SPECIFIC AIMS

We propose to establish the **Center for Genomic Excellence in Human Brain Development and Evolution**. This center would be unique among current NHGRI Centers by focusing on the human brain. Our objective is to develop new approaches and methods for generating and analyzing multi-dimensional genomic data for the developing brain of human and non-human primates (NHP; chimpanzee and macaque). In so doing, we expect to achieve comprehensive and conceptual advances in our understanding of genomic processes underlying human brain development in health and disease, and thereby further our understanding of what makes us uniquely human.

We will directly address several critical issues in genomics and neuroscience: the need (1) to improve single cell transcriptomics and the study of cellular heterogeneity, (2) to improve the annotation and functional characterization of human and NHP non-coding elements, (3) to identify and functionally characterize human-specific features of development, and (4) to identify regulatory mutations and elucidate molecular networks compromised in complex diseases. To achieve these goals we have assembled a group of scientists from diverse fields that share interests in 'omics' technologies and/or neuroscience, and that have a strong history of scientific interactions and involvement in large-scale projects. For the first iteration of the Center, we propose the following 5 aims.

Aim 1: Multi-dimensional genomic analyses of human and NHP brain single cells and tissues. We will apply existing and develop new approaches to generate and integrate multi-dimensional omics data (genome, chromatin modifications, transcriptome, post-transcriptional processes, and proteome) on the level of tissues (**subaim 1.1**), single cells (**subaim 1.2**), and cell types (**subaim 1.3**). To achieve this, we will employ a novel tissue processing protocol we named HC (Hibernation-Cryopreservation), which keeps a human or NHP brain in prolonged hibernation to stabilize nucleic acids and proteins, and allow for concurrent collection and subsequent cryopreservation of (1) live cells, (2) single cell suspensions, and (3) tissue samples from (4) hundreds of regions. This procedure will be complemented by the generation and neural differentiation of human and NHP induced pluripotent stem cells (iPSC) to create enriched populations of two major cell types of the cerebral cortex: neural progenitors and projection neurons.

Aim 2: Integrated analyses of multi-dimensional single cell and tissue-level genomics data. Since no one method is sufficient to fully capture biological complexity, we will (**subaim 2.1**) develop multi-level computational strategies to de-convolute identities and molecular dynamics of single cells, as well as (**subaim 2.2**) deeply analyze and integrate multi-dimensional genomics data across multiple variables.

Aim 3: Elucidation of common and cell type specific regulatory and molecular networks compromised in autism spectrum disorders. We will perform targeted re-sequencing using customized sequence capture probes to screen for mutations in *cis*-regulatory loci in autism spectrum disorders (ASD) quartets (an affected child and his/her unaffected sibling and parents) that do not carry either a *de novo* CNV or *de novo* loss of function mutations (**subaim 3.1**). Target human cis-regulatory loci identified in Aim 1 will be prioritized based on their association with previously implicated as ASD risk genes, spatio-temporal pattern of activity and human specificity. We will characterize the mutation burden of *de novo* or transmitted mutations (**subaim 3.2**) and leverage this data with (**subaim 3.3**) computational network-oriented methods to elucidate associated networks disrupted in ASD. Elements carrying recurrent mutations or rare mutations mapping contiguous with previously implicated genes will be prioritized for functional characterization under Aim 4.

Aim 4: Modeling and functional characterization of human-specific and ASD-associated *cis*-regulatory elements. We will functionally characterize those non-coding elements that are human-specific (**subaim 4.1**) or mutated in developmental disorders (**subaim 4.2**). To elucidate their functional role(s) and model consequences of human-specific and ASD-associated changes on neurodevelopment, we will employ human bacterial artificial chromosome (BAC)-based mouse transgenesis. These experiments will be complemented with in-depth analyses of molecular circuits and cellular processes to reveal the transcriptional regulatory networks mechanisms and functional insights into human- and ASD-specific aspects of brain development.

Aim 5: Creation of an infrastructure to facilitate research and teaching. We will create an infrastructure to facilitate internal and external research activities by disseminating our bio-specimens (i.e., tissues, single cell preparations, iPSC), tools, methods, protocols, and data to the community as well as by soliciting outside ideas through a public web portal (www.neuroCEGS.org) (**subaim 5.1**). We will also establish outreach programs and integrate training in basic and translational neurogenomics (**subaim 5.2**) in all Center activities to advance the understanding and employment of genomics approaches for the study of biomedical problems and human brain development and evolution.

RESEARCH STRATEGY

SIGNIFICANCE

Overview and rationale: The overall objective of this project is to establish a Center with a multi-disciplinary group of investigators that will develop upon several cutting-edge genomic approaches in a unique and innovative way to elucidate molecular networks underlying human brain development and evolution. This will be achieved through the generation and exploration of integrated multi-dimensional genomic scale data from single cells and tissues of developing and adult human and non-human primate (NHP; chimpanzee and macaque) brains. We will use these new sources of information to facilitate the identification of regulatory mutations and to elucidate shared and cell type specific molecular networks compromised in autism spectrum disorders (ASD). Finally we will implement approaches to model and functionally characterize human-specific and ASD-associated regulatory mutations in the context of mouse brain development.

The human brain is arguably our most complex organ, organized into highly functionally distinct regions. Even within each of its functionally distinct regions, there is usually remarkable cellular heterogeneity. This complexity along with the rare availability of human and some NHP brain tissues make the application of genomic technologies particularly challenging. Overcoming these challenges requires synergistic integration of expertise across fields (e.g., genomics and other high-throughput analyses, computational biology, biostatistics, genetics of brain disorders, human developmental neurobiology, evolutionary biology, mouse genetics, stem cell biology, and genomics database creation), as well as the development of new approaches to make rare tissues accessible to wider community. To make human and NHP brains broadly accessible to other genomics experts, we will develop novel approaches and create a unique specimen bio-resource. Our proposed organizational structure combines the expertise of each individual key investigator and establishes a CEGS that is much more than the sum of its parts. Through the integration of multi-dimensional genomic approaches, such as single cell transcriptomics, phased genomes, HITS-CLIP, unbiased cell-type clustering, network based analyses, and integrated “personalized” computational analyses, we will seek to understand our most complex and compelling biological subject: the human brain.

The biological problem and its challenging aspects: Understanding the molecular processes involved in the development and functional organization of biological systems, as well as their alterations in disease states, requires precise measurements of nucleic acid- and protein-level across different cell types and developmental time points. This is particularly difficult in the human brain due to its cellular complexity and limited availability of tissue. As a result, studies of human and NHP brains have been largely limited to anatomical and imaging studies. On the other hand, the use of genetically tractable model organisms has greatly enriched our understanding of brain physiology and development. However, these studies are limited by our evolutionary distance from these organisms, and by the lack of contextual and functional interpretations of disease-associated variations in the human genome, especially those within non-coding regions. For example, null mutations in human and mouse orthologs frequently result in different phenotypes or even the absence of neural phenotype in mouse[1,2]. Thus, deciphering both conserved and human-specific mechanisms involved in brain development is of crucial importance to the understanding human evolutionary history and predisposition to certain diseases[3]. Furthermore, as we advance our knowledge of the differences and the similarities in mammalian brain development, we will be better positioned to perform informative and relevant experiments in model organisms. Genomic analyses provide a unique opportunity to systematically study complex and rare tissues, such as the human and NHP brain. However, previous genomic studies of the human brain have mainly been performed using microarrays, which are limited in sensitivity and inability to detect many important transcriptional and post-transcriptional events. In the proposed CEGS, we will leverage a number of well-focused cutting-edge genomic methods and develop new approaches to address critical problems in human neuroscience. The Center’s contribution to genomic innovations will be broad and the insights gained from proposed activities will reach well beyond genomics and neuroscience at several levels, as described below.

Significance for genomic sciences:

The need for better understanding of genomic processes in complex biological systems. Comparative analyses of genomes from different species have revealed that humans and other mammals have similar coding complexity when it comes to the number of protein-coding genes. These similarities imply that some additional reservoir of transcriptional regulation exists that explains species differences, especially the remarkable cellular

diversity and development of the human brain. Recent advances in next-generation sequencing technology make substantial insight into this complexity possible[4]. In particular, the advent of RNA-sequencing (RNA-seq) has enabled global and unbiased transcriptional profiling, including of rare and non-coding transcripts, in any organism with many fewer prior assumptions than preceding methods[5-7]. In addition, chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) provides unprecedented insights into the complexity of *cis*-regulatory regions and chromatin modifications by generating high-resolution profiles of transcription factors (TFs) binding sites and histone modifications enrichment[8-11].

