**Significance:** The rich data generated by the PsychENCODE Consortium are part of a preeminent resource for studying regulatory mechanisms in the human brain [1]. One of its unique aspects is its coverage of major psychiatric diseases, such as autism spectrum disorder (ASD) and schizophrenia (SCZ). PsychENCODE datasets have been assembled by many investigators over a number of years, and they are housed in a central depository (www.synapse.org). These are complemented by a number of other large-scale genomic resources, such as ENCODE, GTEx, Roadmap, BrainSpan, and CommonMind, which provide valuable contexts for additional human organs and tissues. By using these valuable datasets, we propose to conduct comprehensive, integrative analyses in order to find non-coding functional elements that are specific to the brain (Aims 1 and 2). We also propose to test a prioritized set of these predictions using STARR-seq assays, which provide a direct readout of enhancer activity genome-wide, in addition to CRISPR genome editing, which measures disease-associated variants in their native genomic context. The experimental testing will be performed using primary human neuronal progenitors and their differentiated neurons on 3D cortical forebrain spheroids. Finally, this project will continue to support all PsychENCODE data at Synapse, augmented by a psychSCREEN web engine that allows users to directly search for specific regulatory elements, as well as visualize their annotations and raw data for specific cell types. Thus, this project will leverage PsychENCODE data to improve our knowledge of brain functions and to maximize the impact of PsychENCODE data on the broader research and clinical communities.

**Innovation:** This project includes innovations within each aim and in the overall design. We will develop novel computational and statistical methods for analyzing and integrating genomic and epigenomic data (Aim 1). We will perform integrative analysis of the massive data in PsychENCODE and other consortia, which requires many biology-driven innovations (Aim 2). We will perform single cell transcriptome analyses on the cerebral cortex and cerebellum during four critical epochs of psychiatric disease risk (Aims 1 & 2). These transcriptome maps will enable detailed analysis of the specific cell types associated with these diseases. Our STARR-seq assay will be used to assay all enhancer activities throughout the genome (Aim 3). Furthermore, we will perform STARR-seq and CRISPR on primary human neuronal progenitors, which can be differentiated into neuronal cell types that are implicated in psychiatric diseases. We will also use cutting-edge microfluidic devices to culture 3D cortical forebrain spheroids that accurately recapitulate brain tissues (Aim 4). This integration between computation, large-scale data analysis, single cell analysis, and genome-wide experimental testing using disease-relevant primary cells and organoids are led by investigators who have a history of performing innovation research. Furthermore, our project will continue to be the centralized location for coordinating, housing and sharing all PsychENCODE data. Given that there, To make PsychENCODE data more readily accessible to the broader scientific community, we will develop an innovative and efficient search engine (“psychSCREEN”) to enable users without any programming expertise to visualize these data at the level of individual regulatory elements.

**Aim 1 - Developing methods to find brain-specific enhancers, integrating them into regulatory networks, deconvolving their regulation in a cell-type-specific fashion, and relating them to variation**

**1a. Overview.** Genotypes drive phenotypes and impact psychiatric disorders through complex gene regulatory networks. We aim to unravel gene regulatory networks for various psychiatric disorders, and to investigate biological mechanisms of how genotypes drive the psychiatric phenotypes. In particular, we plan to develop machine learning and pattern recognition methods by integrating various epigenomic signals and enhancer RNA expression patterns to predict active enhancers in different cell types across different brain regions. We will then examine how genetic variations modulate enhancers and regulatory networks to control the expression of genes associated with psychiatric diseases. We will describe each sub-aim in two parts—***Preliminary*** results and research ***Plan***.

**1b. Finding brain-specific enhancers.** ***Preliminary:*** Over the course of our work in the ENCODE and modENCODE projects since 2003 [2, 3], we have gained extensive experience in annotating non-coding DNA. We have developed machine learning methods to integrate signals for histone modifications, DNA methylation, chromatin accessibility, sequence conservation, sequence motifs, and gene annotations to identify enhancers, including those that are distal to their target genes. We have also built robust computational pipelines for processing massive amounts of data and identifying enhancers, transcription factor binding sites, and regulatory modules [4][4], which lay the foundation for this project.

***Plan:*** By leveraging recent advances (from the White Lab) in STARR-seq – a high-throughput assay for directly measuring enhancer activity genome-wide (see Aim 3) – we will develop a new approach for finding enhancers. The White Lab has successfully performed STARR-seq on primary human neuronal progenitors (phNPC) and their differentiated neurons, and more recently, on 3D cortical forebrain spheroids (hFS). We will first call peaks in STARR-seq profiles by extending our MUSIC method for calling histone mark peaks [5] and calculating statistical significance using a Poisson model. We will then estimate the false discovery rate (FDR) by swapping the control library with the STARR-seq library. STARR peaks above the 5% FDR cutoff will be used as the gold standard of neuronal enhancers to develop a machine learning framework for finding such enhancers. We will use matched filters to integrate the ChIP-seq signals of multiple histone modifications—matched filters can identify an enriched peak-trough-peak ("double peak") spatial signal at active enhancers. We will then use a linear support vector machine (SVM) to combine the normalized matched filter scores from different epigenetic marks (e.g., H3K27ac, an enhancer mark, and ATAC-seq or DNase-seq signals, which measure chromatin accessibility). The goal of the SVM is to predict STARR-seq enhancers.

