#####Peccolab Grant text#########

Aim 3 Data processing and bioinformatics analysis

Sub 3.1. Data processing and analysis to identify fetal QTLs

We will analyze all the generated data and integrate the genotype data with other data for QTL analysis. For doing this we will build on our considerable experience in ENCODE, modENCODE, 1000 Genomes, KBase, Brainspan and PsychENCODE in doing integrative and comparative analysis. We will use the standardized RNA-seq processing pipelines including data organization, format conversion, and quality assessment which will then be run in large-scale on the PDC (Protected Data Cloud) to process the RNA-Seq data first. Specifically, we will employ STAR \cite{ 23104886 } to uniquely align the filtered reads to their reference genome and RSEM \cite{21816040} to quantify expression profiles of each type of annotation entry retrieved from the latest release of the GENCODE project. Additional quality control measures will be introduced to assess potential issues including sequencing error rate, ribosomal contamination, DNA contamination and gene coverage uniformity and the correlation between technical and/or biological replicates.

ENC
+ ZW

We will use the ChIP-seq data processing pipeline developed in Gerstein lab to process data in ChIP-seq data. 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. In this pipeline, we added a new peak caller MUSIC \cite{25292436} 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 & Brainspan.

PEAK-
SEQ

Moreover, we developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers based on our past experience in non-coding annotation, as part of our 10-year history with the ENCODE and modENCODE projects. \cite{20126643}. We will use the better enhancer definition

EARLIER
WORK
ON
METATRACK

provided by the Epigenome Roadmap \cite{25693563,25533951,25693566}, and more recently from ENCODE projects. In particular, we will develop a new machine learning framework that combines pattern recognition within the signal of various epigenomic features and transcription of enhancer RNA (eRNA) with sequence-based features to predict active enhancers across different brain regions and other tissues in the Epigenome Roadmap project. The pattern within the signal of different epigenetic datasets will be computed from regulatory regions identified using different massively parallel assays and we will determine to what extent this pattern is conserved across a diverse set of species. This method will be used to predict fetal brain specific active enhancers based on H3K27ac ChIP-Seq datasets generated as part of this grant as well as ChIP-seq generated by the Epigenome Roadmap, ENCODE and PsychENCODE projects.

ADD ME

Moreover, we have developed an eQTL analysis pipeline developed in Gerstein lab based on our experiences on PsychENCODE capstone projects. We will use this pipeline to identify various QTLs including eQTLs for both long and short RNAs, cell type specific eQTLs, splicing QTLs,   ChIP-QTLs and ATAC-QTLs in early human brain development. ~~Specifically, the ge~~notype data will run through the imputation pipeline which is also developed in our lab. 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 use 1000 Genome phase 3 as the general reference panel for imputation. We will also try the recently released HRC Reference Panel for imputation of rare SNPs. The imputed genotypes will be filtered according to imputation confidence score (INFO), minor allele frequency (MAF), SNP missing rate and Hardy-Weinberg Equilibrium (HWE). We will use Matrix eQTL software package for eQTL analysis. The gene expression matrix will be normalized according to gender, Age, RNA Integrity Number (RIN) and library preparation batch (LIB) for eQTL analysis. Probabilistic Estimation of Expression Residuals (PEER) factors, ancestry vectors, age and gender will be used as Covariates input for Matrix eQTL. Based on our sample size, we will

TOO STRONG

calculate both cis-eQTL and trans-eQTL. Finally we will correct for the multiple hypothesis effect of many SNPs in LD for a given gene for eQTL analysis.

SplicingQTLs（reserved for andrew）

Sub 3.2. Early brain expression dynamics from differentially expressed genes

WE HAVE MUCH EXPR IN EXPR DYN& CLUST.

Our previously analysis revealed that the early brain samples have their own specific gene expression patterns. In particular, we ran tSNE clustering based on the similarity of gene expression of the tissue samples in brainspan and gtex projects and plotted them on first three tSNE directions (Fig XXX). The red and orange samples correspond to the early developmental samples (i.e., prenatal samples in brainspan), and forms a separated cluster. The blue and cyan samples correspond to the infant and adult brainspan samples, and also were clustered together. The samples from same gtex tissues also form their specific clusters. This suggests that the early developmental samples share specific gene expression patterns.

WE WILL DEV. A NEW METH WE CALL IT..

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

Sub 3.3. Dynamic modeling of brain developmental gene regulatory networks by integrating adult data from GTEx, PsychENCODE, etc.

