Administrative Supplement Application for P50 MH106934 'Functional Genomics of Human Brain Development'

Public health statement

Autism is a complex developmental syndrome of unknown etiology with high personal and societal costs. Recent epidemiological data supports a population prevalence of just under 3% when the full spectrum is considered. Considerable genetic and phenotypic heterogeneity has complicated efforts to establish the biological substrates of the syndrome. The parent grant aims to understand the molecular mechanisms of normal brain development to understand how dysregulation of these processes lead to Autism Spectrum Disorders and other psychiatric disorders.

Abstract

The P50 center, with a multi-disciplinary group of investigators at Yale and UCSF, will be generating and integrating multi-dimensional genomic scale data from single cells and postmortem tissues of developing and adult human and non-human primate brains, as well as from blood and postmortem specimens of patients affected with Autism Spectrum Disorders (ASD). To elucidate molecular networks underlying human brain development and evolution it is necessary to integrate multi-dimensional data sets generated within P50 center and also by the other members of the PsychENCODE consortium, and this requires developing innovative analytical methods. We are requesting this supplement –to establish a PsychENCODE data analytics core (DAC) to uniformly process and analyze genomic datasets from the PsychENCODE consortium members, and discover the genomic elements and genotypes associated with brain diseases. The proposed supplement activities will increase while preserving the parent award's overall impact within the original scope of award.

Scope of the project

The PsychENCODE consortium consists of several groups that are coming together to map the non-coding elements and epigenetic landscape in control and diseased post-mortem human brain tissues and two model in vitro systems. Although the initial focus will be on ASD, bipolar disorder (BD), and schizophrenia (SCZ), we expect that the knowledge gained will be broadly applicable to the normal development of the central nervous system as well as to other human brain disorders. Each group will apply comprehensive, unbiased, complementary, and non-overlapping approaches to gain insight into the roles of non-coding elements in developing and adult, healthy and diseased human brains. Because of the heterogeneity in the consortium we need to standardize and harmonize datasets across all groups to facilitate data mining and analyses and novel tools are required to integrate data across groups in the psychENCODE and other consortia as well. This will also allow us to provide the data to the public in uniform format.

The objectives of the PsychENCODE DAC is to 1) establish uniform data processing pipelines and calibration resources for PsychENCODE consortium data, 2) discover the brain-specific spliced transcripts and enhancers, and 3) find the psychiatric disease-associated genotypes via the integrative analysis. The expertise of the key investigators (Gerstein, Sestan, State), their role in other large scale consortia (Gerstein in ENCODE, modENCODE, 1000 Genomes, exRNA) and previous collaboration is well suited to creating a DAC sub-component. The proposed supplement activities will increase while preserving the overall impact of parent award's scope and focus on ASD and human brain development. The Yale/UCSF subcomponent of the DAC will closely coordinate its efforts with the analyses done by the other two DAC subcomponents (headed by K White and Z Weng).

Composition of the DAC

The PsychENCODE DAC group will consist of investigators in the labs of Kevin White at the University of Chicago, Zhiping Weng at the University of Massachusetts Medical School and Mark Gerstein and Nenad Sestan at Yale University. Dr. White's lab will provide a platform and infrastructure for uniform processing of the data and running the pipelines. They will also focus on enhancer analysis. Dr. Weng's lab will support the enhancer analysis and also the construction of a number of the pipelines. Dr. Gerstein and Sestan's lab will develop a number of standardized pipelines and quality control metrics, and focus on the discovery of brain specific splicing transcripts, the aggregated quantitative trait locus (QTL) analysis, and integration of the above mentioned data sets with whole genome sequencing data and genotyping data from blood and post-mortem brains of patients affected with autism spectrum disorder (ASD), BD and SCZ. The DAC will report to and take directions from the PsychENCODE analysis working group, supporting the group's activities in integrative analysis.

Specific aims of the supplement

The PsychENCODE DAC will coordinate with CEGS and PsychENCODE investigators to achieve the following project aims.

