “postdictions” already documented in the literature. These results indicate that deep challenges remain in even defining the task of function assignment, with a particular difficulty posed by the problem of defining function in a way that is not dependent on either flawed gold standards or the input data itself. In particular, we suggest that using the gene ontology (or other similar systematizations of function) as a gold standard is unlikely to be the way forward.

Schizophrenia Coexpression to Prioritize Candidate Variants

RNA coexpression data are commonly used to construct gene networks, but it is often considered to be more difficult to interpret than protein interactions. This is in part due to the lack of consensus on methods for constructing networks from expression profiles and the relatively poor performance of coexpression for function prediction, as measured by its ability to recapitulate data with GO, KEGG, or other databases. On the other hand, coexpression affords a major advantage over current large-scale protein interaction databases: It can be used to create “condition-specific” networks. In our preliminary results, we show that by appropriate consideration of data pretreatment, aggregation, and network construction, coexpression networks become a powerful tool for gene function analysis, on par with or better than protein interaction networks in terms of overall properties, while providing condition specificity. We leverage these properties to examine the role of de novo schizophrenia variants in the most comprehensive analysis to date of coexpression patterns in schizophrenia, combining seven studies of prefrontal cortex in affected individuals and unaffected controls.

PUBLICATIONS


In Press

Our laboratory focuses on analyzing human genetic variation and understanding how genetic mutations contribute to severe idiopathic neuropsychiatric disorders. We do this by studying large pedigrees living in the same geographic location, where one can study the expressivity and segregation of variants in a similar environmental background and with fewer population stratification concerns. Toward this end, we collect pedigrees in Utah and elsewhere and then use exome and whole-genome sequencing to find mutations that segregate with syndromes in the pedigrees. We focus on the discovery of families with rare diseases and/or increased prevalence for syndromes such as intellectual disability, autism, and schizophrenia. Once we identify mutations that likely contribute to a disease, we undertake detailed functional studies of these mutations and the biological processes affected. Several projects are still at an early stage, but some of the projects that are sufficiently far along to discuss publicly are listed below.

**Ogden Syndrome and the Amino-Terminal Acetylation of Proteins**

M. Doerfel, Y. Wu [in collaboration with R. Marmorstein, Philadelphia, Pennsylvania; T. Arnesen and N. Reuter, Norway; P. van Damme, Belgium]

We have previously identified a lethal X-linked disorder of infancy comprising a distinct combination of particular craniofacial features producing an aged appearance, growth failure, hypotonia, global developmental delays, cryptorchidism, and acquired cardiac arrhythmias. The first family was identified in Ogden, Utah, with five affected boys in two generations of family members. A mutation was identified as a c.109T>C (p.Ser37Pro) variant in *NAA10*, a gene encoding the catalytic subunit of the major human amino-terminal acetyltransferase (NatA). This same mutation was identified in a second unrelated family, with three affected boys in two generations. This X-linked malformation and infantile lethality syndrome has been named Ogden syndrome, in honor of the hometown where the first family resides. This is the first human disease involving a defect in the amino-terminal acetylation of proteins, a common (yet vastly understudied) modification of eukaryotic proteins carried out by amino-terminal acetyltransferases (NATs). There is significantly impaired biochemical activity of the mutant hNaa10p, suggesting that a reduction in acetylation of some unidentified proteins by hNaa10p might lead to this disease. There is currently very limited knowledge of the functional importance of Nt acetylation at the protein level and at the organismal level.

To understand the detrimental impact of the Naa10 S37P mutation, we performed structural, molecular, and cellular investigations. The recently determined *Schizosaccharomyces pombe* NatA complex crystal structure was used as a template to generate a model of human NatA, revealing a highly conserved complex. The model allowed for comparison of Naa10 wild type and Naa10 S37P within the NatA complex and suggested a decreased flexibility for Naa10 S37P in regions involved in catalysis and at the interface with the auxiliary subunit Naa15. The hydrogen bonding network between Naa10 and Naa15 was also rearranged. In vitro enzyme kinetics of Naa10 S37P demonstrated a reduced catalytic capacity, probably due to impaired peptide substrate binding. In agreement with the structural model, Naa10 S37P displayed a reduced capacity to form a stable NatA complex. Amino-terminal acetylome analyses of patient B cells and fibroblasts provided a survey of Nt acetylation in human noncancer cells. In line with previous NatA knockdown data, patient-derived S37P B cells and fibroblasts have reduced Nt acetylation for a subset of NatA-type substrates compared to cells from healthy family members, demonstrating in vivo perturbation of Naa10 (NatA)-mediated Nt acetylation in Ogden syndrome males. THOC7, one of the affected proteins, was shown to depend on Nt acetylation.
for its stability. Ogden syndrome fibroblasts further displayed abnormal cell proliferation and migration capacity. Therefore, the Ogden syndrome mutant Naai10 is impaired in NatA complex formation and catalytic capacity, and patient cells display reduced in vivo Nε acetylation and cellular phenotypes potentially linked to the defects observed in the males suffering from this disease.

**Low Concordance of Multiple Variant-Calling Pipelines: Practical Implications for Exome and Genome Sequencing**

J. O’Rawe [in collaboration with K. Wang, Los Angeles, California]

To facilitate the clinical implementation of genomic medicine by next-generation sequencing, it will be critically important to obtain accurate and consistent variant calls on personal genomes. Multiple software tools for variant calling are available, but it is unclear how comparable these tools are or what their relative merits in real-world scenarios might be. We sequenced 15 exomes from four families using commercial kits (Illumina HiSeq 2000 platform and Agilent SureSelect version 2 capture kit), with ~120× mean coverage. We analyzed the raw data using near-default parameters with five different alignment and variant-calling pipelines (SOAP, BWA-GATK, BWA-SNVVer, GNUMAP, and BWA-SAMtools). We additionally sequenced a single whole genome using commercial kits (Illumina HiSeq 2000 platform and Agilent SureSelect version 2 capture kit), with ~120× mean coverage. SNV concordance between five Illumina pipelines across all 15 exomes was 57.4%, whereas 0.5%–5.1% of variants were called as unique to each pipeline. Indel concordance was only 26.8% between three indel-calling pipelines, even after left-normalizing and intervalizing genomic coordinates by 20 base pairs. Of CG variants that fall within targeted regions in exome sequencing, 11% were not called by any of the Illumina-based exome analysis pipelines, but only 54.0%, 44.6%, and 78.1% of the GATK-only, SOAP-only, and shared SNVs could be validated. Additionally, our analysis of two families (one with four individuals and the other family with seven), demonstrated additional accuracy gained in variant discovery by having access to genetic data from a multigenerational family. Our results suggest that more caution should be exercised in genomic medicine settings when analyzing individual genomes, including interpreting positive and negative findings with scrutiny, especially for indels. We advocate for renewed collection and sequencing of multigenerational families to increase the overall accuracy of whole genomes.