**1c. Building brain-specific networks. *Preliminary:*** We have previously contributed a large body of work on regulatory networks. Via data integration, we have constructed gene networks of various regulators, including transcription factors (TF) and micro-RNAs (mRNA) and their target genes [6-12]. Upon analyzing the structures of these networks, we found that, relative to centrality, hierarchy levels are better predictors the regulator importance [6, 13-16]. Thus, we developed a general-purpose algorithm to measure the hierarchical structure of any type of regulatory network [17]. Our network analysis software tools include TopNet [18], tYNA[19], and PubNet [20]. In addition to the global attributes of regulatory networks (such as their hierarchy), we also analyzed local topological features, such as network motifs (e.g., feed-forward loops) [6, 9, 12]. We further integrated regulatory networks with gene expression to uncover functional modules [21-24]. We integrated ENCODE data on TF binding, histone modifications, and target gene expression to establish regulatory relationships using a probabilistic model named TIP [25][25]. Identifying potential enhancers from gene-distal regions, we used these modules to characterize the associations between TF binding and gene expression [26-29]. We further integrated these data with protein-protein interaction and transcriptional regulation networks [8, 9, 30, 31]. This enabled us to separate TFs into histone-sensitive and -insensitive classes, which refined the prediction of target gene expression levels. To analyze multiple interconnected networks simultaneously, we constructed co-expression networks from the extensive RNA-seq data in various consortia [3]. We further developed a novel framework consisting of a cross-species multi-layer network (OrthoClust) to analyze co-expression networks in an integrated fashion using orthologous genes across species [29][29].

***Plan:*** We will predict enhancers (Aim 1b) and promoters (from GENCODE annotation) and build gene regulatory networks for different brain regions and psychiatric disorders. Using brain enhancers and other regulatory elements, we will first find the TFs that bind to these regions (using TF ChIP-seq data and sequence motif analyses), and then these with their target genes if their gene expression accurately predicts the target genes’ expression using machine learning methods. We will also use gene loops (as defined by Hi-C) in fetal brain [32] as well as new adult neuronal and glial Hi-C data – developed as part of the PsychENCODE Project (Geschwind, PI) – to help assign distal regulatory regions to genes. We will build a gene regulatory network for each brain region or psychiatric disease. We will then study the structure and dynamics of our inferred regulatory networks and compare them across brain regions or disease types using the arsenal of methods we have developed (e.g., network motif algorithms to discover recurrent patterns of connectivity—specifically, recurrent regulatory feedback and feed-forward motifs). Extending these methods, we will use graph algorithms to discover clusters of highly connected genes within these networks. We expect to find regulatory network structures (such as hub genes, modules, and pathways) that are specific to certain diseases or brain regions, and we will annotate these network structures using enriched biological functions for the associated target genes.

**1d. Developing methods for single cell analysis. *Preliminary:*** The genetic risk for major psychiatric disorders (such as ASD and SCZ) implicate specific cerebral cortical cell types and developmental stages [33], so a goal of our work is to identify regulatory networks active in major cortical cell types. We have produced extensive single cell RNA-seq (scSeq) datasets via a BRAIN Initiative Award, and used them to develop deep single cell transcriptome maps of *in vivo* fetal human brain tissue. We propose to perform and scSeq, single nuclear sequencing (nucSeq), and Dro-Nc-Seq [34, 35] on postnatal human brain (Aim 2). Major cell types can be identified from these single cell data using unsupervised clustering (using the R package Seurat [36])) and confirmed with an alternative hierarchical, non-spectral clustering method, such as reverse graph embedding (e.g., Monocle2; [37]).The presence of known marker genes and Gene Ontology terms are used to further annotate clusters of cells*[38]* and identify reliable clusters.

***Plan:*** By pooling transcripts within cells in a cluster that represent a single major cell type, we will mitigate the inherent variability in single cell RNA-seq and produce a reliable and complete map of cell-type-specific transcriptomes. We will then integrate these maps with tissue-derived regulatory networks (Aim 1c) to infer regulatory relationships in major cell classes, such as deep and superficial excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, and microglia, as well as vascular, immune, and endothelial cells. Moreover, we will perform network-based deconvolution of major cell types from whole tissue transcriptomes [39] and use the results to cross-validate transcriptomes derived from single cell data. We will also apply scImpute [40], with has been shown to produce robust transcriptomes from single cell data, thereby improving identification of cell types and analysis of differential expression. A substantial subset of the inferred regulatory relationships will be validated in Aims 3 and 4, and this validation data can be used to further refine analytic approaches and models.

**1e. Developing approaches for relating regulatory networks to human genomic variation. *Preliminary:*** We have extensive experience in identifying expression quantitative loci (eQTL) and allelic sites. In particular, we have developed the AlleleSeq method, which uses RNA-seq and ChIP-seq data to detect allelic sites, including those associated with gene expression and TF binding [41]. Furthermore, AlleleSeq constructs personal diploid genomes. Using AlleleSeq, we have spearheaded allele-specific analyses as part of our efforts in several major consortia, including ENCODE and the 1000 Genomes Project[3, 12, 42]. We have further developed AlleleSeq and applied the new version to 1,139 RNA-seq and ChIP-seq datasets for 382 samples in the 1000 Genomes Project, which enabled us to annotate the 1000 Genomes Project SNP catalog with allelic information. We constructed a database (AlleleDB) to house all the results as a resource. Both AlleleSeq and AlleleDB have are widely used by the scientific community. Recently, we also developed PrivaSeq, a tool to quantify how much individual-characterizing information is leaked by eQTLs[43].

