Specific Aims

To the Review Committee: The public facing documents of this application (Narrative and Abstract) are less detailed than usual, as this essential line of research has become extremely high-risk over the last two years for the Investigators, and their staff. Please read this Specific Aims page for the greater details on the experimental design and analysis plan we propose.

Many of the psychiatric disorders are thought to have a neurodevelopmental component. This is probably best understood for schizophrenia, which is thought to begin to arise in the second trimester, the developmental period during which the brain is formed. Similarly, alterations in brain development have been hypothesized for Bipolar Disorder, Autism, OCD, Tourette's Syndrome and perhaps most of the psychiatric disorders. In recognition of the potential importance of understanding the molecules and mechanisms of brain development throughout the lifespan, we and others, have begun to examine the transcriptomic and epigenetic profiles of the human brain at different ages. One such large project, the BrainSpan Atlas of the Developing Human Brain (BrainSpan.org; Drs. Knowles and Gerstein), studied up to 16 regions (11 cortical, 5 subcortical) of 43 high quality human brains from 10 weeks post-conception to 40 years of age, with a number of molecular modalities, including RNA-Seq. Initial examination of the data from this project quickly revealed that gene expression in fetal brains was dramatically different from that in non-fetal brains. Dimensional reduction of the gene expression data, as measured by RNA-Seq by either hierarchical clustering or Principle Component Analysis (PCA) clearly demonstrated that pre- and post-natal gene expression varies by age (it is PC1; data not shown).

The Genotype-Tissue Expression (GTEx) Project (gtexportal.org/home) has generated tissue specific expression Quantitative Trait Loci (eQTL) maps (and multi-tissue maps). With 13 CNS and 35 non-CNS tissue types, and over 7,000 samples this is a comprehensive project, but with an average age of the CNS tissue donors of 60-69, the brain eQTL maps do not reflect the state of the prenatal brain.

The PsychENCODE Consortium (psychencode.org; Drs. Knowles, Gerstein and Crawford) is a group of projects that "aims to produce a public resource of multi-dimensional genomic data using tissue and cell-type specific samples from approximately 1,000 phenotypically well-characterized high quality healthy and diseased human post-mortem brains, …" (The PsychENCODE Consortium, 2016). As part of the activities of this Consortium, Drs. Gerstein and Knowles are leading Capstone Project 4, which will generate a high power eQTL map of adult frontal cortex by combining ~2,300 samples from BrainSpan, GTEx, CommonMind, PsychENCODE, and other available sources. Unfortunately, comparable data from fetal samples is limited. We propose to fill this gap of data and knowledge, by collecting samples of fetal cortical brain tissue and performing a set of molecular assays and data analyses to produce maps of eQTLs, chromatin QTLs of ATAC-Seq and the histone mark H3K27Ac (promoter and enhancer mark) and surveys of additional chromatin marks (CTCF and 100 transcription factors expressed in fetal brain), as well as data to determine the 3D organization of the fetal genome across late first trimester to the end of the second trimester. These data will enable testing of the relationship between the genomic elements important for the development of the brain in the second trimester, and the genetic risk for the psychiatric, neurological, and the neurodevelopmental disorders. Specifically, we propose:

Aim 1. Collect a large sample (n=750) of cortical brain tissue from 10-24 post-conception weeks (PCW), and use these tissues for a number of molecular assays.

a) Genotyping of all samples with the Illumina Global Screening Array (GSA), which will contain a backbone of ~660,000 SNPs, which provides LD coverage and imputation accuracy of >0.8, for over 87% of the genome.
b) Perform bulk long RNA-seq (strand-specific ncRNA and mRNA >100bp) of all samples.
c) Perform ChiP-Seq of the chromatin marks H3K27Ac (750 samples) CTCF (24 samples), and of a panel of 100 transcription factors on at least 4 samples for each transcription factor.
d) Perform ATAC-Seq on all samples.
e) Perform deep Hi-C analysis of at least 1 billion reads per sample for 12 samples.

Aim 2. Data processing and bioinformatics analysis

a) Data processing and analysis to identify fetal QTLs
b) Early brain expression dynamics from differentially expressed genes
c) Dynamic modeling of brain developmental gene regulatory networks by integrating adult data from GTEx, PsychENCODE, etc
d) Integrate the fetal data with the genetic, expression and epigenetic data from studies of the psychiatric diseases to provide greater insight into the developmental aspects of the pathology

Aim 3. Provide an easy-to-use, web-based informatics framework for communication of the raw and computed data of this PsychENCODE project to other neuroscientists.

A. Significance

Many of the psychiatric disorders are thought to have a neurodevelopmental component. This is probably best understood for schizophrenia, which is thought to begin to arise in the second trimester (Weinberger 1987, Raedler, Knable et al. 1998, Lewis and Levitt 2002, Schmidt-Kastner, van Os et al. 2006), the developmental period during which the brain is formed. Similarly, alterations in brain development have been hypothesized for Bipolar Disorder, Autism, OCD, Tourette's Syndrome and perhaps most of the psychiatric disorders. In recognition of the potential importance of understanding the molecules and mechanisms of brain development throughout the lifespan, we and others, have begun to examine the transcriptomic and epigenetic profiles of the human brain at different ages. One such large project, the BrainSpan Atlas of the Developing Human Brain (BrainSpan.org; Drs. Knowles and Gerstein), studied up to 16 regions (11 cortical, 5 subcortical) of 43 high quality human brains from 10 weeks post-conception to 40 years of age, with a number of molecular modalities including RNA-Seq. Initial examination of the data from this project quickly revealed that gene expression in fetal brains was dramatically different from that in non-fetal brains. Dimensional reduction of the gene expression data, as measured by RNA-Seq by either hierarchical clustering or Principle Component Analysis (PCA) clearly demonstrated that pre- and post-natal gene expression varies by age (it is PC1; data not shown). A graphic demonstration of this is seen in Figure 1, to the right, where the BrainSpan expression data is shown on a heatmap for the top 10 cortically expressed genes in adults (top 10 rows/genes) and top 10 cortically expressed genes from the P3-P5 fetal samples (bottom 10 rows/genes). It is clear that the highest expressed genes in the cortical samples are expressed at lower levels in the fetus, while the converse is true for the highest expressed cortical genes in the fetus, which are nearly absent in the adult. More difficult to see, is that there is also very little region-to-region variation across the cortical samples, at each time period (more on this below).

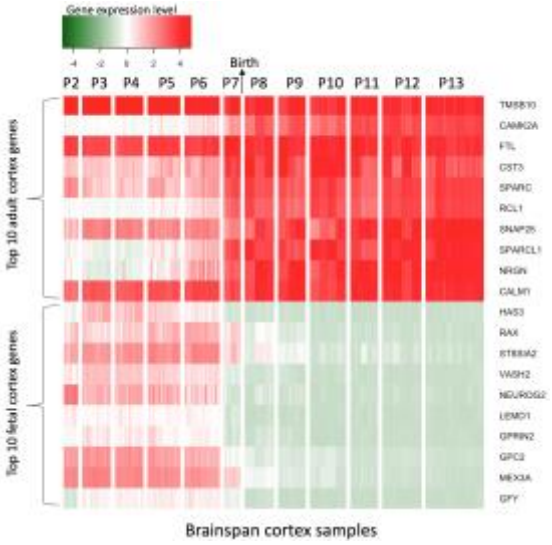

Figure 1

The Genotype-Tissue Expression (GTEx) Project (gtexportal.org/home), which utilizes RNA-Seq to measure gene expression across hundreds of samples of many tissues of the human body, plus genotypes from each individual, combines these data to generate tissue specific expression Quantitative Trait Loci (eQTL) maps (and multi-tissue maps)(Consortium 2015). With 13 CNS and 35 non-CNS tissue types, and over 7,000 samples this is a comprehensive project, but with an average age of the CNS tissue donors of 60-69, neither the GTEx RNA-Seq data, nor the eQTL map, reflect the state of the brain prenatally, when we think the psychiatric disorders may arise. This is crucial; as eQTL maps are powerful tools to link DNA sequence variation to gene expression. Frequently, disease associated GWAS variants are non-coding and it is not obvious what transcript they may alter the expression of, to influence trait or disease. eQTL maps provide this link between genome and transcriptome.

The PsychENCODE Consortium (psychencode.org; Drs. Knowles, Gerstein and Crawford) is a group of projects that "aims to produce a public resource of multi-dimensional genomic data using tissue and cell-type specific samples from approximately 1,000 phenotypically well-characterized high quality healthy and diseased human post-mortem brains, as well as functionally characterize disease-associated regulatory elements and variants in model systems"(Psych, Akbarian et al. 2015). As part of the activities of this Consortium, Drs. Gerstein and Knowles are leading Capstone Project 4, which will generate a high power eQTL map of adult frontal cortex by combining ~2,300 samples from BrainSpan, GTEx, CommonMind, PsychENCODE, and other available sources. Unfortunately, the number of available fetal samples with data across these projects is limited. We propose to fill this gap of data, and hence knowledge, by collecting samples of fetal cortical brain tissue and performing a set of molecular assays and data analyses to produce maps of eQTLs, chromatin QTLs of ATAC-Seq and the histone mark H3K27Ac (promoter and enhancer mark) and surveys of additional chromatin marks (CTCF and 100 transcription factors expressed in fetal brain), as well as data to determine the 3D organization of the fetal genome across late first trimester to the end of the second trimester. These data will enable testing of the relationship between the genomic elements important for the development of the brain in the second trimester, and the genetic risk for the psychiatric, neurological, and the neurodevelopmental disorders.

B. Innovation

- Large sample size of human fetal cortical brain tissue will enable discovery of eQTLs for RNA and chromatin QTLs of ATAC-Seq and H3K27Ac peaks.

- Survey of several other molecular modalities (100 Transcription Factors, CTCF, Hi-C for 3D genome structure) in fetal brain tissue.
- Relatively low-cost, for the valuable information generated.

C. Approach

C.1 Overall Experimental Strategy. As mentioned above, there is a significant gap in our knowledge of the regulatory and epigenetic landscape of fetal brain development, and this knowledge maybe vital to understand, and hence rationally treat the psychiatric, and perhaps other brain disorders. The BrainSpan project (BrainSpan.org) first led this effort by determining the pattern of gene expression in up to 16 regions of 43 brains across the human lifespan, which clearly demonstrated that gene expression in the brain is very different pre- and post-natally. As a survey of gene expression across brain region and developmental time, it also lacked the power to map the eQTLs that regulate this gene expression. The goal of mapping human brain eQTLs has fallen to the GTEx project (Consortium 2015), which has done a wonderful job for adult tissues, but has not studied fetal tissue, despite the knowledge that gene expression, and hence gene regulation is substantially different in the womb. The PsychENCODE Consortium/projects(Psych, Akbarian et al. 2015), have taken on improving the brain eQTL map and extending it to mapping the epigenetic components. Unfortunately, most of the work of the PsychENCODE and other consortiums has been on adult tissue. Hence, as a field, we still lack good maps of eQTLs, enhancers and transcription factor binding sites during the prenatal period. This study is designed to close that gap.

Our overall goal is to generate a large, high-quality, multidimensional dataset of fetal brain development from ages 10-24 weeks post-conception. We will collect 750 fetal cortical brain samples from across this time period and perform high-quality genotyping, RNA-Seq of strand-specific total RNA, ChIP-Seq of H3K27Ac, and ATAC-Seq of all samples. Additionally, we will perform ChIP-Seq of CTCF and 100 transcription factors, and a developmental time series of the 3D structure of the fetal genome, using Hi-C, on a subset of samples. These data will be analyzed by a talented team of biologists to generate eQTLs, splicing QTLs (sQTLs), chromatin QTLs (cQTLs) of H3K27Ac and ATAC-Seq peaks, and transcription factor binding and one kilobase resolution 3D interaction maps of fetal brain. These products will then be analyzed further to determine gene networks and their regulatory elements, as well as the relationship of the genetic regulatory elements in fetal brain to human psychiatric disorders.

We have chosen to focus on fetal cortical brain, as it easily identifiable in the mixture of tissue fragments from the termination procedure. Fetal brains are rarely delivered intact (and we will not alter the treatment of any patient). Within the each fragment of fetal cortex we can standardize dissection by using the dark, cortical plate as a reference, providing a consistent mixture of brain region/cell types, across all 750 samples.

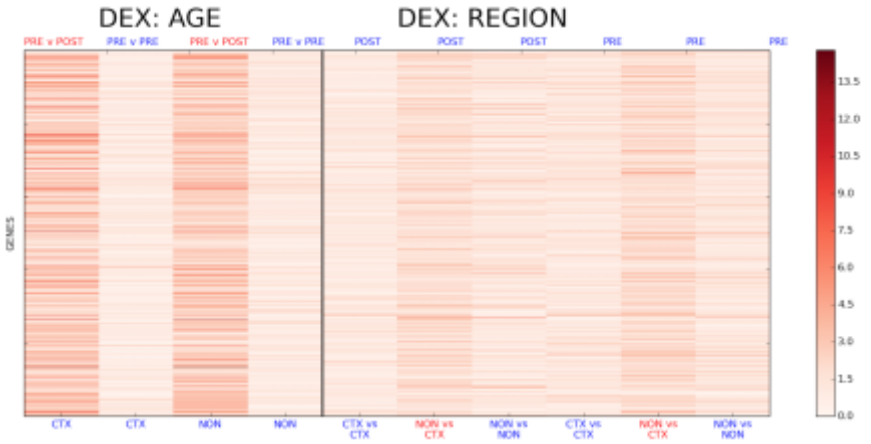

Figure 2

Another source of sample-to-sample variability we have considered is: What part of the cortex did each particular fragment come from (e.g., frontal vs. visual, or motor vs. temporal), and is this variation something we have to worry about? We have observed that there are almost no statistically significant differentially expressed (DEX) genes between cortical samples from different regions in adult brains in the BrainSpan data. There are an intermediate number of DEX genes between cortical and non-cortical regions in adults, and the largest number between adult and fetal brains (remember, age is PC1 and accounts for the largest source of variation). For this proposal we examined the BrainSpan data to determine if fetal cortical regions have the same paucity of truly DEX genes, as the adult brains. As you can see in Figure 2, the CTX vs. CTX comparison in the Pre- samples (third from right, 366 DEX genes at $p<0.001$) is one of the lightest, nearly the same as the Post- CTX vs. CTX comparison, and much less DEX than the Pre-natal vs. Post-Natal CTX vs CTX comparisons (3,873 DEX genes at $p<0.001$). So, regional variation should not be a problem, and if necessary, we can build a model to place each sample into regional space using the DEX genes.

