2/2 Discovery and validation of neuronal enhancers associated with the development of psychiatric disorders

This project uniquely combines the core functions of a Data Analysis Center (DAC) and Data Coordination Center (DCC) in the PsychENCODE project with experimental validation efforts based on analysis performed by those computational efforts. Analysis efforts will also make use of single cell sequencing which is the most advanced method for analysis of both tissue composition and cell type specific signaling pathways. Validations will be performed using genomic-scale enhancer functional assays . Additionally the project will allow for deeper analysis on a subset of validated enhancers that are implicated in neuropsychiatric disease by the DAC/DCC effort. We will use stem cells differentiated into organoids to measure the functional relevance of these enhancers during the development of different neuronal lineages, and we will determine whether there are differences in enhancer activity for alleles associated with affected vs. unaffected individuals. This project will thus provide a comprehensive analysis of neuronal cell enhancer annotation and function associated with human brain neuronal development and psychiatric disorders.

Aim 1 - Develop novel methods to find brain-specific enhancers, build regulatory networks, deconvolve brain-region-specific regulation, and relate enhancers to variation

We will develop a new machine-learning framework that combines pattern recognition of various epigenomic signals with sequence-based features to predict active enhancers across different brain regions. This method will be parameterized based on STARR-seq data generated in the project. We will link the enhancers into larger regulatory networks and circuits. We will develop uniform expression quantitative loci (eQTL) analysis and chromatin QTL (cQTL) pipelines for identifying genetic variants associated with differential activity across individuals. Finally, we develop methods to process brain single-cell RNA-seq data to find cell-type specific patterns of gene expression. We will combine all these methods to build integrative models of gene regulation incorporating, which will enable us to predict gene expression perturbations associated with disease variants in specific tissues in different brain regions.

Aim 2. Apply analytical methods to the psychENCODE data corpus, integrating data from 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.

PsychENCODE datasets will be uniformly processed. We will then apply the methods developed in Aim 1 and other methods we have already developed. We will integrate the PsychENCODE data with data from the GTEx, ENCODE, CommonMind, and BrainSpan consortia. 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). Based on the integrative model developed in Aim 1, we will select candidate enhancers and variants for experimental testing in Aims 3 and 4. One important way to increase the impact of the psychENCODE consortium is to release data to the community. We will build a PsychENCODE Portal to release all curated data, metadata, and analysis results to the broader research community at yearly intervals. The Portal will be powered by a web-based engine psychSCREEN for searching and visualizing the annotations in any specified locus and the disease-associated variants it harbors, and visualize the underlying experimental data via the UCSC Genome Browser.

Aim 3. Systematic genome-wide validation of PsychENCODE Regulatory Elements

The output of Aims 1 and 2 will be a series of predictions of which regulatory elements are functionally relevant for the expression of neuronal genes that are involved in normal development and/or neuropsychiatric disorders that are a focus of the PsychENCODE consortium. Biological validation of these predictions requires testing the regulatory potential of the genomic regions identified, including allele-specific quantification. Using methods we have used in the ENCODE and modENCODE projects, we will systematically functionally validate predicted regulatory elements. Specifically, we have developed a protocol for testing enhancers genome-wide for the whole human genome, based on the STARR-seq methodology. We will test predictions for enhancer activity in cell models, including primary human neuronal precursor cells. For up to 200 validated enhancers we will test effects on nearby gene expression using CRISPR knockout.

Aim 4. Biological Validation of PsychENCODE Regulatory Elements

We have developed novel 3D organoid cultures based using forebrain spheroid and primary human precursor neuronal cells that can be differentiated into different neuronal lineages. Additionally, we have grown such neuronal organoids in microfluidic chips that allow the rapid testing of large numbers of conditions. Choosing from validated enhancers from Aim 3, we will synthesize 100 validated enhancers with polymorphisms predicted to affect function between alleles. We will transfect reporter constructs into cells differentiated into different lineages and under different conditions to study single cells based on droplet RNA sequencing technology and on microfluidic chips.

Significance: The rich data generated by the PsychENCODE Consortium are a preeminent resource for studying regulatory mechanisms in the human brain [1]. One of its unique aspects is the coverage of major psychiatric diseases, such as autism spectrum disorder (ASD) and schizophrenia (SCZ). PsychENCODE datasets have been assembled by many investigators over several years, and they are housed in a central depository (www.synapse.org) and shared with the public. These data 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. Leveraging these valuable datasets, we propose to conduct comprehensive, integrative analyses to find non-coding functional elements in brain neurons (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, and with 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, and on 3D cortical forebrain spheroids. Finally, this project will continue to manage all PsychENCODE data at Synapse, augmented by a psychSCREEN web engine that allows users to directly query specific regulatory elements, as well as visualize their annotations and underlying signal profiles 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 ASD and control cerebral cortex and cerebellum during four critical epochs of psychiatric disease risk (Aims 1 & 2). These transcriptome maps will enable detailed analysis of the cell types associated with these diseases. Our STARR-seq assay will measure 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 primary human neural progenitors (phNPCs) and 3D cortical forebrain spheroids (hFS) 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 are millions of regulatory elements, accessing individual elements in real time requires many software engineering innovations. 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 investigate the biological mechanisms of how genotypes drive gene expression patterns underlying psychiatric phenotypes. In particular, we plan to develop machine learning and pattern recognition methods that integrate 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 [3, 4], 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 [5], 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 a large range of cell types, and in this project will assay cell types that include primary human neuronal progenitors and their differentiated neurons. We will first call peaks in STARR-seq profiles by extending our MUSIC method for calling histone mark peaks [6] and calculating statistical significance using a Poisson model. STARR-seq peaks that score <5% false discovery rate 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) to predict STARR-seq enhancers.

1c. Building brain-specific networks. ***Preliminary:*** We have previously contributed a large body of work on regulatory networks. We have constructed gene networks of various regulators, including transcription factors (TF) and micro-RNAs (mRNA) and their target genes [7-13]. Upon analyzing the structures of these networks, we found that, compared with centrality, hierarchy levels are better predictors the regulator importance [7, 14-17]. Our network analysis software tools include TopNet [18], tYNA [19], and PubNet [20]. In addition to the global attributes of regulatory networks, we also analyzed local topological features, such as network motifs (e.g., feed-forward loops) [7, 10, 13]. 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]. 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 [9, 10, 30, 31]. To analyze multiple interconnected networks simultaneously, we constructed co-expression networks from the extensive RNA-seq data in various consortia [4].

Plan: We will predict enhancers (Aim 1b) and promoters (from GENCODE annotations) 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 connect these TFs with their target genes if the TF gene expression accurately predicts the target genes' expression using machine learning methods. We will also use chromatin loops defined by Hi-C data in fetal brain [32] as well as new adult neuronal and glial Hi-C data—generated as part of the PsychENCODE Project (Geschwind, PI)—to assign distal regulatory regions to genes. We will build a gene regulatory network for each brain region and psychiatric disease. We will then study the structure and dynamics of our inferred regulatory networks and compare them across brain regions and disease types using the arsenal of methods we have developed. Extending these methods, we will use graph algorithms to discover clusters of highly connected genes within these networks. We expect to find 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 scSeq, single nucleus sequencing (nucSeq) and droplet nuclei sequencing (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 AlleleSeq, which uses RNA-seq and ChIP-seq data to detect allelic sites, including those associated with gene expression and TF binding [41]. AlleleSeq also constructs personal diploid genomes. We have spearheaded allele-specific analyses as part of our efforts in several major consortia, including ENCODE and the 1000 Genomes Project [4, 13, 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 SNP catalog with allelic information. We constructed a database (AlleleDB) to house all the results as a resource. Both AlleleSeq and AlleleDB 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 (e.g., 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 for processing genotype data from different projects. We will also use both 1000 Genomes Project and the 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 integrated methods to model gene regulatory mechanisms. For example, we applied statistical models to characterize relationships between TF binding levels 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 to identify temporal expression patterns for various biological processes, such as cell cycle oscillation and degradation expression patterns, embryonic development, and cancer progression [46]. We developed Loregic, a method to characterize the gene regulatory logic in complex systems [47]. We used Loregic to identify the cooperative interactions among TFs at promoters and enhancers by integrating ENCODE and TCGA data. We also have extensive experience using network frameworks to integrate human variation data. Our NetSNP method 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 for prioritizing mutations in non-coding regions that may cause diseases [48].

Plan: We will model gene regulatory networks at the systems level to study how human variations affect psychiatric diseases. Our modeling will include major genomic regulatory elements 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, further expressed as a matrix operation: $G = R \times T + Q \times E$, where *R* and *Q* are two matrices capturing the linear contributions of TF expression and eQTLs to gene expression (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., $R = R_{promoter} + R_{enhancer}$). 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 $R \times T$ and $Q \times E$). 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: $G = C \times S = R \times T + Q \times E$. We will estimate *S* based on *G* and *C* by minimizing total error. Alternatively, we can compute *S* by $\text{pinv}(C) \times (R \times T + Q \times E)$, where $\text{pinv}(C)$ is the pseudo-inverse of matrix *C*, i.e., $\text{pinv}(C) \times C$ = the identity matrix. The final equation reveals that the individual relationship between genotypes (*E*) and phenotypes (*S*), and the matrix $Q' = \text{pinv}(C) \times Q$ 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 data. This resource will support our project's effort of developing new methods for identifying and testing enhancers and variants therein. Furthermore, the resource will be released to the public. **The resource will thus support the research efforts of other members of the PsychENCODE Consortium and substantially increase the impact of the PsychENCODE Project.**

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 [49] to align the quality filtered sequencing reads to the human genome and RSEM to quantify expression profiles of each GENCODE-annotated transcript. Our quality control (QC) measures assess sequencing errors, ribosomal RNA contamination, gene coverage uniformity, and the correlation between technical and biological replicates. The ChIP-seq pipeline includes QC steps, read alignment, peak calling, motif analysis and super-enhancers identification. The Gerstein Lab has developed two peak calling algorithms, PeakSeq [50] and MUSIC [6]. 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 [51].

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.

2c. Single cell capture, RNA-seq library preparation, and sequencing. ***Preliminary:*** We have defined consistent and reproducible molecular signatures in tissues from brain regions that implicate cell-specific transcriptional regulation in ASD [33, 52]. To identify the major cell types with specific transcriptional programs 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 and nucSeq on the postmortem cerebral cortex and cerebellum from at least 15 ASD cases, along with matched 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 [52] 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 comparing 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 in human neural model systems.