Specific aim 1 PsychENCODE data processing pipelines and resources

The current PsychENCODE consortium employs a wide range of assays, including RNA-seq (total, small and long, Iso-seq), epigenetic mappings (WGBS for DNA methylation, ChIP-seq for various histone marks), accessible chromatin mapping (DNase-seq and ATAC-seq), and whole genome sequencing. Each group, however, uses different input materials. To compare data from different groups that will be generated on different control and disease brain tissue samples, it is necessary to apply uniform data processing steps. Specific aim 1 addresses these needs.

Subaim 1.1 Development of uniform data processing pipelines.

The Data Analysis Center of the ENCODE consortium, co-led by Dr. Zhiping Weng and Dr. Mark Gerstein, has established uniform processing pipelines and data quality control (QC) metrics for major data types such as RNA-seq, ChIP-seq of transcription factors and histone marks, DNase-seq, WGBS, CLIP-seq, fRIP-seq and Hi-C. Many of these types of data are being generated by the psychENCODE consortium, and the ENCODE pipelines and QC metrics

will guide the psychENCODE DAC. In addition, the psychENCODE DAC will develop uniform processing pipelines for NoMe-seq and ATAC-seq.

Gerstein lab will help develop a number of standardized pipelines and quality control metrics which will be run in large-scale on the PDC (Protected Data Cloud). This lab has considerable expertise in developing these standardized pipelines and evaluating them in many consortia including ENCODE, exRNA, KBase. We have developed tools like RSeqTools [1], IQSeq [2], FusionSeq [3] for the processing of RNA-seq data, and PeakSeq [4], which was the first peak caller to process ChIP-seq data relative to a correctly normalized input DNA control as well as accounting for variability in genome-wide sequence mappability, and MUSIC [5] for processing ChIP-seq data. The lab also played an important role in developing the so-called IDR method for determining reproducibility of target lists identified from replicate ChIP-seq experiments in order to correctly set thresholds uniformly across different ChIP-seq experiments for different TFs across different labs. These and other standards for performing and analyzing ChIP-seq experiments were published in Landt et al. 2012 [6]. Both Peak-Seq and MUSIC will be applied to the core psychENCODE pipeline.

Subaim1.2 Development of calibration methods.

In this subaim, we will perform integrated analyses of cross-platform data form all psychENCODE projects as well as aggregate data from other relevant consortia and projects.The PsychENCODE project will also avail the data from other consortia including ENCODE, PGC, CommonMind, GTEx, Roadmap epigenomics project and BrainSpan. Consequently we aim to develop calibration methods that will generate unified scoring for all datasets. For example, one can compare parameters for calling enhancers between ENCODE and psychENCODE. We will compare parameters like thresholds for calling peaks in ChIP-seq data, look at discrepancies between the two projects and calibrate parameters to reduce biases due to cross-project analyses. In addition to comparing the parameters in the uniform processing pipelines, we will also directly compare the annotated genomic regions or transcripts called by these pipelines. We will identify the brain cell types studied by ENCODE and other consortia, match them with the most appropriate datasets in psychENCODE, and investigate whether the corresponding pipelines in the two consortia have detected a similar set of genomic 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 either or both pipelines. If we identify major differences, we will investigate whether they are due to the underlying raw data, or the differences in the pipelines.

We plan to build this resource upon our experience on setting the standards within ENCODE and other consortia. For example, capitalizing on the uniformly processed and matched experimental data obtained by mod/ENCODE consortia, we have performed a series of comparative studies across distant metazoan phyla. A comparative analysis of human, worm, and fly revealed remarkable conservation of general properties of regulatory networks [7]. Also, as part of the GENCODE project we carried out a comprehensive annotation of pseudogenes, which was further integrated with ENCODE and 1000 Genomes Project data. All the information

was stored in an online resource called psiDR [8]. Moreover, the Gerstein Lab is in charge of DAC of exRNA consortium to develop the standardized RNA-seq pipelines for exRNAs.

Subaim 1.3 Development of aggregated eQTLs and allele analysis.

PsychENCODE will assemble the largest transcriptomic and epigenomic dataset, from approximately 1000 postmortem brains and will assess population variation of transcriptome and epigenomic features in the human brain. To achieve this, the DAC aims to aggregate all of the RNA-seq and also ChIP-seq datasets and call eQTLs and finding allelic sites. This will be carried out by the Gerstein lab using their Allele related tools, in particular, AlleleSeq and AlleleDB.