2

WE HAVE EXP, IN ENC AS AWG LEAD WE'VE DONE X, 1, & 2

After finishing the fetal QTL analysis, we will perform integrative and comparative analysis of fetal gene and QTL and adult data from GTEx, CommonMind and PsychENCODE project. 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 fetal and adult stages and different tissues \cite{25249401}. 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. We will first normalize [FV1] and correct the batch effects of the gene expression data using COMBAT \cite{16632515 }. 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 and developmental stages. We will then cluster this network into gene co-expression modules and find modules (with associated gene expression signatures) enriched in fetal and adult 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 \cite{25091629}.

EARLIER AIM

We will use this grant combined with other projects like PsychENCODE, CommonMind, Brainspan and GTEX dataset to expand our understanding of the molecular activity of cells in the human brain by identifying genes that predominantly express one allele and exploring the potential clinical relevance of such allelic imbalance. We will focus on quantifying differences in transcription between maternal and paternal alleles using the matched genotype and RNA-sequence data available in the PsychENCODE dataset. We will integrate similar analyses of allelic imbalance performed by our lab using the matched genotype and RNA-seq data produced by the this grant, PsychENCODE and CommonMind project. We expect to be able to generalise results obtained from this grant using the 11 distinct cortical and 5 sub-cortical regions of the healthy adult human brain available in BrainSpan. Integrating data at this scale requires large amounts of RNA expression and matching genotype information from different cell-types, brain regions, developmental stages and/or tissues. To that end we will also incorporate data and results from the GTEx project in order to further broaden our survey of allelic

Y MORE IS BETTER CAT. OF ALLELIC REGIONS & SNPS imbalance to identify potentially brain-specific allelic effects. Once compiled, this allelic survey of unprecedented resolution will be of substantial benefit to the wider research community.

Ultimately, we will build a comparison of fetal QTL and adult QTL maps. We will use these data to compare the relative contribution of pre- and postnatal gene expression to signals in the GWAS, exome and WGS studies of the psychiatric disorders.

###Yale site description####

The Yale site will consist of investigators in the labs of Mark Gerstein at Yale University, Zhiping Weng at the University of Massachusetts Medical School and Daifeng Wang at Stony Brook University to form a data processing group (DAC?). Dr. Gerstein's lab will develop a number of standardized pipelines and quality control metrics,provide a platform and infrastructure for uniform processing of the data and running the pipelines and focus on the discovery of fetal brain specific genes, the aggregated quantitative trait locus (QTL) analysis, and integration of all data sets with whole genome sequencing data and genotyping data. Dr. Weng's lab will support the enhancer analysis and also the construction of a number of the pipelines. Dr. Wang's lab will work on the co-expression networks across developmental stages, brain regions and different tissues. The data processing group will report to and take directions from the analysis working group, provide feedback for data quality to data generation group and support the group's activities in integrative analysis.

MORE

###budget justification#####

Mark Gerstein, Ph.D. is the Albert Williams Professor of Biomedical Informatics. His lab (http://gersteinlab.org) was one of the first to perform integrated data mining on functional genomics data and to do genome-wide surveys. His tools for analyzing motions and packing are widely used. Most recently, he has designed and developed a wide array of databases and computational tools to mine genome data in humans, as well as in many other organisms. He has worked extensively in the 1000 genomes project in the SV and FIG groups. He also worked in the ENCODE pilot project and currently works extensively in the ENCODE and modENCODE production projects. He is also a co-PI in DOE KBase and the leader of the Data Analysis Center for the NIH exRNA consortium. In these roles Dr. Gerstein has designed and developed a wide array of databases and computational tools to mine genomic data in humans as well as in many other organisms.He will be directly involved in all levels of sequencing data analysis and integration of different data modalities. HE WILL DIRECT

Shuang Liu, Postdoctoral Research Associate (XXX calendar months), has a strong background in scientific computation, biomedical data analysis and image processing. Dr. Liu obtained her Ph.D. in in Biomedical Engineering from The University of Texas at Austin. The University of Texas at Austin. Before join Gerstein lab, she was a postdoctoral research associate working on brain MRI data analysis in Neurology at Yale. She has experience on biomedical data and image processing for 10 years. Dr. Liu is very interested in integrative analysis of neuroimaging and neurogenomics data. In Gerstein lab, she participates in Brainspan and PsychENCODE projects. She developed pipelines for ChIP-seq data processing and evaluated association between GWAS SNPs and brain enhancers. She is also working on brain eQTLs analysis. She will work on the genotype data imputation and eQTL analysis in this project.

Gamze Gürsoy , Postdoctoral Research Associate (XXX calendar months) received her PhD in Bioinformatics from University of Illinois at Chicago (UIC) in 2016. She developed softwares for construction of three-dimensional models of genome using various Chromosome Conformation Capture and electron microscopy data from different eukaryotes. She was also involved in studies related to stochasticity in biochemical reaction networks. She will work on using the standard pipeline to processing RNA-seq data in this project.