***Plan:*** We will implement a harmonized pipeline to integrate the analyses of eQTLs, chromatin QTLs (cQTLs) and allelic sites using large datasets (i.e., PsychENCODE, CommonMind, and GTEx). Such large-scale harmonization is acutely sensitive to batch effects, which our pipeline will try to remove. The gene expression matrix will be normalized according to sex, age, RNA Integrity Number (RIN) and library preparation batch for QTL analysis. We will develop a uniform imputation approach when processing genotype data from different projects. We will also use both 1000 Genomes Project and the recently released HRC Reference Panel for imputation on the Michigan imputation server. We will use Matrix eQTL and the FastQTL packages for eQTL identification. Finally, we will correct for multiple hypothesis tests of SNPs in linkage disequilibrium for a given gene during QTL analysis. Using these pipelines, we will harmonize the variation, regulatory and gene expression data from different large consortia and build a comprehensive QTL catalog. We will integrate the QTL catalog with AlleleDB, enhancers and gene expression to create a brain-specific multi-layered human genome variation database. The allelic sites will be used to bolster the power of QTL detection using the open-source WASP tool [44].

**1f. Integrative modeling. *Preliminary:*** Based on machine learning and network approaches, we have developed various integrated methods to model gene regulatory mechanisms. For example, we applied statistical models to characterize the relationships between the extent of TF binding and gene expression by integrating ChIP-seq and RNA-seq data [45]. Recently, we developed DREISS, a method to integrate a state-space model with dimensionality reduction using matrix factorization to identify the temporal expression patterns for various biological processes, such as the oscillation and degradation expression patterns during the cell cycle, embryonic development, and cancer progression [46]. We have also developed Loregic, a method to characterize the gene regulatory logics in complex systems [17]. We used Loregic to identify the cooperative logic among TFs binding to promoters and enhancers in leukemia by integrating ENCODE and TCGA data. We also have extensive experience in using the network framework to integrate human variation data. Our NetSNP method [47] quantifies the indispensability of each gene by incorporating multiple network and evolutionary properties. Based on network properties and other genomic features, we have developed FunSeq [47] and FunSeq2 [49] for prioritizing somatic variants. Using 1000 genomes data, we have prioritized mutations in non-coding regions that may cause diseases [47].

***Plan:*** We will model gene regulatory networks at the systems level to study how human variation affects psychiatric diseases. Our modeling will include major genomic regulatory elements (including enhancers) and will integrate genomic variation and single cell data. We plan to use a matrix formalism that includes several matrices and vectors: the *G* matrix denotes the expression of genes in individual tissues. We can either include all ~20,000 protein-coding genes or just biomarker genes for psychiatric diseases. The *T* vector represents the expression levels of all TF genes. The *E* matrix represents the genotypes at select eQTLs for all individuals. Gene expression can thus be mathematically modeled as a function of TF expression and eQTLs, which is further expressed as matrix operation: *G* = *R* X *T* + *Q* X E, where *R* and *Q* are two matrices capturing the linear contributions of TF expression and eQTLs to gene expression and X denotes matrix multiplication. This formulism can be separately applied to the binding of TFs to the promoters and enhancers of their respective target genes (i.e., *R* = *Rpromoter* + *Renhancer*). By integrating the data on all individuals, our goal is to estimate the *R* and *Q* matrices, and to find the conserved and individual-specific network structures (i.e., by analyzing the network homogeneity and heterogeneity of *R* X *T* and *Q* X E). We will extend this approach to decompose each tissue into its constituent cell types, with the fractions of the cell types dependent on the genotypes of the individual, denoted by the matrix *S*. The single cell RNA-seq data generated in this project will provide cell type-specific gene expression, denoted by the matrix *C*. The equation thus becomes: *G* = *C* X *S* = *R* X *T* + *Q* X E. We will estimate *S* based on *G* and *C* by minimizing total error. Alternatively, we can compute *S* by pinv(*C*) X (*R* X *T* + *Q* X E), where pinv(*C*) is the pseudo-inverse of matrix *C*, i.e., pinv(*C*) X *C* = the identity matrix. The final equation reveals that the individual relationship between genotypes (*E*) and phenotypes (*S*), and the matrix *Q*'= pinv(*C*) X *Q* quantifies how genotypes affect phenotypes.