**Advancing Precision Medicine through Clinical Grade Whole-Genome Sequencing, Return of Results, and Neuromodulation**


For widespread precision medicine to become a reality, many things must be optimized, including clinical-grade sample collection, high-quality sequencing data acquisition, digitalized phenotyping, rigorous generation of variant calls, and online sharing of medical history and genomic data with research participants and others. We report the detailed phenotypic characterization, clinical-grade whole-genome sequencing (WGS), as well as the 2-year outcome of a man with severe obsessive compulsive disorder (OCD) treated with deep brain stimulation (DBS) of the nucleus accumbens/anterior limb of the internal capsule. As part of his integrated medical care, his genome was sequenced, and variants were detected in the Illumina WGS Clinical Laboratory Improvement Amendments (CLIA)-certified laboratory. It is increasingly apparent that mental illness results from a constellation of genetic and environmental factors. Consistent with this, WGS did not reveal any one mutation of large effect, but instead showed that he carries several alleles that have been shown to elevate risk for mental illness. This includes the p.Val66Met variant in brain-derived neurotrophic factor (BDNF), the p.Glu429Ala allele in methylenetetrahydrofolate reductase (MTHFR), and the p.Asp7Asn allele in choline O-acetyltransferase
Genomics 211

We identified thousands of other variants in his genome, including pharmacogenetic variants, and one mutation led to the discovery that he has untreated bilateral cataracts and other visual disturbances. We archived all data in the GVFClin format and also returned many results to this person. Since implantation of the deep-brain stimulator, this man has reported steady improvement, highlighted by a decline in his Yale-Brown obsessive compulsive scale (YBOCS) score from ~38 to a score of ~25. A rechargeable Activa RC neurostimulator battery has been of major benefit in terms of facilitating a degree of stability and control over the stimulation. His psychiatric symptoms reliably become worse within hours of the battery becoming depleted, thus providing confirmatory evidence for the efficacy of DBS for OCD in this person. To our knowledge, this was the first study in the clinical neurosciences that integrates detailed neuropsychiatric phenotyping, deep-brain stimulation for OCD, and clinical-grade WGS, with management and the first return of WGS results to a person with severe mental illness.

PUBLICATIONS


In Press


DEVELOPMENT OF NEXT-GENERATION SEQUENCING TECHNOLOGY

In 2013, we completed important phases of several projects and began new initiatives. A paper done in collaboration with colleagues at the University of Edinburgh on the role of the Disrupted-In-Schizophrenia (DISC1) gene and psychiatric disorders was published in 2013. We continued and expanded that project to study the role of DISC1 by continuing the analysis of ~260 genes that interact with DISC1 in more than 1500 individuals. We further expanded this project under new funding from the National Institute of Mental Health (NIMH) by beginning whole-genome sequence analysis on 39 members of the translocation family.

In collaboration with colleagues at Trinity University, Dublin, we refined our analysis of the data obtained the previous year from sequencing parent-child trios with a child afflicted by schizophrenia. The results showed a link to autism and implicated a class of genes involved in chromatin remodeling. A paper describing this work was submitted in December 2013.

In collaboration with the University of Iowa and Johns Hopkins University, we also completed the sequencing for our ongoing synaptome project. Approximately twice as many exomes as we were funded for were completed. We published a paper on a new analysis method as part of this project and will complete the analyses of these data in the coming year.

In 2013, we initiated the remaining step that we believe to be necessary for a comprehensive, overall plan to understand the genetics of psychiatric disorders by implementing high-throughput functional studies in Caenorhabditis elegans in collaboration with the Hammell lab at CSHL. We also moved ahead our targeting technology to better enable us to sequence small regions of the genome in very large numbers of people. Thus, we now have in place initial genetic studies on exomes or whole genomes to broadly identify possible candidate genes, as well as dual follow-up approaches of ultra-high-throughput-targeted resequencing of these candidates in much larger sample sizes and high-throughput functional tests to further characterize candidates. Although this “pipeline” needs further refinement, with it in place, we look forward to applying it to our many candidates in 2014.

In the area of cancer genetics, we began collaboration on colon cancer with colleagues at Stony Brook University. This initially focused on RNA sequencing and methylation analysis.

Finally, we made considerable progress in our plant genomics and de novo assembly research. This is being driven mostly by adapting new methods to take advantage of the increasing read lengths on the Pacific Biosciences instrument. In collaboration with the Schatz lab at CSHL, we have been working on both pure Pacific Biosciences assemblies and hybrid assemblies with Illumina data included.

Investigating the Biology of DISC1

S. Teng, S.E. McCarthy, M. Kramer, W.R. McCombie [in collaboration with D. Blackwood, D. Porteous, P. Thompson, and I. Deary, University of Edinburgh; A. McRae, University of Queensland]

The DISC1 gene has been implicated in psychiatric disorders such as major depression (UP), bipolar disorder (BD), and schizophrenia (SZ). This gene was originally identified in a Scottish family with a high burden of mental disorders (St Clair et al., Lancet 336: 12–16 [1990]). We have undertaken several projects to study the genetics of DISC1. We previously described targeted sequencing of the 528-kb DISC1 gene region in more than 1500 individuals (221 BP, 240 SZ, 192 UP, and 889 controls). We found a nominal association of noncoding single-nucleotide polymorphisms (SNPs) (rs16856199) with major depression, as well as several extremely rare variants that seem to be present only (or almost only) in affected patients. We believe that the DISC1 locus represents a microcosm of the complexity that has made a better understanding of the genetics of
major psychiatric disorders so difficult to unravel. There are multiple types of variants at this one locus. Some of these are common variants that provide a slight increase in risk for a disorder (depression, in this case). Other variants in the same gene are extraordinarily rare in the population, but they appear to confer an extremely high risk for a disorder, probably approaching 100%. In addition, there are probably interactions among multiple variants in the DISC1 gene and among variants in DISC1 and other genes. This genetic complexity is then amplified across hundreds or even thousands of genes contributing to the disorders. In 2013, we published this work in *Molecular Psychiatry* (Thomson et al., *Mol Psych* doi 10.1038/mp.2013.68 [2013]).

To better understand the potential interaction between DISC1 and other genes, we have been detecting variants in genes thought to interact with DISC1. Last year, we reported on the targeted resequencing of the DISC1 interactome, a network of genes that converge on pathways critical for neuronal signaling and genes important in the treatment of SZ, BD, and UP. Using a custom Nimblegen solution capture of a ~13-Mb target, we resequenced in 1543 samples (using the same sample set as those sequenced in the study above) the exons, promoters, and conserved regions of 264 genes that are known to directly interact with DISC1. In 2013, we performed data quality control analysis to reduce the numbers of false associations in the case-control study. Samples with high missing genotypes, outlying heterozygosity rate, or discordant sex information were removed and individuals of divergent ancestry were identified. We filtered the single-nucleotide variants (SNVs) with excessive missing genotypes or a significant deviation from Hardy–Weinberg equilibrium. The final data set included 1453 samples and 284,699 SNVs. Among the SNVs with high quality, 78% have a minor allele frequency of less than 1%, and 67% of the SNVs have not been previously reported in the 1000 Genomes Project European subset. To detect the statistical associations between the single SNV and psychiatric disorders, we performed a Fisher’s exact test across all SNVs on combined diagnoses: SZ, BP, and UP, respectively. To assess the associations at the genic level, we performed two gene-based association methods including the basic burden test and the sequence kernel association test. In 2014, we aim to expand the analysis to analyze epistatic interactions between genes and to further investigate and prioritize variants and genes for functional follow-up.

