Aim 1

Capstone 1. Cross-Disorder Gene Expression Analyses

Specific Aims of the Capstone 1 [prepared by Dan, Chunyu & Kevin]
Psychiatric disorders are highly heritable, but also have a highly complex, polygenic risk architecture polygenic risk architecture [1, 2]. It is becoming increasingly appreciated that both the rare and common genetic variation that contributes to these disorders often lies outside of coding regions, and because of this it is expected that most of this disease risk will act on gene regulation, rather than protein structure. So, to have mechanistic knowledge of how genetic risk leads to disease it is necessary to connect genetic variation to its functional consequences. This necessitates comprehensive analyses of the regulatory regions, epigenetic modifications, and gene expression patterns present across brain regions and comparing these patterns between healthy and diseased human brain. The PsychEncode project is focused on providing a landmark resource elucidating the landscape of the regulatory genome as it applies to brain development and function, so as to permit mechanistic interpretation of disease associated genetic variation across several major psychiatric diseases.

The consortium is essentially organized around specific techniques and individual disorders, with no single laboratory or group performing transcriptomic, methylomic and chromatin structure analyses across all of the major disorders. The importance of cross-disorder analyses is becoming more appreciated as both a significant proportion of polygenic risk and rare risk variants cross disease boundaries boundaries [1-5]. Since the majority of the original funding has gone into generating data, additional support for integrative cross disorder analyses are necessary. The analysis that is currently funded is primarily geared towards individual diseases, or otherwise limited in scope.

Here we (Geschwind, Liu and White) propose to begin collaborative cross disorder analyses within PsychEncode from data already generated, and build a process that incorporates additional data as it is produced. This is one of four capstone projects that we expect to lead to significant publications within the next one year, so as to increase the profile of the resource within the community, and provide data and analyses that will guide future work. At this point, there will sufficient RNAseq data and genotyping data to perform cross disorder analyses of RNAseq based gene expression, as well as preliminary eQTL analyses. Subsequently as additional data on histone modifications and chromatin structure become available, these will be intergrated into cross disorder analysis. The data that is available for this project is summarized in the Table above.

RNAseq have undergone 50 or 100 bp paired end Ribo-Zero based sequencing at an average depth of about 50M reads per sample. By June 2016, a total of 600 samples representing 113 individuals from UCLA will be completed and quality checked. Combined with the Common Mind (CMC) samples, which after QC and removal of outliers includes BD (n = 44), SCZ (161) and control samples (236) and the BrainGVEX cohort (19 SCZ, 17 BD and 184 controls), this will amount to over 600 SZ, 300 BD, and 65 ASD and a similar number of matched control samples, as well as additional control brains. Genotypes are available from most of the samples, permitting reasonable powered eQTL and sQTL analysis. We will lead the efforts gathering the data so a cross-site analysis can be performed.

Proposed Activities:

1) Identify cross disorder patterns of differential gene expression (DE) and splicing. We will combine all of the RNAseq data and process it accounting for biological and technical covariates using a linear mixed model to define DE genes. We will also perform meta-analysis of logFC values using a random effects model, restricted maximum likelihood estimate with the *metafor* package in R (Viechtbauer W. 2005). In tandem, we will perform the same analysis using a smaller subset of carefully matched samples (age, sex, RIN) to show that the regression does not bias the results.

For splicing we will calculate the percent spliced in (PSI) of each alternative splicing event using MATS, which performs very well for simple splicing events [6, 7]. In the future, we can compare MATS to other methods to identify novel splicing events, which may not be detected by MATs

2) We will create cross disorder gene networks using WGCNA[8] to identify distinct as well as overlapping biology. We will start by using ME correlations (stringently corrected for multiple comparisons) to identify how the transcriptional alterations in ASD, SCZ, and BD cerebral cortex are altered to control brain networks, as well as identify disease specific patterns. In separately funded work, the Geschwind lab has been working with the Battle lab at Hopkins to develop integrating splicing and expression networks. Once these methods are more fully tested, they can also be applied to this problem. eQTL data derived in #3 below will also be used to identify causal relationships, adding causal directionality to the network edges, which is not possible at large scale and with precision using expression data alone.

3) We will perform eQTL and sQTL analyses to permit causal inference. Using the Network Edge Orienting (NEO)[9] with the WGCNA and disease association data, we can calculate probabilities of edges of all the network connections. Subsequently, causal inference can be resolved for some of the regulatory relationships, and disease GWAS signals as well.

In addition to using individual genes and splicing events, we will also use control and disease specific module eigengenes as composite measures to identify possible trans eQTL that may be driving co-expression. This will permit tying causal genetic factors to the observed transcriptional alterations at a network level.

1. Geschwind DH, Flint J. Genetics and genomics of psychiatric disease. Science. 2015;349(6255):1489-94. doi: 10.1126/science.aaa8954. PubMed PMID: 26404826; PMCID: 4694563.