The Gerstein lab has a tremendous amount of experience in developing large databases of QTLs and allelic sites. AlleleSeq [9] is a tool developed specifically for the detection of allelic sites, including those associated with gene expression and transcription factor binding using RNA-seq and ChIP-seq datasets. AlleleSeq has been applied in several publications [10-12]. Notably, we have previously used AlleleSeq in allele-specific analyses associated with gene expression using ENCODE RNA-seq datasets from a single cell line [11]. Recently, we have further developed AlleleSeq and applied the new version to 1,139 RNA-seq and ChIP-seq datasets for 382 cell lines found in the 1000 Genomes Project. We harmonized and aggregated multiple RNA-seq and ChIP-seq datasets separately for each cell line and uniformly reprocessed them using the updated AlleleSeq. This allowed us to annotate the 1000 Genomes Project SNP catalog with allelic information. We constructed a database, AlleleDB, to house all the results. The database can be queried for specific genomic regions and visualized as a track in the UCSC browser [13] and visualizer such as the Integrated Genomics Viewer [14] or downloaded as flat files for downstream analyses for users that are more advanced in bioinformatics training. We continue to maintain and update AlleleSeq as a publicly available resource. It has been utilized considerably by the scientific community, as indicated by the number of citations and publications using our data and tool.

Aim 2 Brain specific spliced transcripts and enhancers

After building the calibration resource, we can compare the PsychENCODE data against non-brain data. We can identify the brain-specific spliced transcripts and enhancers from these comparisons.

Subaim 2.1 Identification of brain specific spliced transcripts.

We aim to identify the brain-specific spliced transcripts and splice sites from the RNA-seq data and collect them in a database. We will map the RNA-seq data to the transcriptome of human genome, compare PsychENCODE brain data to data from non-brain tissue in GTEx and ENCODE projects, and discover the brain-specific spliced transcripts.

The Gerstein and Sestan labs have rich experience in the mapping of RNA-seq data and the construction of transcripts form human brain and non-brain tissues. For example, RSEQtools is a computational package that enables expression quantification of annotated RNAs and identification of splice sites and gene models. IQseq provides a computationally efficient method to quantify isoforms for alternatively spliced transcripts. Both of these tools employ a special sequence read format we developed that can dissociate genome sequence information from

RNA-Seq signal, maintaining the privacy of test subjects. FusionSeq is designed for detecting fusion transcripts generated from either trans-splicing or genomic translocations in paired-end RNA-sequencing.

In particular we will use tools that we have developed such as RSEQtools, IQSeq and FusionSeq. The combination of these three tools allows us to efficiently discover brain specific spliced transcripts and their potential functions. The in-depth analysis will help to reveal the relationship between these specific splicings and the development of human brain. First, we will record all possible splicing events occurring in annotated genes, using RSEQtools package [1]. The reads will be mapped in parallel to the genome reference and the custom splice junction library. Both known and novel exon connections will be recorded. Next, we will use de novo splice site discovery tools (Tophat [15], AGE, BreaksSeq, etc.) to identify novel exon boundaries. Further, we will investigate the landscape of trans-splicing events between different primary transcripts using FusionSeq [3]. Finally, the splice sites identified from the three methods above will be compared to human gene annotation (e.g. GENCODE [16]). We also aim to characterize the brain-specific splicing dynamics under different conditions. We will use IQSeq [2] to assess quantitatively the extent of alternative promoter usage, exon-skipping, intron-retention, and alternative TSS under different conditions. We will characterize the condition-specific usage of alternative splicing events, and identify changes that vary significantly among conditions. The brain specific alternative splicing events will be compared to the cellular RNAs tissue-specific splicing patterns [17].

2.2 Building brain-specific enhancer database.

We will first aggregate all psychENCODE enhancers identified as a result of Specific aim 1and compare them with enhancers developed in other projects such as ENCODE, and Roadmap to determine a list of brain-specific enhancers. We will then do the uniform calling using our own enhancer pipeline for the enhancement.