We also considered, but have not included, the study of the fetal methylome. Ideally, this would be done with whole genome bisulfite sequencing (WGBS) of NeuN+ and NueN- cell fractions, or by using the WGBS variant, NOMe-Seq (Kelly, Liu et al. 2012), which can deconvolute the signal from multiple cell types, but both are quite expensive and would only provide a survey of sites, not a complete map of methylation QTLs (mQTLs), which would be more desirable. The alterative use of Illumina EPIC Methylation arrays for 750 samples would add ~$250,000+ to the project, and only provide an mQTL map of methylation sites derived from studies of oncogenesis. Similarly, we have considered, but removed, study of the fetal gene expression at the single cell level. The Knowles laboratory is one of the leaders of using these technologies (SCAP-T.org) and could employ either Drop-Seq, or other droplet based technologies (e.g., 10X Genomics Chromium), to determine gene expression in thousands of cells per tissue sample, cluster these into consistent groups across samples and determine cell-specific eQTLs, but this is beyond the scope of the present budget. Both of these can be performed with frozen tissue (using nuclei in the case of Drop-Seq or 10X Genomics, personal communication from Kun Zhang, UCSD), hence we will bank unused tissue from the project for future studies.

C.2 Principle Investigators.

James Knowles M.D.-Ph.D. is the newly hired Professor and Chair of the Department of Cell Biology and the Deputy Director of the Genomics Institute at SUNY Downstate. He is both a board-certified psychiatrist and a well-established psychiatric geneticist with years of experience in large-scale collaborations. He studies the genetics of multiple complex disorders (Schizophrenia, Bipolar Disorder, Early-Onset Depression and Obsessive-Compulsive Disorder). He is an expert in neuroscience and Next-Generation Sequencing (NGS) and is one of the PIs of the BrainSpan project (BrainSpan.org) and the Single Cell Analysis Program (SCAP-T.org). His laboratory is also a leader in the analysis of NGS data, and he was one of the co-PIs of the NHGRI funded network to make and distribute software for RNA-Seq analysis (iSeqTools Network, http://iseqtools.org/). He is a member of the PsychENCODE Consortium (psychencode.org) and the Whole Genome Sequencing of Psychiatric Disorders consortium (WGSPD). His site will also include his PsychENCODE Consortium collaborator at USC, **Peggy Farnham, Ph.D.** who is the Chairman and Professor of Department of Biochemistry and Molecular Medicine. She is an international leader in the study of chromatin regulation and its control of transcription factor binding and function. She is a member of an international consortia of genomic scientists working on the ENCODE project and was a member of an NIH Roadmap Reference Epigenome Mapping Center.

Mark Gerstein, Ph.D. is the Albert Williams Professor of Biomedical Informatics at Yale University. His lab (gersteinlab.org) was one of the first to perform integrated data mining on functional genomics data and to do genome-wide surveys. He was a member of the 1000 Genomes Project and is currently a leader in the ENCODE and modENCODE projects. He led the data analysis team of the BrainSpan project (BrainSpan.org). He is also a co-PI in DOE KBase and the leader of the Data Analysis Center for the NIH exRNA Consortium. In these roles Dr. Gerstein has designed and developed a wide array of databases and computational tools to mine genomic data in humans, as well as in many other organisms. Dr. Gerstein has led the Data Analytics Core (DAC) of the PsychENCODE Consortium and will direct the Data Analysis Group of this project. **Zhiping Weng, Ph.D.** is a Professor in Biochemistry and Molecular Pharmacology at University of Massachusetts Medical School. She has worked for the last decade on biological problems ranging from genomic to protein-protein interaction analysis. She has participated in the ENCODE project since its inception in 2003, and she is leading the Data Analysis Center (DAC) for ENCODE Phase III (2011-2017) and will co-lead the DAC with Prof. Gerstein for Phase IV (2017-). She has been a member of the PsychENCODE Data Analysis Center, working on the data analysis pipelines and integrative analysis. **Daifeng Wang, Ph.D.** is an Assistant Professor in the Department of Biomedical Informatics at Stony Brook University. He has ~10 years of research experience developing specialized computational and bioinformatics approaches to analyze next generation sequencing datasets and systematically understand gene expression dynamics, gene regulatory networks and circuits in complex biological processes. He was a key participant in the data analysis centers (DACs) for ENCODE, modENCODE, PsychENCODE and KBase when he worked as postdoctoral associate and associate research scientist in Gerstein Lab at Yale University.

Richard Myers, Ph.D. has been a major contributor to basic and disease-applied human genomics for more than 30 years. He has long studied *cis*- and *trans*-acting components of gene expression, initially at mechanistic, biochemical levels, and in the past 20 years, at genome-wide, network levels. He has also made major contributions to the Human Genome Project, high-throughput genetic technologies and studies, and the ENCODE Project. He also has long-standing interests in the genetics and genomics of neurological diseases, including epilepsy, neurodegenerative diseases, autism, and psychiatric disorders. **Gregory Cooper, Ph.D.** has a strong record in the development and use of strategies for genomic analysis and disease genetics. He was a

lead developer of multiple widely used approaches to identify genomic positions and variants (e.g., GERP, CADD) that have important biological roles and/or disease effects. He was also a key contributor to a number of genetic studies of complex human traits, including neurological diseases and expression variation. Drs. Myers and Cooper have worked together on many of these projects for more than a decade, which have benefitted from their complementary and overlapping expertise.

Greg Crawford, PhD is an Associate Professor of Medical Genetics at Duke University. He trained with Francis Collins at NHGRI and was one of the developers of DNase-seq to identify DNase hypersensitive (DHS) sites genome-wide, and was among the first to compare global chromatin maps across cell types(Crawford, Holt et al. 2004, Crawford, Davis et al. 2006, Crawford, Holt et al. 2006, Boyle, Davis et al. 2008, Song, Zhang et al. 2011). He identified variable DHS sites between individuals(McDaniell, Lee et al. 2010), and participated in a study that showed the existence of cQTLs(Degner, Pai et al. 2012) which explain a large fraction of eQTLs in lymphoblastoid cell lines(Degner, Pai et al. 2012). Dr. Crawford extended DNase-seq and ATAC-seq to intact frozen tissues, which will be valuable for this study. He is a member of the PsychENCODE Consortium and is identifying cQTLs in 300 adult brain samples from controls and individuals with schizophrenia.

C.3 Aim 1. Collect a large sample (n=750) of cortical brain tissue from 10-24 post-conception weeks (PCW), and use these tissues for a number of molecular assays.

C.3.1 Source of samples. Samples will come from Kings County Hospital Center in Brooklyn, NY. Kings County is one of the hospitals run by New York City Health and Hospitals and runs multiple pregnancy termination clinics. Dr. Natalie Ohly, runs a second trimester termination clinic that performs ~200 elective terminations per year. Additional clinics at Kings County perform ~1,000 elective terminations per year. All patients will sign consent for permission to use the fetal tissue for medical research and permit the deposition of the genetic, epigenetic and transcriptomic data into public databases. In our experience at LA County Hospital at USC, over 90% of women sign consent to use of the tissue, with many expressing the thought that 'something good might come from a bad situation'. All tissue will be de-identified of personal information of the mother and fetus with staff excluding samples undergoing termination for known genetic abnormalities from entering the study and only providing the estimated date of conception. All collection and use of tissue samples will undergo review and require approval by the respective Institutional IRBs. We propose to collect approximately 188 high-quality samples per year in each of the first four years of the project. We also have a fetal tissue bank of ~150 tissues, at present, from our work on the SCAP-T project, as a back-up, in case of sample collection difficulties.

C.3.2 Sample Collection and Dissection. Samples will be collected at the OR by a trained technician, placed in 4^{o} hypothermosol solution and transported on ice across the street to the Knowles laboratory at SUNY Downstate. Cortical tissue is easily identifiable due to the curvature of its outer most surface, which also contains remnants of pia mater (often with blood vessels). Fresh tissue will be dissected, sliced into slices using a vibrotome and examined under a microscope to ensure the distinctive cortical plate is present as expected. The region between the subventricular zone and the marginal zone will be dissected, and 100 mg tissue snap frozen in liquid N_2 for RNA extraction, other slices will be dissociated with trypsin and cell number and the cell number determined. Aliquots of 1-10 million, as appropriate will then be fixed in 1% formaldehyde for Hi-C and ChIP-seq of H3K27Ac and CTCF. Additional slices will be frozen in liquid N_2 and subsequently pulverized using a Cellcrusher (cellcrusher.com). This pulverized tissue will then be divided and shipped on dry ice to Duke (ATAC-Seq) and HudsonAlpha (TF ChIP-Seq). DNA will be extracted using AutoGen kits, preferably from non-cortical brain regions.

C.3.3 Genotyping. DNA samples extracted from fetal tissue will be genotyped with Illumina Global Screening Arrays (GSA). The GSA contains a backbone of ~660,000 SNPs, which provides LD coverage and imputation accuracy of 0.88 (Africans) to 0.94 (European) (and intermediate for Native Americans and East and South Asians) for SNPs with MAF>1.0% from the Phase 3 1000 Genomes Project (1KGP). These data will provide imputed genotypes from between 6.5M (East Asian) to 12.6M (African) SNPs with MAF>1.0%, and an accuracy score of 0.80, across the wide ranging ethic populations we expect to sample from in Brooklyn. Additionally, there are over 2,000 curated SNPs for malformations and chromosomal abnormalities which will be used to remove fetal samples with chromosomal aberrations (CNV analysis will be used for the same purpose). Genotyping will be performed at the USC Molecular Genomics Core Facility, which has processed over 100,000 genotyping microarrays. We propose to genotype samples in approximately four batches (~yearly) to minimize batch effects requiring correction during analysis.

C.3.4 RNA-seq. RNA will be extracted from the cortical plate region of each fetus using Direct-zol RNA MiniPrep kit (Zymo Research), which was the best performer of five kits tested in the Knowles laboratory. This kit also

efficiently captures short RNA (miRNA, piRNA, etc.), so a portion of the RNA will be retained for future short RNA-Seq. Total RNA will be QCed with an Agilent BioAnalyzer 2200 TapeStation and samples with RINe>8 will converted to Illumina DNA sequencing libraries using Illumina "TruSeq Stranded Total RNA with Ribo-Zero Human" kits (Illumina #RS-122-2301) as we have for >300 samples from the PsychENCODE, and other projects. This is the sample library construction kit used by all of the PsychENCODE projects, which will enhance analyses across developmental time points and brain regions, for both Poly-A and nonPoly-A RNA molecules. To increase throughput, decrease labor cost and improve consistency, we have automated this protocol on a Hamilton STARlet liquid handling robot, with an integrated thermocycler (TRobot, Biometra), which has yielded libraries of high quality and nearly identical insert size distributions and yields. To minimize the number of batches, 24-48 samples will be run at a time. The mass of the libraries will be determined with picoGreen (using STARlet) and the TapeStation, inconsistencies will be broken using KAPA Library Quantification, and equimolar pools of 12 libraries constructed. Each pool will then sequenced in multiple flow cell channels on Illumina sequencers (one HiSeq2500, two HiSeq2000s) at SUNY with a single-end reads of 101bp, plus an index read, yielding a minimum of 40 million reads per library.

C.3.5 ChIP-seq. We propose to perform ChIP-Seq of H3K27Ac for 750 samples, CTCF of 24 samples and 4 samples each for 100 Transcription Factors. H3K27Ac and CTCF were chosen to provide consistency with the existing PsychENCODE data. We have observed very little variation in CTCF peaks across samples, consistent with the high evolutionary conservation of insulators, however, if we observe greater variability in fetal brain, particularly across fetal age, we will increase the sample size accordingly. We will perform ChIP according to the ENCODE best practices (Landt, Marinov et al. 2012), which were worked out, in part, in the Farnham and Myers laboratories. For H3K27Ac and CTCF, ~400,000 cells will be processed for ChIP by crosslinking with formaldehyde, followed by isolation of nuclei, then sonicated using a Bioruptor 200-UCD (Diagenode, Sparta, NJ), and 10ng of the chromatin will be retained as an input control. The rest of the chromatin from each dish will then be immunoprecipitated (IP) for H3K27Ac (Active Motif #39133) or CTCF (Cell Signaling Technology #3418, lot#1), and enrichment will be checked by qPCR and barcoded DNA sequencing libraries will be constructed using KAPA Hyper Prep Kit from KAPA Biosystems, Inc. Library quality and quantity will be checked on an Agilent Bioanalyzer 2200 TapeStation. Libraries will then be pooled and loaded onto flowcells for bridge amplification using an Illumina cBot. Phi-X will be included in each channel to provide real-time QC during sequencing on HiSeq 2000/2500 sequencers. Pools of barcoded libraries will then be sequenced with single-end reads of 50 bp and we will generate a minimum of 20 million uniquely mapped reads for each sample. Although sonicated input DNA will be saved from each sample, we propose to determine the sequence from only ~2 of these. As we have observed in our PsychENCODE project, we expect the data from the input controls will be very consistent across samples and plan to pool the reads and then use the combined data for input control for the entire set of samples. In the unlikely event that the data are inconsistent across samples, we will sequence input controls for every one.

For the ChIP-Seq of 100 TFs, we have chosen to assess 4 individuals to balance two competing goals. On the one hand, we wish to analyze the largest number of TFs, but want comprehensive and accurate data. Because we are forgoing ChIP-seq technical replicates, testing multiple individuals provides an avenue to (albeit imperfectly) measure assay reproducibility, and provides robustness to potential outliers. Further, testing multiple individuals will dramatically improve discovery of polymorphic elements (sites bound by a TF in a subset of people), which are abundant in human populations(Pickrell, Marioni et al. 2010, Pickrell, Pai et al. 2010). With four individuals, we will capture ~95% of binding events that are present in 50% of individuals and most (~61%) events that exist in only 20%; eliminating even one individual would drop coverage of common polymorphic events considerably. As such, while using more than 4 brains would necessarily drop the number of TFs that we could assay (in a given budget), fewer individuals would reduce data quality and comprehensiveness. The Myers and Farnham labs, and others, in the ENCODE Consortium have successfully performed ChIP-seq on antibodies to a total of 331 distinct TFs (encodeproject.org). Rather than pre-selecting a list of specific TFs to test in each sample, we will follow a strategy that provides a universal dataset across all samples and flexibility over time to account for new antibodies that become available, knowledge gleaned from other groups and other projects, and intersections with other investigators that may be studying TF biology in relation to psychiatric diseases. Furthermore, we will use the BrainSpan poly-A and initial total RNA-Seq data from this project to determine which TFs are most highly expressed during 10-24 PCW. In a preliminary analysis of 96 adult brains from the Pritzker Brain Bank, we found 250 TFs that have a median expression of at least 2 FPKM (a threshold at which TFs typically become "chippable") and 133 of these are highly expressed (>10 FPKM). The tested set of 100 TFs is likely to include RNA Polymerase 2 ("Pol2"), a core component of the machinery that binds to active promoters and generates transcripts; TAF1, a general TF enriched at promoters; and p300, a factor that binds to

enhancers. TF ChIP-Seq will be performed with ENCODE best practices and the resulting libraries will be sequenced to a depth of 20-30 million mapped reads at HudsonAlpha.