**Aim 2. Apply analytical methods to the PsychENCODE data corpus, integrating data from GTEx, ENCODE, and other consortia, annotating GWAS SNPs associated with psychiatric diseases, prioritizing the discovered regulatory elements for validation, and visualizing all data and annotations in an integrated fashion.**

**2a. Overview.** Aim 2 has two related goals. We will first apply the novel methods we have described in Aim 1 to the PsychENCODE data corpus. To eliminate batch effects and to ensure data quality, we will first process all PsychENCODE data uniformly. Data from other public sources will be incorporated into our analyses, including but not limited to data from GTEx, ENCODE, CommonMind, and BrainSpan. We will calculate eQTLs and prioritize GWAS SNPs associated with psychiatric diseases. From single cell analysis of cerebral cortical cell types at several developmental stages, we will build cell type-specific networks and identify the cell types implicated in psychiatric diseases. Combining all these results, we will identify candidate enhancers and SNPs for experimental validation in Aims 3 and 4. Secondly, we will generate a comprehensive, uniformly processed data and annotation resource from PsychENCODE and other relevant consortia. This resource will support our project's effort of developing new methods for identifying and testing enhancers and variants therein. Furthermore, the resource (which includes data, metadata, processing and analysis code, in addition to the resulting annotations), will be shared with the greater PsychENCODE Consortium and released to the research community. **The resource will thus support the research efforts of other members of the PsychENCODE Consortium and substantially increase its impact.**

**2b. We will process all PsychENCODE datasets using uniform processing pipelines.** ***Preliminary:*** We have implemented ENCODE RNA-seq, ChIP-seq, and ATAC-seq uniform processing pipelines in a Protected Data Cloud (PDC) and applied them to the PsychENCODE data that is currently available. The RNA-seq pipeline includes data organization, format conversion, and quality assessment. Specifically, we use STAR [48] to align the sequencing quality filtered reads to the human genome and RSEM [51] to quantify expression profiles of each GENCODE-annotated transcript. Our quality control (QC) measures assess sequencing errors, ribosomal RNA contamination, DNA contamination, gene coverage uniformity, and the correlation between technical and biological replicates. The ChIP-seq pipeline includes QC steps, read alignment, peak calling, peak comparison, peak annotation, motif analysis and super-enhancers identification. The Gerstein Lab has developed two peak calling algorithms, PeakSeq [49] and MUSIC[5] . PeakSeq calls the peaks for transcription factor ChIP-seq data and is used by the ENCODE and modENCODE consortia. MUSIC performs multi-scale decomposition of ChIP signals to enable simultaneous and accurate detection of enrichment at a wide range of peak breadths. MUSIC is particularly applicable to histone modifications and some transcription factors that display both punctate and broad regions of enrichment. The ATAC-seq pipeline has similar QC and processing steps and uses MACS2 as the peak caller [50].

***Plan:*** We will continue to improve these uniform-processing pipelines and build additional pipelines for new data types, e.g., the single cell RNA-seq data that are described below. We will process all PsychENCODE data using these uniform pipelines before integration. Furthermore, we will process all other publicly available datasets that we plan to incorporate into our integrative analysis using these pipelines, including but not limited to the large amounts of data developed by various consortia.

**2c. Single cell capture, RNA-seq library preparation, and sequencing. *Preliminary:*** We have defined a consistent and reproducible molecular pathology in tissues from brain regions that implicate cell-specific transcriptional regulation in ASD [33, 51].To identify the major regulatory cell types in the cerebral cortex and their dysregulation in psychiatric diseases, we have developed a robust and highly parallel technology for profiling single nuclei/cell transcriptomics (scSeq/nucSeq) in frozen postmortem brain tissues both in vitro and in vivo (Fig 1). We have processed over 40,000 cells, providing a demonstration of the methods, and an unprecedented atlas of cell types and their molecular composition in the developing human cerebral cortex.

***Plan:*** We will perform scSeq/nucSeq on the postmortem cerebral cortex and cerebellum from at least 15 ASD cases, along with matched healthy controls. We will cover three major epochs: infancy and childhood (age 2-10), adolescence (10-20), and adulthood (20-40), which not only parallel key changes in ASD-associated gene expression [51] but also represent critical epochs in psychiatric disease risk. We will integrate these single cell data with the ATAC-seq and Hi-C data being produced in the Geschwind Lab in whole tissue or bulk neurons versus glia. In addition to the scSeq method, we will apply a slightly modified version (Dro-Nc-seq), to allow profiling of nuclei from the postnatal brain which, unlike the fetal brain, cannot be easily dissociated for standard scRNAseq [34]. Dro-Nc-seq correlates highly with Drop-seq, thereby enabling the detection of specific cell classes and profiles. It has been shown to work for frozen postmortem human brain tissue [34]. Individual cells are rapidly isolated, captured and processed in nanoliter-sized droplets using microfluidics [34]. Dro-Nc-seq incorporates unique molecular indexes during amplification to allow elimination of PCR amplification artifacts. We will profile over 6,000 cells per sample (30,000 per stage in 5 samples), which is sufficient to detect rare cell classes that comprise only 0.05% of the total cell population (Poisson distribution and empirical results). We will also use Drop-seq in Aim 4 to validate regulatory relationships at the level of specific cell classes after enhancer activation or deletion by genome engineering of *in vitro* human neural model systems.