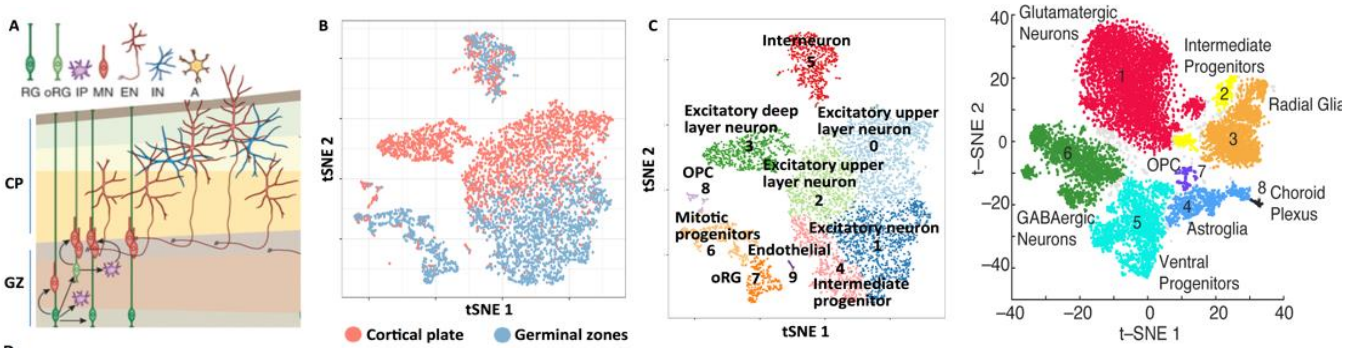


Figure 1: Single cell transcriptome analyses of the developing human cerebral cortex and 3D *in vitro* model of neuronal development using human forebrain spheroids (hFS). (A) Diagram of the cortical anlage and locations of its main cell types. (B, C) t-SNE visualization of 8,000 single cell transcriptomes from the human cortical anlage at **GW17. (D)** Highly parallel, single cell analysis of hFS (n= 11,838 cells; BD™ 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 (TBR2, INSM1 and HES6); dorsal progenitors (LHX2, PAX6, and GLAST1). 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 [2] how 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 **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. We have applied simulation, machine learning, and knowledgebase design for working with multi-layered datasets.

Plan: We plan to perform data harmonization across datasets based on our extensive experience in these consortia [1, 7, 10, 28, 53]. We will uniformly process all the datasets using the pipelines detailed in Aim 2b. We will develop calibration methods to perform unified scoring for all datasets. 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 GWAS SNPs associated with 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 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 use these networks to validate cell type specific transcriptional and enhancer dysregulation that we have identified in ASD by inference from bulk tissue [33, 52]. Given the clear overlap between ASD and other major psychiatric disorders such as SCZ [54], these data will also be valuable for linking sequence variation to 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 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 allow 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 methods 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.

Aim 3. Systematic genome-wide validation of PsychENCODE regulatory elements

3a. Overview. The results of Aims 1 and 2 will constitute a series of predictions of which regulatory elements are functionally relevant for the expression of neuronal genes involved in normal development and/or neuropsychiatric disorders that are a focus of the PsychENCODE Consortium. Biological validation of these predictions requires testing the regulatory potential of these regions, including allele-specific quantification.

3b. Validating enhancers on a genome-wide scale. ***Preliminary:*** Over the last several years, the White Lab has led efforts to functionally validate regulatory elements (as identified by the ENCODE and modENCODE projects) based on predictions made by the Gerstein and Weng Labs [8, 13, 28, 55] (Zhang et al., submitted). Most recently, our labs have worked together with Drs. Geschwind and Liu on data production and analysis for the PsychENCODE Consortium[1]. Also, we have recently developed a protocol for testing enhancers genome-wide, based on the **STARR-seq** (Self-Transcribing Active Regulatory Region sequencing) methodology originally developed in *Drosophila* by the Stark Lab and used by others on limited regions of mammalian genomes [56, 57]. STARR-seq involves the insertion of putative enhancers into the transcript, instead of upstream of promoters in the reporter vector. The enhancer sequence effectively acts as a barcode in high-throughput sequencing (Fig 2). More specifically, genomic DNA is sheared and end-repaired (Fig 2A), optionally captured if targeted regions are to be screened (as opposed to the entire genomeFig 2B), and subsequently cloned into screening vectors containing a promoter, which then expresses a reporter transcript (Fig 2C). The enhancers are cloned into the 3' end of the transcript, whereby the reporter transcript will contain the enhancer sequence. This pool of screening vectors is transfected into cells (Fig 2D) and then cultured under appropriate conditions for recovery, growth, and/or differentiation (Fig 2E). mRNA is purified and reverse transcribed, followed by uniform amplification of the inserts, and then sequenced using high-throughput sequencing (Fig 2F). Abundant copies of the reporter transcripts that contain specific enhancers can identify enhancers that up-regulate transcription. STARR-seq removes the need for expensive array synthesis of enhancers while capturing natural variation for quantitative analysis of enhancer function.

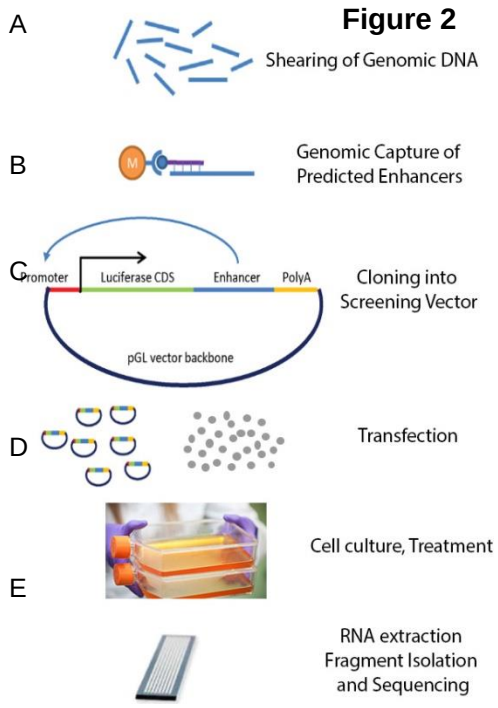


Figure 2

Prior to our optimizations, there have been major limitations to scaling this methodology to the whole human genome. For example, screening the *Drosophila melanogaster* genome required transfecting between 0.5 and 1 billion S2 cells. This makes the direct application of STARR-seq technique to the human genome very difficult and expensive, because the human genome is 20 times larger than the fly genome. Our optimizations of STARR-seq in human cells modifies and builds upon the episomal plasmid library approach, expanding its capabilities. We have overcome the major challenges for scaling STARR-seq up to the entire human genome (namely, the required library complexity, large-scale transfection of cells, and inherent inaccuracy of the assay due to PCR duplicates during the sequencing step introduces significant challenges). First, by optimizing multiple parameters in the candidate element cloning step, we have increased complexity while introducing molecular barcodes that allow for PCR duplicate elimination, resulting in a screening library that covers 2.65 Gb of the human genome. Our typical libraries now have more than 50 fragments covering each base pair, given ~250 million post-filtering fragments. This represents a comprehensive screening library, and allows us to effectively screen genomic fragments with enhancer activity in downstream experiments. Second, using industrial-scale transfection protocols, the White Lab has devised a robust technique to screen either the entire human genome or a fraction of the genome that has been captured using oligonucleotide probes (**CapStarr-seq**). We are producing whole genome STARR-seq datasets at a coverage of more than 10X per expressed base pair, with >300 million paired end 100bp reads, or 50-100M reads for capture STARR-seq. Third, accuracy has been

improved by using single molecular barcodes during the RNA preparation step, along with 160 or more index primers for sequencing, thus allowing for more accurate quantification and elimination of PCR duplicates without removing unique RNA fragments associated with *bone fide* transcriptional activity. Example results are shown in Fig 3. The end result of our optimizations and improvements is that active enhancer regions of the genome are easily identified (Fig 3A, C), the results are highly reproducible (Fig 3B, E), and consistent with traditional reporter assays (Fig 3D), and sequences with enhancer activity overlap with open chromatin marks (Fig 3A, F, G) as well as with RNA expression levels from nearby genes (Fig 3G).

Plan: In years 1 and 2 of the project, we will focus our whole genome and STARR-seq efforts on validating candidate enhancers from the PsychENCODE Project (Aims 1 and 2) using SH-SY5Y neuroblastoma cell lines

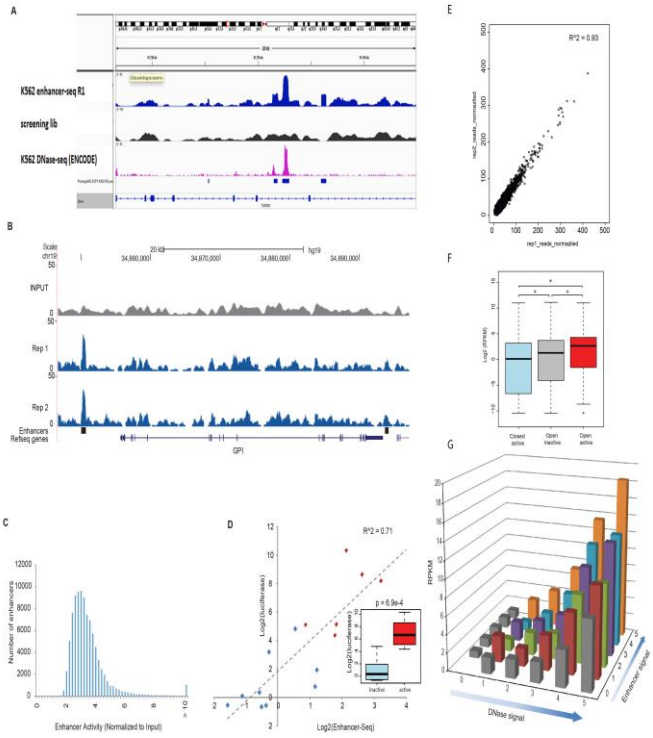


Figure 3. Whole genome STARR-seq (A) Genome browser screen shot shows consistency between the K562 human-STARR-seq signal (1st row) and DNase-seq signal (3rd row). **(B)** Genomic snapshot displaying the *GPI* locus region, as detected by WG-STARR-seq. There is a strong enhancer region approximately 10-kb upstream of *GPI* and another, weaker enhancer regions in the 3'UTR region. Each blue track signifies a normalized enhancer signal of each biological replicate. The gray track represents the normalized input library. **(C)** WG-STARR-seq shows a wide range of enhancer signal strength distributions for all detected enhancers. The median fold change observed was 3.08, with a dynamic range between 1.33 and 119.12. **(D)** The enhancer activity of 6 strong and 9 weak enhancers were validated using traditional luciferase assays in biological triplicates. A strong correlation was observed between luciferase signal and WG-STARR-seq enhancer activity, providing validation of the technique. **(E)** Normalized reads from sequencing were used for reproducibility plots between biological replicates. **(F)** Comparison of expression levels between genes (as measured by RPKM) nearby different groups of enhancers. Statistical significance was calculated using Wilcox Sum Rank test (*p = 2.2e-16). **(G)** Plot comparing expression level of nearby genes in relation to both DNase I signal and enhancer activity. Both DNase I signal and enhancer signals are binned into 6 separate groups according to DNase I signal and enhancer signal rank (0 – 5), respectively.