We also began a project to analyze the complete genomes of 39 members of the DISC1 family in which the original translocation was found in order to search for independently segregating variants that may explain why many of the translocation carriers develop SZ or recurrent UP but others have only minor or no psychiatric diagnoses. We will then be able to compare these potential variants with the variants observed in the DISC1 interactome. On completion of whole-genome sequencing, we were able to produce high-quality libraries with >33× mean coverage and >93% of bases covered at 20× or higher depth for each sample. The data were processed through our standard analysis pipeline that uses Burrows–Wheeler aligner (BWA) mapping (Li et al., *Bioinformatics* 25:1754 [2009]), SAMtools (Li et al., *Bioinformatics* 25:2078 [2009]), BamTools (Barnett et al., *Bioinformatics* 27:1691 [2001]), and Picard for alignment processing, and QC and GATK (DePristo et al., *Nat Genet* 43:491 [2011]) for variant calling. Recently, the variants were recalled using GATK’s multisample calling feature in order to obtain more accurate genotyping across samples. In 2013, we received a grant from the NIMH to continue the study of this important locus.

In 2014, we plan to further investigate the variants and to use analysis methods and software such as the GEMINI software suite and Ingenuity Variant Analysis (www.ingenuity.com) to classify and prioritize variants (Paila et al., *PLoS Comput Biol* 9:e1003153 [2013]). We will expand the analysis from SNPs and small indels to include copy number variants (CNVs) and structural variants. In addition, we plan to perform methylation analysis to scan for differentially methylated regions and to follow up on our findings in a larger Scottish and NIMH case control set.

**Exome-Based Sequencing of Parent–Offspring Trios with Schizophrenia**


In collaboration with Trinity College Dublin, our data analysis of de novo mutations in 57 parent–offspring trios with schizophrenia (42 trios with no family history
of psychosis [sporadic] and 15 trios with a family history of psychosis [familial] was finalized and submitted for publication at Molecular Psychiatry in December 2013. The fine-tuning of this work improved the significance of our findings summarized below.

In addition to the analysis we reported in 2012, we found in our sporadic trios that de novo mutations likely to be disrupting or damaging to protein function in genes were less tolerant to new or rare mutations ($P = 2 \times 10^{-5}$) (Petrovski et al., PLoS Genet 9: e1003709 [2013]). We also refined our disease ontology analysis to improve the specificity of the gene–phenotype relationships. Following this revised analysis using 72 fine-tuned NeuroCarta (Portales-Casamar et al., BMC Genomics 129: doi 10.1186/1471-2164-14-129 [2013]) phenotypes (defined by 10 or more quality-filtered associated genes), the overlap between genes with de novo mutations (DNMs) and genes implicated in autism and intellectual disability remained. The overlap with autism was the greatest of DNMs in genes such as CHD8, MECP2, AUTS2, and MLL2, which supports previous genetic and epidemiological studies that suggest a shared genetic component between schizophrenia and autism.

We also improved our analysis of DNMs in potential chromatin modifier genes. First, we found a significant enrichment of DNMs in genes coding for chromatin modifiers that have been implicated in the etiology of mental illnesses. Second, we expanded the analysis to epigenetic modifying genes based on protein domain specificity and observed a significant enrichment overall of DNMs in these genes.

An important addition to this analysis was determining the consistency of our findings using data from nine exome-sequencing studies of trios with neurodevelopmental disorders, including schizophrenia ($n = 3$) and autism ($n = 4$), ID ($n = 2$), as well as six unaffected cohorts. With our data set included, an increase in nonsense mutations was observed in 40% of the neurodevelopmental data sets compared to just one control data set (de Ligt et al., J Med 367: 1921 [2012]; Gulsuner et al., Cell 154: 1237 [2013]; Iossifov et al., Neuron 74: 285 [2012]; Neale et al., Nature 485: 242 [2012]; O’Roak et al., Nature 485: 246 [2012]; Rauch et al., Lancet 380: 1674 [2012]; Sanders et al., Nature 485: 237 [2012]; Xu et al., Nat Genet 44: 1365 [2012]). More than 66% of the neurodevelopmental data sets had functional de novo mutations affecting protein function in haploinsufficient genes or genes intolerant to novel mutations, in contrast to two control cohorts. No unaffected cohort had an enrichment in genes implicated in autism spectrum disorder (ASD) or intellectual disability (ID). Finally, genes involved in chromatin regulation of gene transcription were enriched in ~60% of the neurodevelopmental data sets, whereas enrichment was only observed in one control data set.

Finally, our submission included work combining the power of orthogonal and complementary genomic analysis to understand the function of genes about which little is known. Using gene expression data, genes with DNMs in our schizophrenia trios presented reduced coexpression with other genes in the prefrontal cortex, in contrast to other brain and nonbrain regions, supporting hypotheses suggesting that genes with loss-of-function mutations express lower connectivity.

We aim to expand this study of de novo variants in more trios with schizophrenia. However, given the role of chromatin remodeling and epigenetic regulation of gene transcription during brain development and in the adult brain, we aim to study how perturbations of genes implicated in chromatin remodeling affect neuronal development and function of the PVD (peripheral vascular disease) neuron in C. elegans and in neurons derived from patient-specific induced pluripotent stem cells, shedding light on the pathogenesis of neurodevelopmental disorders. We also aim to explore the role of chromatin remodeling in the adult mouse amygdala in response to fear conditioning, which may provide insight into the molecular mechanisms of anxiety and posttraumatic stress disorder (PTSD) that could be used to develop novel treatments for these conditions.

Investigation of Bipolar Disorder Genetics Using Exome Capture and Resequencing
J.S. Parla, S. Muller, G. Cheang, M. Kramer, E. Ghiban, S. Goodwin, W.R. McCombie [in collaboration with J. Potash, University of Iowa; P. Zandi, F. Goes, R. Karchin, and A. Chakravarti, Johns Hopkins School of Medicine]

Bipolar disorder (BP) is a complex psychiatric disease for which previous studies have failed to find significant causative variants of large effect. In 2010, we began using exome capture and next-generation sequencing to study bipolar disorder using family samples as well as case-control analysis. By 2012, we had successfully sequenced more than 60 family samples
and ~1000 case-control samples. Although this work allowed us to develop and streamline high-throughput capture and data-processing pipelines, it became quite clear that we would need a much larger sample set in order to investigate this likely polygenic phenotype. In 2013, we scaled up sequencing and brought the total samples completed for the project to 2329. This basically doubled our goal for the project and matched the output of the previous 3 years of the project combined. In 2013, we focused the analysis on variants that were predicted to be damaging by several software tools—SIFT (Ng PC1 et al., *Nucleic Acid Res* 13: 3812 [2003]), PolyPhen (Adzhubei et al., *Nat Methods* 7: 248 [2010]), and VEST (Carter et al., *BMC Genomics* 14: 1–16 [2013]), which is a tool developed by our collaborators—and that segregated with disease in the family samples. We then assessed whether there was any increased burden in case versus control samples for those genes. This analysis led us to focus on 10 possibly significant genes. We then used Sanger sequencing to genotype family members of the probands who were affected cases in the case-control sample set; however, we did not find any further evidence of segregation. In 2014, we plan to further refine our analysis techniques and to focus more on variant analysis in the larger case-control data set.