C.3.6 ATAC-seq. ATAC-Seq will be performed in the laboratory of Dr. Crawford at Duke University, where the procedures are well-established(Crawford, Holt et al. 2004, Crawford, Davis et al. 2006, Crawford, Holt et al. 2006, Birney, Stamatoyannopoulos et al. 2007, Boyle, Davis et al. 2008, Shibata and Crawford 2009, Song and Crawford 2010, Myers, Stamatoyannopoulos et al. 2011, Song, Zhang et al. 2011, Zhang, Wu et al. 2012), having generated over 500 DNase-seq and ATAC-seq libraries, and have successfully generated ATAC-seq from <20 mg of frozen brain tissue. Pulverized tissue will be thawed in glycerol containing nuclear isolation buffer to stabilize structure(Zhang, Wu et al. 2012). After filtering out larger debris, nuclei are washed with RSB buffer and incubated with Tn5 transposase as part of the standard ATAC-seq protocol(Buenrostro, Giresi et al. 2013). We will include a small number of replicate experiments to ensure we pass strict QC standards used for ENCODE project. ATAC-seq allows for longer sequencing read lengths, which will be helpful to identifying variants and chromatin QTLs. Libraries will be sent to SUNY Downstate for sequencing on and sequenced to a depth of 60-80 million reads. Any experiments that fail QC metrics will be repeated.

C.3.7 Hi-C. Since the first introduction of methods to study the three-dimensional chromosomal structure, extensive work has gone into both improving the efficiency and introducing changes that allow for high-throughput methods to study the interactions genome-wide. One such adaptation is *in situ* Hi-C, which allows for crosslinking, digestion, and ligation all to occur within the intact nucleus. In contrast, traditional Hi-C relies on the dilution of the cross-linked and digested material to prevent random ligation products. By performing these reactions within the intact nuclei, fragments that become ligated are also in close proximity within the nuclei. As we have done in our PsychENCODE project, we will use an *in situ* Hi-C protocol (Rao, Huntley et al. 2014) with a 4-cutter restriction enzyme (MboI) to generate high-resolution 3D genomic landscapes from fetal brain. Using ~3M cells per library, we generated two Hi-C libraries (one for schizophrenia patient, one for control) for our present PsychENCODE Project. Each library was sequenced with ~900 million paried-end 100bp reads and QC'ed with HiCUP (Wingett, Ewels et al. 2015). We had a high percentage of paired reads; after removing duplicates we had 689M unique reads, of which 568M reads (82%) were properly paired, which exceeds the 50 to 70% of paired reads reported by other groups. Of these, 490M (86%) were valid pairs (not dangling ends, circularized, internal, re-ligation, contiguous sequence, wrong size); libraries with more than 50% of valid pairs are considered as good libraries. For this project, we will use this same molecular protocol to generate 12 Hi-C libraries across the developmental window we are sampling. Geschwind and colleagues have recently reported the results of the 3D genome structure from 17-18 PCW using Hi-C(Won, de la Torre-Ubieta et al. 2016). We will extend these observations by examining the 3D genome structure in four time windows, 10-13, 14-18, 19-21 and 22-24 PCWs, using 3 samples at each time point. Sixty percent of the peaks between the Hi-C libraries made from the cortical plate and germinal zone overlapped in the Geschwind study, hence we have made the decision to increase our temporal resolution across the our sampling frame, at the expense of regional resolution, as little is known about the temporal dynamics of the 3D structure of the genome in early brain development.

C.4 Aim 2. Data processing and bioinformatics analysis.

C.4.1 Data processing and analysis to identify fetal QTLs. We will analyze all the generated data and integrate the genotype data with other data for QTL analysis. For doing this we will build on our considerable experience in ENCODE, modENCODE, 1000 Genomes, KBase, BrainSpan and PsychENCODE in doing integrative and comparative analysis. We will use the standardized RNA-seq processing pipelines including data organization, format conversion, and quality assessment which will then be run in large-scale on the PDC (Protected Data Cloud) to process the RNA-Seq data first. Specifically, we will employ STAR (Dobin, Davis et al. 2013) to uniquely align the filtered reads to their reference genome and RSEM (Li and Dewey 2011) to quantify expression profiles of each type of annotation entry retrieved from the latest release of the GENCODE project. Additional quality control measures will be introduced to assess potential issues including sequencing error rate, ribosomal contamination, DNA contamination and gene coverage uniformity and the correlation between technical and/or biological replicates.

We will use the ENCODE ChIP-seq data processing pipeline developed by both Gerstein lab and Zhiping Weng's lab. This pipeline includes steps of quality assessment, trimming the contamination, alignment of the fastq files, peak calling and downstream analysis such as peak comparison, peak annotation, motif analysis and super-enhancers identification. The Gerstein lab developed PeakSeq (Rozowsky, Euskirchen et al. 2009), a versatile tool for identification of TF binding sites and a standard peak calling program used by the ENCODE and modENCODE consortia for ChIP-Seq datasets (Rozowsky, Euskirchen et al.). We will also use a new peak caller MUSIC (Harmanci, Rozowsky et al.) recently developed in Gerstein lab. MUSIC performs multiscale

decomposition of ChIP signals to enable simultaneous and accurate detection of enrichment at a range of narrow and broad peak breadths. This tool is particularly applicable to studies of histone modifications and previously uncharacterized transcription factors, both of which may display both broad and punctate regions of enrichment. We have already implemented this pipeline to process ChIP-Seq data from both PsychENCODE and BrainSpan.

Moreover, we have developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers based on our past experience in non-coding annotation, as part of our 10-year history with the ENCODE and modENCODE projects (Yip, Alexander et al.). We will develop a framework using matched filter to aggregate the signal of histone modifications on massively parallel reporter assays (MPRA) peaks flanking enhancers. The method will identify an enriched peak-trough-peak ("double peak") signal at active enhancers in different ChIP-Seq experiments for various histone modifications. We will combine the normalized matched filter scores from our different epigenetic marks (H3K27Ac and ATAC-Seq (similar to DHS) associated with active regulatory regions by the Roadmap Epigenomics Mapping, PsychENCODE project and the ENCODE Consortia, using a linear SVM. The normalized matched filter score for each epigenetic feature in a particular region will be scaled by its optimized weight and added together to form the discriminant function. The sign of the discriminant function will be used to predict whether a region is an enhancer. Features with larger weights (eg. H3K27Ac) are predicted to be more important in discriminating enhancers from non-regulatory regions in the model. We will use the better enhancer definition provided by the Epigenome Roadmap (Leung, Jung et al. 2015, Roadmap Epigenomics, Kundaje et al. 2015, Ziller, Edri et al. 2015), and more recently from ENCODE projects. In particular, we will develop a new machine learning framework that combines pattern recognition within the signal of various epigenomic features and transcription of enhancer RNA (eRNA, some of which, particularly the 1D-eRNA, which we will detect) with sequence-based features to predict active enhancers across different brain regions and other tissues in the Epigenome Roadmap project. The pattern within the signal of different epigenetic datasets will be computed from regulatory regions identified using different massively parallel assays and we will determine to what extent this pattern is conserved across a diverse set of tissues. This method will be used to predict fetal brain specific active enhancers based on H3K27Ac ChIP-Seq datasets generated as part of this grant, as well as ChIP-seq generated by the Epigenome Roadmap, ENCODE and present PsychENCODE projects.

Moreover, we have implemented a standard eQTL analysis pipeline in Gerstein lab for our current PsychENCODE capstone projects where we are generating an eQTL map of adult frontal cortex using ~2,300 samples from the PsychENCODE, CommonMind and GTEx projects, and genomic privacy paper (Harmanci and Gerstein). We will use this pipeline to identify various QTLs, including eQTLs for long RNAs, splicing QTLs, ChIP-QTLs and ATAC-QTLs, in early human brain development. Genotypes will be imputed using the *ricopili* pipeline (Rapid Imputation Consortium Pipeline), in order to streamline quality control, genotype imputation, and statistical analysis of genome-wide single nucleotide polymorphism (SNP) data. *Ricopili* consists of four primary, independent modules: (1) pre-imputation data processing and quality control; (2) principal components analysis (PCA); (3) genotype imputation of untyped variants; and (4) post-imputation statistical analysis. Briefly, in the pre-imputation step, input genotype data (PLINK binary format) is reformatted for downstream analysis, and initial summaries of classic technical parameters (e.g. minor allele frequency, per-individual and per-site missing rates, case/control missingness, Hardy-Weinberg equilibrium) are produced. The second module consists of data filtering and relatedness testing, followed by PCA using EIGENSTRAT(Price, Patterson et al. 2006) to identify ancestry outliers and any detectable population substructure. Prior to imputation, SNP positions, identifiers, and alleles are aligned to the relevant reference genome assembly (using LiftOver), and genotype data is divided into overlapping 5 megabase (Mb) segments (~1000) for subsequent, parallel haplotype pre-phasing and imputation using SHAPEIT2/IMPUTE2 (Delaneau, Marchini et al. 2011, Howie, Fuchsberger et al. 2012). We will use 1000 Genome phase 3 as the general reference panel for imputation. We will also try the recently released HRC Reference Panel for imputation of rare SNPs. It is important to note that phasing/imputation is more difficult in persons of recent African ancestry, as their greater genetic diversity (and lower linkage disequilibrium) reduces the accuracy of haplotype estimation (Howie, Fuchsberger et al. 2012). Several alternative algorithms, including MaCH-admix (Liu, Li et al. 2013, Roshyara, Horn et al. 2016) offer improved imputation accuracy for admixed populations, and have previously been integrated into the *ricopili* pipeline.

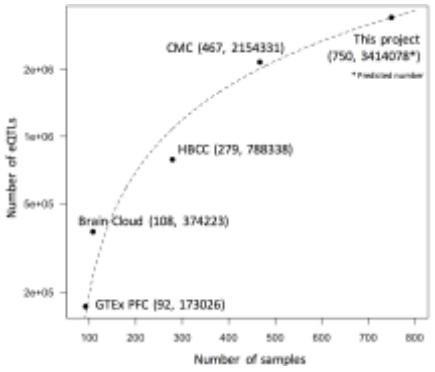

Figure 3 Predication of eQTL discovery

We will use Matrix eQTL and/or fastQTL package for eQTL analysis. The gene expression matrix will be normalized according to gender, Age, RNA Integrity Number (RIN) and library

preparation batch (LIB) for eQTL analysis. Probabilistic Estimation of Expression Residuals (PEER) factors, ancestry vectors, age and gender will be used as covariates input for Matrix eQTL/fastQTL. Based on our sample size, we will calculate both cis-eQTL and trans-eQTL. Finally we will correct for the multiple hypothesis tests of SNPs in LD for a given gene for eQTL analysis. Using the data presented in the CommonMind Consortium paper (Fromer, Roussos et al. 2016), we expect to discover ~3.4M eQTLs (many will be in LD with each other, Figure 3).

Similarly, QTLs for H3K27Ac and ATAC-Seq peaks (cQTLs; chromatin QTLs) will be evaluated using fastQTL, (Ongen, Buil et al. 2016) which utilizes a ß approximation of permutations to determine significance. Peaks will be normalized and regressed on SNP dosage in a 5 kb window, controlling for 10 PCs from PCA of peaks and ancestry PCs from PCA of SNP array data. Only the most significant SNP for each peak will be retained. To control for testing multiple peaks, we will apply the Storey and Tibshirani correction (Storey and Tibshirani 2001) to the ß approximated permutation P-values. To date, we have performed cQTL analysis for ATAC-seq samples generated from 300 adult brain samples from controls and individuals, and have identified over 6,000 cQTLs (**Figure 4**).

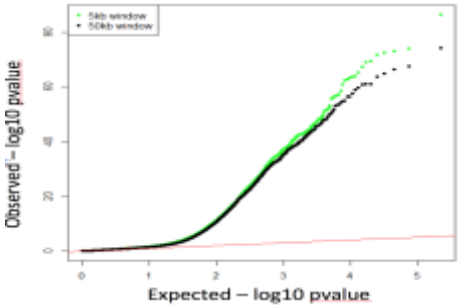

Figure 4 cQTLs identified from psychENCODE ATAC-seq datasets generated from 300 brain samples

Comparing cQTL & eQTLs. eQTL analysis identifies variants associated with expression levels where each locus consists of multiple SNPs in high LD. For each eQTL, we will first determine whether any SNPs within that locus are contained within a regulatory element identified in the same sample. For each ATAC-associated eQTL, we will determine whether: (i) the regulatory element shows tissue- or individual-specific chromatin changes; (ii) the SNP in the regulatory element shows evidence of allelic imbalance; and (iii) the SNP within the regulatory element is a cQTL. These features provide evidence that a SNP might be causal for the linked expression changes. We will also test eQTL SNPs outside of regulatory elements for association with cQTL as they could suggest a mechanism for altering regulatory activity leading to expression change.

Calling 3D Genome Structure from Hi-C data. Hi-C data on fetal tissues will be processed using HiC-Pro (Servant, Varoquaux et al. 2015), and HiCUP (Wingett, Ewels et al. 2015) pipeline tools, which map the raw reads, perform quality control analyses, normalize the interaction frequency, and identify valid pairs. HiCPlotter (Akdemir and Chin 2015) and HiTC (Servant, Lajoie et al. 2012) software programs will be used for plotting to produce normalized interaction frequency heatmaps. To call topologically associating domains and subdomains, Domain callers (Dixon, Selvaraj et al. 2012) and TopDom (Shin, Shi et al. 2016) programs will be used. For identification of significant interactions and differentially connected loopings, GOTHiC (http://bioconductor.org/packages/release/bioc/html/GOTHiC.html) and diffHiC (Lun and Smyth 2015) R packages will be used.