The technical capability of mass-spectrometry for large-scale proteome analyses has also significantly advanced in recent years[12,13], to the point where a small proteome such as one in yeast can be completely quantified in just a few hours[14]. Complete quantification of the human proteome is still out of reach, but the technical advances have improved sensitivity and specificity to the point where we can confidently detect more than 10,000 expressed proteins[15] (our unpublished data; see Preliminary data). Moreover, as we will demonstrate below, this will become a strong analysis tool for correlating and interpreting data from other genomic techniques, such as RNA-seq.

Production of genome-wide maps of chromatin modifications, expressed transcripts and proteins represents a powerful approach for surveying the diversity of regulatory elements, RNA species and proteomes. Indeed recent findings from the ENCODE project[16-18], show that 60% of the human genome is transcribed in a variety of cell lines, with the majority of the transcripts belonging to non-coding RNAs, suggesting a remarkable genomic complexity. Moreover, studies have also uncovered pervasive involvement of *cis*-regulatory DNA variants in common human diseases[18-20]. How these findings regarding coding and non-coding elements in human cell lines relate to the complexity of developing human tissues is still elusive and is a critical barrier to surmount in genomic sciences. While there have been thousands of separately generated and analyzed regulatory, transcriptional, and post-transcriptional datasets, only a handful of studies have generated and integrated multiple data modalities. The generation of such multi-dimensional datasets is still challenging in cell lines and genetically tractable model organisms, let alone the human brain. However, the integration of multi-dimensional genomic-scale datasets into a unified and functionally meaningful model is essential for a predictive interrogation of the genome-to-phenome relationships in the context of normal and abnormal development, and evolutionary specializations. One of the great utilities of such network models is that they allow new insights into cellular and developmental dynamics, species differences, and diseases states, as we will demonstrate here for humans and NHPs.

Genomic variation across human tissues and cells is increasingly evident[21-26]. Besides single nucleotide variation (SNV), dividing cells accumulate larger somatic structural variants (SVs). These include copy number variations (CNVs, i.e., duplications and deletions), as well as inversions and translocations, each affecting hundreds to millions of nucleotides. Both the frequency of somatic genomic variations and their impact on RNA expression and cellular function are poorly understood, at least in part due to the lack of carefully matched and comprehensive tissue and cell-level genomics data.

Taken together, the challenges outlined above illustrate the need for generating multi-dimensional genomics data, for improving genomic techniques and analyses, and for making complex biological systems, such as the human brain, and complex disorders, such as ASD, more accessible to cutting-edge genomics methods and analyses. We expect our proposed integrative collection of multi-level genomics data in matched biological samples across organisms and disease states, including novel integrated data analysis and construction of regulatory network abstractions, will largely surmount these challenges (see innovation section).

The need to improve single cell genomics: The ability to accurately measure gene expression changes in a small number of phenotypically similar cells or single cells is of great importance for resolving a variety of problems in many biological disciplines, especially in neuroscience. As previously reported[27], gene expression may be regulated in opposing directions in different cell types, thereby appearing static in composite data. Despite tremendous advances in genomic technologies, scaling to the level of single-cells has been very challenging. All existing single cell RNA-seq methods require significant amplification of starting material and all available protocols are in some way limited[28-35]. These limitations force one to choose between prioritization of throughput or coverage of single cell transcriptomics. Furthermore, non-coding RNAs are largely under-represented with current single-cell methods. Finally, collecting high quality and well-defined single cells from the tremendously complex brain tissue is a considerable challenge. This is a particularly the case when working with rare human post-mortem brain tissue, as we will demonstrate herein for humans and NHPs.

The need for better functional assays to study non-coding variations: Studies in evolutionary and developmental biology have shown that some phenotypic variations among species are driven by the evolution of gene regulatory circuits, including changes in TFs and cis-regulatory elements. The same regulatory circuits are also affected in many human diseases, but the relationship between regulatory mutations and changes in gene expression and the subsequent functional implications remain poorly understood. The current limitations in functional characterization of genomic sequence variations in regulatory elements can be traced to a couple of major obstacles that will be substantially removed by our proposed analyses of multi-dimensional genomics data and the application of a novel approach by the CEGS (see Innovation section).

Significance for neuroscience:
The complexity of the human brain is reflected in ~86 billion neurons and at least an equal number of glial cells[36], which can be further subdivided into hundreds of different types based on their morphology, connectivity, and molecular and electrophysiological properties. Furthermore, all these different cell types have to be generated and integrated into functional networks within very precise constraints; deviations from this normal course of development can lead to a variety of disorders. Single cell genomics present the most tangible way to resolve this cellular heterogeneity of human and NHP brains.

Significance for evolutionary biology:
What makes us human? Scientists have struggled for centuries to pinpoint the qualities that separate human beings from other primates. Primates share many characteristics; however human-specific specializations seem to lie in the realm of cognition and behavior. Most prominently these include language, propensity for social learning, and the ability to make and use complex tools. Understating the genetic basis of human brain specializations is perhaps one of the most challenging tasks of modern science. Comparative sequence analyses have led to genome-to-phenome findings related to inter-NHP divergence, including human-specific genetic gains and losses[37-39]. However, there are a vast number of genomic sequences for humans and NHPs that are difficult to link directly to the emergence of specific phenotypic traits. As mentioned before, comparative and developmental multi-dimensional genomics studies provide excellent tools to bridge the gap between genetic changes and phenotypic evolution in humans. Moreover, understanding human specializations help us evaluate and improve animal models.

The need for genomic studies of non-human primates: The publication of several NHP genomes (notably the chimpanzee, gorilla, baboon, orangutan, and macaque), has greatly improved our ability to make evolutionary and functional inferences from these species. NHPs are critical biomedical models for many aspects of human physiology and disease states. Yet despite the available genomic data, the genetic basis of phenotypic variations in NHPs remains poorly understood. Unfortunately, this is partly due the poor current annotation quality of chimpanzee and macaque genomes compared to that of human. The full potential of NHPs can only be realized with integrative multi-dimensional genomics data that captures the subtle differences among primates, a requirement that is equally critical for understanding human evolution. Most notably, comparative genomics studies strongly suggest that the main differences between modern humans and chimpanzees are likely due to changes in gene regulation, rather than direct modification of protein coding genes. This hypothesis proposed by King and Wilson in 1975[40], was reinforced by results from the ENCODE project that suggest functional/regulatory roles of much of the genome that does not code for proteins. However, ENCODE-like multi-dimensional genomic data do not exist for NHPs and those for the developing human brain are still in their infancy and are not generated from the same tissue samples[21,41-48]. We will improve the annotations of the NHP genomes by generating a comprehensive and multi-dimensional genomic data in chimpanzee and macaque.

Despite enormous interest in human evolution and NHP genomic studies and the recent availability of a wide range of techniques, little progress has been made in this area. The number of researchers actively engaged in applying genetic, developmental, and comparative approaches to characterize and understand the uniqueness of the human brain is insufficient considering the relative importance of this area. A major impediment is that both humans and chimpanzee are not amenable to experimental studies. A government ban on breeding of great apes for research, and the suspension of all new grants for biomedical and behavioral research on living chimpanzees[49], have pushed the field to a critical point. These limiting factors have led to the scarcity of formative or substantive research on the neural basis of human uniqueness and NHP divergence. Essentially, this may be our last chance to properly undertake these studies by using post-mortem tissue exclusively from

chimpanzees that died from natural causes. Thus, the proposed CEGS will provide a strong platform for building additional experimental programs to investigate the neural basis of human uniqueness and NHP divergence.

The need for better model-based approaches to study human-specific genomic changes: The gradual accumulation of variations in sequences that control gene expression is a major determinant of inter-individual and inter-species differences[20,50-54]. This process has been well documented for a few tractable model organisms[51], but is just beginning to be explored in humans and NHPs. Therefore, better annotation and discovery of functional non-coding elements in developing human and NHP brains, will improve our ability to interpret variations in *cis*-regulatory regions and developmental functional non-coding elements, and is crucial to our understanding of evolution. Human and chimpanzee development is not amenable to direct study using the techniques from experimental developmental biology. Human evolutionary developmental biology must thus be based on inference from work on model organisms combined with human genetics and comparative anatomy. The 'middle-out' approach, which emphasizes the relationship between developmental genomic processes and their phenotypic outcomes, is ideal for this type of inference. These problems can be overcome by making human and NHP brain more accessible to advanced genomics techniques, profiling at a higher spatial resolution (more regions and cell types), studying genomic processes during development when neural circuits are being formed, and modeling of recent human-specific genomic changes in model organisms such as the mouse.