(since billions of cells can easily be grown and transfected, each replicate requires approximately 1 billion cultured cells). While not ultimately an ideal model for testing the disease and developmental relevance of predicted enhancers in a nervous system-specific manner, SH-SY5Y cells share neuronal molecular characteristics with the primary human neuronal precursor cells (phNPCs), and they match *in vivo* fetal brain development to a substantial extent once differentiated in Retinoic Acid and Brain Derived Neurotrophic Factor (see Fig 1, Aim 2 and [58]). They are a more suitable cell line than a non-neural cell when performing unbiased high-throughput screening of the entire human genome for enhancer activity. The data from these experiments will then be used to improve and refine the predictions made in Aim 2.

In Years 2-4 we will turn our focus to phNPCs, and we will use the CapStarr-seq method with specific sets of predictions from Aim 2. Although not genome-wide, CapStarr-seq allows tens of thousands of DNA elements to be tested per experiment. Multiple rounds will be performed as prediction algorithms are iteratively refined in Aims 1 and 2, taking into account the STARR-seq data. For whole genome STARR-seq and for CapStarr-seq, we will perform 3 biological replicates. We will use the phNPCs because they can be grown in 96 well plates, and we have optimized conditions with over 70% transfection efficiency [59-61]. We will analyze at least 2 time points, T0 (48 after being placed in differentiation media, consisting primarily of neural progenitors) and T6 (6 weeks, to capture development and maturation of different neuronal lineages and asytrocytes)[58, 59]. Cell types to be used in Aims 3 and 4 are shown in the Table below.

Cell Name	Tissue of Origin	Cell Type	Obtained from	Media/Culture Conditions
D8R49	Fetal Cortex	phNPC [58]	Primary	Neurobasal + BIT9500 +GF (Proliferation) Neurobasal +B27 + GF (Differentiation)
2242.1	Skin Fibroblast	iPSC [62, 63]	Pasca Lab	E8 Medium + E8 Supp (Proliferation)
2242.1	iPSC	Forebrain Spheroid [2]	Pasca Lab	E6 Medium + E6 Supp + GF (Differentiation)
SH-SY5Y	Bone Marrow Neuroblastoma	Adherent [58]	ATCC	DMEM + 10% FBS

STARR-seq analysis: We will use established STARR-seq data processing pipelines developed by the Gerstein and White groups, including the definitions of signal profiles defined by PeakSeq [50] and MUSIC [4] to generate a set of regions that show significant enrichment. These regions will be highly sensitive but will contain many false positives. Therefore, we will use the large compendium of existing functional genomics datasets from the PsychENCODE, ENCODE and RMEC projects, utilizing peaks from histone marks and transcription factors to build *a priori* probability estimates of the locations of regulatory regions. We will use the activating marks and transcription factors that associate with enhancers (H3K4me1, H3K27ac, H3K9ac, P300, ATAC-seq, DNase/FAIRE) to build these probabilities. We will also utilize transcription factor binding motif and sequence conservation data as variants in the *a priori* location estimates. We will then combine the whole genome STARR-seq results with these probabilities in a Bayesian framework, and we will train generalized linear models for scoring the candidate relaxed list of regions that we identified from STARR-seq. The sorted list of regions will be used for further validation in the CRISPR mutational assays (detailed below) and the single cell transgenic reporter assays (described in Aim 4).

3c. Validating enhancers using CRISPR genomic editing. ***Preliminary:*** A major caveat of the STARR-seq method is that all assays are done on transfected plasmids that lack the genomic context of the loci from which they have been derived. The advantage of this method is that it allows us to systematically test thousands of genomic regions for regulatory potential. However, for many loci it can be only a modest indicator of the actual regulatory function within the native genomic context. Based on previous experience, we expect that data collected will narrow the regions of interest from tens of thousands of candidate enhancer elements down to hundreds or thousands of partially validated enhancer elements. We will prioritize lists of functionally-validated enhancer elements based on the results from Aim 2. We will then test these regulatory elements using CRISPR-mediated genome editing to determine which regulatory elements show differential function when mutated. Fig 4 shows a recent experiment from the Geschwind Lab [32] using two independent CRISPR constructs to demonstrate the effect on *FOXG1* expression (measured by qRT-PCR) of a specific enhancer associated with SCZ risk that they also showed has allele-specific variants (see Won et al. *Nature* 538:523 2016 for more details[32]). We aim to test 200 such candidates associated with normal developmental or disease states, based on the STARR-seq results and predictions from Aim 2.

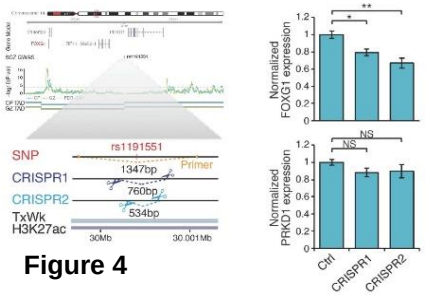


Figure 4

Plan: Using phNPCs, we will functionally validate approximately 200 candidate enhancer regions by endogenously mutating them using the CRISPR/Cas9 targeting system. By utilizing CRISPR technology, we are able to edit the genome using CRISPR and CRISPR-associated (Cas) genes that have been exploited to achieve site-specific DNA recognition and cleavage [64]. In this fashion, not only are we interrogating our target enhancers in their endogenous chromatin context, but we will also be able to obtain a clearer picture of which gene(s) the regulatory element may control. We will generate loss-of-function mutations in putative enhancers using a 96-well plate format and use a qRT-PCR (quantitative reverse transcriptase PCR) assay of nearby gene transcripts to generate quantitative transcriptional read outs. Combining the STARR-seq and CRISPR enhancer mutational data procured in Aim 3, along with the analytical framework to prioritize disease and neuronal subtype/developmental-specific enhancers for each cell type, will enable us to select and target the most disease-relevant enhancers for further testing and evaluation in Aim 4.

It is worth noting that, collectively, these experiments will be performed in parallel with similar experiments being performed on the ENCODE cell lines and on pancreatic tumor cells and organoids that rely on performance of high-throughput gene editing using CRISPR-Cas9 to create a deletion (ENCODE grants UM1 HG009442 and UM1 HG009426), and will thus leverage considerable infrastructure and expertise as the need for any troubleshooting arises. Specific steps that may require optimization include transfection/electroporation efficiency, iteration of library building based on updated predictions from Aim 2, and optimizing timing of harvesting transfected phNPCs after differentiation into neuronal lineages. However, we have used these cells in many studies and regularly obtained over 70% transfection efficiency [58-61].

Aim 4. Deeper biological validation of PsychENCODE regulatory elements

4a. Overview. Geschwind and colleagues have developed novel primary neural progenitor cultures that model brain development *in vitro* (phNPCs) and 3D organoid cultures based on iPS cells differentiated into forebrain cortical spheroids (hFS) that recapitulate all of the major cell classes of the developing brain. Savas Tay has been able to grow such neuronal organoids in microfluidic chips that allow the rapid testing of large numbers of conditions [65]. Choosing from validated enhancers from Aim 3, we will synthesize 100 validated enhancers with

polymorphisms predicted to affect function between alleles, and we will test these in phNPC cells differentiated into various lineages and under different conditions in plate-based assays and microfluidic chips. Additionally for 10-15 enhancers that the criteria of differential alleic expression associated with disease, we will transfect them into phNPCs, followed by neuronal differentiation, and we will perform Drop-Seq based single cell sequencing to associate the functioan of these enhancers with particular cell types and developmental stages.

4b. *In vitro* modeling. *Preliminary*: Human neural stem cells and primary human neuronal progenitors (phNPCs) circumvent a major challenge facing our understanding of human brain function by providing us with access to living tissue representing different human nervous system cell types, developmental stages and diseases. We have developed both 2D and 3D culture systems for modeling human brain development, disease risk variants and synaptic maturation *in vitro* [58, 62, 63, 66], which provide an unprecedented opportunity to experimentally validate the predicted regulatory relationships. We established a high-throughput quantitative framework to compare differentiation in culture to *in vivo* fetal development and demonstrated extensive overlap of our cultures to *in vivo* brain by standard immunocytochemical methods and comprehensive analysis of gene expression [58, 62, 63, 66]**. The clear matching of transcriptomic patterns in our human *in vitro* models to *in vivo* developmental trajectories and cell types provide confidence that these systems provide a valid platform to assess gene regulatory networks** [58, 62, 63, 66]**.** phNPCs generate neurons with stereotypical morphologies similar to what is observed *in vivo*, first forming bipolar migrating cells, followed by axonogenesis and increases in dendritic arborization and exhibit synaptic activity [58], which also permits more advanced cell biological and physiological characterization. We further validated remarkably similar expression of both classical neuroanatomical markers and transcriptomically defined regional markers between differentiated phNPCs and human fetal cortex cells [58]. Recently, we have also developed even more mature and synaptically active 3D cortical forebrain spheroids (hFS; [62]), and are implementing a highly advanced, novel 3D culture system to model ASD risk genes developed by our collaborator, Sergiu Pasca (see letter of collaboration). This includes all major component cell types that can be identified by scRNAseq [2] (see Fig 1 in Aim 1). In addition to single cell transcriptomic analyses, we have used methylation data to assess the epigenetic maturation (epigenetic clock)[67], which demonstrates clear maturation of our 3D cultures. We will use the 2D cultures using phNPCs for first line, high throughput validation, followed by the 3D hFS [2, 62], which also provide more mature cultures, where later acting, postnatal putative cortical enhancers can be tested. Finally, we have developed a microfluidics system (Tay Lab) capable of culturing up to 1,500 NPC experiments in parallel (Figure 5).