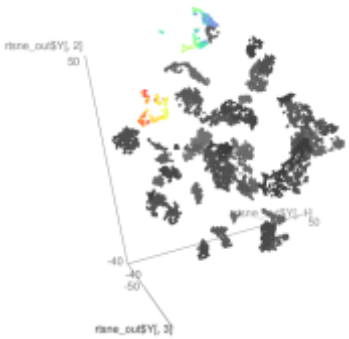

Figure 5 Clustering tissue samples of BrainSpan and GTEx based on their similarity of gene expression over first three tSNE dimensions. The red and orange samples correspond to the early developmental samples (i.e., prenatal samples in BrainSpan), and form a separated cluster. The blue and cyan samples correspond to the infant and adult BrainSpan samples, and are clustered together. The samples from other GTEx tissues also form their specific clusters (grey clusters).

C.4.2 Early brain expression dynamics from differentially expressed genes. As mentioned above, gene expression in early brain development is very different from later in life. As a further example of this, we dimensionally reduced all the BrainSpan and GTEx gene expression data using tSNE and plotted the first three dimensions (Figure 5). The red and orange samples correspond to the early developmental samples (i.e., prenatal samples in BrainSpan), and forms a separate cluster. The blue and cyan samples correspond to the infant and adult BrainSpan samples, and also were clustered together. The samples from other GTEx tissues also form specific clusters (in grey). This suggests that the early developmental samples share specific gene expression patterns, which is different from other brain developmental stages and tissues. Thus, we plan to discover the dynamics of gene expression in early brain development in this sub-aim.

We have substantial experience in developing computational approaches to identify specific dynamic patterns of gene expression. We have developed a novel clustering algorithm, OrthoClust to simultaneously cluster multi-layer networks (Yan, Wang et al.). We applied OrthoClust to developmental gene expression datasets of worm (C. elegans) and fruitfly (D. melanogaster), and discovered the cross-species and species-specific gene

co-expression modules (Figure 6). We also found the modular eigengenes, revealing the systematically gene expression and regulation dynamics during embryonic development. In 2016, we also developed another novel computational method, DREISS to identify the gene expression dynamics driven by internal and external regulatory networks (Wang, He et al.). In particular, we applied DREISS to the time-series gene expression datasets of C. elegans and D. melanogaster during their embryonic development (Figure 7). We analyzed the expression dynamics of the conserved, orthologous genes (orthologs), seeing the degree to which these can be accounted for by orthologous (internal) versus species-specific (external) TFs. We found that between two species, the orthologs have matched, internally driven expression patterns, but very different species-specific, externally driven ones. This is particularly true for genes with evolutionarily ancient functions (e.g. the ribosomal proteins), in contrast to those with more recently evolved functions (e.g., cell-cell communication).

We plan to use the OrthoClust, for analyzing the co-expression networks in an integrated fashion by utilizing the orthology relationships of genes between fetal and adult stages (using the data from the ~2,300 sample PsychENCODE Capstone project) and comparison tissues (from GTEx and ENCODE) (Yan, Wang et al.), which is likely to identify differentially expressed genes in fetal brain versus adult brain and non-brain tissues. These genes can be the biomarkers for distinguishing different tissues. We will first normalize and correct the batch effects of the gene expression data using COMBAT (Johnson, Li et al.). We have developed a number of advanced methods for normalization, analysis, and comparison of RNA-seq profiles. In particular: 1) incRNA, a method that predicts novel ncRNAs using known ncRNAs of various biotypes as a training set (Lu, Yip et al. 2011); 2) FusionSeq, a pipeline to detect transcripts that arise due to trans-splicing or chromosomal translocations (Sboner, Habegger et al. 2010, Pflueger, Terry et al. 2011); 3) IQSeq, a transcript isoform quantification tool that uses an EM algorithm to resolve the maximum likelihood expression level of individual transcript isoforms (Du, Leng et al. 2012); 4) Pseudo-seq which addresses the issue of quantification of pseudogene and repetitive region expression (Sisu, Pei et al. 2014); and 5) the Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genomic signal tracks (Jee, Rozowsky et al. 2011). In addition, we contributed to the development of a classification and analysis scheme for “spike” event patterns in omics data with longitudinal profiles(Chen, Mias et al. 2012).

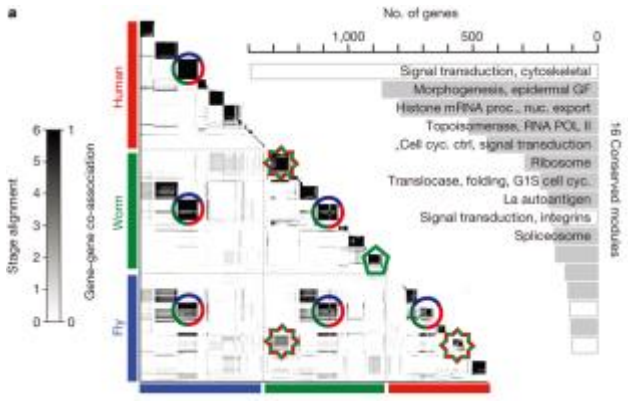

Figure 6 Cross-species gene co-expression network clustering. Left, human, worm and fly gene–gene co-association matrix; darker colouring reflects the increased likelihood that a pair of genes are assigned to the same module. A dark block along the diagonal represents a group of genes within a species. If this is associated with an off-diagonal block then it is a cross-species module (for example, a three-species conserved module is shown with a circle and a worm–fly module, with a star). However, if a diagonal block has no off-diagonal associations, then it forms a species-specific module (for example, green pentagon). Right, the Gene Ontology functional enrichment of genes within the 16 conserved modules is shown. GF, growth factor; nuc., nuclear; proc., processing.

After normalization, we then want to analyze the dynamic patterns of gene expression over our smapling window of early brain development. To identify these patterns, we will develop a new method/pipeline, called DynamBrains. For details, we will first identify the highly expressed genes (DEGs) during brain development and also across other tissues. The DEGs displaying very different expression levels at early stages are potentially regulated by early brain gene regulatory mechanisms. Because individual gene expression might be very noisy, we will further identify the systematic early brain expression patterns from gene co-expression network analysis. Specifically, we will construct a gene co-expression network in which genes are connected if they have high correlated expression profiles during brain development. We will cluster this network into gene co-expression modules using WGCNA. The eigengenes of gene co-expression modules thus represent the systematic developmental expression dynamic patterns. We will also analyze the enriched pathways and functions for each module, and associate them with the module’s eigengene. We will find modules (with associated gene expression signatures) enriched in fetal and adult brain. The modules whose eigengenes showing different pre-natal and post-natal expression are defined as early brain modules. The enriched pathways and functions of early brain modules are potentially related to the early brain development.

C.4.3 Dynamic modeling of brain developmental gene regulatory networks by integrating adult data from GTEx, PsychENCODE, etc. After finishing the fetal QTL analysis and discovering early brain expressed genes, we will perform integrative and comparative analysis of fetal gene and QTL and adult data from

ENCODE, GTEx, CommonMind and PsychENCODE project. We will develop or use our existing tools to form further interrogative analysis to model and identify how eQTLs influence early brain gene expression via the gene regulatory networks. We will also build a comparison of fetal QTL and adult QTL maps.

We have comprehensive experience integrating transcriptomic, metabolomics, and proteomic data. We integrated unknown metabolites, which can constitute as much as 50% of spectral features (Chen, Mias et al. 2012), with transcriptomics profiles from different experimental conditions (Gianoulis, Griffin et al. 2012). By defining statistics to correlate the co-occurrence patterns of metabolites and genes we generated hypotheses about the identities of unannotated biosynthetic pathways. In addition, we have experience with the analysis of proteomic data and its integration with transcriptomics (Smith, Cheung et al. 2007, Wu, Hwang et al. 2007, Sboner, Karpikov et al. 2009, Kitchen, Rozowsky et al. 2014). This allowed us to identify previously uncharacterized proteins in a temporally and spatially resolved manner(Wu, Hwang et al. 2007).

We also have made extensive use of machine-learning to generate models from integrated datasets. For example, we integrated ENCODE data on transcription factor (TF) binding, histone modifications, and target gene expression to establish regulatory relationships using a probabilistic model we named TIP (Target Identification from Profiles) (Cheng, Min et al. 2011). We identified potential enhancers from distal gene regions and we used these modules to quantify the relationship between TF binding and gene expression(Cheng and Gerstein 2011, Cheng, Alexander et al. 2012, Consortium 2012, Yan, Wang et al. 2014). We integrated these data types with protein-protein interaction and transcriptional regulation networks (Gerstein, Lu et al. 2010, Cheng, Shou et al. 2011, Cheng, Yan et al. 2011, Dong, Greven et al. 2012). This allowed us to group TFs into histone-sensitive and -insensitive classes that refined the prediction of gene-regulation targets and effects. Finally, we were able to build cross-organism integrative chromatin models (Yan, Wang et al.).

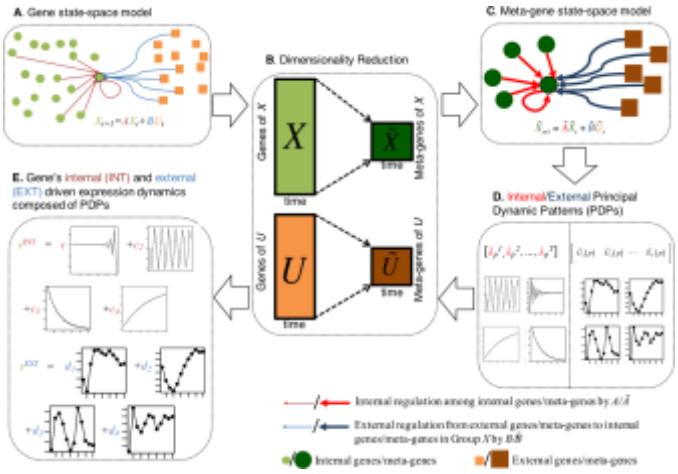

Figure 7 DREISS: Using State-Space Models to Infer the Dynamics of Gene Expression Driven by External and Internal Regulatory Networks. (A) DREISS models temporal gene expression dynamics using state-space models in control theory. The "state" refers to the expressions for a large group of genes of interest, such as the worm-fly orthologous genes investigated here. The "control" refers to any other group of genes that contribute to gene expressions of the "state", such as the species-specific TF studied here. (B) it then projects high-dimensional gene expression space to lower-dimensional meta-gene expression spaces using dimensionality reduction techniques. (C) it derives the effective state-space models for meta-genes so that model parameters can be estimated. (D) it then identifies the meta-gene expression dynamic patterns; i.e., canonical temporal expression trajectories driven by "state" (internal) and by "control" (external) based on the analytic solutions to estimated models. (E) it finally calculates the coefficients of genes for the dynamic patterns of linear transformations between genes and meta-genes.

We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets (Mu, Lu et al. 2011, Gerstein, Kundaje et al. 2012, Yip, Cheng et al. 2012). We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations (Mu, Lu et al.). In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region (Consortium). Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery (Johnson, Li et al. 2007, Khurana, Fu et al. 2013). In recent studies (Khurana, Fu et al. 2013, Fu, Liu et al. 2014), we have integrated and extended these methods to develop a prioritization pipeline called FunSeq. It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). It then identifies potentially deleterious variants in many non-coding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. It also detects their disruptiveness to TF binding sites (both loss-of and gain-of function events). Integrating large-scale data from various resources (including ENCODE and The 1000 Genomes Project) with cancer genomics data, our method is able to prioritize the known TERT promoter driver mutations, and it scores somatic recurrent mutations higher than those that are non-recurrent. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast, and prostate cancer samples (Khurana, Fu et al.). Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.

We will use the data generated in this project, combined with other projects like PsychENCODE, CommonMind, BrainSpan and GTEX dataset to expand our understanding of the molecular activity of cells in the human brain by identifying genes that predominantly express one allele and exploring the potential clinical relevance of such allelic imbalance, by examining the GWAS and sequencing data of the brain disorders. We will focus on quantifying differences in transcription between maternal and paternal alleles using the matched genotype and RNA-sequence data available in the PsychENCODE dataset. We will integrate similar analyses of allelic imbalance performed by our lab using the matched genotype and RNA-seq data produced by this grant, PsychENCODE and CommonMind project. We expect to be able to generalize results obtained from this grant using the 11 distinct cortical and 5 sub-cortical regions of the healthy adult human brain available in BrainSpan. Integrating data at this scale requires large amounts of RNA expression and matching genotype information from different cell-types, brain regions, developmental stages and/or tissues. To that end we will also incorporate data and results from the GTEx project in order to further broaden our survey of allelic imbalance to identify potentially brain-specific allelic effects. Once compiled, this allelic survey of unprecedented resolution will be of substantial benefit to the wider research community. By integrating these large-scale projects data for meta-analysis, we will get have much larger sample size and enables more powerful analysis. Larger sample size will increase the number of cis-eQTLs could be detected and improve accuracy of trans-eQTL detection. Moreover, larger sample size will increase the possibility of discovering more allelic regions. We will also be able to conduct better analysis for Gender-Specific Gene Expression by using more sample size.

It is known that gene regulatory factors work cooperatively, forming a complex regulatory circuit controlling gene expression. We developed Loregic, a general-purpose method to characterize the cooperatively of such regulatory factors (Wang, Yan et al. 2015). Finally, we will identify the gene regulatory logics using Loregic that drive the tissue types such as the biomarker genes associated with specific tissues (Huffman, Koves et al.).

C.4.4 Integration with knowledge of the psychiatric diseases. The enrichment of GWAS signals in regulatory elements has been reported in previous studies (Schizophrenia Working Group of the Psychiatric Genomics 2014, Andreassen, Thompson et al. 2015, Neale and Sklar 2015, Fromer, Roussos et al. 2016). In the BrainSpan project, we found signals for Schizophrenia, Bipolar disorders and Parkinson's disease are enriched in both developmental and adult brain regulatory elements in disease-specific patterns, but the same was not observed for Type 2 diabetes, Coronary artery disease (CAD) and Asthma. To further investigate the enrichment pattern we observed, we will conduct an association analysis between regulatory elements, eQTLs and loci implicated through genome-wide association studies on several prevalent psychiatric diseases. We will use the fetal data in this grant and adult data with schizophrenia, bipolar disorder and autism spectrum disorders from PsychENCODE for this analysis. We will use these data to compare the relative contribution of pre- and postnatal gene expression to signals in the GWAS, exome and WGS studies of the psychiatric disorders. We will try to identify eQTLs that could explain associations with psychiatric diseases. Previous studies showed that most of currently available eQTL maps do not yet provide enough power or developmental diversity to provide clear hypotheses for associations between eQTLs and psychiatric diseases. We will merge the GWAS signals with eQTL maps from our PsychENCODE Capstone projects combined with the fetal eQTL catalogues in this proposed grant. We expect that by using these two more powerful eQTL analysis we will be able to better understanding the eQTL associations with psychiatric diseases.