Significance for medicine: The advent of whole exome sequencing (WES) has enabled rapid identification of causative mutations in protein-coding regions in many Mendelian disorders[55-59]. However, WES has been much less effective for complex disorders. Moreover, many disease associated mutations fall within non-coding regions. Thus, novel systems biology approaches are needed to narrow down candidate disease genes and facilitate the functional interpretation of non-coding mutations by integrating genetic findings with multi-dimensional genomic data in the same biological system.

INNOVATION

Innovation for genomic sciences: *Technology development with a focus on single cell transcriptomics*. The ability to determine gene expression pattern of a small number of cells or single cells is of great importance for resolving a variety of problems in many biological disciplines, as we will demonstrate in humans and NHPs. Ideally, a comprehensive single cell transcriptional profile requires determination of the full-length species of all expressed mRNAs. However, existing methods do not have the sensitivity to detect all expressed genes and have either a 3' or 5'-biased or variable transcript representation. Non-coding RNAs are not at all well represented with current methods. The small starting amounts of RNA and difficulties in collecting cells acutely isolated from a great number of tissue samples present considerable limitations.

A further critical issue is how to apply and improve single genomic techniques for rare tissues, such as human and chimpanzee brains, that are often in undermined conditions when obtained. To overcome these problems, we will implement several measures, including improvements (1) in cell and RNA preservation using a novel protocol we named Hibernation and Cryopreservation (HC); (2) in RNA amplification and sequencing length, and (3) in the development of agnostic and integrative computational methods.

Generation and network-based exploration of multi-dimensional genomics data. While there are thousands of separately generated and analyzed regulatory, transcriptional, post-transcriptional, and protein-protein interaction networks, integration of such data for a comprehensive analysis is still challenging even in cell lines and genetically tractable model organisms. This CEGS will provide needed resource and a unique opportunity to generate a multi-dimensional dataset to study human and NHP brains. For example, our collection of macaque and human parental DNA samples provides an unprecedented opportunity to identify parent-of-origin allelic biases and genomic loci undergoing genomic imprinting. Furthermore, we have access to brains of fetal monozygotic human twins and parental DNA, which provide a unique opportunity to study these processes in monozygotic twins. In addition, we will develop computational methods to generate a unified and functionally meaningful network model from this multi-dimensional genomics data, thus to predict the genome-to-phenome relationships in the context of normal development, disease states, and evolutionary specializations.

Facilitating genomic studies of human and NHPs: This proposal goes beyond traditional model systems by studying genomic processes in human, chimpanzee, and macaque neurodevelopment. Examination of closely

related NHPs (particularly the chimpanzee) facilitates meaningful evolutionary comparisons, allowing for investigation of genomic and phenotypic differences. Macaque serves two important roles in the comparison. First, it is the out-group species that allows sequence divergence to be put into an evolutionary context (e.g., the identification of changes as human-specific). Second, it enables comprehensive coverage of neurodevelopment in NHP to an extent that will never be possible with chimpanzee due to ethical and legal restrictions. However, such analyses are rare and unique, especially when it comes to developing or implementing cutting-edge genomics methods.

Improvements in functional interpretation of cis-regulatory sequence variations: One of the main obstacles is our inability to prioritize from hundreds of possible TFs that bind to regulatory elements in a specific cell type, at a particular time point, or under different physiological conditions. Currently the most reliable method for detecting these TFs is to perform ChIP-seq using a well-characterized antibody for a specific TF in a cell type, at a known developmental time point, and physiological condition of interest. While this approach has revealed many functionally important trans-interactions, it cannot easily be scaled to accommodate thousands of TFs in our proteome, and myriad developmental and physiological conditions. We will overcome this by applying a co-expression network approach based on our spatially and temporally rich multi-dimensional dataset, to rank TFs by their correlation to the expression pattern of genes near the putative regulatory elements. The validity of this approach will be tested in transgenic mice expressing human BACs (see below). This approach can also be adapted to other genomic applications.

A second major obstacle to the functional characterization of sequence variations in genomic regulatory elements is our inability to reliably interrogate their normal functionality. Current methods mainly rely on luciferase-based assays in cultured cells or transgenic reporter assays of a small putative regulatory region, which may be too small to show proper regulation and thus may lead to spurious gene expression that bears no relation to the function of these regions *in vivo*[60]. Moreover, the standard transgenic approach does not allow for testing the whether the region is necessary to endogenous transcription, and can be complicated by positional effects or shadow enhancers[61]. To overcome this limitation we have developed an approach that will harness human bacterial artificial chromosomes (BACs) to model human expression patterns in transgenic mice and then systematically mutate putative human-specific regulatory elements or generate disease associated variants to examine the consequence on transcription. This approach will allow us to identify *cis*-regulatory sequence variations required for normal or disease-related human-specific expression and to decipher possible functional consequences of mis-expression of the human protein in the context of whole body development.

Improvements in cross species comparisons and primate genome annotations: We will improve the annotations of the NHP genomes and cross species comparisons by generating a comprehensive and multi-dimensional genomics dataset in humans and NHPs.

Innovation for neuroscience: *Genomic analyses of cellular heterogeneity directly in human and NHP brains*. Cellular heterogeneity in the human brain is a critical issue in neuroscience. It is still difficult to isolate specific cell types from the human brain using methods like fluorescence-activated cell sorting (FACS) on human cells, and cell cultures do not accurately represent tissue complexity. The necessity of using post-mortem tissue, uncertainties in when tissue will be available and the circumstances of the tissue-donors' death can dramatically affect the stability of nucleic acids and proteins, further challenging the field. Despite the fact that single-cell genomics is an active field, existing technologies have only been applied to a small number of studies on the mammalian brain, including one recent study of somatic mutations in human fetal neurons[21]. To make post-mortem human brain tissue more amenable to advanced genomics methods, the Sestan lab has developed a number of innovative approaches to prolong post-mortem cell viability and the stability of nucleic acids and proteins using a novel HC (Hibernation and Cryopreservation) protocol, and applied it to single cell transcriptomics.

Using multi-dimensional genomic data to gain insights into neurobiological processes. Multi-dimensional genomic analyses will provide a unique opportunity to systematically study biological processes in the context of human brain development and evolution, by analyzing the expression patterns of genes associated with specific biological processes. For example, one can query for transcriptional signatures of recent neural activity by

profiling different cells, tissue, and time points for expression patterns of activity-dependent genes and correlate these with other co-expressed genes, epigenetic signatures, and cis-regulatory loci.

Innovation for evolutionary biology: *Creation of a resource of a unique human and NHP bio-specimens and genomics data*. Currently, there are around 1,884 chimpanzees kept in captivity in the US, with 864 in biomedical laboratories (chimpcare.org). As elaborated earlier, this might be one of the last opportunities to perform comprehensive multi-dimensional analyses, genomic or otherwise, between humans and chimpanzees. This project will create a unique resource consisting of high quality brain specimens [tissues, single cells, induced pluripotent stem cells (iPSCs)] of humans, chimpanzees, and macaques. Our collection currently includes, to the best of our knowledge, the only two fresh frozen fetal chimpanzee brains (see Resources for details).

Improving systems approaches to evolutionary biology. We will generate multi-dimensional data in NHP and develop computational tools to improve cross-species mapping and analyses of these data.

Innovation for medicine: The CEGS will improve our understanding of a gene-driven systems biological approach to complex disorders, as we will demonstrate here for ASD. We will complement ongoing WES efforts with targeted mutational screening of putative cis-regulatory elements and improve analytic methods to construct and integrate multi-dimensional data in order to elucidate common and cell type specific regulatory/molecular networks compromised in ASD. These will greatly enhance our ability to gain insight into disease etiologies from multi-dimensional genomics data. The principles developed here can obviously be applied to other datasets and complex disorders.