Additionally, on the analysis front, our collaborator Rachel Karchin proposed a new hybrid likelihood model to assess functional variants. The method is called BOMP (burden or mutation position test) and is a combination of a burden test with a test of the positional distribution of variants. This work was published in *PLoS Genetics* in 2013 (Chen et al., *PLoS Genet* 9: e1003224 [2013]). In 2013, we also submitted a paper to the *Journal of Human Genetics* describing our work comparing data-processing techniques, quality filters, and variant detection software tools in order to produce an accurate call set.

**Functional Screening in the PVD Neuron of C. elegans to Unravel Neuropsychiatric Disorders**

O. Mendivil Ramos, S.E. McCarthy, M. Kramer, W.R. McCombie [in collaboration with C. Aguirre-Chen and C. Hammell, Cold Spring Harbor Laboratory]

*C. elegans* is an important model system for the identification and functional mapping of genes involved in complex biological processes, including those of the nervous system. In particular, it has been a useful model for a variety of neuropsychiatric disorders including schizophrenia, Alzheimer’s disease, and neurodegeneration. The number of neurons in the *C. elegans* nervous system is known. Moreover, most of the connections among neurons are known, as well as the lineage of every cell from a fertilized embryo to the adult animal. The PVD neuron forms several levels of well-characterized dendrites and is competent for bacterial ingestion-mediated RNA interference (RNAi). This unique feature enables functional testing for a neural developmental phenotype at high throughput.

Late in 2013, we established a collaboration with the Hammell lab at CSHL to functionally assess some of the most prominent variants that we have identified in human genetics studies as being likely associated with neuropsychiatric disorders. We assembled a list of 117 human candidate genes identified in various ongoing genetic studies and found that 47 of them had good homologs in *C. elegans*. The Hammell lab performed a preliminary screen of these genes, obtaining six phenotypes of altered branching of the PVD neuron. In 2014, both labs will be expanding this research line by extending the list of homologs to additional candidates, as well as doing controlled studies to better estimate the efficacy of this model in assessing the large number of genetic variants being identified for major psychiatric disorders.

**Highly Multiplexed Targeted Resequencing of Submillion Base Genomic Targets**

J.S. Parla, P. Deshpande, S. Ethe-Sayers, M. Kramer, E. Ghiban, S. Goodwin, W.R. McCombie [in collaboration with R. Karchin, Johns Hopkins School of Medicine]

Improvements in next-generation sequencing technology allow users to produce enormous amounts of sequencing data in increasingly short times, but this is predominantly directed at a large amount of sequence from each sample (such as the entire genome). There remains a significant and growing need to efficiently sample small genomic regions (on the order of 0.5 Mb) in extremely large cohorts of samples. In 2012, we worked to establish a technique using Nimblegen solution capture probes to barcode and capture highly multiplexed samples. We created a custom Nimblegen probe set targeting five genes that cover ~637 kb of the genome. We then successfully validated multiplexing up to 96 samples per capture, followed by sequencing of the 96plex set on one
Illumina HiSeq lane. We were able to achieve ~90% of target coverage at ≥20× sequence depth and ~200× mean target coverage with this method.

In 2013, we began testing methods to increase the level of sample multiplexing to take advantage of the high mean coverage that we were getting on the 96 individually barcoded samples along with increased sequence output from the Illumina instruments. We investigated sample-pooling techniques where five or 10 individuals were pooled together, and that entire pool was subsequently barcoded and sequenced up to a 96plex of 10 samples per pool (960plex). These pools were then run across three HiSeq lanes. We began testing a modified GATK pipeline (DePristo et al., Nat Genet 43: 491 [2011]) for variant calling in the anonymous pools. We found that we were able to detect the correct variants for the individuals within those pools to a high degree (only ~1%–3% of variants that were called in the samples when they were sequenced individually were not found in the pools on average); however, we found we also call a large number of extraneous false-positive variants with this method of ascertainment. This led as to begin studying the effect of modifying several sequence quality filters to obtain a high degree of sensitivity while reducing the number of false positives. More work is needed to refine the filters and methods used to call variants accurately in these anonymous pools.

Further investigation to explore complex pooling designs will be performed in 2014. These investigations will include simulations and bench work to evaluate the performance of various pooling designs for identifying rare and common variants. In particular, we will begin testing sample pooling using a two-dimensional matrix such that individuals will be pooled by a unique row/column location before barcoding.

Genomics of Colorectal Cancer
E. Antoniou, W.R. McCombie [in collaboration with E. Li and J. Williams, Stony Brook University]

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and is the second leading cause of cancer-related deaths in the Western world. The mortality from colorectal cancer in Caucasian Americans (CAs) has been declining, but, in contrast, it has been rising in African Americans (AAs). The response to chemopreventive agents differs in terms of the biological/genetic heterogeneity of the tumors of AAs and CAs. Because chemopreventive agents exert their effect through a molecular target, this disparity would suggest differences at the genetic level. This project aims to determine the attribution of genetic and epigenetic differences to the racial/ethnic health disparity of colon cancer.

In 2012, to validate our laboratory and analysis pipelines, we used RNA-Seq on a cell line (HT29) treated with 5-aza-deoxy-cytidine. In 2013, we compared our results to the data from previous microarray studies. We published an article on this comparison, showing the superiority of the RNA-Seq technique (Xu et al., BMC Bioinformatics 14; S1: doi 10.1186/1471-2105-14-S9-S1 [2013]). We also used RNA-Seq to obtain gene expression measurements on 18 tumor/normal paired AA tissue samples and as many samples from CA patients. Analysis is ongoing.

In addition, we prepared 57 reduced representation bisulfite sequencing (RRBS) libraries, 52 of them from paired tumor/normal tissues. In 2014, we will prepare another 30 RRBS libraries and more RNA-Seq libraries from the same paired samples. We will then proceed with the combined analysis of gene expression and DNA methylation. Additional biopsies are also being collected on an ongoing basis at Stony Brook University Hospital.

Plant Genomics

Rice (Oryza sativa) is one of the most important food crops in the world, especially in the developing world. Relatively little effort has been made to explore the nature of structural variation within and between sub-populations of domesticated O. sativa. The best way to fully understand the genomic diversity of rice is to perform whole-genome shotgun sequencing and de novo assembly. The difficulties in assembling the short reads initially provided by next-generation sequencing have resulted in efforts to sequence other strains of rice that rely on mapping the sequence read back to a reference sequence. This results in the loss of information where there are structural differences between the two strains. Often, these differences constitute biologically salient information. Recent advances in sequencing
technology and in computational approaches to sequence assembly have significantly improved the power and reliability of de novo assembly. In this project, we used these new tools to develop de novo assemblies of three divergent rice genomes representing the indica (IR64), aus (DJ123), and japonica (Nipponbare) subpopulations and to determine the extent and distribution of structural variation among them.