C.5 Aim 3. Provide an easy-to-use, web-based informatics framework for communication of the raw and computed data of this PsychENCODE project to other neuroscientists.
This project, like the parent PsychENCODE Consortium will share information among consortium members and the broader research community through a website (www.psychENCODE.org) and a knowledge portal. The website will provide descriptive information about this and every project, news about the Consortium, and up-to-date information on tissue banks, protocols, and sample sizes. The knowledge portal, developed by the PsychENCODE Data Coordinating Center at Sage Bionetworks (synapse.org/ - !Synapse:syn2787333), is designed to provide a centralized environment for accessing data, protocols, and analytical output to enable collaboration among and beyond consortium members. Data from this project will be released to the broader research community at yearly intervals, after sufficient QC of the data to ensure it is of high quality. Access to this human data will be shared using a controlled access mechanism that complies with regulatory requirements and governance policies regarding protections of personal information. We, like all PsychENCODE investigators, will ensure that data can be visualized through Genome Browsers such as the UCSC Genome Browser and/or IGV.

C.6 Elements unique to this site (Yale University).
The Yale site will consist of investigators in the labs of Mark Gerstein at Yale University, Zhiping Weng at the University of Massachusetts Medical School and Daifeng Wang at Stony Brook University to form a Data

Analysis Group. Dr. Gerstein's lab will develop a number of standardized pipelines and quality control metrics,provide a platform and infrastructure for uniform processing of the data and running the pipelines and focus on the discovery of fetal brain specific genes, the aggregated quantitative trait locus (QTL) analysis, and integration of all data sets for meta-analysis. Dr. Weng's lab will support the enhancer analysis and annotating disease-associated enhancers and discovering functional genomic elements associated with psychiatric diseases using an integrative approach. Dr. Wang's lab will work on the analysis to identify early brain gene expression dynamics, gene co-expression network analysis, and model the gene regulatory networks driving early brain development. The bioinformatics group will give feedback to data production groups on data quality and support all groups for integrative data analysis.

C.7 Timeline.

	Year 1	Year 2	Year 3	Year 4	Year 5
Samples	188	188	188	188	
Genotyped	188	188	188	188	
RNA-Seq	188	188	188	188	
H3K29Ac	188	188	188	188	
CTCF	6	6	6	6	
TF ChIP-Seq	100	100	100	100	
ATAC-Seq	188	188	188	188	
Hi-C	3	3	3	3	
Data Analyses	+++	+++++	+++++	+++++	+++++

References

Akdemir, K. C. and L. Chin (2015). "HiCPlotter integrates genomic data with interaction matrices." Genome Biol **16**: 198.

Andreassen, O. A., et al. (2015). "Correction: Improved Detection of Common Variants Associated with Schizophrenia and Bipolar Disorder Using Pleiotropy-Informed Conditional False Discovery Rate." PLoS Genet **11**(11): e1005544.

Birney, E., et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Boyle, A. P., et al. (2008). "High-resolution mapping and characterization of open chromatin across the genome." Cell **132**: 311-322.

Buenrostro, J. D., et al. (2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." Nat Methods **10**(12): 1213-1218.

Chen, R., et al. (2012). "Personal omics profiling reveals dynamic molecular and medical phenotypes." Cell **148**(6): 1293-1307.

Personalized medicine is expected to benefit from combining genomic information with regular monitoring of physiological states by multiple high-throughput methods. Here, we present an integrative personal omics profile (iPOP), an analysis that combines genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles from a single individual over a 14 month period. Our iPOP analysis revealed various medical risks, including type 2 diabetes. It also uncovered extensive, dynamic changes in diverse molecular components and biological pathways across healthy and diseased conditions. Extremely high-coverage genomic and transcriptomic data, which provide the basis of our iPOP, revealed extensive heteroallelic changes during healthy and diseased states and an unexpected RNA editing mechanism. This study demonstrates that longitudinal iPOP can be used to interpret healthy and diseased states by connecting genomic information with additional dynamic omics activity.

Cheng, C., et al. (2012). "Understanding transcriptional regulation by integrative analysis of transcription factor binding data." Genome Research **22**(9): 1658-1667.

Cheng, C. and M. Gerstein (2011). "Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells." Nucleic Acids Research **40**(2): 553-568.

Cheng, C., et al. (2011). "TIP: A probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles." Bioinformatics **27**(23): 3221-3227.

Cheng, C., et al. (2011). "Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors." Genome Biology **12**(11): R111.

Cheng, C., et al. (2011). "A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets." Genome Biology **12**(2): R15.

Consortium, E. P. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

Consortium, G. T. (2015). "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans." Science **348**(6235): 648-660.

Understanding the functional consequences of genetic variation, and how it affects complex human disease and quantitative traits, remains a critical challenge for biomedicine. We present an analysis of RNA sequencing data from 1641 samples across 43 tissues from 175 individuals, generated as part of the pilot phase of the Genotype-Tissue Expression (GTEx) project. We describe the landscape of gene expression across tissues, catalog thousands of tissue-specific and shared regulatory expression quantitative trait loci (eQTL) variants, describe complex network relationships, and identify signals from genome-wide association studies explained by eQTLs. These findings provide a systematic understanding of the cellular and biological consequences of human genetic variation and of the heterogeneity of such effects among a diverse set of human tissues.

Crawford, G. E., et al. (2006). "DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays." Nature Methods **3**: 503-509.

Crawford, G. E., et al. (2004). "Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites." Proceedings of the National Academy of Sciences of the United States of America **101**(4): 992-997.

Crawford, G. E., et al. (2006). "Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)." Genome Research **16**(1): 123-131.

Degner, J. F., et al. (2012). "DNase I sensitivity QTLs are a major determinant of human expression variation." Nature **482**(7385): 390-394.

Delaneau, O., et al. (2011). "A linear complexity phasing method for thousands of genomes." Nat Methods **9**(2): 179-181.

Human-disease etiology can be better understood with phase information about diploid sequences. We present a method for estimating haplotypes, using genotype data from unrelated samples or small nuclear families, that leads to improved accuracy and speed compared to several widely used methods. The method, segmented haplotype estimation and imputation tool (SHAPEIT), scales linearly with the number of haplotypes used in each iteration and can be run efficiently on whole chromosomes.

Dixon, J. R., et al. (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." Nature **485**(7398): 376-380.

Dobin, A., et al. (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.

MOTIVATION: Accurate alignment of high-throughput RNA-seq data is a challenging and yet unsolved problem because of the non-contiguous transcript structure, relatively short read lengths and constantly increasing throughput of the sequencing technologies. Currently available RNA-seq aligners suffer from high mapping error rates, low mapping speed, read length limitation and mapping biases. RESULTS: To align our large (>80 billon reads) ENCODE Transcriptome RNA-seq dataset, we developed the Spliced Transcripts Alignment to a Reference (STAR) software based on a previously undescribed RNA-seq alignment algorithm that uses sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure. STAR outperforms other aligners by a factor of >50 in mapping speed, aligning to the human genome 550 million 2 x 76 bp paired-end reads per hour on a modest 12-core server, while at the same time improving alignment sensitivity and precision. In addition to unbiased de novo detection of canonical junctions, STAR can discover non-canonical splices and chimeric (fusion) transcripts, and is also capable of mapping full-length RNA sequences. Using Roche 454 sequencing of reverse transcription polymerase chain reaction amplicons, we experimentally validated 1960 novel intergenic splice junctions with an 80-90% success rate, corroborating the high precision of the STAR mapping strategy. AVAILABILITY AND IMPLEMENTATION: STAR is implemented as a standalone C++ code. STAR is free open source software distributed under GPLv3 license and can be downloaded from http://code.google.com/p/rna-star/.

Dong, X., et al. (2012). "Modeling gene expression using chromatin features in various cellular contexts." Genome Biology **13**(9): R53.

Du, J., et al. (2012). "IQSeq: integrated isoform quantification analysis based on next-generation sequencing." PLOS ONE **7**(1): e29175.

With the recent advances in high-throughput RNA sequencing (RNA-Seq), biologists are able to measure transcription with unprecedented precision. One problem that can now be tackled is that of isoform quantification: here one tries to reconstruct the abundances of isoforms of a gene. We have developed a statistical solution for this problem, based on analyzing a set of RNA-Seq reads, and a practical implementation, available from archive.gersteinlab.org/proj/rnaseq/IQSeq, in a tool we call IQSeq (Isoform Quantification in next-generation Sequencing). Here, we present theoretical results which IQSeq is based on, and then use both simulated and real datasets to illustrate various applications of the tool. In order to measure the accuracy of an isoform-quantification result, one would try to estimate the average variance of the estimated isoform abundances for each gene (based on resampling the RNA-seq reads), and IQSeq has a particularly fast algorithm (based on the Fisher Information Matrix) for calculating this, achieving a speedup of ~ 500 times compared to brute-force resampling. IQSeq also calculates an information theoretic measure of overall transcriptome complexity to describe isoform abundance for a whole experiment. IQSeq has many features that are particularly useful in RNA-Seq experimental design, allowing one to optimally model the integration of different sequencing technologies in a cost-effective way. In particular, the IQSeq formalism integrates the analysis of different sample (i.e. read) sets generated from different technologies within the same statistical framework. It also supports a generalized statistical partial-sample-generation function to model the sequencing process. This allows

one to have a modular, "plugin-able" read-generation function to support the particularities of the many evolving sequencing technologies.

Fromer, M., et al. (2016). "Gene expression elucidates functional impact of polygenic risk for schizophrenia." Nat Neurosci **19**(11): 1442-1453.

Over 100 genetic loci harbor schizophrenia-associated variants, yet how these variants confer liability is uncertain. The CommonMind Consortium sequenced RNA from dorsolateral prefrontal cortex of people with schizophrenia (N = 258) and control subjects (N = 279), creating a resource of gene expression and its genetic regulation. Using this resource, approximately 20% of schizophrenia loci have variants that could contribute to altered gene expression and liability. In five loci, only a single gene was involved: FURIN, TSNARE1, CNTN4, CLCN3 or SNAP91. Altering expression of FURIN, TSNARE1 or CNTN4 changed neurodevelopment in zebrafish; knockdown of FURIN in human neural progenitor cells yielded abnormal migration. Of 693 genes showing significant case-versus-control differential expression, their fold changes were </= 1.33, and an independent cohort yielded similar results. Gene co-expression implicates a network relevant for schizophrenia. Our findings show that schizophrenia is polygenic and highlight the utility of this resource for mechanistic interpretations of genetic liability for brain diseases.

Fu, Y., et al. (2014). "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer." Genome Biol **15**(10): 480.

Identification of noncoding drivers from thousands of somatic alterations in a typical tumor is a difficult and unsolved problem. We report a computational framework, FunSeq2, to annotate and prioritize these mutations. The framework combines an adjustable data context integrating large-scale genomics and cancer resources with a streamlined variant-prioritization pipeline. The pipeline has a weighted scoring system combining: inter- and intra-species conservation;loss- and gain-of-function events for transcription-factor binding; enhancer-gene linkages and network centrality; and per-element recurrence across samples. We further highlight putative drivers with information specific to a particular sample, such as differential expression. FunSeq2 is available from funseq2.gersteinlab.org.

Gerstein, M. B., et al. (2012). "Architecture of the human regulatory network derived from ENCODE data." Nature **489**(7414): 91-100.

Transcription factors bind in a combinatorial fashion to specify the on-and-off states of genes; the ensemble of these binding events forms a regulatory network, constituting the wiring diagram for a cell. To examine the principles of the human transcriptional regulatory network, we determined the genomic binding information of 119 transcription-related factors in over 450 distinct experiments. We found the combinatorial, co-association of transcription factors to be highly context specific: distinct combinations of factors bind at specific genomic locations. In particular, there are significant differences in the binding proximal and distal to genes. We organized all the transcription factor binding into a hierarchy and integrated it with other genomic information (for example, microRNA regulation), forming a dense meta-network. Factors at different levels have different properties; for instance, top-level transcription factors more strongly influence expression and middle-level ones co-regulate targets to mitigate information-flow bottlenecks. Moreover, these co-regulations give rise to many enriched network motifs (for example, noise-buffering feed-forward loops). Finally, more connected network components are under stronger selection and exhibit a greater degree of allele-specific activity (that is, differential binding to the two parental alleles). The regulatory information obtained in this study will be crucial for interpreting personal genome sequences and understanding basic principles of human biology and disease.

Gerstein, M. B., et al. (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science **330**(6012): 1775-1787.

We systematically generated large-scale data sets to improve genome annotation for the nematode Caenorhabditis elegans, a key model organism. These data sets include transcriptome profiling across a developmental time course, genome-wide identification of transcription factor-binding sites, and maps of chromatin organization. From this, we created more complete and accurate gene models, including alternative splice forms and candidate noncoding RNAs. We constructed hierarchical networks of

transcription factor-binding and microRNA interactions and discovered chromosomal locations bound by an unusually large number of transcription factors. Different patterns of chromatin composition and histone modification were revealed between chromosome arms and centers, with similarly prominent differences between autosomes and the X chromosome. Integrating data types, we built statistical models relating chromatin, transcription factor binding, and gene expression. Overall, our analyses ascribed putative functions to most of the conserved genome.

Gianoulis, T. A., et al. (2012). "Genomic Analysis of the Hydrocarbon-Producing, Cellulolytic, Endophytic Fungus Ascocoryne sarcoides." PLOS Genetics **8**(3): e1002558.

Harmanci, A. and M. Gerstein (2016). "Quantification of private information leakage from phenotype-genotype data: linking attacks." Nat Methods **13**(3): 251-256.
Studies on genomic privacy have traditionally focused on identifying individuals using DNA variants. In contrast, molecular phenotype data, such as gene expression levels, are generally assumed to be free of such identifying information. Although there is no explicit genotypic information in phenotype data, adversaries can statistically link phenotypes to genotypes using publicly available genotype-phenotype correlations such as expression quantitative trait loci (eQTLs). This linking can be accurate when high-dimensional data (i.e., many expression levels) are used, and the resulting links can then reveal sensitive information (for example, the fact that an individual has cancer). Here we develop frameworks for quantifying the leakage of characterizing information from phenotype data sets. These frameworks can be used to estimate the leakage from large data sets before release. We also present a general three-step procedure for practically instantiating linking attacks and a specific attack using outlier gene expression levels that is simple yet accurate. Finally, we describe the effectiveness of this outlier attack under different scenarios.