We have much experience with developing enhancer calling pipelines in the framework of ENCODE, which we will utilize here. For example, we have applied machine-learning methods that integrate multiple genomics features to classify human regulatory regions from ENCODE data of more than 100 transcription factor binding sites. A computational pipeline was developed to identify potential enhancers from regions classified as gene-distal regulatory modules [18]. Making use of the potential enhancers, we developed the Function-based Prioritization of Sequence Variants (FunSeq) tool [12] for identification of candidate drivers in tumor genomes, and more recently, a more elaborate and flexible framework, FunSeq2, integrating various genomic and cancer resources to prioritize cancer somatic variants, especially regulatory noncoding mutations [19]. We are currently coordinating the enhancer prediction and validations for ENCODE 3. As the numbers of enhancer predictions are typically much larger than the small number of experimental validations performed with transgenic mice, we prioritized the predictions that are currently being experimentally tested and will also assess the accuracy of different enhancer prediction strategies once the experimental results are available.

Moreover, our algorithms and pipelines can be also used to find enhancer-target pairs on the genome wide. Due to the enhancer-gene interaction exhibit higher tissue specificity than

enhancer [20], the enhancer and gene interactome will be validated and re-defined using the enhancers called from psychENCODE data. Moreover, we will define new 3D enhancer-gene interactome based on the physical subregions of brain. Finally, the database will be a compendium of pyschENCODE enhancers, including all the enhancers from brain and other sources [20], and enhancer-gene interactome defined by different subregions of brain and other cell-line/tissues.

Aim 3 Integrative analyses to discover functional genomic elements and genotypes associated with psychiatric diseases

Finally after doing all of the unified scoring in building the integrated transcripts and chromatin resources we will attempt to do integrative cross disease analyses across projects. This will attempt to look at various genomic features common to all disorders. For doing this we will build on our considerable experience in ENCODE, modENCODE, 1000 Genomes, KBase, Brainspan in doing large scale cross project integrative analysis.

We have extensive experience in performing large scale integrative analysis in various consortia like ENCODE, modENCODE, 1000 Genomes, KBase and Brainspan. First, using the machine-learning approaches we developed method for identifying individual proximal and distal edges together with miRNA target prediction (and other) algorithms, we have completed the highly ambitious goal of constructing highly integrated regulatory networks for humans and model organisms based on the ENCODE [10] and modENCODE datasets [21,22]. These integrated networks consist of three major types of regulation: TF-gene, TF-miRNA and miRNA-gene, showing rich statistical patterns. For instance, the human regulatory network uniquely displays distinct preferences for binding at proximal and distal regions. The distal binding preference is possibly due to the intergenic space in the human genome, which is much larger relative to the genomes of other model organisms. More recently, we have constructed co-expression networks from the extensive amount of RNA-Seq data generated by ENCODE and modENCODE consortia [23]. We have developed a novel cross-species multi-layer network framework, OrthoClust, for analyzing the co-expression networks in an integrated fashion by utilizing the orthology relationships of genes between species [24]. OrthoClust revealed conserved modules across human, worm and fly that are important for development. We also introduced a framework to quantify differences between networks and by comparing matching networks across organisms, found a consistent ordering of rewiring rates of different network types [25].

We have extensive experience in using network framework to integrate data of human variation. We have developed NetSNP [26], an approach to quantify indispensability of each gene in the genome by incorporating multiple network and evolutionary properties. Based on network properties, as well as many other genomics features, we have developed FunSeq [12], and more recently FunSeq2 [19] for prioritizing variants. Using 1000 genomes variants, our pipeline has demonstrated great potential in prioritizing mutations in non-coding regions that are related to cancer [12].

Finally, laboratories of Nenad Sestan (Yale) and Matthew State (UCSF) have extensive experience in using network-level analyses to integrate multi-dimensional genomic scale data form postmortem human brains and cell types with whole exome sequencing data to generate new insights into pathogenesis of ASD [27-30]. We have also developed [31] XSAnno, a framework for building ortholog models in cross-species transcriptome comparisons. We will improve above tools to develop a standardized framework to integrate and contextualize genetic findings from ASD, SCZ and BD along spatial, temporal, and cellular axes of human and non-human primate brain development. This approach should bridge the gap between high-throughput functional genomic data generated by psychENCODE consortium, gene and GWAS SNP discovery and testable pathophysiological hypotheses on the origin and progression or psychiatric disorders.