|  |
| --- |
|  |
| **Figure 1:** Single cell transcriptome analyses of the developing human cerebral cortex and 3D *in vitro* model of neuronal development using human forebrain spheroids (hFS). (A) Diagram of the cortical anlage and locations of its main cell types. (B, C) t-SNE visualization of 8,000 single cell transcriptomes from the human cortical anlage at **GW17. (D)** Highly parallel, single cell analysis of hFS(n= 11,838 cells; BDTM Resolve system). **Major hFS single cell clusters are identified by the enriched genes that match the *in vivo* clusters in C**. The hFS clusters represent glutamatergic neurons (VGLUT1+) expressing the cortical layer markers TBR1, FEZF2, CTIP2; intermediate progenitors expressing TBR2, INSM1 and HES6; dorsal progenitors expressing LHX2, PAX6, and GLAST1 that also encompass HOPX+ outer radial glia-like cells (oRG). hSS included a small group of oligodendrocyte progenitors (OLIG2, SOX10), ventral progenitors, as well as a group of GABA-ergic cells expressing GAD1, SLC32A1, SCG2, and SST. The data in (D) were obtained from the Pasca Lab at Stanford (see letter). These and the data in a recently published manuscript {PMID 28445565} show that the *in vitro* model contains all of the major neuronal and glial cell classes defined in fetal brain *in vivo* and demonstrate our ability to use scSeq to profile their transcriptomes. |

**2d. Integrate data from other consortia. *Preliminary:*** To increase the power for meta-analysis, we will incorporate data from the ENCODE, GTEx, CommonMind, BrainSpan and Roadmap consortia. We have extensive experience in performing large[-scale](http://mccarrolllab.com/dropseq/%29) **integrative analyses.** We have played key or lead roles in the DOE KBase, Brainspan, ENCODE, modENCODE, 1000 Genomes, PCAWG, and exRNA consortia. We work in multi-disciplinary teams and interact with scientists and physicians of highly diverse backgrounds within these consortia. We have applied simulation, machine learning, and knowledgebase design for working with multi-layered datasets.

***Plan:*** We plan to calibrate the standard for integrating the data based on our extensive experience in these consortia [1, 6, 9, 28, 52]. We will uniformly process all the datasets using the pipelines detailed in Aim 2b. We will develop calibration methods to generate unified scoring for all datasets. For example, to unify the gene expression data in the same brain region by GTEx and PsychENCODE, we will directly compare the identified transcripts and their expression levels and look at discrepancies between the two projects and calibrate normalization method to reduce any bias that is introduced by to batch effects. We will parse the brain cell types in the whole-tissue data using our single cell studies (Aim 2c), match them with the most appropriate datasets in PsychENCODE, and investigate whether using the same uniform pipeline in these consortia detects a comparable set of regulatory elements. While performing this comparison, we will take into account the differences in cell sources and the inherent variation among biological replicates, and focus on the regions and transcripts deemed most significant by all datasets. If we identify major differences, we will investigate whether they are due to the underlying raw data, or to differences in data processing (such as the parameters used, for instance).

**2e. Perform integrative analysis to identify enhancers and prioritize their GWAS SNPs for psychiatric diseases.** ***Preliminary*:** By now, we have acquired massive datasets from PsychENCODE and other consortia (Aim 2d), produced our own single cell transcriptome data (Aim 1d and 2c), built uniform pipelines for processing all these data types (Aim 2b), and developed a battery of methods and pipelines for their analysis (Aim 1). Here we will put these all together and perform the most comprehensive and biology-driven analysis to predict enhancers and prioritize GWAS SNPs for psychiatric diseases. These predictions will be tested in Aims 3 and 4.

***Plan:*** We will first process all data using the uniform processing pipelines detailed in Aim 2b. We will then call enhancers in specific cell and tissue types using the SVM and matched filter approach (Aim 1b). 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. Features with larger weights are predicted to be more important in discriminating enhancers from non-regulatory regions in the model. We will determine the extent to which this pattern is conserved across a diverse set of tissues. We will build cellular regulatory networks based on single cell cerebral cortical transcriptomes from four major periods of brain development (fetal, early childhood; adolescence and adult) to capture major developmental epochs relevant to psychiatric disease (Aim 2c). We will compare these relationships to validate cell type specific transcriptional and enhancer dysregulation that we have identified in ASD by inference from bulk tissue[33, 51] Given the clear transcriptomic and polygenic overlap between ASD and other major psychiatric disorders such as SCZ [53], these data will also be valuable for linking sequence variation to cell class-specific gene regulatory networks across psychiatric conditions. We will analyze the structures of these networks (Aim 1c) and identify the eQTLs, allelic sites, and GWAS SNPs for the enhancers and map them to the networks (Aim 1e). We will identify those enhancers and TFs that are most influential within the network hierarchy. Finally we will perform matrix-based integration on all data (Aim 1f). We plan to compare the *Q*' matrices across different psychiatric diseases and identify disease-specific genotype-phenotype relationships. We will also analyze the gene regulatory circuits, such as the cooperative logic between multiple regulatory factors or enhancers in the regulatory networks of brain regions, cell types, and psychiatric disorders. Based on these analyses, we will produce a set of enhancers and SNPs to be tested in Aims 3 and 4. The testing results will be incorporated into refinements to the model for further development, thereby improving the next round of predictions.