Harmanci, A., et al. (2014). "MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework." Genome Biol **15**(10): 474.
We present MUSIC, a signal processing approach for identification of enriched regions in ChIP-Seq data, available atmusic.gersteinlab.org. MUSIC first filters the ChIP-Seq read-depth signal for systematic noise from non-uniformmappability, which fragments enriched regions. Then it performs a multiscale decomposition, using median filtering, identifying enriched regions at multiple length scales. This is useful given the wide range of scales probed in ChIP-Seq assays. MUSIC performs favorably in terms of accuracy and reproducibility compared with other methods.In particular, analysis of RNA polymerase II data reveals a clear distinction between the stalled and elongating forms of the polymerase.

Howie, B., et al. (2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." Nat Genet **44**(8): 955-959.
The 1000 Genomes Project and disease-specific sequencing efforts are producing large collections of haplotypes that can be used as reference panels for genotype imputation in genome-wide association studies (GWAS). However, imputing from large reference panels with existing methods imposes a high computational burden. We introduce a strategy called 'pre-phasing' that maintains the accuracy of leading methods while reducing computational costs. We first statistically estimate the haplotypes for each individual within the GWAS sample (pre-phasing) and then impute missing genotypes into these estimated haplotypes. This reduces the computational cost because (i) the GWAS samples must be phased only once, whereas standard methods would implicitly repeat phasing with each reference panel update, and (ii) it is much faster to match a phased GWAS haplotype to one reference haplotype than to match two unphased GWAS genotypes to a pair of reference haplotypes. We implemented our approach in the MaCH and IMPUTE2 frameworks, and we tested it on data sets from the Wellcome Trust Case Control Consortium 2 (WTCCC2), the Genetic Association Information Network (GAIN), the Women's Health Initiative (WHI) and the 1000 Genomes Project. This strategy will be particularly valuable for repeated imputation as reference panels evolve.

Huffman, K. M., et al. (2014). "Metabolite signatures of exercise training in human skeletal muscle relate to mitochondrial remodelling and cardiometabolic fitness." Diabetologia **57**(11): 2282-2295.

AIMS/HYPOTHESIS: Targeted metabolomic and transcriptomic approaches were used to evaluate the relationship between skeletal muscle metabolite signatures, gene expression profiles and clinical outcomes in response to various exercise training interventions. We hypothesised that changes in mitochondrial metabolic intermediates would predict improvements in clinical risk factors, thereby offering novel insights into potential mechanisms. METHODS: Subjects at risk of metabolic disease were randomised to 6 months of inactivity or one of five aerobic and/or resistance training programmes (n = 112). Pre/post-intervention assessments included cardiorespiratory fitness ([Formula: see text]), serum triacylglycerols (TGs) and insulin sensitivity (SI). In this secondary analysis, muscle biopsy specimens were used for targeted mass spectrometry-based analysis of metabolic intermediates and measurement of mRNA expression of genes involved in metabolism. RESULTS: Exercise regimens with the largest energy expenditure produced robust increases in muscle concentrations of even-chain acylcarnitines (median 37-488%), which correlated positively with increased expression of genes involved in muscle uptake and oxidation of fatty acids. Along with free carnitine, the aforementioned acylcarnitine metabolites were related to improvements in [Formula: see text], TGs and SI ($R = 0.20$-0.31, $p < 0.05$). Muscle concentrations of the tricarboxylic acid cycle intermediates succinate and succinylcarnitine ($R = 0.39$ and 0.24, $p < 0.05$) emerged as the strongest correlates of SI. CONCLUSIONS/INTERPRETATION: The metabolic signatures of exercise-trained skeletal muscle reflected reprogramming of mitochondrial function and intermediary metabolism and correlated with changes in cardiometabolic fitness. Succinate metabolism and the succinate dehydrogenase complex emerged as a potential regulatory node that intersects with whole-body insulin sensitivity. This study identifies new avenues for mechanistic research aimed at understanding the health benefits of physical activity. Trial registration ClinicalTrials.gov NCT00200993 and NCT00275145 Funding This work was supported by the National Heart, Lung, and Blood Institute (National Institutes of Health), National Institute on Aging (National Institutes of Health) and National Institute of Arthritis and Musculoskeletal and Skin Diseases (National Institutes of Health).

Jee, J., et al. (2011). "ACT: aggregation and correlation toolbox for analyses of genome tracks." Bioinformatics **27**(8): 1152-1154.

We have implemented aggregation and correlation toolbox (ACT), an efficient, multifaceted toolbox for analyzing continuous signal and discrete region tracks from high-throughput genomic experiments, such as RNA-seq or ChIP-chip signal profiles from the ENCODE and modENCODE projects, or lists of single nucleotide polymorphisms from the 1000 genomes project. It is able to generate aggregate profiles of a given track around a set of specified anchor points, such as transcription start sites. It is also able to correlate related tracks and analyze them for saturation--i.e. how much of a certain feature is covered with each new succeeding experiment. The ACT site contains downloadable code in a variety of formats, interactive web servers (for use on small quantities of data), example datasets, documentation and a gallery of outputs. Here, we explain the components of the toolbox in more detail and apply them in various contexts. AVAILABILITY: ACT is available at http://act.gersteinlab.org/ CONTACT: pi@gersteinlab.org.

Johnson, W. E., et al. (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods." Biostatistics **8**(1): 118-127.

Non-biological experimental variation or "batch effects" are commonly observed across multiple batches of microarray experiments, often rendering the task of combining data from these batches difficult. The ability to combine microarray data sets is advantageous to researchers to increase statistical power to detect biological phenomena from studies where logistical considerations restrict sample size or in studies that require the sequential hybridization of arrays. In general, it is inappropriate to combine data sets without adjusting for batch effects. Methods have been proposed to filter batch effects from data, but these are often complicated and require large batch sizes (> 25) to implement. Because the majority of microarray studies are conducted using much smaller sample sizes, existing methods are not sufficient. We propose parametric and non-parametric empirical Bayes frameworks for adjusting data for batch effects that is robust to outliers in small sample sizes and performs comparable to existing methods for large samples. We illustrate our methods using two example data sets and show that our methods are

justifiable, easy to apply, and useful in practice. Software for our method is freely available at: http://biosun1.harvard.edu/complab/batch/.

Kelly, T. K., et al. (2012). "Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules." Genome Res **22**(12): 2497-2506.

DNA methylation and nucleosome positioning work together to generate chromatin structures that regulate gene expression. Nucleosomes are typically mapped using nuclease digestion requiring significant amounts of material and varying enzyme concentrations. We have developed a method (NOMe-seq) that uses a GpC methyltransferase (M.CviPI) and next generation sequencing to generate a high resolution footprint of nucleosome positioning genome-wide using less than 1 million cells while retaining endogenous DNA methylation information from the same DNA strand. Using a novel bioinformatics pipeline, we show a striking anti-correlation between nucleosome occupancy and DNA methylation at CTCF regions that is not present at promoters. We further show that the extent of nucleosome depletion at promoters is directly correlated to expression level and can accommodate multiple nucleosomes and provide genome-wide evidence that expressed non-CpG island promoters are nucleosome-depleted. Importantly, NOMe-seq obtains DNA methylation and nucleosome positioning information from the same DNA molecule, giving the first genome-wide DNA methylation and nucleosome positioning correlation at the single molecule, and thus, single cell level, that can be used to monitor disease progression and response to therapy.

Khurana, E., et al. (2013). "Interpretation of genomic variants using a unified biological network approach." PLoS Comput Biol **9**(3): e1002886.

The decreasing cost of sequencing is leading to a growing repertoire of personal genomes. However, we are lagging behind in understanding the functional consequences of the millions of variants obtained from sequencing. Global system-wide effects of variants in coding genes are particularly poorly understood. It is known that while variants in some genes can lead to diseases, complete disruption of other genes, called 'loss-of-function tolerant', is possible with no obvious effect. Here, we build a systems-based classifier to quantitatively estimate the global perturbation caused by deleterious mutations in each gene. We first survey the degree to which gene centrality in various individual networks and a unified 'Multinet' correlates with the tolerance to loss-of-function mutations and evolutionary conservation. We find that functionally significant and highly conserved genes tend to be more central in physical protein-protein and regulatory networks. However, this is not the case for metabolic pathways, where the highly central genes have more duplicated copies and are more tolerant to loss-of-function mutations. Integration of three-dimensional protein structures reveals that the correlation with centrality in the protein-protein interaction network is also seen in terms of the number of interaction interfaces used. Finally, combining all the network and evolutionary properties allows us to build a classifier distinguishing functionally essential and loss-of-function tolerant genes with higher accuracy (AUC = 0.91) than any individual property. Application of the classifier to the whole genome shows its strong potential for interpretation of variants involved in mendelian diseases and in complex disorders probed by genome-wide association studies.

Khurana, E., et al. (2013). "Integrative annotation of variants from 1092 humans: application to cancer genomics." Science **342**(6154): 1235587.

Interpreting variants, especially noncoding ones, in the increasing number of personal genomes is challenging. We used patterns of polymorphisms in functionally annotated regions in 1092 humans to identify deleterious variants; then we experimentally validated candidates. We analyzed both coding and noncoding regions, with the former corroborating the latter. We found regions particularly sensitive to mutations ("ultrasensitive") and variants that are disruptive because of mechanistic effects on transcription-factor binding (that is, "motif-breakers"). We also found variants in regions with higher network centrality tend to be deleterious. Insertions and deletions followed a similar pattern to single-nucleotide variants, with some notable exceptions (e.g., certain deletions and enhancers). On the basis of these patterns, we developed a computational tool (FunSeq), whose application to ~90 cancer genomes reveals nearly a hundred candidate noncoding drivers.

Kitchen, R. R., et al. (2014). "Decoding neuroproteomics: integrating the genome, translatome and functional anatomy." Nature Neuroscience **17**(11): 1491-1499.

Landt, S. G., et al. (2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." Genome Res **22**(9): 1813-1831.

Leung, D., et al. (2015). "Integrative analysis of haplotype-resolved epigenomes across human tissues." Nature **518**(7539): 350-354.

Allelic differences between the two homologous chromosomes can affect the propensity of inheritance in humans; however, the extent of such differences in the human genome has yet to be fully explored. Here we delineate allelic chromatin modifications and transcriptomes among a broad set of human tissues, enabled by a chromosome-spanning haplotype reconstruction strategy. The resulting large collection of haplotype-resolved epigenomic maps reveals extensive allelic biases in both chromatin state and transcription, which show considerable variation across tissues and between individuals, and allow us to investigate cis-regulatory relationships between genes and their control sequences. Analyses of histone modification maps also uncover intriguing characteristics of cis-regulatory elements and tissue-restricted activities of repetitive elements. The rich data sets described here will enhance our understanding of the mechanisms by which cis-regulatory elements control gene expression programs.

Lewis, D. A. and P. Levitt (2002). "Schizophrenia as a disorder of neurodevelopment." Annu Rev Neurosci **25**: 409-432.

A combination of genetic susceptibility and environmental perturbations appear to be necessary for the expression of schizophrenia. In addition, the pathogenesis of the disease is hypothesized to be neurodevelopmental in nature based on reports of an excess of adverse events during the pre- and perinatal periods, the presence of cognitive and behavioral signs during childhood and adolescence, and the lack of evidence of a neurodegenerative process in most individuals with schizophrenia. Recent studies of neurodevelopmental mechanisms strongly suggest that no single gene or factor is responsible for driving a highly complex biological process. Together, these findings suggest that combinatorial genetic and environmental factors, which disturb a normal developmental course early in life, result in molecular and histogenic responses that cumulatively lead to different developmental trajectories and the clinical phenotype recognized as schizophrenia.

Li, B. and C. N. Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." BMC Bioinformatics **12**: 323.

BACKGROUND: RNA-Seq is revolutionizing the way transcript abundances are measured. A key challenge in transcript quantification from RNA-Seq data is the handling of reads that map to multiple genes or isoforms. This issue is particularly important for quantification with de novo transcriptome assemblies in the absence of sequenced genomes, as it is difficult to determine which transcripts are isoforms of the same gene. A second significant issue is the design of RNA-Seq experiments, in terms of the number of reads, read length, and whether reads come from one or both ends of cDNA fragments. RESULTS: We present RSEM, an user-friendly software package for quantifying gene and isoform abundances from single-end or paired-end RNA-Seq data. RSEM outputs abundance estimates, 95% credibility intervals, and visualization files and can also simulate RNA-Seq data. In contrast to other existing tools, the software does not require a reference genome. Thus, in combination with a de novo transcriptome assembler, RSEM enables accurate transcript quantification for species without sequenced genomes. On simulated and real data sets, RSEM has superior or comparable performance to quantification methods that rely on a reference genome. Taking advantage of RSEM's ability to effectively use ambiguously-mapping reads, we show that accurate gene-level abundance estimates are best obtained with large numbers of short single-end reads. On the other hand, estimates of the relative frequencies of isoforms within single genes may be improved through the use of paired-end reads, depending on the number of possible splice forms for each gene. CONCLUSIONS: RSEM is an accurate

and user-friendly software tool for quantifying transcript abundances from RNA-Seq data. As it does not rely on the existence of a reference genome, it is particularly useful for quantification with de novo transcriptome assemblies. In addition, RSEM has enabled valuable guidance for cost-efficient design of quantification experiments with RNA-Seq, which is currently relatively expensive.

Liu, E. Y., et al. (2013). "MaCH-admix: genotype imputation for admixed populations." Genet Epidemiol **37**(1): 25-37.

Imputation in admixed populations is an important problem but challenging due to the complex linkage disequilibrium (LD) pattern. The emergence of large reference panels such as that from the 1,000 Genomes Project enables more accurate imputation in general, and in particular for admixed populations and for uncommon variants. To efficiently benefit from these large reference panels, one key issue to consider in modern genotype imputation framework is the selection of effective reference panels. In this work, we consider a number of methods for effective reference panel construction inside a hidden Markov model and specific to each target individual. These methods fall into two categories: identity-by-state (IBS) based and ancestry-weighted approach. We evaluated the performance on individuals from recently admixed populations. Our target samples include 8,421 African Americans and 3,587 Hispanic Americans from the Women' Health Initiative, which allow assessment of imputation quality for uncommon variants. Our experiments include both large and small reference panels; large, medium, and small target samples; and in genome regions of varying levels of LD. We also include BEAGLE and IMPUTE2 for comparison. Experiment results with large reference panel suggest that our novel piecewise IBS method yields consistently higher imputation quality than other methods/software. The advantage is particularly noteworthy among uncommon variants where we observe up to 5.1% information gain with the difference being highly significant (Wilcoxon signed rank test P-value < 0.0001). Our work is the first that considers various sensible approaches for imputation in admixed populations and presents a comprehensive comparison.