2. Gratten J, Wray NR, Keller MC, Visscher PM. Large-scale genomics unveils the genetic architecture of psychiatric disorders. Nat Neurosci. 2014;17(6):782-90. doi: 10.1038/nn.3708. PubMed PMID: 24866044; PMCID: PMC4112149.

3. Cross-Disorder Group of the Psychiatric Genomics C, Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, Perlis RH, Mowry BJ, Thapar A, Goddard ME, Witte JS, Absher D, Agartz I, Akil H, Amin F, Andreassen OA, Anjorin A, Anney R, Anttila V, Arking DE, Asherson P, Azevedo MH, Backlund L, Badner JA, Bailey AJ, Banaschewski T, Barchas JD, Barnes MR, Barrett TB, Bass N, Battaglia A, Bauer M, Bayes M, Bellivier F, Bergen SE, Berrettini W, Betancur C, Bettecken T, Biederman J, Binder EB, Black DW, Blackwood DH, Bloss CS, Boehnke M, Boomsma DI, Breen G, Breuer R, Bruggeman R, Cormican P, Buccola NG, Buitelaar JK, Bunney WE, Buxbaum JD, Byerley WF, Byrne EM, Caesar S, Cahn W, Cantor RM, Casas M, Chakravarti A, Chambert K, Choudhury K, Cichon S, Cloninger CR, Collier DA, Cook EH, Coon H, Cormand B, Corvin A, Coryell WH, Craig DW, Craig IW, Crosbie J, Cuccaro ML, Curtis D, Czamara D, Datta S, Dawson G, Day R, De Geus EJ, Degenhardt F, Djurovic S, Donohoe GJ, Doyle AE, Duan J, Dudbridge F, Duketis E, Ebstein RP, Edenberg HJ, Elia J, Ennis S, Etain B, Fanous A, Farmer AE, Ferrier IN, Flickinger M, Fombonne E, Foroud T, Frank J, Franke B, Fraser C, Freedman R, Freimer NB, Freitag CM, Friedl M, Frisen L, Gallagher L, Gejman PV, Georgieva L, Gershon ES, Geschwind DH, Giegling I, Gill M,

Gordon SD, Gordon-Smith K, Green EK, Greenwood TA, Grice DE, Gross M, Grozeva D, Guan W, Gurling H, De Haan L, Haines JL, Hakonarson H, Hallmayer J, Hamilton SP, Hamshere ML, Hansen TF, Hartmann AM, Hautzinger M, Heath AC, Henders AK, Herms S, Hickie IB, Hipolito M, Hoefels S, Holmans PA, Holsboer F, Hoogendijk WJ, Hottenga JJ, Hultman CM, Hus V, Ingason A, Ising M, Jamain S, Jones EG, Jones I, Jones L, Tzeng JY, Kahler AK, Kahn RS, Kandaswamy R, Keller MC, Kennedy JL, Kenny E, Kent L, Kim Y, Kirov GK, Klauck SM, Klei L, Knowles JA, Kohli MA, Koller DL, Konte B, Korszun A, Krabbendam L, Krasucki R, Kuntsi J, Kwan P, Landen M, Langstrom N, Lathrop M, Lawrence J, Lawson WB, Leboyer M, Ledbetter DH, Lee PH, Lencz T, Lesch KP, Levinson DF, Lewis CM, Li J, Lichtenstein P, Lieberman JA, Lin DY, Linszen DH, Liu C, Lohoff FW, Loo SK, Lord C, Lowe JK, Lucae S, MacIntyre DJ, Madden PA, Maestrini E, Magnusson PK, Mahon PB, Maier W, Malhotra AK, Mane SM, Martin CL, Martin NG, Mattheisen M, Matthews K, Mattingsdal M, McCarroll SA, McGhee KA, McGough JJ, McGrath PJ, McGuffin P, McInnis MG, McIntosh A, McKinney R, McLean AW, McMahon FJ, McMahon WM, McQuillin A, Medeiros H, Medland SE, Meier S, Melle I, Meng F, Meyer J, Middeldorp CM, Middleton L, Milanova V, Miranda A, Monaco AP, Montgomery GW, Moran JL, Moreno-De-Luca D, Morken G, Morris DW, Morrow EM, Moskvina V, Muglia P, Muhleisen TW, Muir WJ, Muller-Myhsok B, Murtha M, Myers RM, Myin-Germeys I, Neale MC, Nelson SF, Nievergelt CM, Nikolov I, Nimgaonkar V, Nolen WA, Nothen MM, Nurnberger JI, Nwulia EA, Nyholt DR, O'Dushlaine C, Oades RD, Olincy A, Oliveira G, Olsen L, Ophoff RA, Osby U, Owen MJ, Palotie A, Parr JR, Paterson AD, Pato CN, Pato MT, Penninx BW, Pergadia ML, Pericak-Vance MA, Pickard BS, Pimm J, Piven J, Posthuma D, Potash JB, Poustka F, Propping P, Puri V, Quested DJ, Quinn EM, Ramos-Quiroga JA, Rasmussen HB, Raychaudhuri S, Rehnstrom K, Reif A, Ribases M, Rice JP, Rietschel M, Roeder K, Roeyers H, Rossin L, Rothenberger A, Rouleau G, Ruderfer D, Rujescu D, Sanders AR, Sanders SJ, Santangelo SL, Sergeant JA, Schachar R, Schalling M, Schatzberg AF, Scheftner WA, Schellenberg GD, Scherer SW, Schork NJ, Schulze TG, Schumacher J, Schwarz M, Scolnick E, Scott LJ, Shi J, Shilling PD, Shyn SI, Silverman JM, Slager SL, Smalley SL, Smit JH, Smith EN, Sonuga-Barke EJ, St Clair D, State M, Steffens M, Steinhausen HC, Strauss JS, Strohmaier J, Stroup TS, Sutcliffe JS, Szatmari P, Szelinger S, Thirumalai S, Thompson RC, Todorov AA, Tozzi F, Treutlein J, Uhr M, van den Oord EJ, Van Grootheest G, Van Os J, Vicente AM, Vieland VJ, Vincent JB, Visscher PM, Walsh CA, Wassink TH, Watson SJ, Weissman MM, Werge T, Wienker TF, Wijsman EM, Willemsen G, Williams N, Willsey AJ, Witt SH, Xu W, Young AH, Yu TW, Zammit S, Zandi PP, Zhang P, Zitman FG, Zollner S, Devlin B, Kelsoe JR, Sklar P, Daly MJ, O'Donovan MC, Craddock N, Sullivan PF, Smoller JW, Kendler KS, Wray NR, International Inflammatory Bowel Disease Genetics C. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nature genetics. 2013;45(9):984-94. doi: 10.1038/ng.2711. PubMed PMID: 23933821; PMCID: 3800159.