**2f. Coordinating and sharing PsychENCODE data, metadata, and annotations. *Preliminary:*** Sage Bionetworks (under the leadership of Mette Peters) will incorporate all PsychENCODE data into the Synapse system (www.synapse.org), a platform developed by Sage Bionetworks to support scientific collaborations centered on shared biomedical data. Sage Bionetworks has functioned as a data-coordinating center and data analysis core for several dozen different consortia, where the focus has been on creating open, collaborative cultures supported by the Synapse system. We support the NCI-funded TCGA Pan Cancer Consortium and the NIA-funded Accelerating Medicines Partnership. Five years ago, we co-funded the CommonMind Consortium (CMC) in collaboration with partners in industry, academia, and the NIMH. The CMC is a pre-competitive partnership that grew out of the pressing need for data on neuropsychiatric disorders, and Synapse is used to capture and share information about every step in the research process (www.synapse.org/cmc). The success of the CMC model prompted the NIMH to support the use of Synapse in additional consortia, including the first phase of the PsychENCODE Consortium (www.synpase.org/pec) and the Brain Somatic Mosaicism Network (BSMN; www.synapse.org/bsmn). Together, these NIMH-supported consortia include over 150 researchers from 16 institutions that have collectively generated the largest molecular dataset from brain tissue of individuals diagnosed with neuropsychiatric disorders. Synapse tracks samples and stores content in a coordinated, centralized manner. The data are initially shared with other consortium members, followed by dissemination to the broader research community. Several Synapse features promote reproducibility. A ‘Provenance' system describes the connections between the workflow steps. Versioning of content automatically increments with each update, thereby allowing data freezes. Metadata tools capture multiple aspects of the data, including its provenance, a time stamp, depositor, etc.

***Plan:* 1. Integration of phase I and phase II PsychENCODE data.** We currently house the data generated by the 18 grants in Phase I of the PsychENCODE Project. We will continue to support consortium and public access to the data. In addition, we have expanded to accommodate additional data, protocols, and analysis results generated by the phase II grants. To maximize the utility of the data and other Synapse resources (such as the contents by CMC and BSMN), we have created standardized metadata (as defined by ontologies) and will apply them to all datasets. This will make it easy to discover and perform data analysis across diseases, tissues, cell types and assays. As part of the integration effort, we will track datasets with their subject de-identified samples. For example, tissue samples from over 1,000 donors in the brain tissue collections at Mount Sinai, University of Pennsylvania, the University of Pittsburgh and the NIMH Human Brain Collection Core have been assayed by multiple studies within PsychENCODE, CMC, and BSMN. Our system allows automatic identification of the data from the same samples across projects, which empowers integrative analyses of pan-omic data. **2. Support of the uniformly processed data resource**. Starting with the raw input data, we will manage and disseminate the uniformly processed data resource by building infrastructure in Synapse that tracks all processing steps and analysis output. The output from the method developed in Aim 1, data processing in Aim 2, and regulatory element validation in Aims 3 and 4 will be loaded back to Synapse, thereby providing full transparency of the analytical processes. All collaborating teams will have access to raw and processed data, metadata, code used in the pipelines, and the analytical results. Public access will be given per established data release schedule. **3. Integrating the data resource in Synapse with psychSCREEN.** One important way to increase the impact of the PsychENCODE Consortium is to release all datasets to the community. We will build a PsychENCODE Portal to release all curated data, metadata, and analysis results to the broader research community annually. The Portal will be powered by a web-based engine (psychSCREEN) for searching and visualizing the entire registry of candidate regulatory elements in the human genome, along with their associated activities across all PsychENCODE samples. psychSCREEN will be modeled after the SCREEN tool built by the Weng lab for ENCODE (http://screen.umassmed.edu). The user can search the annotations in any specified locus and the disease-associated variants it harbors, and visualize the underlying experimental data via the UCSC Genome Browser. We will develop a framework to integrate psychSCREEN with data in Synapse. The model for this will be inspired by similar initiatives, such as what had been done in the Progenitor Cell Biology Consortium (PCBC), where there was interest in interactive visual explorers of genomic data. We built several tools that allowed people to explore expression data and regulatory mechanisms of this expression. Additionally, we integrated this with GTEx expression data, thereby enabling users to compare signatures of expression between stem cells characterized in the PCBC and tissue-specific expression as captured by GTEx.

**Elements unique to this site:** Our site will be the computational and analytical component of the proposal (**Aims 1 and 2**), consisting of investigators in the labs of Zhiping Weng at the University of Massachusetts Medical School, Mark Gerstein at Yale University, Daifeng Wang at Stony Brook University and Mette Peters at Sage Bionetworks. The Gerstein Lab will develop a number of standardized pipelines and quality control metrics, provide a platform and infrastructure for uniform processing of the data, run the pipelines, focus on the discovery of brain-specific genes, perform aggregated quantitative trait locus (QTL) analysis and single cell deconvolution, as well as integrate all of the datasets for meta-analysis. The Weng Lab will support the enhancer analysis, annotate disease-associated enhancers, and discover functional genomic elements associated with psychiatric diseases using an integrative approach. The Weng Lab will also develop the psychSCREEN tool for searching the ~2 M predicted regulatory elements and visualizing all annotations and underlying raw data associated with individual elements. The Wang Lab will work to identify brain gene expression dynamics, perform gene co-expression network analysis, and model the gene regulatory networks. Sage Bionetworks will develop a collaborative space (Synapse) for centralized storage of data, protocols, analysis methods, and results generated by this project, in addition to implementing a data release process for the collection and verification of data from the various production centers in PsychENCODE. This group will interact frequently with the experimental component of the proposal, provide them with enhancer and genetic variant predictions and use their testing data to further improve the computational methods.