Lu, Z. J., et al. (2011). "Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data." Genome Res **21**(2): 276-285.

We present an integrative machine learning method, incRNA, for whole-genome identification of noncoding RNAs (ncRNAs). It combines a large amount of expression data, RNA secondary-structure stability, and evolutionary conservation at the protein and nucleic-acid level. Using the incRNA model and data from the modENCODE consortium, we are able to separate known C. elegans ncRNAs from coding sequences and other genomic elements with a high level of accuracy (97% AUC on an independent validation set), and find more than 7000 novel ncRNA candidates, among which more than 1000 are located in the intergenic regions of C. elegans genome. Based on the validation set, we estimate that 91% of the approximately 7000 novel ncRNA candidates are true positives. We then analyze 15 novel ncRNA candidates by RT-PCR, detecting the expression for 14. In addition, we characterize the properties of all the novel ncRNA candidates and find that they have distinct expression patterns across developmental stages and tend to use novel RNA structural families. We also find that they are often targeted by specific transcription factors (approximately 59% of intergenic novel ncRNA candidates). Overall, our study identifies many new potential ncRNAs in C. elegans and provides a method that can be adapted to other organisms.

Lun, A. T. and G. K. Smyth (2015). "diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data." BMC Bioinformatics **16**: 258.

McDaniell, R., et al. (2010). "Heritable individual-specific and allele-specific chromatin signatures in humans." Science **328**(5975): 235-239.

Mu, X. J., et al. (2011). "Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project." Nucleic Acids Res **39**(16): 7058-7076.

In the human genome, it has been estimated that considerably more sequence is under natural selection in non-coding regions [such as transcription-factor binding sites (TF-binding sites) and non-coding RNAs (ncRNAs)] compared to protein-coding ones. However, less attention has been paid to them. To study selective pressure on non-coding elements, we use next-generation sequencing data from the recently completed pilot phase of the 1000 Genomes Project, which, compared to traditional methods, allows for the characterization of a full spectrum of genomic variations, including single-nucleotide polymorphisms (SNPs), short insertions and deletions (indels) and structural variations (SVs). We develop a framework for combining these variation data with non-coding elements, calculating various population-based metrics to compare classes and subclasses of elements, and developing element-aware aggregation procedures to probe the internal structure of an element. Overall, we find that TF-binding sites and ncRNAs are less selectively constrained for SNPs than coding sequences (CDSs), but more constrained than a neutral reference. We also determine that the relative amounts of constraint for the three types of variations are, in general, correlated, but there are some differences: counter-intuitively, TF-binding sites and ncRNAs are more selectively constrained for indels than for SNPs, compared to CDSs. After inspecting the overall properties of a class of elements, we analyze selective pressure on subclasses within an element class, and show that the extent of selection is associated with the genomic properties of each subclass. We find, for instance, that ncRNAs with higher expression levels tend to be under stronger purifying selection, and the actual regions of TF-binding motifs are under stronger selective pressure than the corresponding peak regions. Further, we develop element-aware aggregation plots to analyze selective pressure across the linear structure of an element, with the confidence intervals evaluated using both simple bootstrapping and block bootstrapping techniques. We find, for example, that both micro-RNAs (particularly the seed regions) and their binding targets are under stronger selective pressure for SNPs than their immediate genomic surroundings. In addition, we demonstrate that substitutions in TF-binding motifs inversely correlate with site conservation, and SNPs unfavorable for motifs are under more selective constraints than favorable SNPs. Finally, to further investigate intra-element differences, we show that SVs have the tendency to use distinctive modes and mechanisms when they interact with genomic elements, such as enveloping whole gene(s) rather than disrupting them partially, as well as duplicating TF motifs in tandem.

Myers, R. M., et al. (2011). "A user's guide to the Encyclopedia of DNA Elements (ENCODE)." PLoS Biology **9**(4).

Neale, B. M. and P. Sklar (2015). "Genetic analysis of schizophrenia and bipolar disorder reveals polygenicity but also suggests new directions for molecular interrogation." Curr Opin Neurobiol **30**: 131-138.

Over the last few years, genetics research has made significant strides in identifying many risk factors for schizophrenia and bipolar disorder. These risk factors include inherited common single nucleotide polymorphisms, copy number variants, and rare single nucleotide variants, as well as rare de novo variants. For all variants, the common theme has been that of polygenicity, meaning that many small genetic risk factors influence risk in the population and that no gene or variant on its own has been shown to be fully deterministic of schizophrenia or bipolar. When taken together, biological themes that have emerged including the importance of synaptic function and calcium signaling. This has implications for our understanding of the biological underpinnings of these diseases.

Ongen, H., et al. (2016). "Fast and efficient QTL mapper for thousands of molecular phenotypes." Bioinformatics **32**(10): 1479-1485.

Pflueger, D., et al. (2011). "Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing." Genome Res **21**(1): 56-67.

Half of prostate cancers harbor gene fusions between TMPRSS2 and members of the ETS transcription factor family. To date, little is known about the presence of non-ETS fusion events in prostate cancer. We used next-generation transcriptome sequencing (RNA-seq) in order to explore the whole transcriptome of 25 human prostate cancer samples for the presence of chimeric fusion transcripts. We generated more than 1 billion sequence reads and used a novel computational approach (FusionSeq) in order to identify

novel gene fusion candidates with high confidence. In total, we discovered and characterized seven new cancer-specific gene fusions, two involving the ETS genes ETV1 and ERG, and four involving non-ETS genes such as CDKN1A (p21), CD9, and IKBKB (IKK-beta), genes known to exhibit key biological roles in cellular homeostasis or assumed to be critical in tumorigenesis of other tumor entities, as well as the oncogene PIGU and the tumor suppressor gene RSRC2. The novel gene fusions are found to be of low frequency, but, interestingly, the non-ETS fusions were all present in prostate cancer harboring the TMPRSS2-ERG gene fusion. Future work will focus on determining if the ETS rearrangements in prostate cancer are associated or directly predispose to a rearrangement-prone phenotype.

Pickrell, J. K., et al. (2010). "Understanding mechanisms underlying human gene expression variation with RNA sequencing." Nature **464**(7289): 768-772.

Understanding the genetic mechanisms underlying natural variation in gene expression is a central goal of both medical and evolutionary genetics, and studies of expression quantitative trait loci (eQTLs) have become an important tool for achieving this goal. Although all eQTL studies so far have assayed messenger RNA levels using expression microarrays, recent advances in RNA sequencing enable the analysis of transcript variation at unprecedented resolution. We sequenced RNA from 69 lymphoblastoid cell lines derived from unrelated Nigerian individuals that have been extensively genotyped by the International HapMap Project. By pooling data from all individuals, we generated a map of the transcriptional landscape of these cells, identifying extensive use of unannotated untranslated regions and more than 100 new putative protein-coding exons. Using the genotypes from the HapMap project, we identified more than a thousand genes at which genetic variation influences overall expression levels or splicing. We demonstrate that eQTLs near genes generally act by a mechanism involving allele-specific expression, and that variation that influences the inclusion of an exon is enriched within and near the consensus splice sites. Our results illustrate the power of high-throughput sequencing for the joint analysis of variation in transcription, splicing and allele-specific expression across individuals.

Pickrell, J. K., et al. (2010). "Noisy splicing drives mRNA isoform diversity in human cells." PLoS Genet **6**(12): e1001236.

While the majority of multiexonic human genes show some evidence of alternative splicing, it is unclear what fraction of observed splice forms is functionally relevant. In this study, we examine the extent of alternative splicing in human cells using deep RNA sequencing and de novo identification of splice junctions. We demonstrate the existence of a large class of low abundance isoforms, encompassing approximately 150,000 previously unannotated splice junctions in our data. Newly-identified splice sites show little evidence of evolutionary conservation, suggesting that the majority are due to erroneous splice site choice. We show that sequence motifs involved in the recognition of exons are enriched in the vicinity of unconserved splice sites. We estimate that the average intron has a splicing error rate of approximately 0.7% and show that introns in highly expressed genes are spliced more accurately, likely due to their shorter length. These results implicate noisy splicing as an important property of genome evolution.

Price, A. L., et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat Genet **38**(8): 904-909.

Population stratification--allele frequency differences between cases and controls due to systematic ancestry differences-can cause spurious associations in disease studies. We describe a method that enables explicit detection and correction of population stratification on a genome-wide scale. Our method uses principal components analysis to explicitly model ancestry differences between cases and controls. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. Our simple, efficient approach can easily be applied to disease studies with hundreds of thousands of markers.

Psych, E. C., et al. (2015). "The PsychENCODE project." Nat Neurosci **18**(12): 1707-1712.

Raedler, T. J., et al. (1998). "Schizophrenia as a developmental disorder of the cerebral cortex." Curr Opin Neurobiol **8**(1): 157-161.

The hypothesis that schizophrenia results from a developmental, as opposed to a degenerative, process affecting the cerebral cortex has become popular in current thinking about the disorder. While many of the data gathered in support of this hypothesis do not in themselves represent conclusive proof, an intriguing picture is emerging from a variety of research approaches. These approaches include the observation of minor physical anomalies, premorbid neuropsychological and social deficits, obstetrical complications, and exposure to adverse intrauterine events. Morphometric brain measurement techniques and neuropathological studies have perhaps provided more substantial support.

Rao, S. S., et al. (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." Cell **159**(7): 1665-1680.

Roadmap Epigenomics, C., et al. (2015). "Integrative analysis of 111 reference human epigenomes." Nature **518**(7539): 317-330.

The reference human genome sequence set the stage for studies of genetic variation and its association with human disease, but epigenomic studies lack a similar reference. To address this need, the NIH Roadmap Epigenomics Consortium generated the largest collection so far of human epigenomes for primary cells and tissues. Here we describe the integrative analysis of 111 reference human epigenomes generated as part of the programme, profiled for histone modification patterns, DNA accessibility, DNA methylation and RNA expression. We establish global maps of regulatory elements, define regulatory modules of coordinated activity, and their likely activators and repressors. We show that disease- and trait-associated genetic variants are enriched in tissue-specific epigenomic marks, revealing biologically relevant cell types for diverse human traits, and providing a resource for interpreting the molecular basis of human disease. Our results demonstrate the central role of epigenomic information for understanding gene regulation, cellular differentiation and human disease.

Roshyara, N. R., et al. (2016). "Comparing performance of modern genotype imputation methods in different ethnicities." Sci Rep **6**: 34386.

A variety of modern software packages are available for genotype imputation relying on advanced concepts such as pre-phasing of the target dataset or utilization of admixed reference panels. In this study, we performed a comprehensive evaluation of the accuracy of modern imputation methods on the basis of the publicly available POPRES samples. Good quality genotypes were masked and re-imputed by different imputation frameworks: namely MaCH, IMPUTE2, MaCH-Minimac, SHAPEIT-IMPUTE2 and MaCH-Admix. Results were compared to evaluate the relative merit of pre-phasing and the usage of admixed references. We showed that the pre-phasing framework SHAPEIT-IMPUTE2 can overestimate the certainty of genotype distributions resulting in the lowest percentage of correctly imputed genotypes in our case. MaCH-Minimac performed better than SHAPEIT-IMPUTE2. Pre-phasing always reduced imputation accuracy. IMPUTE2 and MaCH-Admix, both relying on admixed-reference panels, showed comparable results. MaCH showed superior results if well-matched references were available (Nei's GST </= 0.010). For small to medium datasets, frameworks using genetically closest reference panel are recommended if the genetic distance between target and reference data set is small. Our results are valid for small to medium data sets. As shown on a larger data set of population based German samples, the disadvantage of pre-phasing decreases for larger sample sizes.

Rozowsky, J., et al. (2009). "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls." Nat Biotechnol **27**(1): 66-75.

Chromatin immunoprecipitation (ChIP) followed by tag sequencing (ChIP-seq) using high-throughput next-generation instrumentation is fast, replacing chromatin immunoprecipitation followed by genome tiling array analysis (ChIP-chip) as the preferred approach for mapping of sites of transcription-factor binding and chromatin modification. Using two deeply sequenced data sets for human RNA polymerase II and STAT1, each with matching input-DNA controls, we describe a general scoring approach to address unique challenges in ChIP-seq data analysis. Our approach is based on the observation that sites of

potential binding are strongly correlated with signal peaks in the control, likely revealing features of open chromatin. We develop a two-pass strategy called PeakSeq to compensate for this. A two-pass strategy compensates for signal caused by open chromatin, as revealed by inclusion of the controls. The first pass identifies putative binding sites and compensates for genomic variation in the 'mappability' of sequences. The second pass filters out sites not significantly enriched compared to the normalized control, computing precise enrichments and significances. Our scoring procedure enables us to optimize experimental design by estimating the depth of sequencing required for a desired level of coverage and demonstrating that more than two replicates provides only a marginal gain in information.

Sboner, A., et al. (2010). "FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data." Genome Biol **11**(10): R104.

We have developed FusionSeq to identify fusion transcripts from paired-end RNA-sequencing. FusionSeq includes filters to remove spurious candidate fusions with artifacts, such as misalignment or random pairing of transcript fragments, and it ranks candidates according to several statistics. It also has a module to identify exact sequences at breakpoint junctions. FusionSeq detected known and novel fusions in a specially sequenced calibration data set, including eight cancers with and without known rearrangements.

Sboner, A., et al. (2009). "Robust-Linear-Model Normalization To Reduce Technical Variability in Functional Protein Microarrays." Journal of Proteome Research **8**(12): 5451-5464.

Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). "Biological insights from 108 schizophrenia-associated genetic loci." Nature **511**(7510): 421-427.

Schizophrenia is a highly heritable disorder. Genetic risk is conferred by a large number of alleles, including common alleles of small effect that might be detected by genome-wide association studies. Here we report a multi-stage schizophrenia genome-wide association study of up to 36,989 cases and 113,075 controls. We identify 128 independent associations spanning 108 conservatively defined loci that meet genome-wide significance, 83 of which have not been previously reported. Associations were enriched among genes expressed in brain, providing biological plausibility for the findings. Many findings have the potential to provide entirely new insights into aetiology, but associations at DRD2 and several genes involved in glutamatergic neurotransmission highlight molecules of known and potential therapeutic relevance to schizophrenia, and are consistent with leading pathophysiological hypotheses. Independent of genes expressed in brain, associations were enriched among genes expressed in tissues that have important roles in immunity, providing support for the speculated link between the immune system and schizophrenia.

Schmidt-Kastner, R., et al. (2006). "Gene regulation by hypoxia and the neurodevelopmental origin of schizophrenia." Schizophr Res **84**(2-3): 253-271.