APPROACH

Overview: Our overall goal in this CEGS to elucidate genomic processes underlying human brain development, evolution, and dysfunction. Our projects will complement other ongoing projects studying human genomic variations (1000 Genomes) and functions (ENCODE) by generating and analyzing multi-dimensional genomic data in the human brain. While there have been other genomic based projects (Allen Brain Atlas, BrainSpan) and studies[41-44] of the human brain, this CEGS would be unusual for (1) using complementary multifaceted genomic approaches on the same brain samples and single cells from both humans and NHPs (chimpanzee and macaque), and (2) using this information to identify *cis*-regulatory mutations in ASD, and (3) modeling human-specific and ASD-associated regulatory changes. In doing this, we will also (1) improve single cell transcriptomics and the study of cellular heterogeneity, (2) improve the annotation and characterization of human and NHP functional genomic elements, (3) identify and functionally characterize human-specific features of development, and (4) advance the understanding of common and cell type specific regulatory and molecular networks compromised in ASD. In support of our approach, we first provide basic background and general preliminary data pertaining to our ongoing efforts and expertise, and subsequently integrate additional preliminary data specific to each aim.

BACKGROUND AND PRELIMINARY DATA

Spatio-temporal dynamics of human brain transcriptome. The development of the human brain is an extraordinarily complex process, which is likely reflected in the complexity of the underlying transcriptome. To address this problem, the Sestan lab has curated an extensive collection of post-mortem human brains across development and adulthood (see the Resource section for details), and in collaboration with the Gerstein and State labs, has applied different genome-wide platforms to characterize the transcriptome. Using exon arrays, we have analyzed[42-43] gene expression and alternative exon usage in 16 brain regions comprising 1,340 samples from male and female brains spanning the entire course of human brain development and adulthood. We determined that the vast majority of genes are differentially regulated across regions and/or time, with the bulk of the differences found in prenatal development. Also, Weighted Gene Co-Expression Network Analysis[62] and other analyses revealed that the developing transcriptome is organized in distinct co-expression networks and shows robust sex-biased gene expression and splicing.

In addition, we have also participated in the BrainSpan consortium (www.brainspan.org) by performing RNA-seq on the same human samples. This public resource generated over 9 billion uniquely mapped mRNA and small RNA reads (manuscripts under preparation). The analysis of the dataset has revealed a tremendous complexity of the human brain transcriptome, by identifying a substantial number of novel protein-coding and non-coding transcripts, exons, transcriptionally active regions, and splicing and RNA-editing events. However, these findings also highlighted three major problems; (1) the need to profile transcriptional dynamics at cell type or single cell level, (2) the current limitations in extracting much of the information that would be possible if a complementary multi-dimensional genomic dataset were available, and (3) the lack of such comprehensive datasets in closely related NHPs. Motivated to solve these problems, we have also made substantial effort in developing new approaches to expand the use of genomic analyses in the human and NHP brain. Some of these technical improvements, such as the novel protocol for preserving human tissue and application of single cell transcriptome, are described in the specific aims.

Insights into primate and human evolutionary biology from the brain transcriptome. We have also recently performed comparative mRNA and small RNA-seq on the homologous regions of the adult chimpanzee and macaque brains. Briefly, these preliminary analyses (manuscript under preparation) uncover a new set of novel transcripts and remarkable inter-primate divergence and human-specific expression patterns. Just within the 16 adult regions, we have identified over 3,000 coding and non-coding mRNA transcripts that have human-specific expression pattern in one or more regions (Figures 1 and 2). These preliminary data illustrate the richness of human-specific transcriptional differences and provide a framework for human and NHP analyses proposed in this application. Our proposed research will tackle major remaining issues: (1) What is the extent of human-specific (also inter-primate differences) across development, cell types and single cells? and (2) What are the transcriptional and post-transcriptional regulatory mechanisms driving these species differences?

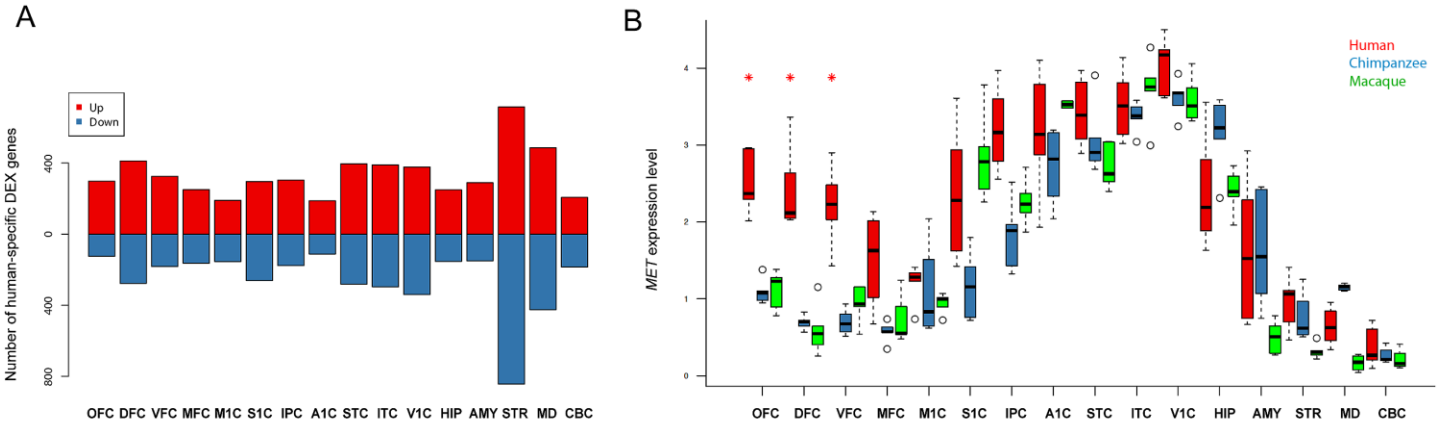

Figure 1. Inter-primate differences in the expression of coding and long non-coding RNAs. ***A)*** *Number of genes that have up- (h>p=m; red) or down-regulated (h<p=m; blue) expression selectively in at least one of 16 regions in the human brain.* ***B)*** *Expression [log_2 (RPKM+1)] pattern of MET, a gene previously linked to ASD (28), which has significantly differential expression in human prefrontal cortex. Significant differences are labeled with an asterisk. Human (h) is colored red, chimpanzee (p) in blue, and macaque (m) in green.*

Insights into complex disorders from the human brain transcriptome. Another realization that has emerged from the above studies is that even spatially and temporally diverse transcriptome data can be used to generate new insights into the etiology of complex disorders. For example, the Gunel lab, together with the State and Sestan labs, has recently determined that null mutations of *LAMC3*, which is expressed in the human, but not mouse, developing cortex, cause severe malformations of cortical development in humans but not mice[1]. Another study[63] uncovered that human, but not mouse, *NOS1* mRNA is regulated by FMRP, an RNA-binding protein altered in Fragile X syndrome (FXS), and absent in developing post-mortem human FXS neocortex. Moreover, mice expressing a functioning human *NOS1* transcript reproduced the phenotype and provided a tool for identifying a specific human sequence mediating this molecular interaction with FMRP.In addition, most recently, the State and Sestan labs have taken a bottom-up approach to identifying convergence among a set of ASD-risk genes discovered byWES, by interrogating the multi-dimensional transcriptional dataset of human brain[43] and mouse neocortical layers (manuscript under review). We conducted co-expression analyses using