4. Cross-Disorder Group of the Psychiatric Genomics C. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013;381(9875):1371-9. doi: 10.1016/S0140-6736(12)62129-1. PubMed PMID: 23453885; PMCID: PMC3714010.

5. Malhotra D, Sebat J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell. 2012;148(6):1223-41. doi: 10.1016/j.cell.2012.02.039. PubMed PMID: 22424231; PMCID: 3351385.

6. Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(51):E5593-601. doi: 10.1073/pnas.1419161111. PubMed PMID: 25480548; PMCID: 4280593.

7. Liu R, Loraine AE, Dickerson JA. Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. BMC bioinformatics. 2014;15:364. doi: 10.1186/s12859-014-0364-4. PubMed PMID: 25511303; PMCID: 4271460.

8. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article17. doi: 10.2202/1544-6115.1128. PubMed PMID: 16646834.

9. Aten JE, Fuller TF, Lusis AJ, Horvath S. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Syst Biol. 2008;2:34. doi: 10.1186/1752-0509-2-34. PubMed PMID: 18412962; PMCID: PMC2387136.

Capstone 2. Adult and Disease - Epigenetic map

Subaim 1.3 Transcriptome and eQTL analyses across human brain development (Capstone 3a)

Several studies have pointed to mid-fetal brain development as the vulnerable period for ASD and schizophrenia risk. It is therefore critical to map and study and the effect of DNA variations on gene expression during the prenatal and postnatal brain development. Capstone project 3a proposes to map eQTLs in the human neocortex across prenatal and postnatal development. We are performing RNA-seq and whole genome sequencing (WGS) data from 200 neocortical samples across multiple periods of development including the entire course of the prenatal development, infancy and adolescence.

In continuation of our previous efforts, for the detection of genetic variants by WGS, we will utilize the GATK pipeline for local realignment for indels, base quality score recalibration, SNP and indel calling, and variant quality score recalibration. Alongside the variant calling pipeline we will run extensive quality metrics that include assessment of sequencing quality, read coverage, duplicate reads, Ti/Tv ratio, and number of detected variants. To accurately measure gene expression levels, several first order QC analyses will be implemented, followed by employing RSEQtools to quantify expression. Two types of gene expression values- read counts and RPKM (reads per kilobase of exon model per million mapped reads)- will be generated and the R package DESeq will be used to identify differentially expressed (DEX) genes. *Cis*-eQTLs will be identified through linear regression between normalized gene-level expression values and individual genotypes, focusing on all SNPs within ±1 Mb of the transcriptional start site (TSS) of each gene, as done by the Genotype-Tissue Expression (GTEx) project. Any batch effects, technical/biological covariates, or other important covariates for gene expression will be included in the regression analysis. Due to our modest sample size for eQTL discovery, we expect that the effect size of *trans*-eQTLs will make detection difficult in this cohort. However, we would explore detecting *trans*-eQTLs by correcting only for SNPs and not for genes.