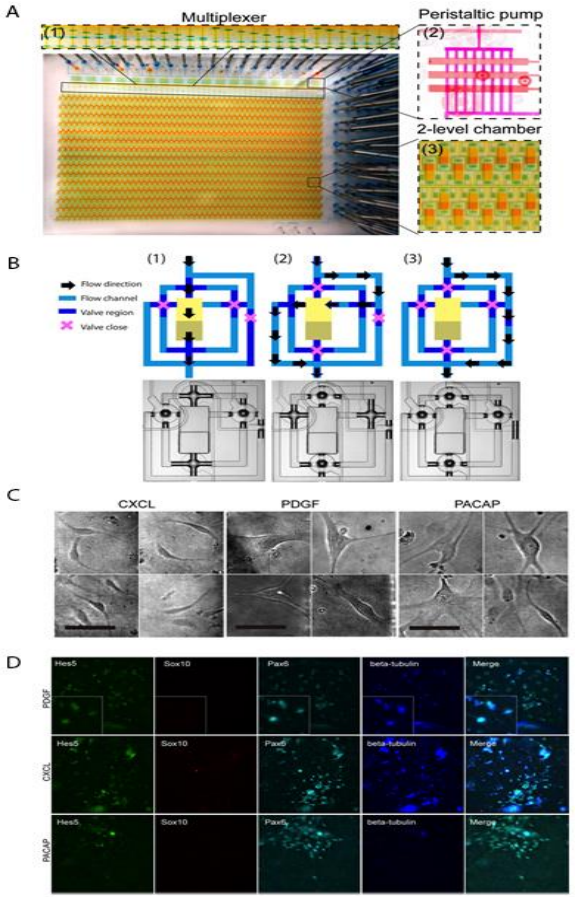


Figure 5. Microfluidic culture system for high-throughput, dynamical analysis of neuronal cell models. **(A)** This microfluidic system performs automated cell culture processes such as cell seeding, stimulation with growth factors, time-lapse imaging and cell tracking, and cell retrieval. An on-chip multiplexer measures several fluids containing signaling molecules or drugs, and mixes them at predetermined ratios, creating complex chemical inputs. A peristaltic pump delivers these inputs to 1,500 independent cell culture chambers for dynamical cell stimulation. Each of the 1,500 chambers can be programmed to receive a different chemical stimulus. The system automatically tracks individual cells, 2-D populations or organoids via time-lapse microscopy. Cells can be immune-stained during or at the end of the experiments, and image processing reveals protein expression and morphology information at the single cell level, allowing quantitative analysis. **(B)** Schematic drawings (top row) and optical images (bottom row) of three distinct flow modes. (1) Fluid is directed to flow over the culture chamber directly (cell loading and retrieval mode); (2) Fluid is guided through the buffering region from the side (stimulation mode); (3) Fluid can be directed to bypass the chamber unit to avoid cross-contamination or perform other fluid manipulation. **(C)** Bright field images of neuronal stem cells (NSCs) cultured on chip in media containing (from left to right) 1000 ng/ml CXCL, 50 ng/mL PDGF and 100 nM PACAP. The scale bars are 50 μm in all images. **(D)** Immunostaining images of NPCs exposed to (top row to bottom) 50 ng/mL PDGF, 1000 ng/ml CXCL and 100 nM PACAP. Markers used for determining NSCs differentiation states are Hes5, Sox10, Pax6 and beta-tubulin. Insertions in the top row are selected NSCs cells with distinct morphology.

Plan. Choosing from validated enhancers from Aim 3, we will synthesize 100 validated enhancers with polymorphisms predicted to affect function between alleles. These will be transfected into engineered phNPC cells (see below) and assayed using an automated microfluidic culture system and associated integrated platform that we have developed for dynamic stimulation, cell manipulation, and time-lapse microscopy (Fig 5). This system allows multi-mode cell culture (single cell, 2-D monolayer and in 3-D organoids) and dynamic stimulation across 1,500 individually addressable cell culture units for high-throughput quantitative studies on mammalian cells (Fig 5A). Each of the 1,500 culture chambers can be programmed to receive a different set of reagents (Fig 5B). Coupled with custom software for chip control and computational data processing, the system can perform programmed delivery of thousands of formulated fluids to any designate on-chip culture unit, while monitoring and analyzing corresponding cellular responses via live cell microscopy. We have thus far used this system to investigate dynamic signaling in the differentiation of Neural Stem Cells (NSCs). Our experiments using primary embryonic (NSCs) and neuronal organoids demonstrated that NSCs proliferation, differentiation and lineage programming can be efficiently assessed at the single cell level via tracking the expression level of self-renewal (Hes5) and differentiation (Dcx) markers in response to dynamic growth factor inputs (Fig 5 C&D). Using this microfluidics system, we will assess the functional differences between alleles for each enhancer tested in our GFP reporter constructs transfected into phNPC cells. This will give us an unprecedented opportunity to test the output of a large number of neuronal enhancers during development and differentiation. This also permits delivery of patterning molecules (e.g. Wnt, Shh, Bmp, Smad inhibitors, and RA) to assess the relationships between these enhancers and the signals involved in regionalization and maturation.

4c. Validation of target gene expression effects. While reporter assays are powerful for testing sufficiency of an enhancer sequence to drive expression via a minimal promoter, we also wish to test the ability of a subset of 10-15 enhancer sequences to affect nearby gene expression. For this purpose we have engineered fluoresecent reporter phNPC-based lines of specific progenitor classes or postmitotic neuron classes using the CRISPR/Cas9 technology that can be used to screen for quantitative changes in cell fate or class, as well as purify specific cell lineages for profiling. **We will use CRISPR/Cas9-mediated transcriptional activation (dCas9-VP64) and enhancer deletion (e.g [32]) to validate its activity, as we have previously shown for long range distal brain enhancers (as identified using Hi-C data) using the same system that we propose to use here [32].** We have developed VP64 transcriptional activation to validate a predicted human enhancer region that we predicted was a distal enhancer for Gli-3, a forebrain patterning gene (Fig 6), supporting the feasibility of this sub-aim.

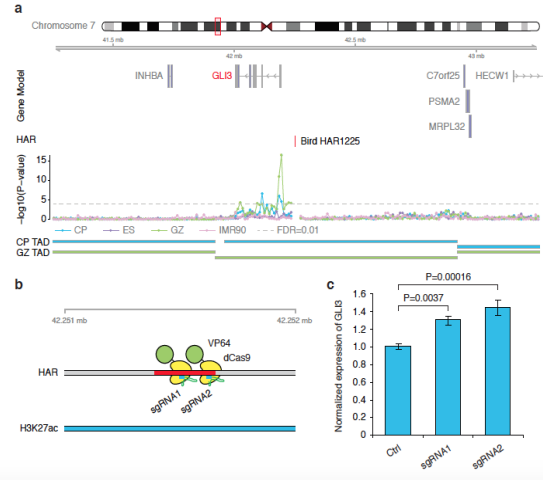


Figure 6: CRISPR/Cas9-mediated transcriptional activation of a human accelerated enhancer to functionally validate a target gene. (A) Hi-C Interaction map of a HAR/enhancer that is predicted to interact with *GLI3*. **(B)** Targeted binding sites for two guide RNAs (gRNAs). This HAR is located in a predicted active enhancer (H3K27ac) in fetal brain. **(C)** Targeting dCas9-VP64 to the HAR results in a 30-40% increase in the expression level of *GLI3* in phNPCs differentiated for 4 weeks. Two sets of gRNAs targeting different regions were cloned into EF1a-dCas9-VP64-2A-GFP-sgRNA. An empty vector without any inserted gRNA was used as control. Viruses were generated by co-transfection of CRISPR vectors with pVSVg and psPAX2 in HEK293 cells. Primary human neural progenitor cells (phNPC) were infected with viruses (empty vectors, gRNA1, gRNA2) on the day of split and differentiated as previously described. After 2.5 weeks of differentiation, cells that are infected (GFP^+) were sorted by FACS. *GLI3* expression levels were measured by qRT-PCR (LightCycler 480 SYBR Green I

4d. Enhancer single cell Drop-seq for mapping enhancer activity to specific neuronal lineages. We have optimized plasmid delivery via electroporation using the Nucleofector II device (Lonza) to achieve ~70% efficiency with minimal toxicity. 10 to 15 isolated engineered reporter phNPCs lines will be mixed with our standard non-recombined cultures and differentiated for 2 to 10 weeks to model major stages of cortical neurogenesis. We will harvest cells for single cell isolation, performing each experiment in quadruplicate, and pool cells to obtain 6000 for Drop-seq (which we show in Aim 2; Fig 1 is sufficient to identify and profile all major cell classes *in vitro* and *in vivo* in developing brain) to assess changes in cell fate or changes in transcription due to enhancer activation in specific cell classes sorted based on reporter activity. We can identify cell classes derived from each distinct progenitor type and infer lineages by analyzing cells generated at each time point. Here, because we have cell class definitions from *in vivo* developing brain (e.g., Aims 1 and 2), we can perform a supervised analysis. We will quantify the diversity of cells generated in each progenitor class and characterize

the lineages to those classes, as well as any enhancer activities that do not affect cell proliferation or fate (cell type composition), but that affect other cellular processes by single cell profiling.
Pitfalls. 2D cultures may not recapitulate all of the cell types and regulatory events, as well as our 3D systems. In addition, for enhancers that are acting post-natally, the 3D hFS system may be preferred due to its ability to achieve a more mature state matching post-natal development. So as an alternative, we will use 3D cultures to test a subset of those enhancers that passed through screens in Aim 2, but do not show activity in Aim 3, even though the more time-consuming culturing of 3D FS is limiting relative to the microfluidic and 2D methods. These 3D hFS can be transfected using the methods described above. We can also engraft progenitor reporter lines into the germinative layers of 3D cultures, and use immunocytochemistry or FISH for markers of the identified descendants of those progenitors and sort them by the reporter expression (following culturing for transcriptomic analysis). Given its high throughput for assessment of enhancer activity and effects on cell type transcriptomes, we will rely primarily on transcriptomics. Neuronal morphology, location, and connectivity are intimately related with function over a wide set of inhibitory and excitatory neurons. In future work, we aim to characterize the morphology, localization, and connectivity that is regulated by enhancers in molecularly defined cell classes.

1/2 Elements unique to this site (Weng/Gerstein): Our site will be the computational and analytical component of the proposal, 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 (**Aim 1**). 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 (**Aim 2**). 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 (**Aim 2**). The Wang Lab will work to identify brain gene expression dynamics, perform gene co-expression network analysis, and model the gene regulatory networks (**Aim 1**). 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 (**Aim 2**). 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.