Specific need for the supplement

The data analytics core (DAC) was not funded in the parent grant and is need to accommodate and integrate a much larger dataset that will be generated as part of the psychENCODE consortium.

Relationship to parent grant and how the supplement fits within the scope of parent grant

The goal of the parent grant is to develop new approaches and methods for generating and analyzing multi-dimensional genomic data for the developing brain of human and ASD. To accomplish this we will produce transcriptome data, chromatin architecture data and genome-wide maps of chromatin modifications at the level of tissue, cell types and single cells in control and ASD brains. These maps define the locations and activation states of diverse functional elements including protein-coding genes, their regulatory elements (promoters, enhancers), and non-coding transcripts. The data will be integrated with other data sets generated by other members of PsychENCODE consortium including those in control and two other psychiatric diseases- schizophrenia and bipolar disorder. The purpose of this supplement is to create the DAC to standardized and harmonized data and analyses across all psychENCODE projects.

References

1. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, et al. (2011) RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics 27: 281-283.
2. Du J, Leng J, Habegger L, Sboner A, McDermott D, et al. (2012) IQSeq: integrated isoform quantification analysis based on next-generation sequencing. PLoS One 7: e29175.
3. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, et al. (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol 11: R104.
4. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, et al. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27: 66-75.

5. Harmanci A, Rozowsky J, Gerstein M (2014) MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability-Corrected Multiscale Signal Processing Framework. Genome Biol 15: 474.
6. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22: 1813-1831.
7. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, et al. (2014) Comparative analysis of regulatory information and circuits across distant species. Nature 512: 453-456.
8. Pei B, Sisu C, Frankish A, Howald C, Habegger L, et al. (2012) The GENCODE pseudogene resource. Genome Biol 13: R51.
9. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, et al. (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol 7: 522.
10. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489: 91-100.
11. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. (2012) Landscape of transcription in human cells. Nature 489: 101-108.
12. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342: 1235587.
13. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.
14. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nat Biotechnol 29: 24-26.
15. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105-1111.
16. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, et al. (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7 Suppl 1: S4 1-9.
17. Li G, Bahn JH, Lee JH, Peng G, Chen Z, et al. (2012) Identification of allele-specific alternative mRNA processing via transcriptome sequencing. Nucleic Acids Res 40: e104.
18. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, et al. (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol 13: R48.
19. Fu Y, Liu Z, Lou S, Bedford J, Mu X, et al. (2014) FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. Genome Biol 15: 480.
20. He B, Chen C, Teng L, Tan K (2014) Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci U S A 111: E2191-2199.
21. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, et al. (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 330: 1775-1787.
22. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, et al. (2011) A cis-regulatory map of the Drosophila genome. Nature 471: 527-531.
23. Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, et al. (2014) Comparative analysis of the transcriptome across distant species. Nature 512: 445-448.
24. Yan KK, Wang D, Rozowsky J, Zheng H, Cheng C, et al. (2014) OrthoClust: an orthology-based network framework for clustering data across multiple species. Genome Biol 15: R100.
25. Shou C, Bhardwaj N, Lam HY, Yan KK, Kim PM, et al. (2011) Measuring the evolutionary rewiring of biological networks. PLoS Comput Biol 7: e1001050.
26. Khurana E, Fu Y, Chen J, Gerstein M (2013) Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol 9: e1002886.

27. Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, et al. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell 155:997-1007.

28: Cotney J, Muhle RA, Sanders SJ, Liu L, Willsey AJ, et al., (2015). The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. Nat Commun 6:6404.

29. Liu L, Lei J, Sanders SJ, Willsey AJ, Kou Y, et al. (2014). DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. Mol Autism 5:22.

30. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, et al., (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485: 237-241.

31. Zhu Y, Li M, Sousa AM, Sestan N (2014) XSAnno: a framework for building ortholog models in cross-species transcriptome comparisons. BMC Genomics15:343.