During the last quarter of 2012, Pacific Biosciences Inc. (PacBio) released a new polymerase and associated chemistry (P4/C2) that greatly increased the read length on this machine, up to an average of 4.5 kbp. A hardware upgrade also doubled the throughput of the instrument in mid 2013. Taking advantage of these improvements, we produced 19× coverage of the Nipponbare genome using the new P4/C2 enzyme/chemistry combination. We also obtained 28× genome coverage of longer Illumina MiSeq reads (paired end 250 bp). A new assembly of the Nipponbare genome was made using the PacBio and Illumina 250-bp reads. The N50 contig size doubled from 21 kbp to 58 kbp. Furthermore, a new error-correction algorithm for the PacBio sequences was written by Mike Schatz’s group. After applying this new error-correction pipeline, the N50 contig size grew to 155 kbp. In conclusion, we can now generate a de novo assembly of a complex genome such as rice using next-generation sequencing platforms at a fraction of the cost and time that it took to make the bacterial artificial chromosome (BAC)-based rice reference genome.

In the last quarter of 2013, Pacific Biosciences introduced yet another polymerase and chemistry as well as new library preparation protocols that have doubled the average read length on this platform. We are planning to generate enough coverage of at least one rice genome with these new reagents to assemble a de novo rice genome using only PacBio sequences, with the hope of achieving even larger contig sizes.

De Novo Assembly of Yeast Genomes Using Long Insert Libraries and the Pacific Biosciences RS II Instrument


Next-generation sequencing has proven to be an invaluable tool for the understanding of genetics. However, the relatively short reads generated by instruments such as the Illumina HiSeq pose a limitation to the de novo assembly of larger genomes. Pacific Biosciences provides an instrument that can generate reads that are thousands of bases in length. These long reads provide the architecture required for de novo assembly and also provide superior retention of structural elements important to the understanding of genetic variation.

In 2013, we began exploring the use of long read sequencing for the de novo assembly of yeast strains. We selected a large fragment (>7 kb) library from the Saccharomyces cerevisiae W303 genome using a Blue-Pippin from Sage Sciences. Sequences were generated using the Pacific Biosciences RS II instrument with p5-c3 chemistry. The results yielded >7x5× coverage across the genome with reads in excess of 10 kb. These sequences were assembled with HGAP and the Celera Assembler (Chin et al., Nat Methods 10: 563 [2013]; Myers et al., Science 287: 2196 [2014]). The resulting contig N50 length approached one million bases, essentially only limited by the chromosome lengths of the organism. Only one chromosome, containing a very long tandem repeat, was not represented by a single contig. In 2014, we plan to continue to optimize these protocols and test them on additional yeast genomes.

PUBLICATIONS


PLANT GENOME ORGANIZATION, EVOLUTION, AND FUNCTION

Our lab has two goals: (1) research into plant genomics and (2) developing tools and resources for use by the genomics research community. Our research includes a broad range of physical, statistical, and functional genomics in model plant systems and agriculturally important crop plants. We also contribute to two large-scale, multi-institutional, cyberinfrastructure collaboratives that are designed to serve broad research and educational communities. During the past year, we were joined by visiting scientist Fazhan Qi and postdoctoral researcher Bo Wang. Shiran Pasternak has advanced his career, taking a position in the industry sector.

Gramene


The Gramene project provides online reference resources for plant genomes and curated pathways to aid functional genomics research in crops and model plant species. Our website, www.gramene.org, allows the study of gene function by combining genome annotation and experimental data with cross-species comparisons. This past year witnessed remarkable growth of the website as the project accomplished several major milestones culminating in our 40th release since the inception of this project. Highlights include the addition of six new species, the premiere of the Plant Reactome database, and a total redesign of Gramene webpages and search functions.

The Gramene team develops comparative genomics databases in collaboration with the Ensembl Plants project at the European Bioinformatics Institute (EBI). Thanks to this teamwork, we added six new complete reference genomes to the collection, bringing the total to 28. Perhaps the most significant of these was the genome of bread wheat, Triticum aestivum, released by the International Wheat Genome Sequencing Consortium. This hexaploid genome is the largest (~17 Gbp) and most complex angiosperm genome ever sequenced. The first draft, although gene-rich, is still a highly fragmented assembly; nevertheless, its addition to Gramene completes the triumvirate of major cereal grains in our collection: wheat, rice, and maize. These three crops constitute 90% of world-wide grain production and are keystones to future agricultural sustainability. Bread wheat’s unusually large genome can be attributed in part to allohexaploidy (AABBDD), which arose during domestication. Understanding the evolution and interaction among subgenomes will be greatly aided by study of its diploid progenitors. Two of these, Triticum urartu (AA) and Aegilops tauschii (DD), were also added this past year to the Gramene/Ensembl Plants collection, following their release by the Beijing Genomics Institute (BGI). These, along with the barley genome (Hordeum vulgare), make Gramene a preeminent site for study of Triticeae genomes.

Gramene continues to build a genus-level phylogenomics resource for rice and related Oryza species. Rice emerged ~15 million years ago. The Oryza genus includes two cultivated species (Asian and African rice) and 21 wild species adapted to a broad range of tropical and subtropical habitats around the world. The genomics research wishing to exploit such diversity for crop improvement is also establishing Oryza as a model system for genus-level study of trait evolution, speciation, and domestication. Working with the National Science Foundation (NSF)-funded Oryza Genome Evolution (OGE) project (principal investigator: Rod Wing, University of Arizona) and with the international consortium known as I-OMAP, Gramene will soon host high-quality genome assemblies for more than half of these species. Last year,
Gramene added complete reference assemblies for two new species and will include six additional genomes in the coming release, bringing the total to 12. One of the challenges of comparative genomics research is that different genome projects apply different protocols to identify genes, leading to methodological bias that confounds comparative analysis. To remedy this, Gramene staff worked with the OGE project to develop a consistent annotation protocol (using the MAKER-P pipeline) that was applied both to the new genomes and to the *Oryza* genomes previously sequenced and annotated by other projects. Gramene will continue to support both sets of annotations by creating a new website to represent I-OMAP genomes and phylogenetic trees.

**PLANT GENOME RESEARCH**

In the last decade, the decoding of complete plant genomes has helped scientists understand plant function and evolution and how to alter economically important traits. Many disciplines are required to generate reference genomes. This starts with laboratory scientists who generate the raw sequence data and goes to computational biologists and bioinformaticians, such as those in our lab, who interpret the output. This interpretation includes the assembly of raw sequence reads into overlapping segments (“contigs”) that are used to create a scaffold to discern their order and orientation within chromosomes. Another step is annotation—the discovery and description of genes and other functional elements, as well homologies to other genomes. In addition, this information must be faithfully communicated and visualized, such as in web-based platforms like Gramene.

All of these activities are rapidly evolving with the advent of new sequencing technologies, algorithms, and data-handling requirements. For example, high-depth sequencing of RNA transcripts at low cost is providing new evidence that informs genome annotation and is spurring the development of new software to model and perform this task. Low-cost sequencing is also transforming the types of questions that can be asked, moving beyond the generation of a single reference for a given species. Ongoing projects within the maize, rice, and *Arabidopsis* research communities are now sequencing hundreds or thousands of genotypic backgrounds within species, gathered from carefully constructed populations, wild populations, and breeding germplasms. Information on genetic variation is helping scientists to understand the genetic basis of phenotypic traits and the origins of domesticated crop species. New technologies are also driving research in epigenetics, the study of heritable traits that are not caused by changes in underlying genome sequence. Epigenetic mechanisms include modification of DNA by methylation and various forms of histone modification that can cause changes in gene expression and other phenotypes. Both of these modification types can be studied with new sequence technologies and analysis methods.