2/2 Elements Unique to This Site (White/Geschwind): The University of Chicago (White, Tay) and UCLA (Geschwind) together provide the entire experimental component of this proposal, including the biological models and samples, single cell sequencing, genome scale enhancer validation (STARR-seq), CRISPR engineered enhancer characterization (using mutation and targeted VP64), and microfluidic-based single cell and organoid-based quantitative analyses of enhancer reporter assays. Specifically, for **Aim 1 and 2,** the Geschwind Lab (UCLA) will provide Hi-C and ATAC-seq maps from fetal human tissue to aid in assigning distal regulatory elements to genes and derive single cell gene expression profiles from DroNc-Seq on frozen human brain from controls and ASD across 3 broad developmental periods for building control and disease-relevant regulatory networks at the single cell level. For **Aim 3,** the White Lab will work closely with the Geschwind Lab to instantiate their well-validated phNPC protocols into the genome-wide STARR-seq and CapStarr-seq enhancer assays. For **Aim 4,** the Geschwind Lab will work closely with Tay and White Labs to perform functional validation experiments using well-characterized *in vitro* systems, hpNPC and hFS, and Crispr/CAS9-mediated deletion or CAS9-VP64-mediated activation of enhancers. They will leverage engineered reporter lines carrying lineage markers, as well as Drop-seq to characterize the functional outcomes of enhancer activation or repression. Finally, The University of Chicago will also issue a small subcontract to cover the salaries of a postdoc and Dr. Liu (at The University of Illinois, Chicago). Dr. Liu is a participant in the broader PsychENCODE Consortium and is an expert in brain eQTL analysis, and will therefore participate in **Aim 2**.

References

1. Psych, E.C., et al., *The PsychENCODE project.* Nat Neurosci, 2015. **18**(12): p. 1707-12.
2. Birey, F., et al., *Assembly of functionally integrated human forebrain spheroids.* Nature, 2017. **545**(7652): p. 54-59.
3. 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.
4. Gerstein, M.B., et al., *Comparative analysis of the transcriptome across distant species.* Nature, 2014. **512**(7515): p. 445-8.
5. 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.
6. 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.
7. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.
8. Negre, N., et al., *A cis-regulatory map of the Drosophila genome.* Nature, 2011. **471**(7339): p. 527-31.
9. 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.
10. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science, 2010. **330**(6012): p. 1775-87.
11. 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.
12. 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.
13. Boyle, A.P., et al., *Comparative analysis of regulatory information and circuits across distant species.* Nature, 2014. **512**(7515): p. 453-6.
14. 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.
15. 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.
16. 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.
17. 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.
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. Wang, D., et al., *Loregic: a method to characterize the cooperative logic of regulatory factors.* PLoS Comput Biol, 2015. **11**(4): p. e1004132.
48. Khurana, E., et al., *Integrative annotation of variants from 1092 humans: application to cancer genomics.* Science, 2013. **342**(6154): p. 1235587.
49. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
50. Rozowsky, J., et al., *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.* Nat Biotechnol, 2009. **27**(1): p. 66-75.
51. 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.
52. Sun, W., et al., *Histone Acetylome-wide Association Study of Autism Spectrum Disorder.* Cell, 2016. **167**(5): p. 1385-1397 e11.
53. Pei, B., et al., *The GENCODE pseudogene resource.* Genome Biol, 2012. **13**(9): p. R51.
54. Gandal, M.J., et al., *Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap.* bioRxiv, 2016.

55. Kittler, R., et al., *A comprehensive nuclear receptor network for breast cancer cells.* Cell Rep, 2013. **3**(2): p. 538-51.
56. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.
57. Vanhille, L., et al., *High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq.* Nat Commun, 2015. **6**: p. 6905.
58. Stein, J.L., et al., *A quantitative framework to evaluate modeling of cortical development by neural stem cells.* Neuron, 2014. **83**(1): p. 69-86.
59. Konopka, G., et al., *Modeling the functional genomics of autism using human neurons.* Mol Psychiatry, 2012. **17**(2): p. 202-14.
60. Rosen, E.Y., et al., *Functional genomic analyses identify pathways dysregulated by progranulin deficiency, implicating Wnt signaling.* Neuron, 2011. **71**(6): p. 1030-42.
61. Wexler, E.M., et al., *Genome-wide analysis of a Wnt1-regulated transcriptional network implicates neurodegenerative pathways.* Sci Signal, 2011. **4**(193): p. ra65.
62. Pasca, A.M., et al., *Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture.* Nat Methods, 2015. **12**(7): p. 671-8.
63. Pasca, S.P., et al., *Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome.* Nat Med, 2011. **17**(12): p. 1657-62.
64. Mali, P., et al., *RNA-guided human genome engineering via Cas9.* Science, 2013. **339**(6121): p. 823-6.
65. Zhang, C., et al., *Universal Microfluidic System for Analysis and Control of Cell Dynamics.* bioRxiv, 2017. **http://www.biorxiv.org/content/early/2017/06/28/157057**.
66. Martinez, R.A., et al., *Genome engineering of isogenic human ES cells to model autism disorders.* Nucleic Acids Res, 2015. **43**(10): p. e65.
67. Horvath, S., *DNA methylation age of human tissues and cell types.* Genome Biol, 2013. **14**(10): p. R115.

References

1. Psych, E.C., et al., *The PsychENCODE project.* Nat Neurosci, 2015. **18**(12): p. 1707-12.
2. Birey, F., et al., *Assembly of functionally integrated human forebrain spheroids.* Nature, 2017. **545**(7652): p. 54-59.
3. 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.
4. Gerstein, M.B., et al., *Comparative analysis of the transcriptome across distant species.* Nature, 2014. **512**(7515): p. 445-8.
5. 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.
6. 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.
7. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.
8. Negre, N., et al., *A cis-regulatory map of the Drosophila genome.* Nature, 2011. **471**(7339): p. 527-31.
9. 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.
10. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science, 2010. **330**(6012): p. 1775-87.
11. 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.
12. 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.
13. Boyle, A.P., et al., *Comparative analysis of regulatory information and circuits across distant species.* Nature, 2014. **512**(7515): p. 453-6.
14. 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.
15. 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.
16. 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.
17. 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.
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. Wang, D., et al., *Loregic: a method to characterize the cooperative logic of regulatory factors.* PLoS Comput Biol, 2015. **11**(4): p. e1004132.
48. Khurana, E., et al., *Integrative annotation of variants from 1092 humans: application to cancer genomics.* Science, 2013. **342**(6154): p. 1235587.
49. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
50. Rozowsky, J., et al., *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.* Nat Biotechnol, 2009. **27**(1): p. 66-75.
51. 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.
52. Sun, W., et al., *Histone Acetylome-wide Association Study of Autism Spectrum Disorder.* Cell, 2016. **167**(5): p. 1385-1397 e11.
53. Pei, B., et al., *The GENCODE pseudogene resource.* Genome Biol, 2012. **13**(9): p. R51.
54. Gandal, M.J., et al., *Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap.* bioRxiv, 2016.

55. Kittler, R., et al., *A comprehensive nuclear receptor network for breast cancer cells.* Cell Rep, 2013. **3**(2): p. 538-51.
56. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.
57. Vanhille, L., et al., *High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq.* Nat Commun, 2015. **6**: p. 6905.
58. Stein, J.L., et al., *A quantitative framework to evaluate modeling of cortical development by neural stem cells.* Neuron, 2014. **83**(1): p. 69-86.
59. Konopka, G., et al., *Modeling the functional genomics of autism using human neurons.* Mol Psychiatry, 2012. **17**(2): p. 202-14.
60. Rosen, E.Y., et al., *Functional genomic analyses identify pathways dysregulated by progranulin deficiency, implicating Wnt signaling.* Neuron, 2011. **71**(6): p. 1030-42.
61. Wexler, E.M., et al., *Genome-wide analysis of a Wnt1-regulated transcriptional network implicates neurodegenerative pathways.* Sci Signal, 2011. **4**(193): p. ra65.
62. Pasca, A.M., et al., *Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture.* Nat Methods, 2015. **12**(7): p. 671-8.
63. Pasca, S.P., et al., *Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome.* Nat Med, 2011. **17**(12): p. 1657-62.
64. Mali, P., et al., *RNA-guided human genome engineering via Cas9.* Science, 2013. **339**(6121): p. 823-6.
65. Zhang, C., et al., *Universal Microfluidic System for Analysis and Control of Cell Dynamics.* bioRxiv, 2017. **http://www.biorxiv.org/content/early/2017/06/28/157057**.
66. Martinez, R.A., et al., *Genome engineering of isogenic human ES cells to model autism disorders.* Nucleic Acids Res, 2015. **43**(10): p. e65.
67. Horvath, S., *DNA methylation age of human tissues and cell types.* Genome Biol, 2013. **14**(10): p. R115.

RESOURCE SHARING PLAN

This project will generate single-cell RNA-seq, single-nuclear RNA-seq, Dro-nuc-Seq, and STARR-seq data. As soon as they are analysis ready, they will be shared with all members of the PsychENCODE consortium and other qualified investigators according to the NIH Genomic Data Sharing Policy through the Sage Bionetworks Synapse system. Sage has worked with independent ethical advisors, legal counsels and an IRB to develop appropriate governance policies and procedures to support Synapse operations. These efforts enable the contribution and use of human data for research purposes, while protecting personal information and respecting individuals' expectations for privacy. All Synapse governance policies are described here: http://docs.synapse.org/articles/governance.html. The study will also be registered in dbGaP. And submited to other relevant NIH-designated data repository (e.g., dbGaP, GEO, SRA, the Cancer Genomics Hub54) if necessary. All other data will be submitted to the ENCODE DCC as instructed by NHGRI and DCC staff. A copy of all data will be kept on the Bionimbus Protected Data cloud maintained by the Institute for Genomics and Systems Biology.

Public sharing of tools and pipelines will enhance the reproducibility and broaden the impact of our research. We will release our analysis results in uniform formats of metadata, including versions of the programs or pipelines used to generate the results and all tools and data processing methods, using public version-control repositories such as GitHub. We will use the Sage Bionetworks Synapse platform for management of data, metadata and analytical results as generated through the grant. The Synapse web portal is an environment for sharing data, results, methods and tools that enables the tracking of analysis steps and publication of analysis results to collaborators and eventually the broader community. The Synapse portal is used as the collaboration space and public data release portal of all current psychENCODE related grants, for which Sage Bionetworks functions as the Data Coordinating Center. We will establish clear guidelines and standards in preparing the datasets for integration and preparing the analysis results for public release. All data shared with the larger community will be shared through the Synapse servers and browsers, and general genome browsers (such as NCBI, ENSEMBL, and NCBI) whenever possible.