References

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

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

3. Gerstein, M.B., et al., *Comparative analysis of the transcriptome across distant species.* Nature, 2014. **512**(7515): p. 445-8.

4. Yip, K.Y., et al., *Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors.* Genome Biol, 2012. **13**(9): p. R48.

5. Harmanci, A., J. Rozowsky, and M. Gerstein, *MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework.* Genome Biol, 2014. **15**(10): p. 474.

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

7. Negre, N., et al., *A cis-regulatory map of the Drosophila genome.* Nature, 2011. **471**(7339): p. 527-31.

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

9. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science, 2010. **330**(6012): p. 1775-87.

10. Yan, K.K., et al., *Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks.* Proc Natl Acad Sci U S A, 2010. **107**(20): p. 9186-91.

11. Cheng, C., et al., *Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data.* PLoS Comput Biol, 2011. **7**(11): p. e1002190.

12. Boyle, A.P., et al., *Comparative analysis of regulatory information and circuits across distant species.* Nature, 2014. **512**(7515): p. 453-6.

13. Yu, H. and M. Gerstein, *Genomic analysis of the hierarchical structure of regulatory networks.* Proc Natl Acad Sci U S A, 2006. **103**(40): p. 14724-31.

14. Bhardwaj, N., P.M. Kim, and M.B. Gerstein, *Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators.* Sci Signal, 2010. **3**(146): p. ra79.

15. Bhardwaj, N., et al., *Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets.* PLoS Comput Biol, 2010. **6**(5): p. e1000755.

16. Bhardwaj, N., K.K. Yan, and M.B. Gerstein, *Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels.* Proc Natl Acad Sci U S A, 2010. **107**(15): p. 6841-6.

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

18. Yu, H., et al., *TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics.* Nucleic Acids Res, 2004. **32**(1): p. 328-37.

19. Yip, K.Y., et al., *The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks.* Bioinformatics, 2006. **22**(23): p. 2968-70.

20. Douglas, S.M., G.T. Montelione, and M. Gerstein, *PubNet: a flexible system for visualizing literature derived networks.* Genome Biol, 2005. **6**(9): p. R80.

21. Luscombe, N.M., et al., *Genomic analysis of regulatory network dynamics reveals large topological changes.* Nature, 2004. **431**(7006): p. 308-12.

22. Qian, J., et al., *Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data.* Bioinformatics, 2003. **19**(15): p. 1917-26.

23. Yu, H., et al., *Genomic analysis of gene expression relationships in transcriptional regulatory networks.* Trends Genet, 2003. **19**(8): p. 422-7.

24. Cheng, C., et al., *mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer.* Genome Biol, 2009. **10**(9): p. R90.

25. Cheng, C., R. Min, and M. Gerstein, *TIP: A probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles.* Bioinformatics, 2011. **27**(23): p. 3221-3227.

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

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

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

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

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

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

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

33. Parikshak, N.N., et al., *Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism.* Cell, 2013. **155**(5): p. 1008-21.

34. Habib, N., et al., *DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq.* bioRxiv, 2017.

35. Lake, B., et al., *Integrative Single-Cell Analysis By Transcriptional And Epigenetic States In Human Adult Brain.* bioRxiv, 2017.

36. Macosko, E.Z., et al., *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.* Cell, 2015. **161**(5): p. 1202-14.

37. Qiu, X., et al., *Reversed graph embedding resolves complex single-cell developmental trajectories.* bioRxiv, 2017.

38. Pollen, A.A., et al., *Molecular identity of human outer radial glia during cortical development.* Cell, 2015. **163**(1): p. 55-67.

39. Oldham, M.C., et al., *Functional organization of the transcriptome in human brain.* Nat Neurosci, 2008. **11**(11): p. 1271-82.

40. Li, W.V. and J.J. Li, *scImpute: Accurate And Robust Imputation For Single Cell RNA-Seq Data.* bioRxiv, 2017.

41. Rozowsky, J., et al., *AlleleSeq: analysis of allele-specific expression and binding in a network framework.* Mol Syst Biol, 2011. **7**: p. 522.

42. Chen, J., et al., *A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals.* Nat Commun, 2016. **7**: p. 11101.

43. Harmanci, A. and M. Gerstein, *Quantification of private information leakage from phenotype-genotype data: linking attacks.* Nat Methods, 2016. **13**(3): p. 251-6.

44. van de Geijn, B., et al., *WASP: allele-specific software for robust molecular quantitative trait locus discovery.* Nat Methods, 2015. **12**(11): p. 1061-3.

45. Cheng, C., et al., *Understanding transcriptional regulation by integrative analysis of transcription factor binding data.* Genome Res, 2012. **22**(9): p. 1658-67.

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

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

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

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

50. Liu, T., *Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells.* Methods Mol Biol, 2014. **1150**: p. 81-95.

51. Sun, W., et al., *Histone Acetylome-wide Association Study of Autism Spectrum Disorder.* Cell, 2016. **167**(5): p. 1385-1397 e11.