In addition to these efforts, we will seek to identify structural variants (SVs), including copy number variants (CNVs), to identify svQTLs in the human brain across development. There is substantial evidence that large, rare CNVs contribute risk to psychiatric disorders, including ASD and schizophrenia and the GTEx consortium has recently demonstrated profound regulatory impact of SVs and ascribed common CNVs to previously implicated risk loci from GWAS of several disorders. Using WGS data from 2,000 samples from ASD families in the Simons Simplex Collection and in collaboration with the Talkowski lab at Harvard/MGH we have developed a reliable pipeline for SV detection. This combines seven prediction algorithms to capture both changes in read-depth (Genome STRiP, CNVnator, CN.mops) and split reads/read-pairs indicating SV breakpoints (Lumpy, Delly, Wham, Manta). svQTLs will be identified using methods described in the GTeX paper, specifically assessing gene-level expression normalized by gene length. We will identify common and rare svQTLs using FastQTL to assess the 1Mbp upstream and downstream of the SV breakpoint. We will assess overlap with cell type specific markers to identify svQTLs that are potentially specific to certain cell types.

We also propose to assess whether there is evidence of enrichment for noncoding de novo mutations identified in whole genome sequencing (WGS) data from over 5,000 samples in autism spectrum disorder (ASD) families. Our prior work on 519 ASD families has identified 69 de novo mutations per child. Comprehensive analysis of about 60,000 annotation categories did not identify a significant excess of mutations in cases vs. sibling controls for any independent category; the strongest signal came from de novo missense mutations. We are working on extending this analysis to over 1,300 ASD families. Along with increasing the sample size, we need to identify noncoding regions likely to have functionally important impact in neurodevelopment. The QTLs from the data of this capstone project provide an independent method to identify such regulatory regions. We will assess whether there is evidence of enrichment of de novo mutations in ASD cases under all QTLs using 10,000 label-swapping permutations to assess significance. In addition, we will consider subsets of these QTLs based on gene target (e.g. brain expressed genes, ASD genes), species conservation, and QTL specificity (e.g. evidence of QTL effect limited to specific developmental stages or cell types).

Capstone 3b. Construction of a Developmental EpiMap

ABSTRACT

The goal of PsychENCODE project is to map and characterize transcriptomes and functional genomic elements from the postmortem human brain as well as cell culture systems in normal human brain and across disorders. Within the PsychENCODE, the aim of our parent project is to compare transcriptomes and regulatory genomic elements mapped by acetylated and methylated histones between mid-fetal human brains and brain "organoids" derived from induced pluripotent stem cells (iPSC) derived from the same fetal specimen. The scope is to validate the iPSC model by directly comparing isogenic neurons that are iPSC-derived with those derived from postmortem human brains. The scope of this supplement is two-fold: (1) to complete data analyses in order to achieve the scientific goals of **Capstone 3b**, and (2) to contribute to the activities of the **DAC/DCC**. In **Capstone 3b** (**Construction of a Developmental EpiMap**), we will achieve the goal of understanding differences across age (early development through adulthood) with respect to gene expression and regulatory elements; we will accomplish this goal by integrating our datasets with those generated by the entire PsychENCODE project. The data will be also compared to those obtained from external datasets (BranSpan, GTEX, Roadmap). The scope of this capstone is to understand differences in gene regulation across human development and how regulatory genomic regions control the transcriptome as cortical development unfolds dynamically. The scope of the **DAC/DCC** is to complete processing of all data according to common RNA-seq and ChIP-seq pipelines and ensure sharing and dissemination of all data to consortium members and the scientific community. Thus, the current supplement will allow us to (1) contribute to the overall activities of the DAC/DCC; (2) finish RNAseq and Chipseq data analyses pertinent to Capstone 3b; (3) perform cross-comparative and integrative analyses pertaining to Capstone 3B using both PsychENCODE and external datasets (3) submit manuscripts for wide dissemination to the scientific community.

SPECIFIC AIMS

Specific Aim 1. Capstone 3B.

Aim **1A. Analyses of RNAseq datasets.** Perform comparative analyses between fetal, early postnatal and adult brain transcriptomes as well as cellular model systems (iPSC-derived organoids at different stages of development and CNON cells) from Psychencode consortium datasets; correlate with external datasets such as BrainSpan and other consortia.

Aim 1B. Analyses of Chip-seq datasets. Comparison of acetylated and methylated histone marks and maps of open/closed chromatin states across normal development, and focusing on three stages: (1) fetal; (2) childhood; (3) adulthood. Using Psychencode consortium data, specifically assess genome-wide enrichment for the histone marks H3K4me3, H3K27ac, H3K27me3 in mid-fetal development, childhood and adulthood brain samples. The aim is to find specific elements whose activity is increased or decreased at specific stages to map their overall spatio-temporal dynamics across epochs of human brain development.