Neurodevelopmental changes may underlie the brain dysfunction seen in schizophrenia. While advances have been made in our understanding of the genetics of schizophrenia, little is known about how non-genetic factors interact with genes for schizophrenia. The present analysis of genes potentially associated with schizophrenia is based on the observation that hypoxia prevails in the embryonic and fetal brain, and that interactions between neuronal genes, molecular regulators of hypoxia, such as hypoxia-inducible factor 1 (HIF-1), and intrinsic hypoxia occur in the developing brain and may create the conditions for complex changes in neurodevelopment. Consequently, we searched the literature for currently hypothesized candidate genes for susceptibility to schizophrenia that may be subject to ischemia-hypoxia regulation and/or associated with vascular expression. Genes were considered when at least two independent reports of a significant association with schizophrenia had appeared in the literature. The analysis showed that more than 50% of these genes, particularly AKT1, BDNF, CAPON, CCKAR, CHRNA7, CNR1, COMT, DNTBP1, GAD1, GRM3, IL10, MLC1, NOTCH4, NRG1, NR4A2/NURR1, PRODH, RELN, RGS4, RTN4/NOGO and TNF, are subject to regulation by hypoxia and/or are expressed in the vasculature. Future studies of genes proposed as candidates for

susceptibility to schizophrenia should include their possible regulation by physiological or pathological hypoxia during development as well as their potential role in cerebral vascular function.

Servant, N., et al. (2012). "HiTC: exploration of high-throughput 'C' experiments." Bioinformatics **28**(21): 2843-2844.

Servant, N., et al. (2015). "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing." Genome Biol **16**: 259.

Shibata, Y. and G. E. Crawford (2009). "Mapping regulatory elements by DNaseI hypersensitivity chip (DNase-Chip)." Methods in Molecular Biology **556**: 177-190.

Shin, H., et al. (2016). "TopDom: an efficient and deterministic method for identifying topological domains in genomes." Nucleic Acids Res **44**(7): e70.

Sisu, C., et al. (2014). "Comparative analysis of pseudogenes across three phyla." Proc Natl Acad Sci U S A **111**(37): 13361-13366.

Pseudogenes are degraded fossil copies of genes. Here, we report a comparison of pseudogenes spanning three phyla, leveraging the completed annotations of the human, worm, and fly genomes, which we make available as an online resource. We find that pseudogenes are lineage specific, much more so than protein-coding genes, reflecting the different remodeling processes marking each organism's genome evolution. The majority of human pseudogenes are processed, resulting from a retrotranspositional burst at the dawn of the primate lineage. This burst can be seen in the largely uniform distribution of pseudogenes across the genome, their preservation in areas with low recombination rates, and their preponderance in highly expressed gene families. In contrast, worm and fly pseudogenes tell a story of numerous duplication events. In worm, these duplications have been preserved through selective sweeps, so we see a large number of pseudogenes associated with highly duplicated families such as chemoreceptors. However, in fly, the large effective population size and high deletion rate resulted in a depletion of the pseudogene complement. Despite large variations between these species, we also find notable similarities. Overall, we identify a broad spectrum of biochemical activity for pseudogenes, with the majority in each organism exhibiting varying degrees of partial activity. In particular, we identify a consistent amount of transcription (approximately 15%) across all species, suggesting a uniform degradation process. Also, we see a uniform decay of pseudogene promoter activity relative to their coding counterparts and identify a number of pseudogenes with conserved upstream sequences and activity, hinting at potential regulatory roles.

Smith, A., et al. (2007). "Leveraging the structure of the Semantic Web to enhance information retrieval for proteomics." Bioinformatics **23**(22): 3073-3079.

Song, L. and G. E. Crawford (2010). "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells." Cold Spring Harbor Protocols **2010**(2).

Song, L., et al. (2011). "Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity." Genome Research **21**(10): 1757-1767.

Storey, J. D. and R. Tibshirani (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays (Technical Report 2001-28). Palo Alto, CA, Department of Statistics, Stanford University.

Wang, D., et al. (2016). "DREISS: Using State-Space Models to Infer the Dynamics of Gene Expression Driven by External and Internal Regulatory Networks." PLoS Comput Biol **12**(10): e1005146.

Gene expression is controlled by the combinatorial effects of regulatory factors from different biological subsystems such as general transcription factors (TFs), cellular growth factors and microRNAs. A subsystem's gene expression may be controlled by its internal regulatory factors, exclusively, or by external subsystems, or by both. It is thus useful to distinguish the degree to which a subsystem is regulated internally or externally-e.g., how non-conserved, species-specific TFs affect the expression of conserved, cross-species genes during evolution. We developed a computational method (DREISS, dreiss.gerteinlab.org) for analyzing the Dynamics of gene expression driven by Regulatory networks, both External and Internal based on State Space models. Given a subsystem, the "state" and "control" in the model refer to its own (internal) and another subsystem's (external) gene expression levels. The state at a given time is determined by the state and control at a previous time. Because typical time-series data do not have enough samples to fully estimate the model's parameters, DREISS uses dimensionality reduction, and identifies canonical temporal expression trajectories (e.g., degradation, growth and oscillation) representing the regulatory effects emanating from various subsystems. To demonstrate capabilities of DREISS, we study the regulatory effects of evolutionarily conserved vs. divergent TFs across distant species. In particular, we applied DREISS to the time-series gene expression datasets of C. elegans and D. melanogaster during their embryonic development. We analyzed the expression dynamics of the conserved, orthologous genes (orthologs), seeing the degree to which these can be accounted for by orthologous (internal) versus species-specific (external) TFs. We found that between two species, the orthologs have matched, internally driven expression patterns but very different externally driven ones. This is particularly true for genes with evolutionarily ancient functions (e.g. the ribosomal proteins), in contrast to those with more recently evolved functions (e.g., cell-cell communication). This suggests that despite striking morphological differences, some fundamental embryonic-developmental processes are still controlled by ancient regulatory systems.

Wang, D., et al. (2015). "Loregic: a method to characterize the cooperative logic of regulatory factors." PLoS Comput Biol **11**(4): e1004132.

The topology of the gene-regulatory network has been extensively analyzed. Now, given the large amount of available functional genomic data, it is possible to go beyond this and systematically study regulatory circuits in terms of logic elements. To this end, we present Loregic, a computational method integrating gene expression and regulatory network data, to characterize the cooperativity of regulatory factors. Loregic uses all 16 possible two-input-one-output logic gates (e.g. AND or XOR) to describe triplets of two factors regulating a common target. We attempt to find the gate that best matches each triplet's observed gene expression pattern across many conditions. We make Loregic available as a general-purpose tool (github.com/gersteinlab/loregic). We validate it with known yeast transcription-factor knockout experiments. Next, using human ENCODE ChIP-Seq and TCGA RNA-Seq data, we are able to demonstrate how Loregic characterizes complex circuits involving both proximally and distally regulating transcription factors (TFs) and also miRNAs. Furthermore, we show that MYC, a well-known oncogenic driving TF, can be modeled as acting independently from other TFs (e.g., using OR gates) but antagonistically with repressing miRNAs. Finally, we inter-relate Loregic's gate logic with other aspects of regulation, such as indirect binding via protein-protein interactions, feed-forward loop motifs and global regulatory hierarchy.

Weinberger, D. R. (1987). "Implications of normal brain development for the pathogenesis of schizophrenia." Arch Gen Psychiatry **44**(7): 660-669.

Recent research on schizophrenia has demonstrated that in this disorder the brain is not, strictly speaking, normal. The findings suggest that nonspecific histopathology exists in the limbic system, diencephalon, and prefrontal cortex, that the pathology occurs early in development, and that the causative process is inactive long before the diagnosis is made. If these findings are valid and not epiphenomena, then the pathogenesis of schizophrenia does not appear to fit either traditional metabolic, posttraumatic, or neurodegenerative models of adult mental illness. The data are more consistent with a neurodevelopmental model in which a fixed "lesion" from early in life interacts with normal brain maturational events that occur much later. Based on neuro-ontological principles and insights from animal research about normal brain development, it is proposed that the appearance of diagnostic

symptoms is linked to the normal maturation of brain areas affected by the early developmental pathology, particularly the dorsolateral prefrontal cortex. The course of the illness and the importance of stress may be related to normal maturational aspects of dopaminergic neural systems, particularly those innervating prefrontal cortex. Some implications for future research and treatment are considered.

Wingett, S., et al. (2015). "HiCUP: pipeline for mapping and processing Hi-C data." F1000Res **4**: 1310.

Won, H., et al. (2016). "Chromosome conformation elucidates regulatory relationships in developing human brain." Nature **538**(7626): 523-527.

Three-dimensional physical interactions within chromosomes dynamically regulate gene expression in a tissue-specific manner. However, the 3D organization of chromosomes during human brain development and its role in regulating gene networks dysregulated in neurodevelopmental disorders, such as autism or schizophrenia, are unknown. Here we generate high-resolution 3D maps of chromatin contacts during human corticogenesis, permitting large-scale annotation of previously uncharacterized regulatory relationships relevant to the evolution of human cognition and disease. Our analyses identify hundreds of genes that physically interact with enhancers gained on the human lineage, many of which are under purifying selection and associated with human cognitive function. We integrate chromatin contacts with non-coding variants identified in schizophrenia genome-wide association studies (GWAS), highlighting multiple candidate schizophrenia risk genes and pathways, including transcription factors involved in neurogenesis, and cholinergic signalling molecules, several of which are supported by independent expression quantitative trait loci and gene expression analyses. Genome editing in human neural progenitors suggests that one of these distal schizophrenia GWAS loci regulates FOXG1 expression, supporting its potential role as a schizophrenia risk gene. This work provides a framework for understanding the effect of non-coding regulatory elements on human brain development and the evolution of cognition, and highlights novel mechanisms underlying neuropsychiatric disorders.

Wu, L., et al. (2007). "Global Survey of Human T Leukemic Cells by Integrating Proteomics and Transcriptomics Profiling." Molecular & Cellular Proteomics **6**(8): 1343-1353.

Yan, K. K., et al. (2014). "OrthoClust: an orthology-based network framework for clustering data across multiple species." Genome Biol **15**(8): R100.

Increasingly, high-dimensional genomics data are becoming available for many organisms.Here, we develop OrthoClust for simultaneously clustering data across multiple species. OrthoClust is a computational framework that integrates the co-association networks of individual species by utilizing the orthology relationships of genes between species. It outputs optimized modules that are fundamentally cross-species, which can either be conserved or species-specific. We demonstrate the application of OrthoClust using the RNA-Seq expression profiles of Caenorhabditis elegans and Drosophila melanogaster from the modENCODE consortium. A potential application of cross-species modules is to infer putative analogous functions of uncharacterized elements like non-coding RNAs based on guilt-by-association.

Yip, K. Y., et al. (2010). "Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data." PLOS ONE **5**(1): e8121.

We performed computational reconstruction of the in silico gene regulatory networks in the DREAM3 Challenges. Our task was to learn the networks from two types of data, namely gene expression profiles in deletion strains (the 'deletion data') and time series trajectories of gene expression after some initial perturbation (the 'perturbation data'). In the course of developing the prediction method, we observed that the two types of data contained different and complementary information about the underlying network. In particular, deletion data allow for the detection of direct regulatory activities with strong responses upon the deletion of the regulator while perturbation data provide richer information for the identification of weaker and more complex types of regulation. We applied different techniques to learn the regulation from the two types of data. For deletion data, we learned a noise model to distinguish real signals from

random fluctuations using an iterative method. For perturbation data, we used differential equations to model the change of expression levels of a gene along the trajectories due to the regulation of other genes. We tried different models, and combined their predictions. The final predictions were obtained by merging the results from the two types of data. A comparison with the actual regulatory networks suggests that our approach is effective for networks with a range of different sizes. The success of the approach demonstrates the importance of integrating heterogeneous data in network reconstruction.

Yip, K. Y., et al. (2012). "Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors." Genome Biol **13**(9): R48.
BACKGROUND: Transcription factors function by binding different classes of regulatory elements. The Encyclopedia of DNA Elements (ENCODE) project has recently produced binding data for more than 100 transcription factors from about 500 ChIP-seq experiments in multiple cell types. While this large amount of data creates a valuable resource, it is nonetheless overwhelmingly complex and simultaneously incomplete since it covers only a small fraction of all human transcription factors. RESULTS: As part of the consortium effort in providing a concise abstraction of the data for facilitating various types of downstream analyses, we constructed statistical models that capture the genomic features of three paired types of regions by machine-learning methods: firstly, regions with active or inactive binding; secondly, those with extremely high or low degrees of co-binding, termed HOT and LOT regions; and finally, regulatory modules proximal or distal to genes. From the distal regulatory modules, we developed computational pipelines to identify potential enhancers, many of which were validated experimentally. We further associated the predicted enhancers with potential target transcripts and the transcription factors involved. For HOT regions, we found a significant fraction of transcription factor binding without clear sequence motifs and showed that this observation could be related to strong DNA accessibility of these regions. CONCLUSIONS: Overall, the three pairs of regions exhibit intricate differences in chromosomal locations, chromatin features, factors that bind them, and cell-type specificity. Our machine learning approach enables us to identify features potentially general to all transcription factors, including those not included in the data.

Zhang, W., et al. (2012). "High-resolution mapping of open chromatin in the rice genome." Genome Research **22**(1): 151-162.

Ziller, M. J., et al. (2015). "Dissecting neural differentiation regulatory networks through epigenetic footprinting." Nature **518**(7539): 355-359.
Models derived from human pluripotent stem cells that accurately recapitulate neural development in vitro and allow for the generation of specific neuronal subtypes are of major interest to the stem cell and biomedical community. Notch signalling, particularly through the Notch effector HES5, is a major pathway critical for the onset and maintenance of neural progenitor cells in the embryonic and adult nervous system. Here we report the transcriptional and epigenomic analysis of six consecutive neural progenitor cell stages derived from a HES5::eGFP reporter human embryonic stem cell line. Using this system, we aimed to model cell-fate decisions including specification, expansion and patterning during the ontogeny of cortical neural stem and progenitor cells. In order to dissect regulatory mechanisms that orchestrate the stage-specific differentiation process, we developed a computational framework to infer key regulators of each cell-state transition based on the progressive remodelling of the epigenetic landscape and then validated these through a pooled short hairpin RNA screen. We were also able to refine our previous observations on epigenetic priming at transcription factor binding sites and suggest here that they are mediated by combinations of core and stage-specific factors. Taken together, we demonstrate the utility of our system and outline a general framework, not limited to the context of the neural lineage, to dissect regulatory circuits of differentiation.