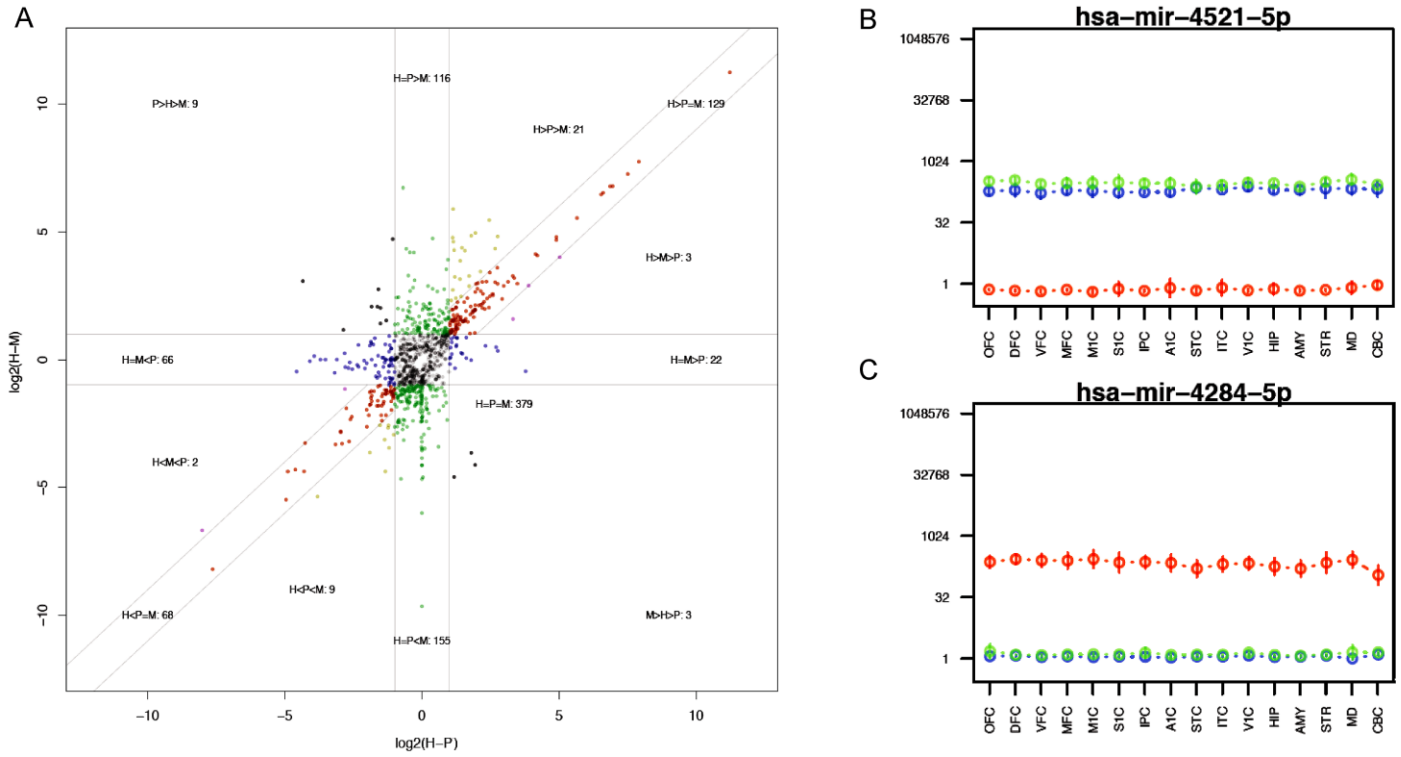

Figure 2. Inter-primate differences in the expression of micro RNAs*. **A**) Scatter plot highlighting miRNAs that exhibit statistically significant (q<0.05) species-level (log2-transformed) differences between human and chimp (x-axis) and human and macaque (y-axis). miRNAs >2-fold increased or decreased between pairs of species (also highlighted by the grey guidelines) are colored red for human-specific difference, blue for chimp-specific, and green for macaque-specific. Counts of the number of miRNAs in each category are provided in the relevant region. **B** and **C**) Expression profiles for two miRNAs specifically altered in human (red) compared to chimp (blue) and macaque (green). The normalized average expression (log2-transformed) for each species is plotted, along with the standard deviation between the 5-6 biological replicates for each of the 16 brain regions.*

nine genes strongly associated with ASD (high confidence (hc) ASD genes) based on WES derived from our labs and additional published datasets[64-69] (Figure 3). An independent set of 116 probable ASD genes (pASD) was also selected, including all genes carrying a single *de novo* loss of function (LoF) mutation. We then evaluated for evidence of statistically significant enrichment of pASD genes within spatio-temporal networks defined by hc ASD genes. This approach demonstrated that ASD risk gene co-expression networks converge during mid-fetal development in the frontal cortex, especially within layer 5/6 glutamatergic projection neurons (see preliminary data and Figure 3). Despite recent progress in genetics and imaging, the molecular and cellular pathophysiology of ASD has remained largely uncharacterized. The systems-biological approach we have undertaken has leveraged a tremendously heterogeneous genetic architecture and a foundational human brain gene expression resource to identify one critical point of spatio-temporal convergence, setting the stage for further progress in elaborating mechanisms of disease. However, several additions are now required to advance these studies, namely: (1) genomic data with finer cellular resolution, (2) complementary non-coding and cis-regulatory maps, (3) the identification of additional (especially non-coding) risk loci, and (4) integrative methods to incorporate these with our current findings.

Together, these preliminary studies highlight how investigating only transcriptome data, that is spatially and temporally rich, can be a powerful tool for understanding human development and evolution, and how some of the discoveries would not be possible using traditional model organisms. At the same time, they also illuminate that multi-dimensional data and computational methods to achieve points 1-4 above would substantially improve the ability of the above approach to understand any complex disorder, psychiatric or otherwise. Actionable steps to achieve these goals are outlined in Aims 1- 3.