Aim 1C. **Define the location and activity of putative promoters and enhancers** to construct a developmental epigenomic brain map focusing on three stages: (1) fetal; (2) childhood; (3) adulthood, and the correspondence of this epimap with neural cells derived from the *in vitro* model systems, iPSCs-derived telencephalic organoids and cultured neuronal cells derived from olfactory neuroepithelium (CNON cells). Use histone Chip-Seq data and ChromHMM for each tissue type to segment chromatin into distinct states corresponding to particular biochemical functions, for example, promoters enhancers, insulators. Call putative enhancers and promoters from Chip-Seq and ChromHMM analyses; perform comparative analyses of active enhancer and promoters between fetal, early postnatal and adult cortical brain samples as well as cellular model to identify the brain-specific gene regulatory elements that are active in different stages of development, in models,

Aim 1D. Perform higher order analyses. Using published Hi-C datasets available for fetal and adult brain, integrate the transcriptome and chromatin/TF-binding datasets to link transcripts with likely promoters. Intersect the developmental EpiMap constructed above with datasets of putative mutations in developmental disorders (autism, intellectual disabilities) that are becoming publically available, such as MISSNG and SFARI. Identify and catalogue those noncoding transcripts that are in loci previously implicated in developmental disorders such as autism.

Specific Aim 2. Participation to DAC/DCC activities.

Aim 2A. DAC. Processing of RNA-seq and ChIP-Seq through a common pipeline and implement a standard genotype imputation pipeline.

Aim 2B. DCC. Coordinate data analyses generated through the psychENCODE project grants, maintain the online, publicly accessible record of data, identify long term storage solutions and promote their quick dissemination to the research community.

EXPERIMENTAL DESIGN AND METHODS:

Aim 1A. We will analyze PsychENCODE consortium datasets at different stages of development including fetal, early postnatal, and adult brain tissue as well as cellular model systems (iPSC-derived organoids at different stages of development and CNON cells) to identify and annotate transcriptionally active region to reference to all transcripts (expressed genes, small, long and regulatory RNAs). We will use edgeR to compare transcripts expression levels, from counts data, at different stages of development as well as cellular model systems. Pearson's rank correlation coefficients of gene expression and number of differentially expressed genes will be used to make judgments about similarity of iPSC-derived and dissected cells.

Aim 1B. From histone Chip-Seq data we will identify histone peaks that can be related to actively transcribed or repressed genes/regions. For each histone mark, we will cluster samples by peak concordance (the fraction of common peaks between a pair of samples). This clustering analysis will include PsycENCODE consortium datasets, including fetal, early postnatal, and adult brain tissue and iPSC-derived neurons and CNON cells, and external datasets from projects such as BrainSpan and Roadmap Epigenomics (Roadmap Epigenomics et al., 2015). We expect that clusters should distinguish different developmental stages and cell types, for instance, fetal brains, early postnatal, adult brains, and cellular models, and that neuron-enriched samples will cluster separately from glial samples. We will also detect differential

peaks among the above time points from brains and cellular models. We expect to identify common and differential usage of active/poised/repressed promoters (H3K4me3 and H3K27me3) and active enhancers (H3K27ac) across developmental stages and cell types. This analysis is complementary to chromatin states described below.

Aim 1C. We will segment the genome into distinct chromatin states using ChromHMM (Ernst and Kellis, 2012). To compare segmented chromatin states we will calculate a few metrics (reflecting various differences that can be observed in chromatin): fraction of genome with the same/similar/different states, quantitative values (i.e., phi coefficient) describing patters of changes in contingency table, distribution of boundary shifting between neighboring segments. We will calculate the values by these metrics for iPSC-derived vs VZ/SVZ progenitor and cortical neurons and compare values across cellular models (i.e., iPSC-derived neurons and CNON cells) and across different brain samples from PsychENCODE at different stages of development (fetal, early postnatal, and adult brain tissue) as well as external datasets such as BrainSpan and the roadmap epigenomic dataset of ENCODE (http://egg2.wustl.edu/roadmap/web_portal/).

We will compare our results with those obtained by ENCODE and for non-brain related cell lines/tissues to define novel elements (e.g., TF peaks, expressed transcripts) and classify elements into broad classes: i) likely specific to brain; ii) pertinent to brain; iii) differentially active during brain development; iv) generic. We will define likely brain specific elements as those for which we have evidence of biochemical activity only in developing brain tissues (and, possibly, in brain cancer tissues analyzed by ENCODE) but not in other tissues/cell lines.

Aim 1D. Having identified elements (both transcribed and regulatory) that are likely specific or pertinent to the embryonic, early postnatal and adult brain, we will correlate those with Hi-C published data for embryonic and adult human brain (Rao et al., 2014; Won et al., 2016) to infer likely coupling between regulatory and transcribed regions during human brain development.