Acknowledging that the psychENCODE Project is a community resource project, we will strictly follow the psychENCODE policy on data release. We will also actively participate in the activities directed by the NIMH on updating the data release policy, and we will accept the updated policy whenever available. Furthermore, we will contribute to the development of a consortium software and data analysis sharing plan as part of the psychENCODE Consortium, and comply with such a policy as it applies to our proposed work. We will collaborate closely with other projects in the psychENCODE consortium to ensure that all our analysis results, as soon as they are stable and are of value to the broader community, are rapidly released to the community to further the advancement of research.

PROJECT MANAGEMENT PLAN

Personnel

Dr. Kevin White at the University of Chicago will serve as the overall PI of the program, and he will be responsible for the oversight and coordination of the Center. His research team will be primarily responsible for developing and refining the STARR-seq and CRISPR Cas9 approaches in the Center. His team will make all the constructs and libraries for STARR-seq and CRISPR Cas9 mediated genomic editing assays. His team will also work closely with the Gerstein and Weng groups in the identification of candidate enhancers, and will act as the experimental hub for the Center. His experimental team will coordinate and collaboratively perform the enhancer validation experiments described in Aims 3 and 4 that utilize phNPC cells developed with protocols from the Geschwind group and microfluidics devices developed by the Tay group.

Dr. Zhiping Weng is a Professor and Director of Program in Bioinformatics and Integrative Biology at University of Massachusetts Medical School. She has worked for the last two decades on developing computational and statistical methods and applying them to biological problems ranging from genomics to protein-protein interactions. She has led projects in the ENCODE project since its inception in 2003. She led the Data Analysis Center (DAC) for ENCODE Phase III (2011-2017) and is co-leading the DAC with Prof. Gerstein for Phase IV (2017–). She collaborated with Profs. Kevin White and Mark Gerstein in leading the PsychENCODE Data Analysis Center, implementing data analysis pipelines and performing integrative analyses. She will focus on Aim 2 of this project and direct the Data Coordination Center (DCC). The Weng lab will also perform enhancer analysis, annotate GWAS SNPs associated with psychiatric diseases, prioritize the discovered regulatory elements for validation, and develop the psychSCREEN platform to allow the broad community to visualize all data generated in this project and all psychENCODE data in an integrated fashion.

Dr. Dan Geschwind will be Co-Investigator at UCLA. Dr. Geschwind will work closely with U Chicago investigators in implementation and performance of the experimental validation Aims 3 and 4, and with Gerstein and Weng to integrate brain Hi-C and scSEQ data into their regulatory networks. Geschwind is currently a PI on another PsychENCODE project that is developing genome wide Hi-C contact maps in human neurons and glia that match the developmental stages that will be profiled by scSeq in this proposal. He has established a close working relationship withini psychENCODE with the DCC (Gerstein/Weng) and White labs to perform cross disorder transcriptome integration. His lab also has expertise in in vitro modeling using human neural stem cells, and has developed and validated the primary human neural progenitor lines and other lines and protocols to be used for biological validation of predicted enhancers. Recently, his lab has worked with Sergiu Pasca's lab to use transcriptional networks and methylation (epigenetic clock) to show that the novel 3D culture system for hFS matches in vivo development and maturation into the postnatal period for the first time. This will be invaluable for validating the later acting enhancers (post natal), which may not be active in prenatal brain, the period that other IPSC culture systems represent.

Dr. Mark Gerstein will be Co-Investigator and serve as PI at the Yale. He will be responsible for performing data analysis tasks using the data generated for mapping regulatory regions on the genome as well as publically available datasets. His group will develop statistical models for identifying the regulatory regions in the genome, combine existing datasets to the identified candidate elements and characterize these elements in relation to existing annotations. The knockout and knockdown datasets will be used to prioritize the variants in the elements with respect to how much effect they have on the function of elements. He will tune the prioritization models and perform large-scale analysis of how variants affect the noncoding regulatory elements.

Dr. Mette Peters will be Co-Investigator at Sage Bionetworks. She will be working with Dr. Weng to oversee the Data Coordination Center. Dr Peters is PI on Sage Bionetworks CommonMind Consortium, psychENCODE and Brain Somatic Mosaicism Network collaborations where Sage plays an integral role in project development, oversight of data generation efforts, data management, curation and dissemination of data among collaborative teams and the public. Dr Peters will be the primary liaison with the psychENCODE investigators and the DCC in regards to data management and sharing, to ensure that data, metadata, analytical output and code is tracked and rapidly disseminated in a transparent manner.

Dr. Daifeng Wang will be Co-Investigator at Stony Brook University. He will be responsible for regulatory data integration, modeling and analyzing the gene regulatory networks in Aim 1, working with Dr. Mark Gerstein and Dr. Zhiping Weng. He will identify the specific regulatory network structures associated with psychiatric phenotypes, especially for the disorders and single cell deconvolution. He will also develop computational approaches to predict the regulatory mechanisms how genomic variants drive specific psychiatric phenotypes; e.g., the eQTLs affect the disorder biomarker gene expression via breaking the TF binding sites on the enhancers.

Dr. Savas Tay will be a Co-Investigator at the University of Chicago. He will provide the microfluidics platform for single cell and single organoid live analysis of the reporter constructs tested during the project. Dr. Tay will has a track record of developing novel microfluidics devices for cellular analyses, most recently demonstrating a device capable of culturing neuronal stem cells and monitoring their differentiation. Tay has also successfully worked with Dr. White in the past, developing a tumor stem cell/organoid microfluidics platform for pancreatic and breast cancers.

Dr. Chunyu Liu will be a Co-Investigator in Chicago at University of Illinois, Chicago. He will be responsible for supervision of a postdoc dedicated to eQTL and regulatory analysis associated with Aim 2. Dr. Liu is the PI of two PsychENCODE grants and has coordinated closely with Dr. White on analysis of RNAseq, ATAC-seq and proteomic data to identify eQTL, chromatin-QTL and pQTL associated with human brains and neuropsychiatric disorders.

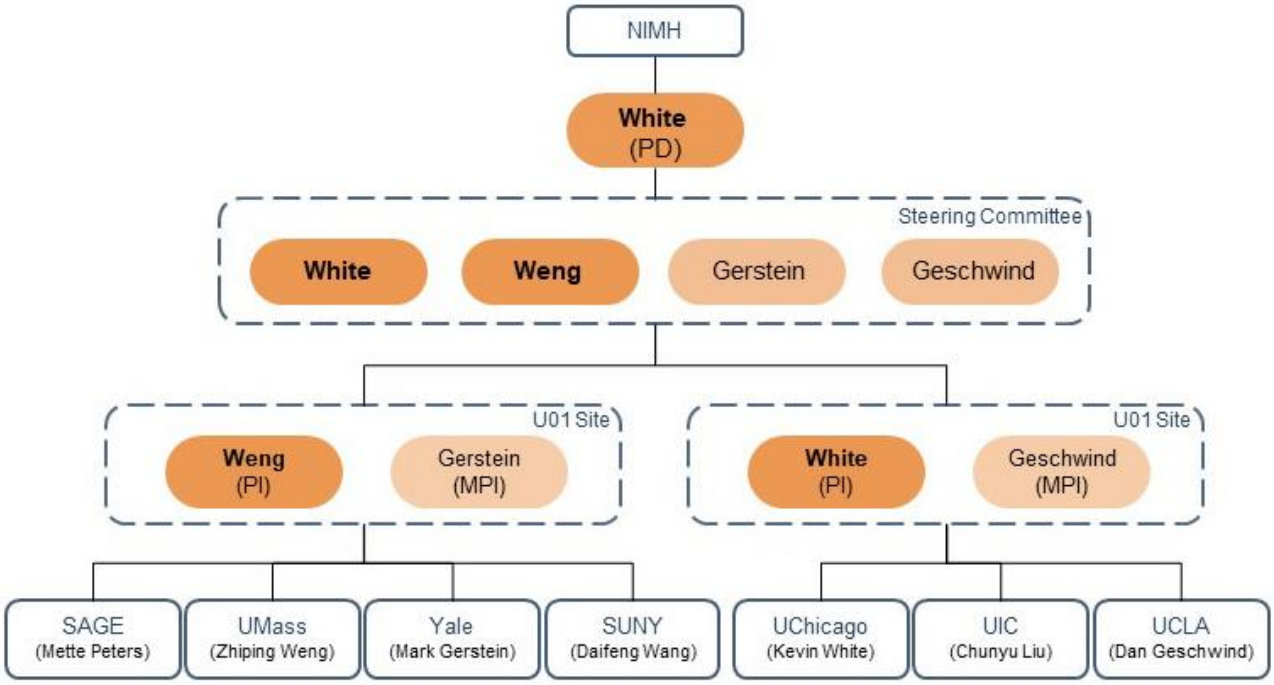
This group of investigators has worked together in the context of past PsychENCODE efforts as well as other genomics and genetics efforts. In addition to all of the investigators working closely together during the PsychENCODE project to date, Drs. White, Weng and Gerstein have collaborated extensively in PsychENCODE, ENCODE and modENCODE, as well as in numerous other initiatives and projects where they have published extensively together over the last 15 years. With these highly productive past interactions, most of the members of the team have a high level of communication, experience and familiarity working with one another.

Management structure

Administrative management:

As Program Director, Dr. White will serve as contact PI and be responsible for submission of progress reports to NIH and all substantive communication regarding the functioning of the Program. His decisions and communications will reflect consensus from a steering committee team composed of himself and the multi-PIs from the collaborative U01 grants, namely Drs. Weng, Gerstein and Geschwind. Dr. White, along with the multi-PIs on the steering team will be responsible for all aspects of the work in the projects including to ensure successful completion at the collaborating institutions.

Dr. White will serve as PI for the Chicago labs, Dr. Weng for UMass, Dr. Gerstein will serve as PI for the Yale/Stony Brook research team, Dr. Peters will be the PI at SAGE and Dr. Geschwind will serve as PI for studies at UCLA. Each of these PIs will act as the point people at their respective institutions, where they will be responsible for their own local fiscal and research administration while keeping the steering team abreast of any developments. The administrative reporting structure of the Center will therefore be as shown in Figure 1. Drs. White, Weng and Gerstein have a long track record of working together in various “PI configurations” on ENCODE and other consortia project grants and contracts, and more recently Dr. Geschwind has worked together with all three of these PIs in the PsychENCODE consortium. Over the last three years all of the investigators in the collaborative U01s have successfully worked together as part of the PsychENCODE consortium.