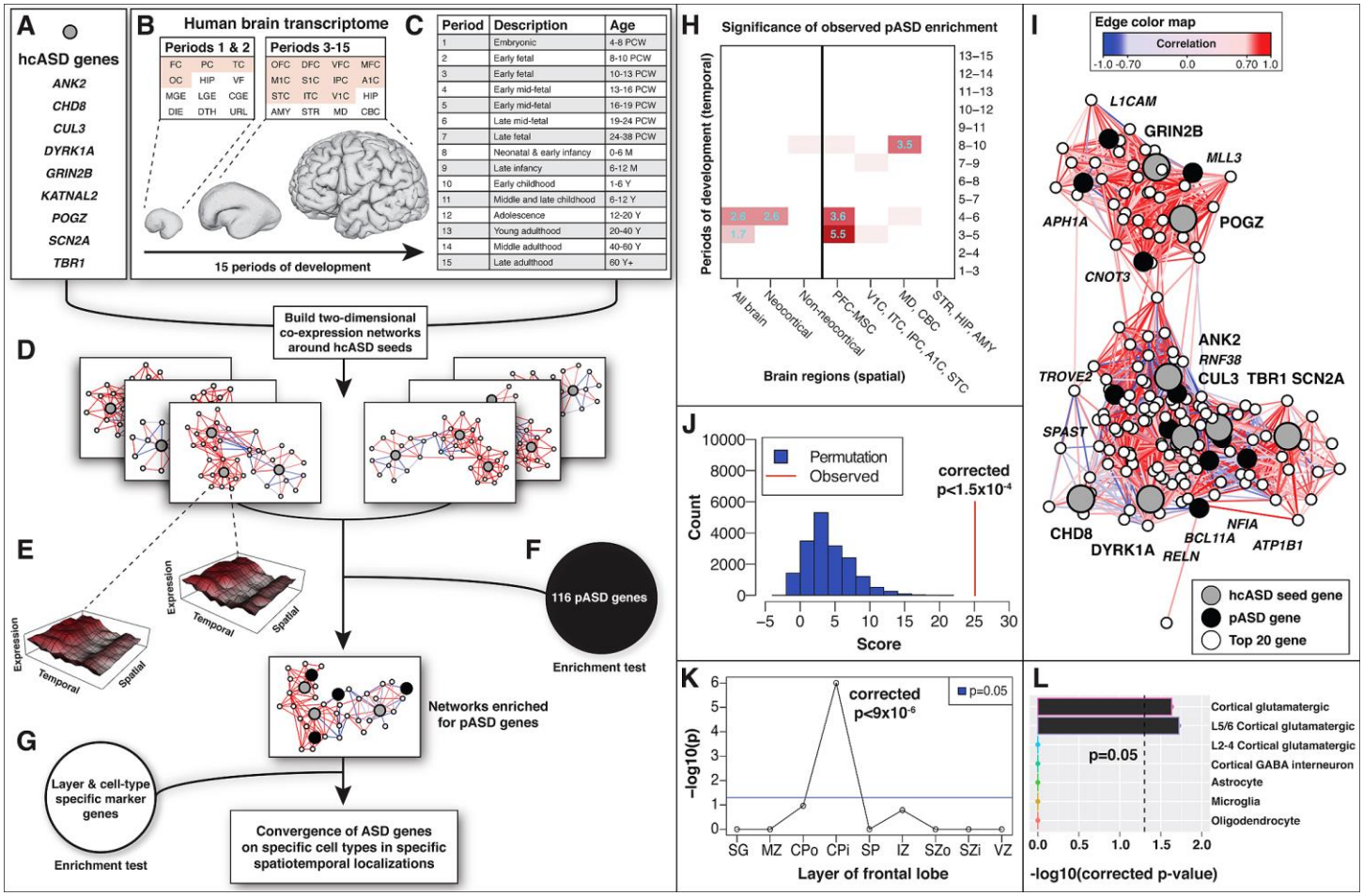

Figure 3. Co-expression networks implicate human mid-fetal deep cortical projection neurons in the pathogenesis of ASD. ***A****) Nine high confidence ASD (hcASD) seed genes. A comprehensive dataset of spatio-temporal gene expression spanning* ***B****) 16 regions of the human brain and* ***C****) 15 periods of brain development (43) informed co-expression analysis.* ***D****) The hcASD genes are used as 'seeds' to build co-expression networks.* ***E****) Mean expression levels for two genes are plotted as a function of period of development and region of the brain. Highly correlated genes will have similar expression profiles. Networks are interrogated for enrichment of* ***F****) an independent set of 116 probable ASD genes (pASD genes) and* ***G****) layer and cell-type specific genes.* ***H****) Spatio-temporal convergence is observed in the prefrontal and primary motor/somatosensory cortex during mid-fetal development (periods 3-6). Co-expression networks were formed from subsets of the expression data based on developmental stage and brain region. Each of the networks was tested for enrichment of 116 pASD genes. This heatmap shows the negative log_{10}(corrected p-value) of this enrichment for each network. Networks that are significant after correction for multiple comparisons have the negative log_{10}(p-value) overlaid with cyan text.* ***I****) The period 3-5 PFC-MSC co-expression network from (H). The lines (edges) link genes with co-expression correlation >|0.7| and the color and intensity represent the strength and type of correlation.* ***J****) The pASD genes enriched within this network represent those with the highest probability of being true ASD genes. The TADA score assesses which pASD genes are likely to be true ASD genes (He et al., PLoS Genetics in press). The combined TADA score in the period 3-5 network is highly significant by permutation test (corrected $p < 1.5x10^{-4}$).* ***K)*** *To improve the spatial resolution of the analysis, the connectivity of the period 3-5 network was assessed in a separate prenatal transcriptome. Significance of the observed connectivity was assessed with a permutation test and is greatest in the CPi region (inner cortical plate; layers 5-6; corrected $p<9x10^{-6}$).* ***L****) Enrichment of cell-type specific marker genes is specific to deep layer cortical glutamatergic neurons (permutation test with 20,000 iterations; corrected for multiple comparisons).*

RESEARCH DESIGN

Aim 1: Multi-dimensional genomic analyses of human and NHP brain single cells and tissues. *Rationale:* The remarkable diversity and developmental dynamics of human neural cell types have left a significant level of unsolved questions in their functional orchestration. Delaying post-mortem deterioration of human brain tissue and preserving the stability of its nucleic acids and proteins, have been particularly challenging yet crucial for the application many of the cutting edge genomics methods. To overcome this problem, we have recently developed a novel approach we named HC (Hibernation and Cryopreservation) protocol, which maintains a post-mortem fetal or adult human brain in a state of prolonged hibernation, thereby stabilizing nucleic acids and proteins and allowing concurrent collection and subsequent cryopreservation of (1) live cells, (2) single cell suspensions, and (3) tissue samples from hundreds of regions of the same brain. We will use this approach for prospective collection of human and NHP tissue and single cells, in addition to our existing collection of human and NHP brain specimens (see Resources). The advent of iPSCs[70] provides a unique opportunity to study the biology of neural cells in species such as human and NHP that are not amenable to experimentation, especially during early stages of their development. Also, iPSCs can be differentiated into a pure population of single cell type that can be profiles by multiple techniques. This project will leverage this technology by generating human, chimpanzee, and macaque iPSCs, which will subsequently undergo neural differentiation to create enriched populations of two major cell types of the cerebral cortex: neural progenitors (NPs) and glutamatergic projection neurons (PNs).

HC protocol (Hibernation and Cryopreservation) protocol for post-mortem human and NHP brains: This protocol was recently developed by the Sestan lab to delay post-mortem deterioration of human brain tissue and preserve the stability of nucleic acids and proteins by combining tissue hibernation and cryopreservation (manuscript under preparation). To optimize this hibernation solution, we extensively tested with mouse tissue prior to applying to human tissue. We then tested with one adult and four mid-fetal human brains in the past four months. In the first hibernation step, the whole hemisphere of a fetal brain or 0.5 cm thin slabs of an adult brain were immersed into an ice cold hibernation solution and transferred to the Sestan lab. Of the five tested brains three were shipped overnight by a major courier delivery service company and processed the next day in our lab. Remarkably, the hibernation solution allows live brain tissue to be stored in a refrigerator for up to five days. In fact, only one fetal brain did not yield live cells after dissociation and culture following this treatment. Prolonged hibernation allows unprecedented opportunity to precisely dissect any number of samples dissectible within 5 days, provided the brain sample is kept at low temperature and in sterile conditions. After dissection, tissue samples were further divided into three portions for cryopreservation procedures: (1) frozen tissue samples (used for tissue analyses in subaim 1.1), (2) minced live tissue samples archived in liquid nitrogen, and (3) single cell suspensions (used for analyses in subaim 1.3). To achieve the goal of preserving RNA and cell morphology, yet maintaining easy retrieval of RNA from each cell, we have tested a number of fixation and preservation reagents based on existing literature and in-house innovation. As a result, we have greatly improved preservation of not only RNA but other genetic material or even proteins, which is crucial in our multi-dimensional approach to decipher molecular dynamics of the human brain. Following this treatment, single cell suspensions are stored in an antifreeze solution. As we have confirmed that cryopreserved suspensions do not freeze at -40C, we will keep cells at -40C in dedicated deep freezers in order to further improve the long-term viability of the cells. Cryopreserved human single cells maintained normal gross morphology (size and shape) and 90% of the preserved cells did not pick up trypan blue dye, commonly used to mark damaged cells. We are now in the process of testing this in preserved single cells stored over a long period of time (months). This protocol will permit dissection of hundreds of regions as well as isolation of both single cells and live cells from the same brain. Therefore, we can now establish a unique tissue and single cell bank for rare human and NHP brain specimens. This will ensure specimens will be utilized in a most timely and efficient manner. Moreover, as single cell technologies develop over the next decade, we will be able to apply new methods to analyze single cells from our reservoir simply by aliquoting the required amount of cells only, without concern for human tissue availability. Finally, archived live cells from the same tissue sample can be subsequently cultured to follow up on genomic findings (e.g., CNVs, RNA editing, protein localization). This would be a transformative bio-resource for genomic studies of human and NHPs. In this Aim we will use this approach, existing frozen samples, and iPSCs to create

multidimensional genomic data at the level of brain regions (subaim 1.1), enriched population of specific cell types derived from iPSCs (subaim 1.2), and single brain cells (subaim 1.3).

Subaim 1.1: Multi-dimensional genomic analyses of regional brain tissues. This project takes advantage of an already existing tissue collection consisting of high-quality fresh-frozen post-mortem macaque brains covering all major periods of development, as well as a limited number of chimpanzee brains including rare fetal and early postnatal specimens in the Sestan lab (see Resources). These will be subjected to mRNA and small RNA transcriptome profiling by paired-end RNA-seq in 16 regions and 15 periods as done for the human brain[43] to build a unique spatio-temporal transcriptome dataset for NHPs, which will also complement the existing human RNA-seq dataset (www.brainspan.org). Also, the higher availability of frozen post-mortem chimpanzee brains will significantly complement the comparative analysis of 3 species transcriptome, rather than with limited number of fresh chimpanzee brains available.

In addition to the existing frozen tissue, the prospective human and NHP repository will be collected using HC protocol. We will collect proband and parental DNA for macaques, humans and chimpanzees, when available. We will also strive to have multiple male and female specimens for each of the 15 periods as described in Kang et al., 2011[43]. Todd Preuss will implement HC protocol in his lab if fresh chimpanzee brain tissue becomes available at the Yerkes Center due to natural death (i.e., only period 14 will be procured for the chimpanzee). Age of NHPs equivalent to human developmental periods will be estimated using *www.translatingtime.net* model[71]. The Sestan lab will dissect all brains and distribute them to other teams. Even though we plan to dissect around 80 regions, based on their biomedical importance, this project will focus on the dorsolateral prefrontal cortex (DFC), primary visual cortex (V1C), striatum (STR) and cerebellar cortex (CBC) of human and macaque representing early-fetal (period 4), mid-fetal (period 6), infancy (period 8) and adult period (period 14) (see Budget Justification for tables of samples). Due to the rarity of available fresh tissue, chimpanzee will be sampled only for adult period (period 14). The four regions were selected based on their functional importance and size, since we will need enough tissue for multiple analyses. Two males and two females will be analyzed per each period, which will allow for detection of sex-based differences. All samples will be subjected to the same set of multi-dimensional -omics analyses (DNA-seq, RNA-seq, ChIP-seq, HITS-CLIP, proteomics, etc.). All DNA-seq, RNA-seq and ChIP-Seq libraries will be sequenced at the Yale Center for Genome Analysis (YCGA), proteome samples will be analyzed at Yale's Keck Center by the Nairn lab, and HITS-CLIP will be done by the Darnell team at the New York Genome Center (NYGC).