52. Pei, B., et al., *The GENCODE pseudogene resource.* Genome Biol, 2012. **13**(9): p. R51.

53. Gandal, M.J., et al., *Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap.* bioRxiv, 2016.

References

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

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

3. Gerstein, M.B., et al., *Comparative analysis of the transcriptome across distant species.* Nature, 2014. **512**(7515): p. 445-8.

4. Yip, K.Y., et al., *Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors.* Genome Biol, 2012. **13**(9): p. R48.

5. Harmanci, A., J. Rozowsky, and M. Gerstein, *MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework.* Genome Biol, 2014. **15**(10): p. 474.

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

7. Negre, N., et al., *A cis-regulatory map of the Drosophila genome.* Nature, 2011. **471**(7339): p. 527-31.

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

9. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science, 2010. **330**(6012): p. 1775-87.

10. Yan, K.K., et al., *Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks.* Proc Natl Acad Sci U S A, 2010. **107**(20): p. 9186-91.

11. Cheng, C., et al., *Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data.* PLoS Comput Biol, 2011. **7**(11): p. e1002190.

12. Boyle, A.P., et al., *Comparative analysis of regulatory information and circuits across distant species.* Nature, 2014. **512**(7515): p. 453-6.

13. Yu, H. and M. Gerstein, *Genomic analysis of the hierarchical structure of regulatory networks.* Proc Natl Acad Sci U S A, 2006. **103**(40): p. 14724-31.

14. Bhardwaj, N., P.M. Kim, and M.B. Gerstein, *Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators.* Sci Signal, 2010. **3**(146): p. ra79.

15. Bhardwaj, N., et al., *Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets.* PLoS Comput Biol, 2010. **6**(5): p. e1000755.

16. Bhardwaj, N., K.K. Yan, and M.B. Gerstein, *Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels.* Proc Natl Acad Sci U S A, 2010. **107**(15): p. 6841-6.

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

18. Yu, H., et al., *TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics.* Nucleic Acids Res, 2004. **32**(1): p. 328-37.

19. Yip, K.Y., et al., *The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks.* Bioinformatics, 2006. **22**(23): p. 2968-70.

20. Douglas, S.M., G.T. Montelione, and M. Gerstein, *PubNet: a flexible system for visualizing literature derived networks.* Genome Biol, 2005. **6**(9): p. R80.

21. Luscombe, N.M., et al., *Genomic analysis of regulatory network dynamics reveals large topological changes.* Nature, 2004. **431**(7006): p. 308-12.

22. Qian, J., et al., *Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data.* Bioinformatics, 2003. **19**(15): p. 1917-26.

23. Yu, H., et al., *Genomic analysis of gene expression relationships in transcriptional regulatory networks.* Trends Genet, 2003. **19**(8): p. 422-7.

24. Cheng, C., et al., *mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer.* Genome Biol, 2009. **10**(9): p. R90.

25. Cheng, C., R. Min, and M. Gerstein, *TIP: A probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles.* Bioinformatics, 2011. **27**(23): p. 3221-3227.

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

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

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

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

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

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

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

33. Parikshak, N.N., et al., *Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism.* Cell, 2013. **155**(5): p. 1008-21.

34. Habib, N., et al., *DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq.* bioRxiv, 2017.

35. Lake, B., et al., *Integrative Single-Cell Analysis By Transcriptional And Epigenetic States In Human Adult Brain.* bioRxiv, 2017.

36. Macosko, E.Z., et al., *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.* Cell, 2015. **161**(5): p. 1202-14.

37. Qiu, X., et al., *Reversed graph embedding resolves complex single-cell developmental trajectories.* bioRxiv, 2017.

38. Pollen, A.A., et al., *Molecular identity of human outer radial glia during cortical development.* Cell, 2015. **163**(1): p. 55-67.

39. Oldham, M.C., et al., *Functional organization of the transcriptome in human brain.* Nat Neurosci, 2008. **11**(11): p. 1271-82.

40. Li, W.V. and J.J. Li, *scImpute: Accurate And Robust Imputation For Single Cell RNA-Seq Data.* bioRxiv, 2017.

41. Rozowsky, J., et al., *AlleleSeq: analysis of allele-specific expression and binding in a network framework.* Mol Syst Biol, 2011. **7**: p. 522.

42. Chen, J., et al., *A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals.* Nat Commun, 2016. **7**: p. 11101.

43. Harmanci, A. and M. Gerstein, *Quantification of private information leakage from phenotype-genotype data: linking attacks.* Nat Methods, 2016. **13**(3): p. 251-6.

44. van de Geijn, B., et al., *WASP: allele-specific software for robust molecular quantitative trait locus discovery.* Nat Methods, 2015. **12**(11): p. 1061-3.

45. Cheng, C., et al., *Understanding transcriptional regulation by integrative analysis of transcription factor binding data.* Genome Res, 2012. **22**(9): p. 1658-67.

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

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

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

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

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

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

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

53. Liu, T., *Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells.* Methods Mol Biol, 2014. **1150**: p. 81-95.

54. Sun, W., et al., *Histone Acetylome-wide Association Study of Autism Spectrum Disorder.* Cell, 2016. **167**(5): p. 1385-1397 e11.

55. Pei, B., et al., *The GENCODE pseudogene resource.* Genome Biol, 2012. **13**(9): p. R51.

56. Gandal, M.J., et al., *Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap.* bioRxiv, 2016.