**Updating the Maize B73 Reference Genome and Developing Pan-Genome Variation Pipelines**

This work was done in collaboration with J. Glaubitz, Cornell University; E. Buckler, Cornell University; R. Fulton, Washington University; R. Wilson, Washington University; and the Maize Genome Sequencing Consortium. Work continues on refining the assembly and annotation of the maize B73 reference sequence. The third version (“RefGen_v3”) of the B73 assembly and annotations is hosted on Gramene and has been deposited in GenBank along with contigs assembled from Roche/454 sequencing of a whole-genome shotgun library. RefGen_v3 incorporates some of these contigs in order to increase coverage of missing gene space. The contigs were selected based on alignment to FLcDNA sequences and inserted into gaps in the RefGen_v2 assembly guided by a genetic map combined with synteny to rice and sorghum. Approximately 500 genes were added or improved with this method, and many unplaced physical map contigs were also anchored on the basis of these maps. The RefGen_v3 assembly was released following acceptance by GenBank.

Work on RefGen_v4 is under way. To improve the maize genome sequence, we are adding additional depth to the existing capillary sequence using the Illumina 2000 platform, as well as generating long reads with PacBio to span repeats. We have resequenced 17,173 bacterial artificial chromosomes (BACs) that constituted the minimal BAC tiling path from the original B73 sequencing project by pooling the BACs in pools of 96 and are generating Illumina data to greater than 150× coverage per clone. This increased depth, density of read pairs, and differential
bias compared to capillary sequencing will serve as an excellent resource to further improve the maize sequence. Long reads from PacBio will be used to improve the order and orientation of assembled contigs and close gaps wherever possible. These data, once completed, will be made available through public databases before completion of the RefGen_v4 assembly.

In *Zea mays*, it is estimated that only 50% of genomic content is held in common between lines due to tremendous haplotype diversity. A single reference assembly for maize cannot serve as a sufficient backbone to capture and describe this variation and is a limiting factor in understanding the genetic variation controlling traits. The maize pan-genome under development will be composed of the B73 reference assembly and novel sequences from other maize inbred lines. Our initial efforts are to identify novel haplotypes among a broad sampling of sequenced germplasm by conducting whole-genome assembly of 30× Illumina reads. At this depth, high-repeat content in maize will challenge existing methods for de novo assembly. Simulations using the ALLPATHS-LG recipe to assemble 30× reads sampled from the B73 reference resulted in ~82% coverage of annotated genes. We are using sequence data for the inbred B97 to further prototype pipeline construction. Because different assembly algorithms have different advantages, we have adopted a meta-assembly strategy. Using this strategy, ~30% of the genome could be assembled with ~72% of genic space covered in scaffolds. After aligning to the BAC-based B73 reference, ~19.78 Mb of novel sequences was identified. We also used a genetic mapping method to anchor novel sequences relative to the B73 physical map. The accuracy rate of this genetic mapping method was tested by anchoring 1000 B73 genes, and the results demonstrated 77.6% accuracy. As for B73, PacBio long reads will be generated and used for scaffolding and gap closing.

**Discovery and Application of Epigenetic Variation in Sorghum and Maize**


DNA methylation has an important role in the regulation of gene expression and control of transposable elements. The patterns of DNA methylation, referred to as the “methylome,” must be faithfully propagated for proper development in plants and animals. In collaboration with the Martienssen lab and DuPont Pioneer, we sequenced the methylome of two maize inbred lines, B73 and Mo17. The genomic DNA is treated with bisulfite, which converts unmethylated cytosine to thymine. Sequencing using next-generation Illumina GA2 paired-end reads, followed by mapping back to the maize genome, resulted in identification of the methylome in single-base resolution. We have generated 20×–30× coverage over the mappable portion of the maize genome. Alignment with RNA sequences indicates that the methylation patterns are correlated with gene expression, small RNA, and alternate splicing. Diversity in cytosine methylation patterns was observed in transposable elements and especially in genes and was found to be largely heritable in recombinant inbred lines (RILs); however, significant deviations from heritability were observed, many of which were conserved in different RILs. This will help us to reveal the roles of DNA methylation in gene regulation and other biological functions in the future.

In *Sorghum*, we used shallow sequencing (2×–4×) of bisulfite-treated root tissues in order to identify regions of open chromatin and patterns of methylation associated with high-confidence gene models. These patterns, together with measures of gene expression and sequence conservation, were used to train a classifier to predict functional gene models. A new set of gene models was produced using the Gramene/EnsEMBL GeneBuilder by combining transcript fragments assembled from RNA-sequencing data with alignments of protein, expressed sequence tags (ESTs), and cDNA sequences from sorghum and other monocots. The classifier, trained on Sbi1.4 gene models, was run on the new set of gene models and predicted a total of 56,128 functional genes and transcriptionally active regions in sorghum.

**Characterization and Analysis of Core Promoter Elements in Plant Genomes**

S. Kumari

Where and when a plant will express or transcribe a gene is controlled by information in the plant genome and is influenced by environmental signals. Transcription
initiation involves the recruitment of the basal transcription factors to the core promoter elements (CPEs) of genes. Very little is known about CPEs in plants, and this study offers a new insight in the field. To better understand the core promoter architecture in plant genomes, we studied a comprehensive set of known CPEs in various monocot and dicot plant genomes.

We developed high-throughput in silico methods for motif prediction using the publicly available CREAD (comprehensive regulatory element analysis and discovery) suite of tools. Our computational pipeline systematically identified transcription regulation motifs at the whole-genome level including TATA-box, CCAAT, INR, BRE, GC-box, DPE, MTE, and Y-Patch motifs containing genes across eight plant genomes. This is the first large-scale genome-wide computational study that has compared and contrasted the distribution profiles of known CPEs across four monocots (*Brachypodium distachyon, Oryza sativa japonica, Sorghum bicolor, Zea mays*) and four dicots (*Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera, Glycine max*). The relative abundance signal of CPEs with respect to transcription start sites (TSSs) was found to be conserved within monocots and dicots with slight differences across monocots and dicots. The TATA-box motif was found to be present in 16%–22% of the plant promoters, whereas the GC-box, XCPE1, and Y-patch motifs were found to be the most prevalent CPEs across all eight genomes.

The structural properties in the regulatory part of genomes distinctly differ from those of the nonregulatory part. The core promoter region’s free-energy profiles (Fig. 1) segregate into two distinct clusters: one clustering monocots together and the other clustering dicots together. Free-energy profiles have the potential to be used for delineating the promoter region as well as for computational TSS identification. It could help build better computational models for predicting the TSS in the promoter region, which remains a challenging problem.