Subaim 1.2: Multi-dimensional genomic analyses of iPSC-derived cell populations. To obtain matching brain tissues and iPSCs, iPSCs will be generated from the meningeal or skin fibroblast of the same human and NHP donors from which the brain will be harvested using HC protocol. We will employ non-integrating episomal vectors expressing human *OCT4*, *SOX2*, *KLF4*, and *c-MYC* genes[72]. We will only include in our study iPSC lines that fulfill the following standard criteria of successful reprogramming: (1) immunohistochemical labeling for pluripotency markers (e.g., NANOG, SSEA4, TRA1-60), and (2) expression of known hES/iPSC markers (*SOX2*, *NANOG*, *LIN-28*, *GDF3*, *OCT4*, *DNMT3B*) by ddPCR.

The iPSCs will be directed to a NP cell fate using a standard, published protocol for ES cell and iPSC differentiation already implemented in the Sestan lab[73,74]. iPSCs are cultured in a neural-inducing media with Noggin, a BMP/SMAD signaling inhibitor, and retinoids for ~14 days. We have used this protocol on H9 ES cells and a line of human iPSCs and confirmed that cells of this lineage express NPC markers SOX2 and NESTIN. We will also confirm their cerebral cortical NPC identity with ddPCR for other marker genes such as *PAX6*, *FOXG1*, and *FEZF2*, but absence of non-cortical markers. *Bona fide* NPs will then be propagated by replacing Noggin and retinoids with FGF2 and EGF, which allows us to greatly expand each cell line. Removal of both factors from the neural media relieves the maintenance of the progenitor state and begins the cell-intrinsic differentiation of cortical PNs. In cell culture, early born cortical PNs acquire their molecular identity between 20-25 days, which we will confirm by the expression of marker genes *FEZF2*, *TBR1*, *BCL11B*, *ZFPM2* and *FOXP2*. After ~60-70 days of continued culture of NPs with early PNs, late PNs are then generated, which can similarly be validated with marker genes *CUX1/2*, *SATB2*, and *POU3F2*. Based on our preliminary network analyses implicating mid-fetal PNs in ASD, we will focus on creating neural NPs and early born PNs. However, with an extended time frame or budget, this project can easily be expanded to include late born PNs for a more comprehensive picture of neural development. We will apply the same techniques as proposed for tissue analysis (subaim 1.1).

Sequencing of genomic DNA: WGS of each individual will be performed on CBC DNA at 30x coverage to ensure accurate and robust variant calls.

mRNA and small RNA-seq: All protocols are standardized and data analysis and quality control will be done using the existing RNA-seq pipeline implemented by the Sestan and Gerstein labs to generate 100 and 50 bp long reads for mRNA (PE) and small RNA species, respectively.

ChIP-seq analysis: While we can perform ChIP-seq analysis on isolated nuclei (not shown), we have opted to use whole tissue for two reasons. First, the amount of embryonic tissue needed for the procedure is a limiting factor. Second, using the same brain samples used for RNA-seq allows a fruitful integration of the ChIP-seq data with the human brain transcriptome that was generated from the same samples. The selected histone marks (H3K4me3, H3K27ac, and H3K27me3) as well as the DNA-binding protein CTCF, identify a large fraction of active enhancers, promoters, and repressors, as well as insulators.The commercially available antibodies were succesfully used for ChIP-seq by the Sestan lab using frozen human and NHP brain tissues. Sheared chromatin will be immunoprecipitated, multiplexed, and sequenced at the 75 bp length.

DNA-Methylation: Due to the cost constraints, we will not profile all three species by bisulfite sequencing. Instead, only human brain samples will be assessed with Illumina Infinium HumanMethylation450 bead chips. These experiments will be funded by pledge from the Kavli Institute (see letters of support from Drs. Rakic and Slayman). We will apply our in-house pipeline for data normalization in order to remove batch effects and to correlate DNA methylation with gene expression and epigenetic modifications.

HITS-CLIP: The Darnell lab will undertake Ago HITS-CLIP and Nova HITS-CLIP on tissue and iPSC-derived neural cells (see the table of samples in Budget Justification). Studies on tissues will include a cross-species comparison of homologous regions between human, chimpanzee and macaque. Our in-house HITS-CLIP protocol works on as few as 10^4-10^5 cells, so that small amounts (10's of milligrams) of tissues will be sufficient for analysis. Additionally, we will continue to improve HITS-CLIP, with developments in the past year, using methods adapted from ribosomal profiling[75], offering up to ~1,000-fold improvements in sensitivity.

We have developed HITS-CLIP to maximize information content – to sequence as many unique RNA binding protein (RBPs)-cross-linked RNA tags per experiment as possible. Strategies include optimizing the biochemistry for maximum signal:noise and minimal PCR amplification. We will consider direct RNA sequencing technologies as they evolve, since eliminating PCR would retain maximum complexity.

These improvements will allow us to compare tissue results with those from iPSC-derived neural cells (see subaim 1.2). Such experiments can readily generate 10^4 cells for study, and in preliminary studies with motor neuron-differentiated iPSCs, we have been able to generate HITS-CLIP data with other RBPs (data not shown). Integration between the tissue and iPSC HITS-CLIP experiments will be aided by orthogonal studies being done in mouse single cell types (not part of this proposal; funded by NIH R01 NS34389; see Fig. 3), using Cre-lox mediated generation of individual epitope tagged RBPs for single cell-type HITS-CLIP analysis of Ago and Nova in mouse brain. Such data will provide a bridge for bioinformatic overlay between the single cell-type advantage of iPSC HITS-CLIP and the more complex but physiologically relevant analysis of brain tissue.

Unbiased proteome analysis: We propose to subject tissue samples and iPSCs to a proteome-wide assessment of protein abundance by label-free liquid-chromatography tandem-mass-spectrometry (LC-MS/MS)[76] and subsequently perform a deeply integrated computational analysis of protein- and RNA-level data derived from the developmental and adult human, chimp, and macaque tissue samples. Protein expression is often of most interest and relevance to molecular-biological research although, due mainly to technological limitations, expression at the RNA-level continues to be used as a proxy for genome-wide gene-expression estimates. By assembling both known and novel transcripts from RNA-seq data[7], we aim to create an extremely accurate and highly tissue-specific reference from which to improve the analysis of LC-MS/MS based protein analyses.

As the name implies, label-free MS/MS requires no chemical or biological modifications be made to the samples under investigation, affording an unbiased sampling of reasonably abundant cellular proteins[77-79]. Briefly, proteins are enzymatically cleaved (typically using Trypsin) and the resulting mixture of peptides loaded onto a liquid-chromatography column, ionized, and injected by electrospray into the mass-spectrometer. Inside the mass-spectrometer, abundance of intact peptides (a.k.a. 'parent ions') is monitored and those with a reported

intensity that exceeds lower-threshold are selected for collision-induced dissociation and the fragmented 'product ions' recorded in a second mass-spectrometer step. The product ions are used to identify the amino-acid sequence of the parent ion and the intensity of the parent ion over time is used to calculate its abundance.

The recommended peptide input to label-free proteomic profiling is 1-10 fmol (http://keck.med.yale.edu/proteomics/technologies/proteinidentification/lcmsms.aspx). Assuming a total of somewhere in the order of 10^8 proteins per cell, this corresponds to a requirement of ~10-100 cells per LC-MS/MS run using current technology[80,81]. We will therefore be able to apply proteomic profiling to the study of the iPSCs at the various stages of their differentiation.