Project workflow management:
Figure 2 summarizes the workflow for the Center. The workflow can be organized conceptually and task-wise into four main components. The first component, shown in blue, is to perform computational predictions of candidate enhancers based on PsychENCODE and other relevant data sets as outlined in Aims 1 and 2. Additionally UCLA (Geschwind) will perform single cell sequencing in control brain representing 3 major human developmental periods and in a psychiatric disorder, ASD, as a comparison proof of principle, that will be used to infer cell type enhancers and regulatory networks. Drs. Weng and Gerstein will be responsible for this component, and they will coordinate with Dr. White who is primarily responsible for the second component, shown in purple. Taking the predictions from the computational team, the physical reagents such as STARR-seq libraries and gRNA constructs will be coordinately designed by White, Weng and Gerstein, and then White's group will be responsible for constructing the physical reagents necessary for the experiments. The third component is the maintenance of the biological model systems, whereby each the three experimental groups will coordinate to analyze enhancers in the biological models. White is responsible for the whole genome STARR-seq experiments in SH-SY5Y neuroblastoma cells, while Geschwind and White will coordinate for large scale cap-STARR-seq and CRISPR experiments in the phNPCs and 2242.1 iPSC cells and derived 3D forebrain spheroids (hFS). Tay will be responsible for building the microfluidics devices and performing single cell and organoid experiments with reporter assays, in close coordination with White (whose lab is co-located in the Institute for Genomics and Systems Biology at the University of Chicago) and with Geschwind's team at UCLA. Accordingly, the White lab is responsible for distributing the appropriate reagents to the investigators experimenting on each biological model. The fourth component is the assays that will be done on each biological model. The main logistical issue we anticipate at this point is exchanging protocols between UChicago and UCLA for culturing and differentiating stem cells into neuronal lineages and neuronal organoids/spheroids. This will be addressed by exchange of personnel between labs at the two sites. We expect to exchange personnel, having those from Chicago come to UCLA for several weeks to learn the culturing methods, and period UCLA visits to Chicago to update protocols and assure similar procedures across sites..etc. Personnel from UCLA will spend 3-4 months in Chicago teaching White and Tay lab personnel their culturing methods and learning how to perform experiments using microfluidics devices.

Additionally, the PIs and lab and computational scientists will communicate frequently to coordinate experimental efforts at the three institutions, organize data transfers and plan synchronize the ongoing workflow. There will be monthly "all hands on deck" teleconference calls using the Blujeans format that allows multiple centers to connect, to show primary data results, address experimental problems, and deal with administrative issues. There will also be additional weekly calls for small working groups that will handle issues at a more granular level and coordinate close interactions, for instance with STARR-seq library transfer, interpretation of data, etc. We will use the Sage Bionetworks Synapse platform for management of data, metadata and analytical results as

generated through the grant. The Synapse web portal is an environment for sharing data, results, methods and tools that enables the tracking of analysis steps and publication of analysis results to collaborators and eventually the broader community. The Synapse portal is used as the collaboration space and public data release portal of all current psychENCODE related grants, for which Sage Bionetworks functions as the Data Coordinating Center. On a day-to-day basis we will also use a collaborative workspace called Slack (https://slack.com) that permits real-time communication and decision making via multi-channel communication. The PIs will have an executive and private communication area for leadership and administration, and each aim (and where relevant, sub aim) will have specific dedicated Slack conversations, facilitating daily communication among collaborators PIs, students and trainees on specific scientific and analytic issues.

Finally, a Publication and Presentation (P&P) policy will be established, a P&P Committee appointed, and basic principles related to such issues established early in the collaboration, drawing upon previous experiences in the PsychENCODE consortium, ENCODE consortium and the large-scale GWAS studies.

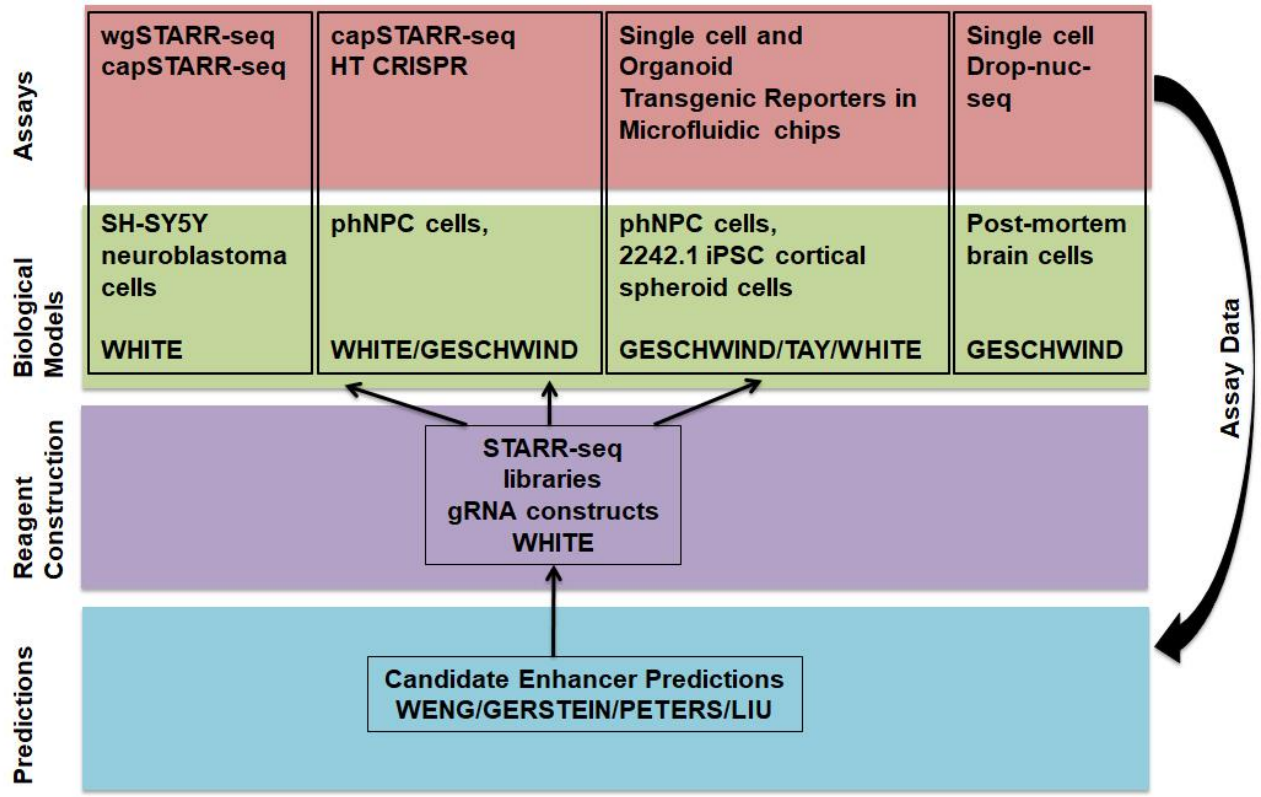


Figure 2. Organization of work for prediction and validation of PsychENCODE Enhancer regions. The collaborative U01s have four key areas of activity; predictions (blue), reagent construction (purple), biological models (green) and assays (red). Teams led by the Weng, Gerstein, Peters and Liu lab will provide data analysis, management for the Psych ENCODE consortium and enhancer predictions for reagent construction activities by the White lab. The predictive analytics will receive feed-back on findings from the biological model validation experiments, allowing them to refine predictive algorithms and approaches. The White lab will create the library reagents and gRNA reagents to support all cell based applications. The White, Geschwind and Tay labs will coordinate to identify the best disease-related candidate enhancer to test in cells and neuronal 3D organoid models.

Intellectual Property

The Technology Transfer Offices at the represented Institutions will be responsible for preparing and negotiating an agreement for the conduct of the research, including any intellectual property. An Intellectual Property Committee composed of representatives from each institution that is part of the grant award, will be formed to work together to ensure the intellectually property developed by the PIs is protected according to the policies established in the agreement.

Conflict Resolution

Given the deep relationships between the investigators at the various institutions, we do not anticipate any major

conflicts. If a potential conflict develops, the PIs shall meet and attempt to resolve the dispute. If they fail to resolve the dispute, the disagreement shall be referred to an arbitration committee consisting of one impartial senior executive from each PI's institution and a third impartial senior executive mutually agreed upon by both PIs. No members of the arbitration committee will be directly involved in the research grant or disagreement. Each of the investigators has agreed to abide by the results of such resolution.

Change in PI Location

If a PI moves to a new institution, attempts will be made to transfer the relevant portion of the grant to the new institution. In the event that a PI cannot carry out his/her duties, a new PI will be recruited as a replacement at one of the participating institutions.

Data submission to the NIMH

Analysis ready data will be shared with all members of the psychENCODE consortium and other qualified investigators according to the NIH Genomic Data Sharing Policy through the Sage Bionetworks Synapse system. Sage has worked with independent ethical advisors, legal counsels and an IRB to develop appropriate governance policies and procedures to support Synapse operations. These efforts enable the contribution and use of human data for research purposes, while protecting personal information and respecting individuals' expectations for privacy. All Synapse governance policies are described here: http://docs.synapse.org/articles/governance.html. The study will also be registered in dbGaP. And submited to other relevant NIH-designated data repository (e.g., dbGaP, GEO, SRA, the Cancer Genomics Hub54) if necessary. All other data will be submitted to the ENCODE DCC as instructed by NHGRI and DCC staff. A copy of all data will be kept on the Bionimbus Protected Data cloud maintained by the Institute for Genomics and Systems Biology.

Outreach:

Our center expects to be in the position to study samples from other funded groups as well as from biology and disease experts within the community to enable the study of specific cell lineages, biologically relevant conditions and diseases that are of high value for discovery of new candidate functional elements but that might not be readily available. In order to accomplish this we will engage the research community to obtain such samples for the mapping center pipelines. Additionally, we will present talks and posters at the annual meeting, hold center wide (and consortia wide calls, when needed) to disseminate our techniques and findings to the community. We plan to participate fully in PsychENCODE led consortium activities. We plan to publish our findings in peer-reviewed journals.