We classified core promoters into three types (TATA-containing, TATA-less, and Core-less) and looked at the functional enrichment of genes associated with these types. TATA-containing genes showed significant overrepresentation of biological processes that included response to abiotic and biotic stress, hormonal stimuli, and regulation of carbohydrate and nucleic acid metabolic processes. TATA-less genes were mainly involved in housekeeping, including transferase activities, hydrolase activities, and various nucleotide-related binding activities. These genes were mainly enriched in biological processes related to nitrogen and phosphorous metabolism, which are key genes involved in plant-yield-associated traits. The Core-less genes showed evidence of unique enrichment in genes involved in ATP binding, signal transduction activities, apoptosis, etc. Our study expands the CPE repertoire in plants, providing impetus for future genome annotation projects and inspiring research efforts aimed at better understanding transcriptional regulation mechanisms and the relationship for crop yield improvement.

**Figure 1.** Free-energy profile of the promoter region genes from eight plant genomes.
PLANT SYSTEMS BIOLOGY
Exploring the Gene Regulatory Network
Guiding Arabidopsis Stele miRNA Expression

Plant roots not only serve to physically anchor plants to the soil but are also responsible for uptake of water and critical nutrients; therefore, they must respond quickly to various environmental stresses such as drought, waterlogging, and heavy metal pollution. microRNAs (miRNAs) have a central role in plant development and response to environmental stress. We are interested in studying the gene regulatory network (GRN) that regulates miRNA expression in the root. Most miRNAs are transcribed by RNA polymerase II, and their transcription process is regulated by transcription factors (TFs). To systematically resolve points of cross-talk among TFs, miRNAs, and their targets in a comprehensive GRN, we make use of a gene-centered yeast one-hybrid (Y1H) experimental system that allows us to monitor protein binding to DNA in yeast.

We have extended the library that initially focused on transcription factors from the root stele (657), the central part of the root important for nutrient transfer, to the complete root (952). Using this system, we screened 172 promoters of stele-expressed miRNAs, their targets, and some highly connected TFs. We have obtained 3768 protein–DNA interactions (PDIs). The resulting network was analyzed with publicly available spatiotemporal expression data in the root that allowed us to examine whether the TFs act as activators or inhibitors of expression. To validate the PDIs found in our yeast work, we screened more than 250 genetic insertion mutants from Arabidopsis and obtained 80 lines for TFs, miRNAs, and miRNA targets in our network. From these, we characterized their molecular phenotype by quantifying TF expression. Our results showed that perturbation in planta of ~80% of TFs and 70% of miRNAs tested displayed molecular phenotypes in the Arabidopsis root.

In addition to molecular phenotypes, the plants were also scored for observable morphological differences, and two mutants with a short-root phenotype were identified. To understand the genes that were bound by one transcription factor, we used a chromatin immunoprecipitation sequencing (ChIP-Seq) approach that allows us to identify plant DNA that is bound by a transcription factor. Using this approach, we identified both genic and intergenic regions bound by the transcription factor and used this to predict a DNA motif that is likely bound by the transcription factor. Further genetic investigation of this transcription factor suggested that it may be involved in stem cell maintenance in the root.

We have begun to use the network to predict genes that may be important in development based on the topology of the network. Within the network, some genes are more highly connected, which suggested that these genes may have an important role in the network. The zinc finger homeodomain (ZF-HD) TFs were found to be highly connected in our network. These genes are conserved across the plant kingdom. We have managed to identify them in more than 300 species. The number of ZF-HD varies from species to species. In Arabidopsis, there are a total of 16 members. Genetic analysis of these TF families indicates that their functions are not restricted to the root and that they are important in flower development. Although our network was focused on root-expressed genes, more than 80% of these genes are expressed in other tissues and will be more broadly applicable beyond root development and response to stress. The GRN provides a framework for modeling adaptive responses to environmental conditions at the whole-plant level. We can use this network to identify candidate genes for improving germplasm that can withstand more detrimental conditions, thus addressing global food security and growing demand for renewable energy resources.

Developmental Networks Controlling Inflorescence Architecture in Maize

The goal of this work is to integrate genetics and genomics data sets to find molecular networks that influence the morphology (architecture) of maize inflorescences (flowers). Because inflorescences bear the fruits and grains that we eat, understanding the genetic and regulatory basis for how these structures are formed has clear relevance to important agronomic traits such as grain yield and harvesting ability. Our data sets represent maize inflorescence primordia sampled during
key developmental transitions and in perturbed genetic backgrounds. The latter includes loss-of-function mutants in three important regulators of the RAMOSA (RA) pathway, which controls stem-cell-fate decisions and ultimately the decision to branch. We have established a robust system to investigate the networks that modulate branching, including characterization of precise timing of developmental events and associated spatiotemporal changes in gene expression. We integrated genome-wide mRNA-Seq data to resolve coexpression networks during key stages of maize inflorescence development and are working to expand these networks by incorporating additional data sets, such as genome-wide transcription factor (TF) occupancy profiles and cis-regulatory information.

The primary objectives for this project included the following: (1) Establish a comprehensive pipeline for mRNA-Seq and ChIP-Seq data analysis in maize by evaluating and optimizing available software for mapping and quantification. This also included testing various statistical methods to extract biological relevance. (2) Characterize genome-wide expression signatures specific to a given developmental event or branching phenotype. We made use of known developmental marker genes and their spatiotemporal transcriptional responses to genetic perturbation in order to test our experimental system and establish a proxy for developmental staging. (3) Evaluate and implement clustering approaches to identify candidate genes that are coexpressed with key regulators and/or coincide with specific developmental events. Our results from this included identification of candidate genes, specifically developmentally regulated TFs, and novel genes of unknown function that are potentially involved in stem-cell maintenance and determinacy. We continue to examine coexpression clusters for enrichment of functional processes and cis-regulatory motifs that lie within proximity of the transcriptional start site of coexpressed genes (with Sunita Kumari). (4) Identify targets of the RA1 TF using ChIP-Seq and integrate results with data from parallel mRNA-Seq experiments. On the basis of this approach, we showed that one-third of the genes with altered expression levels in the ra1 mutant are also bound by RA1. We are incorporating additional ChIP-Seq data sets as they become available to investigate combinatorial binding of TFs associated with the branching pathway. The ChIP-Seq data also provide in vivo confirmation for binding sites of developmental regulators in maize, information that is being leveraged in efforts to resolve cis-regulatory modules across the maize genome.

In addition to maize, we have begun work on sorghum, which is an important emergent bioenergy crop used for human consumption in sub-Saharan Africa. We are using next-generation sequencing approaches to identify the single-nucleotide mutations that are associated with an increase in seed of the sorghum plants in a sorghum EMS (ethylmethanesulfonate) population. Using this approach, we have identified two genes that can change the structure of flowers and generate more seeds. We intend to apply this strategy to perform a large-scale sorghum mutant study to support candidate genes associated with developmental traits in the roots, shoots, and flowers.

In the next phase, we will further prioritize candidates from this work by overlaying the Arabidopsis regulatory network information (see previous section). The resulting hypotheses can be tested in Arabidopsis, for example, for responses to stress, and ultimately translated to agronomic systems. Additionally, candidate genes that are maize and/or grass specific are of high priority because they may contribute to the unique morphology of maize inflorescences and/or features shared among other grasses. We will further use comparative genomics approaches, including both computational and integration of analogous RNA-Seq data sets from closely related grasses, such as sorghum, to identify candidate genes that may contribute to grass-specific aspects of inflorescence architecture.