Timur R. Galeev, Ph. D., Postdoctoral Associate (XXX calendar months). Dr. Galeev has a strong expertise in scientific computation. Before joining the Gerstein lab at Yale University, he obtained his Ph.D. degree (2014) from Utah State University. His Ph.D. research was in the field of theoretical and computational physical chemistry, focused both on applications of modern electronic structure methods and development of new theoretical tools and models of molecular structure and bonding. He is currently working on analysis of functional genomics data. Dr. Galeev will work on allelic expression analysis.

Dr. Daifeng Wang is a tenure-track Assistant Professor in the Department of Biomedical Informatics at Stony Brook University. He has ~10 years of research experience on developing specialized computational and bioinformatics approaches to analyze next generation sequencing datasets and systematically understand gene expression dynamics, gene regulatory networks and circuits in complex biological processes. As a doctoral candidate at The University of Texas at Austin, he developed novel computational methods to identify principal expression dynamic patterns from highly dimensional gene expression datasets, and won the Graduate Student Professional Development Award. He joined Dr. Mark Gerstein's lab as a postdoctoral associate in the Department of Molecular Biophysics and Biochemistry at Yale University in January 2012 and was promoted to be associate research scientist in 2015. He was a key participant in the data analysis centers (DACs) for large scientific consortia including ENCyclopedia Of DNA Elements (ENCODE), modENCODE, PsychENCODE and DOE Systems Biology Knowledgebase (KBase). In these projects, he led teams and developed bioinformatics tools to uniformly process RNA-seq and ChIP-seq datasets, and designed novel computational approaches in a multi-scale modeling framework to study temporal gene expression dynamics and gene regulation for model organisms, cancer and human brain development.

HZ WILL

ZW 2

####Call minutes on Jan 8 2017###

title: multi omics compl. to psychencode?

Yale-suny sb/umass as subs for Data analysis
SUNY-USC as sub/mssm?
Liberer
UNC
RICK

11 pages the same for everyone
same title with 1/5,2/5….
1 page for job assignment at each site
3 pages for our aim (0.5 page for splicing qtls?)
send to mark draft by tue mid day,

have a gdoc link
1 page spec aim
11 page
1 site page
budget justification
google doc
we will send Jim everything by Friday

Jim will work on:
title
overall budget
wait for jim's email about title, # of pages, overall
Jim send budget tonight or tmr morning

Have a call with Jim next Friday 2:00pm ET

#####Jim's original draft###

Specific Aims

Our overall goal is to delineate the ……

As part of this larger goal, we propose:

Aim 1. Collect a large sample (n=500-1,000) of cortical brain tissue from 10-24 weeks post-conception, and use these tissues for a number of molecular assays.

a) Genotyping of all samples with the Illumina Global Screening Array (GSA), which will contain a backbone of ~660,000 SNPs, which provides LD coverage and imputation accuracy of >0.8, for over 87% of the XXX genome.
b) Perform bulk long RNA-seq (strand-specific ncRNA and mRNA >100bp) and small RNA-seq (piRNA and miRNA) of all samples.
c) Perform single-cell RNA-Seq on XXX samples.
d) Perform ChiP-Seq of the chromatin marks CTCF, H3K4me3, and H3K27Ac, and of a panel of 100 transcription factors, on XXX samples.
e) Perform ATAC-Seq on XXX samples.
f) EPIC arrays for CH3? Targeted assay?
g) RNA Binding Proteins?

Aim 2. Create a reference cortical map of bulk gene expression, single cell gene expression and ChiP-seq data from 10 regions dissected from five intact fetal brains (10-24 weeks post-conception).
from lieber

Aim 3. Place the data from Aim 1 into the cortical framework determined in Aim 2 and analyze these data to produce:
a) eQTLs for both long and short RNAs
b) Cell type specific eQTLs
c) SpliceQTLs (sQTL)
d) ChIP-QTLs
e) ATAC-QTLs
f) mQTLs (if we do EPIC arrays)

Aim 4. Compare these QTL maps to those of adult brain using the data from the PsychENCODE, CommonMind, NIH Roadmap and ENCODE projects. (based on CAPSTONE4)
a) Use these data to compare the relative contribution of pre- and postnatal gene expression to signals in the GWAS, exome and WGS studies of the psychiatric disorder.
b) Use these data to weight the transcripts in each GWAS peak and build conditional expression network models

Aim 5. Provide an easy-to-use, web-based informatics framework for communication of the raw and computed data of this PsychENCODE project to other neuroscientists (Synapse?).