Here, we will link differentially expressed genes during brain development (Aim 1A) with differential regulatory regions (Aim 2) by measures of long range chromosome interactions derived from Hi-C data, in order to accurately match regulatory elements with their target genes. This will identify changes in chromatin interactions at those loci showing differential gene expression and differential epigenetic state during development. We are first aiming at assigning pairs of differentially active enhancers and genes. For differentially active enhancers (identified in Aim 1C) we will call genes that are in contacts with them using the Hi-C data. Identified genes will be tested for differential expression. We expect that a large fraction of the target genes (but not all of them, as due to the relatively low resolution of Hi-C contact map some genes may not be detected as targets of differential enhancers) will show differential expression in fetal vs adult brains, thereby validating the functional effect of the enhancers. Next we will conduct pathway enrichment using Ingenuity Pathway Analysis (http://www.ingenuity.com) and DAVID (Dennis et al., 2003) (https://david.ncifcrf.gov/). We expect enrichment of target genes in certain pathways, like GABAergic neuronal differentiation, which would also validate the differential enhancers.
Additionally, we will conduct more refined and precise analysis of enhancer-gene interactions. For this, from Hi-C data we will identify TADs using Juicer (Durand et al., 2016) and TADtool (Kruse et al., 2016). For differential enhancers and genes within a TAD we will compare their interactions relative to a background inferred from other interactions within the TAD. Significant interactions will make a list of enhancer-gene pairs. Then we will test for TADs that are enriched with differential enhancers, by identifying TADs in which the ratio of differential to all enhancers within the TAD is significantly above genome-wide average. Genes in the TADs enriched for differential enhancers will be subject to differential expression analysis. Given that now we are not testing for all genes in the human genome, we will drastically reduce multiple testing and expect to find additional differentially expressed genes. Enhancer-gene pairs will be defined as

before. We will also extend our analysis by selecting 1% of TADs showing the smallest (but may be not significant) p-values of enrichment of differential enhancers. For those TADs we will perform differential gene expression and enhancer-gene matching as just described. As a result we will construct a list of differentially active enhancer-gene pairs.

Analysis of mutations in ASD subject vs controls using the MSSNG and Simons collections. Personal variants for over 2,000 ASD subjects are available in each database. Each database contains variants for familiar trios and quartets rather than for single ASD subjects, and each family has a phenotypically normal parents and a child with ASD (almost always male). To match for sex, fathers will be used as controls. We will compare counts of rare variants defined from their allele frequency in the human population, i.e., <5% AF (set 1), <1% AF (set 2), and <0.1% (set 3). We will test for an enrichment of variants in ASD subjects vs controls in: i) differential enhancers and flanking regions of a few hundred base pairs; ii) TADs where differential enhancers were found (we hypothesize that pathogenic variants in these TADs may alter chromatin organization and affect enhancer activity; iii) TADs where differentially expressed genes were found (we hypothesize that pathogenic variants in these TADs may alter chromatin organization and affect gene expression).

EXPECTED DELIVERABLES AND MILESTONES:

Capstone 3B:

(1) We will provide a comparison between transcripts expressed in cortical brain tissue and cellular models (i.e., iPSC-derived organoids and CNON cells) and transcripts that are brain specific and expressed at various stages of prenatal and postnatal development by comparing our datasets with the Psychencode, ENCODE and other external databases.

(2.) We will produce a list of genomic elements (e.g. enhancers, promoters, TF binding sites, repressed regions) active in cortical brain tissue that are brain specific and expressed at specific stages of prenatal and postnatal development (from PsychENCODE datasets) and in cellular models (i.e., iPSC-derived organoids and CNON cells) to begin to construct a developmental epimap.

(3) We will define the correspondence of the developmental epigenomic brain map with neural cells derived from the *in vitro* model systems and construct a list of differentially active regulatory elements and differentially expressed genes. Analysis of personal variants in public developmental disorders (MISSING, SFARI) will identify a, perhaps overlapping, list genomic elements that are enriched in variants associated with ASD.

Milestones:

Quarter 1: Completing RNA-seq analyses. Completing QC analyses for histone ChipSeq data.

Quarter 2: Repeating some histone ChipSeq data. Completing analyses for existing ChipSeq data. Conducting data integration (regulome with expression).

Quarter 3: Finalizing histone ChipSeq data analyses. Completing analyses for all ChipSeq data. Conducting data integration (regulome with expression).

Quarter 4: Finalizing data integration. Completing analyses for capstone. Writing manuscripts.

LITERATURE CITED

Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45:1113-1120. PMCID: 3919969.

Delaneau O, Marchini J, Zagury JF (2011) A linear complexity phasing method for thousands of genomes. Nat Methods 9:179-181. PMCID:

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4:P3. PMCID: 3720094.
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15-21. PMCID: 3530905.
Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL (2016) Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell systems 3:95-98. PMCID:
Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9:215-216. PMCID: 3577932.
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. PMCID: 4234855.
Kruse K, Hug CB, Hernandez-Rodriguez B, Vaquerizas JM (2016) TADtool: visual parameter identification for TAD-calling algorithms. Bioinformatics 32:3190-3192. PMCID: 5048066.
Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323. PMCID: 3163565.
Margolin AA, Bilal E, Huang E, Norman TC, Ottestad L, Mecham BH, Sauerwine B, Kellen MR, Mangravite LM, Furia MD, Vollan HK, Rueda OM, Guinney J, Deflaux NA, Hoff B, Schildwachter X, Russnes HG, Park D, Vang VO, Pirtle T, Youseff L, Citro C, Curtis C, Kristensen VN, Hellerstein J, Friend SH, Stolovitzky G, Aparicio S, Caldas C, Borresen-Dale AL (2013) Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Sci Transl Med 5:181re181. PMCID: 3897241.
Omberg L, Ellrott K, Yuan Y, Kandoth C, Wong C, Kellen MR, Friend SH, Stuart J, Liang H, Margolin AA (2013) Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. Nat Genet 45:1121-1126. PMCID: 3950337.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575. PMCID: 1950838.
Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159:1665-1680. PMCID:
Roadmap Epigenomics C et al. (2015) Integrative analysis of 111 reference human epigenomes. Nature 518:317-330. PMCID: 4530010.
Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27:66-75. PMCID:
Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, Lee C, Eskin E, Voineagu I, Ernst J, Geschwind DH (2016) Chromosome conformation elucidates regulatory relationships in developing human brain. Nature 538:523-527. PMCID:

Capstone 4. Integrative analysis with CommonMind, Brainspan, GTEx, ENCODE and ROADMAP

We will perform integrative and comparative analysis using GTEx, roadmap, CommonMind, BrainSpan and PsychENCODE data. For doing this we will build on our considerable experience in ENCODE, modENCODE, 1000 Genomes and KBase in doing large scale cross

project integrative and comparative analysis. In order to integrate data from different projects, we will use the data uniformly processed by PsychENCODE pipeline, then normalize the data to correct the batch effects.

After we have normalized all the uniformly processed data from different projects, we will do more integrative analysis. In particular, 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 (Yan, Wang et al. 2014). We would like to identify the greatly differentially expressed genes in brain versus the other tissues in ENCODE and GTEx Project. These genes can be the biomarkers for distinguishing different tissues. After normalization, we then want to analyze their temporal expression dynamic patterns. We want to identify these expression patterns associated with specific brain regions and specific tissues. In particular, we will construct the gene co-expression networks where genes are connected with correlated expression profiles across different tissues. We will then cluster this network into gene co-expression modules and find modules (with associated gene expression signatures) enriched in brain. Finally, we will identify the gene regulatory logics using Loregic that drive the tissue types such as the biomarker genes associated with specific tissues (Huffman, Koves et al. 2014).

We will use the better enhancer definition provided by the Epigenome Roadmap (Leung, Jung et al. 2015, Roadmap Epigenomics, Kundaje et al. 2015, Ziller, Edri et al. 2015), and more recently from ENCODE projects. We will predict brain specific active enhancers based on H3K27Ac ChIP-Seq datasets generated by the Epigenome Roadmap, ENCODE, brainspan and PsychENCODE projects.

We will use Matrix eQTL and/or fastQTL package for eQTL analysis by integrating all genotype and gene expression data from GTEx, roadmap, CommonMind, BrainSpan and PsychENCODE data. The gene expression matrix will be normalized according to gender, Age, RNA Integrity Number (RIN) and library preparation batch (LIB) for eQTL analysis (Fromer, Roussos et al. 2016). The genotypes from different projects will be imputed using the same reference panel and the same imputation pipeline. Probabilistic Estimation of Expression Residuals (PEER) factors, ancestry vectors, age and gender will be used as covariates input for Matrix eQTL/fastQTL. Based on our sample size, we will calculate both cis-eQTL and trans-eQTL. Finally we will correct for the multiple hypothesis tests of SNPs in LD for a given gene for eQTL analysis.

Aim 2

DAC

We will use the standardized RNA-seq processing pipelines including data organization, format conversion, and quality control metrics to process the RNA-Seq data first. Specifically, we will employ STAR (Dobin, Davis et al. 2013) to align the reads to their reference genome and RSEM (Li and Dewey 2011) to quantify expression profiles of each type of annotation entry retrieved from the latest release of the GENCODE project. Additional quality control measures will be introduced to assess potential issues including sequencing error rate, ribosomal contamination and DNA contamination.

We will use the ENCODE ChIP-seq data processing pipeline developed by both Gerstein lab and Zhiping Weng's lab. This pipeline includes steps of quality assessment, trimming the contamination, alignment of the fastq files, peak calling and downstream analysis such as peak comparison, peak annotation, motif analysis and super-enhancers identification. The Gerstein lab developed PeakSeq (Rozowsky, Euskirchen et al. 2009), a versatile tool for identification of TF binding sites and a standard peak calling program used by the ENCODE and modENCODE consortia for ChIP-Seq datasets (Rozowsky, Euskirchen et al.). We will also use a new peak caller MUSIC (Harmanci, Rozowsky et al.) recently developed in Gerstein lab. MUSIC performs multiscale decomposition of ChIP signals to enable simultaneous and accurate detection of enrichment at a range of narrow and broad peak breadths. This tool is particularly applicable to studies of histone modifications and previously uncharacterized transcription factors, both of which may display both broad and punctate regions of enrichment. We have already implemented this pipeline to process ChIP-Seq data from PsychENCODE.