CYBERINFRASTRUCTURE PROJECTS
The iPlant Collaborative
L. Wang, J. Lu, C. Noutsos, J. Stein, Y.K. Lee

This project was done in collaboration with Cold Spring Harbor Laboratory and employs more than 100 staff. It is headquartered at the University of Arizona under the direction of principal investigator Stephen Goff and director Dan Stanzione. Dozens of collaborators are located at more than 20 institutions. The iPlant Collaborative (http://iplantcollaborative.org) is an NSF-funded cyberinfrastructure project that provides public access to high-performance computing, data storage, and tools via customized web-based interfaces. Having completed the first five-year grant, the iPlant Collaborative has made extensive
progress toward meeting these goals and has been recommended for renewal for another five years. Work in the last year has culminated in the development of multiple platforms that are now being used by the research and educational communities. Staff at CSHL directly contributed to some of these cyberinfrastructure platforms or have built upon them to provide ready access to needed software and analysis tools by scientists and educators. Within the Ware lab, these platforms include the Discovery Environment (DE), Atmosphere, and the Taxonomic Name Resolution Service (TNRS).

The DE is the most visible portal to iPlant tools and services. This web-based platform supports an “app store” model of user-extensible tools, automated workflows, and data storage. Users can take advantage of existing tools integrated by iPlant staff and user community or add their own tools to use privately or share. Users may not be aware that the underlying infrastructure provides access to iPlant’s massive data store at the University of Arizona and the Texas Advanced Computing Center (TACC). Computationally intensive tasks are handled by supercomputers located at TACC and other centers within the Extreme Science and Engineering Discovery Environment (XSEDE). So far, well over 400 tools have been integrated into the DE that enable a broad range of research activities, including genome/transcriptome assembly, annotation, RNA-Seq quantification, variant detection, genome-wide association studies (GWAS), and phylogenetics. Members of our lab have had important roles in contributing to workflow design, tool integration, validation, science tutorial, and documentation.

Atmosphere is iPlant’s configurable and cloud-enabled computational resource for the plant research community. From Atmosphere’s web interface, users can launch a virtual machine (VM) with preconfigured working environments and ready-to-use software precustomized. Users can also create their own applications and environments as VMs and share with others via Atmosphere. As with the DE, Atmosphere is a gateway to access iPlant’s core infrastructure resources, such as the high-performance and grid-computing environment and big data-storage system. Using the Atmosphere platform, we created a VM to be used in the fields of ecological and functional genomics. The VM includes various binary tools and R statistics packages used in ecology and genetics research and for plotting complex data in graphs.

The success of genome research depends on our ability to accurately assemble, annotate, and derive meaning from sequence data; however, extremes of genome size, polyploidy, diversity, and repeat content push the limits on current algorithms, expertise, and computational power needed by today’s researchers. In response, iPlant is fostering a community effort to identify best practices and state-of-the-art tools, install and optimize their performance on the nation’s most powerful supercomputers, and make these available as a free online resource. During the last two years, the iPlant Discovery Environment has matured to provide a comprehensive set of tools and services for sequence handling, performing read alignments, RNA-Seq profiling, and de novo genome and transcriptome assembly. To extend these capabilities, we are working to incorporate MAKER-P, a standardized, portable, and easy-to-use plant-genome annotation engine with built-in methods for quality control. As part of this effort, MAKER-P was specifically optimized to take advantage of the parallel computing environment of the TACC Lonestar cluster and is now a supported module. Performance testing showed that MAKER-P can perform high-quality, full-fledged annotation pipelines on even the largest plant genomes in a matter of hours. Incorporation of this resource into the DE fits into an overall strategy that includes downstream functional annotation of protein-coding genes and visualization. MAKER-P is currently available for use as an Atmosphere image.

Our lab contributed directly to the development of the Taxonomic Name Resolution Service (TNRS), a platform to help standardize taxonomic names for all plant species—a nontrivial task. Erroneous and synonymous taxonomic names are a major challenge for virtually every field of plant biology. Large organismal databases (GBIF, SpeciesLink, VegBank, SALVIAS, TraitNet, GenBank, TreeBASE) are plagued by taxonomic error and uncertainty. In some databases, up to 30% of names do not match any published name; furthermore, 5%–20% of published names may be synonymous. Correcting and harmonizing taxonomy are usually the time-consuming and ad hoc responsibility of the individual researcher. The TNRS tool overcomes this barrier, enabling higher-quality comparative biodiversity science. The TNRS is available to other investigators who wish to perform similar taxonomic name resolutions on their data sets, enabling
a wider community to expand the public scientific knowledge base. The TNRS is currently available at http://tnrs.iplantcollaborative.org and is described in a recent publication in BMC Bioinformatics.

A major mission of iPlant is to promote adoption of the cyberinfrastructure through training workshops and outreach at academic institutions and scientific meetings. In 2013, members of the Ware lab participated as instructors in several tools and services workshops focused on transcriptomics, annotation, GWAS, and phylogenetics using the DE, Atmosphere, and Data Store platforms.

KBase, the Department of Energy Systems Biology Knowledgebase
S. Kumari, S. Pasternak, J. Thomason

This project was done in collaboration with the Department of Energy national laboratories and is led by principal investigator Adam Arkin of Lawrence Berkeley National Laboratory (LBNL), with co-principal investigators Rick Stevens, Argonne National Laboratory (ANL); Robert Cottingham, Oak Ridge National Laboratory (ORNL); and Sergei Maslov, Brookhaven National Laboratory (BNL). In addition to Doreen Ware at CSHL, participating investigators included Mike Scharz, CSHL; Pamela Ronald, University of California, Davis; Matthew DeJongh, Hope College, Michigan; Gary Olsen, University of Illinois, Urbana-Champaign; and Mark Gerstein, Yale University.

The DOE Systems Biology Knowledgebase (KBase, www.kbase.us) has two central goals. The scientific goal is to produce predictive models, reference data sets and analytical tools, and to demonstrate their utility in DOE biological research relating to bioenergy, the carbon cycle, and the study of subsurface microbial communities. The operational goal is to create the integrated software and hardware infrastructure needed to support the creation, maintenance, and use of predictive models and methods in the study of microbes, microbial communities, and plants.

Doreen Ware serves as the Plants Science Team Lead for KBase. In addition to providing domain expertise in plant genomics, members of the Ware lab are making significant contributions to software development for this project. The Plant component of KBase allows users to model genotype-to-phenotype relationships using metabolic and functional networks as well as phenotype measurements and high-throughput experiment data. It also supports the reconstruction of new metabolic and functional networks based on expression profiles, protein–DNA, and protein–protein interactions. To accomplish this, we have provided interactive, data-driven analysis and exploration across multiple experiments and diverse data types. We have built narratives that can capture analyses, including rich annotations, visualization widgets, reusable workflows, and custom scripts.

Through KBase resources and infrastructure, users have access to workflows for analyzing data such as those from genotyping assays and expression profiles as well as metabolic models. Users can integrate these with public resources for plant genomes. Using the KBase application programming interface (API), users can create complex queries across disparate data sources and types.

PUBLICATIONS


In Press