In order to create a transparent and reproducible data resource we will use the Synapse Provenance system to track the relationships among raw and processed data and analysis results, including all pipelines and code, which will be shared through a public GitHub repository. All data and associated resources will be made available for the community and beyond

Progress reporting to the NIMH

The PIs will provide annual and overall project period milestones for activity and throughput to the NIMH as requested. It is expected that this will be at the outset of the award and annually thereafter. Additional information will be provided at any time when requested by NIMH program staff.

MILESTONES AND TIMELINES PLAN

The deliverables for all Aims will span 5 years (see GANTT chart attached). For **Aim 1**, the first 6 months will be devoted to developing and refining new methods for finding enhancers (STARR-seq assays). The following 6 months will be devoted to developing new methods for brain-specific enhancer identification. The first half of Year II will entail building brain-specific regulatory networks, followed by the development of methods to perform the needed single cell analysis. For Year III, we will use the single-cell data to cellular regulatory networks, and the second half of Year III will entail running a harmonized pipeline to integrate eQTL, chromatin QTL (cQTL) and allelic sites analysis. In the first half of Year IV (now months 36-42) will use allelic sites provide greater power for QTL identification, followed by integrated modeling based on major genomic regulatory elements including enhancers the genomic variation and single cell data (months 42-48). Finally, in Year V, we will extend and refine Loregic to predict the gene regulatory circuits driving psychiatric disorders (over the course of months 48-54), and during the final 6 months of Year V, we will further refine the integrated modeling on the basis of experimental results.

For **Aim 2**, the first 6 months of the project will include the first round of enhancer predictions, as well as the implementation of uniform processing pipelines applied to psychENCODE data. During the following 6-mongh period, we will complete 40% of our Dro-nuc-Seq analysis, apply uniform processing pipelines to data from other consortia, and incorporate new psychENCODE data into Synapse. The first half of Year II will include a) integrative analysis on all uniformly processed data, and b) public release of psychENCODE data through Synapse. The second half of Year II (months 18-24) will involve the completion of 70% of our Dro-nuc-Seq analysis, in addition to building psychSCREEN. During the same 6-month period, we'll also incorporate new psychENCODE data into Synapse and apply uniform processing pipelines to new data. The first half of Year III will include the second round of enhancer predictions. We'll also connect psychSCREEN to Synapse, and we'll publicly release psychENCODE data via Synapse during this time. The second half of Year III (months 30-36) will see the total completion of Dro-nuc-Seq, and we'll also incorporate psychENCODE data into Synapse and continue to apply uniform processing pipelines to new data as they are released. In the first half of Year IV (months 36-42), there will be the public release of psychENCODE data via Synapse and psychSCREEN. In the second half of Year IV (months 42-48), we'll perform the 3rd round of enhancer predictions and we'll also incorporate additional psychENCODE data into Synapse. For the first half of Year V, we'll publicly release psychENCODE data via Synapse and psychSCREEN. Finally, during the final six months of the project, we'll a) incorporate psychENCODE data into Synapse, b) apply uniform processing pipelines to the new data, and c) carry out the final round of integrative analysis and enhancer prediction.

For **Aim 3**, we'll spend the first 6 months training personnel in multiple sites, and throughout all of the first year, we'll perform genome-wide STARR-seq analysis on SH-SY5Y neuroblastoma cells. For Year II, we will first perform capSTARR seq on SH-SY5Y neuroblastoma cells (months 12-18), and we will do capSTARR seq on phNPCs round 1 predictions. In the first half of Year III, CRISPR-cas9 will be used to validate phNPCs on round 1 predictions, whereas the second half of Year III (months 30-36) will entail capSTARR seq on phNPCs round 2 predictions. In Year IV, CRISPR-cas9 will be used to validate phNPCs on round 2 predictions. In the first half of Year V (months 48-54), we will do capSTARR seq on phNPCs round 3 predictions. The final 6 moths of Year V will entail CRISPR validation in phNPCs on round 3 predictions.

For **Aim 4**, we'll spend the first 6 months training personnel in multiple sites. The following 6 months will involve calibration of *in vitro* human progenitor and stem cell systems in microfluidics devices. In the first half of Year II, we will select and synthesize the first round of 25 bi-allelic enhancers. In the following 6 months

(i.e., months 18-30), functional validation of round 1 enhancers using phNPCs and hFS in microfluidics systems will be performed. In months 30-42, we will perform functional validation of round 2 enhancers using phNPCs and hFS in microfluidics systems. In months 36-52, we will examine 10-15 enhancers in Drop-seq experiments. In months 42-48, we will choose and synthesize third round of 50 bi-allelic enhancers. In months 48-60, we will perform functional validation of round 3 enhancers using phNPCs and hFS in microfluidics systems. In months 54-60, we will examine 10-15 enhancers using VP64 accelerated endogenous gene expression assay.


Aim 1
Aim 2
Aim 3
Aim 4
Develop a new approach for enhancer finding using STARR−seq assays
Develop a new approach for brain specific enhancer identification
Build brain−specific networks
Develop methods for single cell analysis
Build cellular regulatory networks based on single cell data
Implement a harmonized pipeline to integrate eQTL, chromatin QTL (cQTL) and allelic sites analysis
Use allelic sites to bolster the power for QTL finding
Integrated modeling based on major genomic regulatory elements including enhancers the genomic variation and single cell data
Extend our method, Loregic to systematically predict the gene regulatory logical circuits driving psychiatric disorders
Refining the integrated modeling according to experimental validation
Round 1 enhancer predictions; implement uniform processing pipelines and apply them to all psychENCODE data.
Complete 40% of Dro−nuc−Seq, apply uniform processing pipelines to data from other consortia; incorporate new psychENCODE data into Synapse.
Perform integrative analysis on all uniformly processed data; public release of psychENCODE data via Synapse.
Complete 70% of Dro−nuc−Seq; build psychSCREEN; incorporate new psychENCODE data into Synapse; apply uniform processing pipelines to new data.
Round 2 enhancer predictions; connect psychSCREEN to Synapse; public release of psychENCODE data via Synapse.
Complete 100% of Dro−nuc−Seq; incorporate psychENCODE data into Synapse; apply uniform processing pipelines to new data.
Public release of psychENCODE data via Synapse and psychSCREEN
Round 3 enhancer predictions; incorporate psychENCODE data into Synapse.
Public release of psychENCODE data via Synapse and psychSCREEN.
Incorporate psychENCODE data into Synapse; apply uniform processing pipelines to new data. Perform the final round of integrative analysis and enhancer prediction.
Cross−site training of personnel
Whole genome STARR seq on SH−SY5Y neuroblastoma cells
capSTARR seq on SH−SY5Y neuroblastoma cells
capSTARR seq on phNPCs round 1 predictions
CRISPR validation in phNPCs on round 1 predictions
capSTARR seq on phNPCs round 2 predictions
CRISPR validation in phNPCs on round 2 predictions
capSTARR seq on phNPCs round 3 predictions
CRISPR validation in phNPCs on round 3 predictions
Cross−site training of personnel
Calibration of in vitro human progenitor and stem cell systems in microfluidics device
Choose and synthesis first round of 25 bi−allelic enhancers
Functional validation of round 1 enhancers using phNPCs and hFS in microfluidics system
Choose and synthesis second round of 25 bi−allelic enhancers
Functional validation of round 2 enhancers using phNPCs and hFS in microfluidics system
Examine 10−15 enhancers in Drop−seq experiments
Choose and synthesize third round of 50 bi−allelic enhancers
Functional validation of round 3 enhancers using phNPCs and hFS in microfluidics system
Examine 10−15 enhancers using VP64 accelerated endogenous gene expression assay
0
12
24
36
48
60
Time (months)

DATA AND SAMPLE PLAN

This project will leverage the following data as part of its integrative analysis:

1. All data ***currently produced by the PsychENCODE Consortium***. These are provided in the form of a number of deep sequencing data, such as RNA-seq, ChIP-seq (of histone modifications), ATAC-seq for chromatin accessibility, etc. specifically, the raw data come in FASTQ-formatted files, as well as processed BAM files, BigWig files, etc. These are stored at Synapse and the current PsychENCODE Data Coordination Center, which is led by the co-I Mette Peters. The Weng, Gerstein, Wang labs have already downloaded these data and have used them for computations in a secure (human subjects data compliant) environment.
2. All ***future datasets generated by the PsychENCODE Consortium*** will be submitted to Synapse. They will be downloaded by the Weng, Gerstein, and Wang labs via a secure protocol for computation.
3. ***Data from related consortia (e.g., GTEx, ENCODE, BrainSpan, CommonMind, Epigenome Roadmap, 1000 Genomes Project etc)***. The Weng, Gerstein, and Wang labs have individually been granted access to dbGaP access to these datasets. Currently, each of these labs already has access to a copies of the datasets. For example, all existing ENCODE data amount to 300 TB. The data that contain personal information (e.g., GTEx FASTQ files) are stored in a secure file server, in compliance with dbGaP regulations.
4. The Geschwind Lab will generate single-cell RNA-seq, single-nuclear RNA-seq, and Dro-nuc-Seq data for is project. They will be shared with the Weng, Gerstein, and Wang labs for data analysis, and they will also be submitted to Synapse.
5. The White Lab will generate STARR-seq data for this project. They will be shared with the Weng, Gerstein, and Wang labs for data analysis, and they will also be submitted to Synapse.

This project will leverage the following samples as part of its integrative analysis:

1. The Geschwind lab will produce single-cell RNA-seq, single-nuclear RNA-seq, and Dro-nuc-Seq data using post-mortem human brains from donors with appropriate consent. Samples are already on-site as part of the PsychENCODE consortium efforts. Sources include Sun Health Research Institute brain donation program and the Stanley Medical Research Institute.
2. SH-SY5Y cells are currently cultured in the White lab and were obtained from ATCC. These cells will be used for the initial whole genome STARR-seq experiments.
3. Primary human neuronal precursor cells (phNPCs), cell line D8R49, was developed from fetal cortext by the Geschwind laboratory. These cells will be used extensively for the STARR-seq, CRISPR mutagenesis, transgenic reporter assays in microfluidics chips, CRISPR Cas9-VP64 and DropSeq enhancer mapping experiments in Aims 3 and 4.
4. Forebrain cortical spheroid cells are derived from iPSC cell line 2242.1 originally developed from skin fibroblast cells and are differentiated in vitro into neuronal lineages. These cells are generously shared by our collaborator Sergiu Pasca (Stanford) and will be used in the microfluidics reporter assays, and in the CRISPR Cas9-VP64 and DropSeq enhancer mapping experiments in Aim 4.