We will implement a standard genotype imputation pipeline. Genotype imputation will enable us to evaluate the evidence for association at genetic markers that are not directly genotyped and increases the power of eQTL analysis. Moreover, genotype imputation is very important for combining data from studies using different genotyping platforms. Firstly, for the Sample level quality control, we will exam the call rate, heterozygosity and relatedness between genotyped individuals correspondence between sex chromosome genotypes and reported gender of the raw genotype calling using PLINK \cite{22138821}. Then we will perform ancestry analysis on the QCed genotype data to identify the ancestry vectors. In order to improve the genotype imputation accuracy, SHAPEIT2 \cite{22138821} will be used to estimate haplotypes from genotype data. The estimated haplotypes will be used as input for IMPUTE2 for imputation using the selected reference panel. We will also use both 1000 Genome phase 1 or the recently released HRC Reference Panel for imputation on Michigan Imputation Server. The imputed genotypes will be filtered according to imputation confidence score (INFO), minor allele frequency (MAF), SNP missing rate and Hardy-Weinberg Equilibrium (HWE).

DCC

Sage Bionetworks has demonstrated success in enabling broad distribution of data and collaborative analyses across diverse consortia through the use of centralized repositories, analytical tracking and provenance, communication forums and collaborative work across consortium members. These tools allows analyses and research outcomes to be shared across a distributed network of scientists working on a common set of data. It is our goal to continue to employ this approach to coordinate data release and analyses generated through the psychENCODE project grants. Within the scope of this project, Sage Bionetworks proposes to enhance the psychENCODE effort as follows: (1) to ensure that data and knowledge are quickly disseminated, (2) to facilitate project-wide analyses that leverage combined resources across partnering institutions to more effectively answer pressing scientific questions and and to maintain the online, publicly accessible record of data and the research performed, allowing others to freely use the generated knowledge for new purposes, (3) to identify long term storage solutions for the PsychENCODE data. The goal of this proposal is to continue these efforts through further development of the framework and infrastructure the psychENCODE project depends on.

Deliverables

Sage Bionetworks will continue to coordinate with the NIMH and the PsychENCODE investigators to provide the following deliverables in support of the project aims.

(1) Development of a collaborative space for centralized storage of data, protocols, analysis methods, and results generated by psychENCODE.

Sage Bionetworks has developed the Synapse software platform, an open computational platform for research collaboration, where integration, analysis, and publication of data-intensive science occur in real time as the research is performed. Synapse has been used to successfully support large-scale distributed team efforts such as the analysis of TCGA data by the Pan Cancer analysis consortium[1,2] and hosting of the DREAM challenges in computational biology[3]. For this project, we use the Synapse infrastructure to create a centralized repository for storage and sharing of PsychENCODE resources. We will continue the development of a psychENCODE project space within Synapse for data dissemination and analysis. Development of this resource initially focused on data directly generated through the PsychENCODE project but will be expanded to include ancillary data of value to downstream analysis.

(2) Continued development and implementation of a data release process for collection and verification of data from the data production centers.

Systemized processing for data release will enable rapid and facile data access. All data is made available through both web and programmatic interfaces with capabilities for fine-grain tracking of data updates through a versioning system and for association with metadata describing data sources and protocols.

(3) Identify long term storage and cloud computing solutions for psychENCODE data.
Synapse uses Amazon Web Services for data storage, a solution that has been used for data storage throughout the first years of this project. We are able to provide both distribution of the data to investigators to work on their own systems, and direct access to the data on the Amazon cloud as an option. We view the co-location of data and compute capability would provide significant benefits to researchers working the volumes of data expected in this project. We will explore options for long-term data storage and archival to include public cloud providers, academic cloud providers and local storage options. We anticipate that use of public cloud providers will be the safest, most stable, most cost effective option and will need to explore appropriate plans for funding of data storage in perpetuity beyond the time frame of these funding mechanisms.
All data transfer over the public Internet brokered by Synapse is encrypted and occurs over https. Data stored at Amazon through Synapse is maintained in a secure subnetwork within the Amazon cloud; Amazon infrastructure isolates Synapse from all other traffic within the cloud. Data access is only granted to those given permission to access resources by project admins, and we have mechanism to place access restrictions on certain data sets to ensure requests for access are reviewed and approved before data can be accessed. Synapse maintains an access log of all data download requests such that inadvertent data releases can be traced, and Sage's Access and Compliance Team conducts quarterly audits of the system to ensure appropriate data use. For this project, the NIH could be involved in or even solely responsible for managing this system.
Works Cited:

1. Omberg, L. *et al.* Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat. Genet.* **45,** 1121–6 (2013).
2. Chang, K. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–20 (2013).
3. Margolin, A. A. *et al.* Systematic Analysis of Challenge-Driven Improvements in Molecular Prognostic Models for Breast Cancer. *Sci. Transl. Med.* **5,** 181re1–181re1 (2013).