Targeted proteomics analysis: The Nairn lab will perform large-scale targeted proteomic analysis of peptides and proteins corresponding to genes identified for follow-up based on the results of other analyses, such as those exhibiting consistent allele-specific expression or RNA-editing. To achieve this we will employ a variant of the label-free LC-MS/MS approach that monitors, in real-time, pre-defined peptide and product-ion masses[82-84]. The advantage of this approach is highly reliable detection of selected peptides across diverse tissues or experimental conditions and by targeting multiple peptides capable of supporting or rejecting multiple hypotheses, novel or interesting genomic events can be validated with-reduced bias and variability compared to Western blotting.

Subaim 1.3: Integrated analyses of single cells of human and NHP brains: To overcome the problem of cellular heterogeneity in dissected tissue and generate much needed transcriptome data at the level of single cells, we will also use RNA-seq on single cells isolated from DFC, V1C, STR, and CBC of human and macaque at early-fetal (period 4), mid-fetal (period 6), infancy (period 8) and adult period (period 14). At the moment chimpanzee tissue is not available for these analyses. However, this aim is not dependent on access to chimpanzee tissue, but in the fortunate case that it becomes available, it will be a valuable addition to the analyses. We will take three related innovative approaches to address complexity of transcriptome at single cell levels: (1) unbiased high-throughput short read RNA-seq approach to deconvolute cellular composition, (2) long read RNA-seq on selected cell types to depict the landscape of alternative splicing, and (3) targeted proteomics analysis to decipher the relationship between mRNA and protein expression.

Unbiased high-throughput short read RNA-seq of single cells: We have adopted a single-cell RNA-seq protocol from the recently published single-cell tagged reverse transcription (STRT) method[29], which consists of a template-switching reverse-transcription step with a distinct oligonucleotide, thus introducing barcodes and multiplexing of up to 96 single cells. Because this protocol sequences 5' ends of mRNA, it precludes analysis of splice isoforms; however, it gives several other advantages. The major advantage is the great reduction in cost and time, which is crucial in the analysis of highly complex cellular architecture of human brain. Also, it retains strand information and thus permits the distinction of overlapping genes transcribed from opposite strands, it identifies the actual transcription start site (TSS), which is often lost in methods that have a 3' bias. Additionally, it facilitates accurate mRNA quantitation, since each mRNA molecule results in a single cDNA. We have applied this protocol to profile transcriptomes of isolated single cells from mouse neocortical ventricular zone NPs. We detected significant stochastic gene expression within each cell, but this might be due to the nature of our single cell samples in which various types of ventricular zone neural progenitors have been subjected to the analysis, changes in RNA stability due to single cell manipulation, or other technical errors. We will address these issues by implementing several improvements to the technique to increase

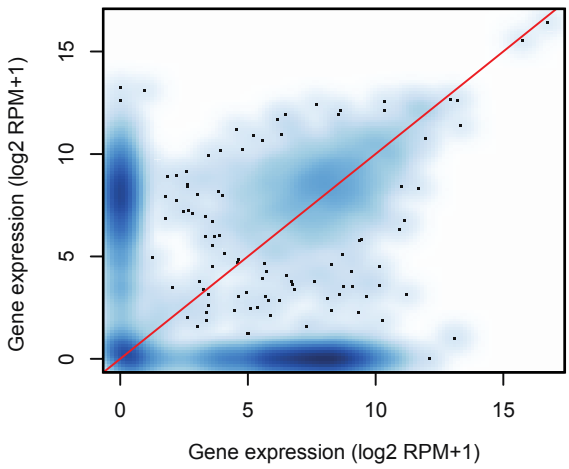

Figure 4. Comparison of two cells from identical tissue. *Smoothed scatter plot comparing (log2 transformed) gene expression estimates in two single cells obtained from the same species, brain, and region. A large fraction of the genes exhibit a binary 'off-on' signal (dark horizontal and vertical regions).*

robustness and sensitivity, such as (1) minimizing degradation and artificial change in RNA levels by applying the HC protocol proposed in subaim 1.1; (2) applying multiple-round linear amplification strategy[28,85-88] to the existing STRT protocol to mitigate possible skew in the amplification of certain kinds of mRNA sequences; and (3) applying a suppression PCR[89] to avoid non-specific amplification. Another critical observation from our preliminary results and other published studies is that the gene numbers detected per cell can be increased with the sequencing depth – however, the increase is not linear but is influenced tremendously by fluctuations in shallow sequencing (a few million reads per cell). Nonetheless, the influence of sequence depth will plateau once the read number reaches approximately 10 million per cell.

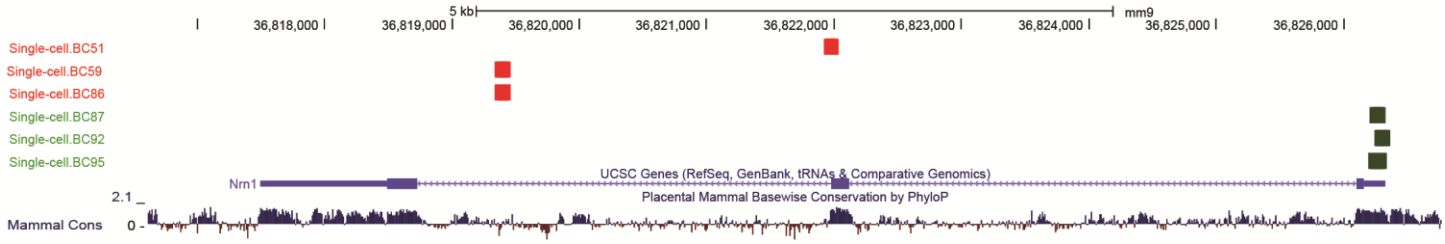

Figure 5. Single cell strand-specific transcriptome mapping. *The RNA-seq reads were mapped to both strands of Nrn1 genes. In a group of single cells (shown in green) we observed 5' enriched mapping in its sense strand, while in some other cells (shown in red) had distinct read mapping to its antisense strand.*

To facilitate efficiency and less harmful single cell capturing, we have been using a PASCA system for high-throughput single cell capturing (done in collaboration with Dr. Peter Koltay at IMTEK (Freiburg, Germany)). The basic technology for single cell handling applied in the PASCA project is based on inkjet-like printing of single cells confined in picoliter sized microdroplets. We utilized the PASCA platform and successfully conducted the STRT-method for single cell RNA-seq profiling described in the above experiments. However, since the PASCA system is still not available for purchase, we have opted to go with the Fluidigm system for high-throughput single cell capturing. The Fluidigm platform will be combined with observation under the microscope to take images of the cell after capture. We have been in contact with scientists at Fluidigm to adapt the cost-effective STRT-seq protocol to their platform and do not anticipate any issues. However, if we experience technical difficulties with Fluidigm platform we could go back to the PASCA, which may be commercialized by then. In fact, we have been offered to house a PASCA prototype in the Sestan lab.

We will analyze cells from the same four regions (DFC, V1C, STR, and CBC) in all three species using the HC protocol in subaim 1.1. Within each region we will sequence 480 single cells (96 X 5 sets of sequencing runs). Based on our preliminary data in which we sequenced multiple types of mouse NPs, we estimate we will need at least 10 cells for effective expression-based clustering and classification. This will allow clustering of multiple different cell types within each dissected region as well as constructing meta-network of cell clusters across different brain regions. Single cell suspensions for a brain will be first tested by a single run on MiSeq to make sure that libraries are of sufficient quality to run all of the samples on more costly HiSeq.

Long read RNA-seq on selected cell types: The major downside of the proposed STRT protocol is the lack of whole transcript coverage by prioritizing the accurate and strand specific 5' biased gene expression quantitation. The fine-tuned full-length RNA-seq information is crucial for splice site analysis. The PacBio RS II sequencing technology resolves single molecules in real time, allowing observation of structural variation not accessible with other technologies. We find these unique capabilities of the PacBio RS II system are ideally suited for a variety of applications, including full-length cDNA profiling. Recent upgrade of the instrument at least doubles the throughput and increases the raw read accuracy significantly. Furthermore, there are some software developments that allow full-length cDNA analysis (https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Transcriptome-Analysis). Moreover, it is possible to use as little as 1 nanogram or less DNA input for the library prep, which would make it feasible to sequence particular cell types with limited input material. We estimate our current throughput will allow us to profile 5-10 X coverage of 2,000-3,000 transcripts at 4-5 Kb median size using two smart cells of sequencing (the total cost for one library prep and two cells would be $1,000), which has prompted us to apply this technology to profile full-length RNA-seq transcriptomes of selected cell types that were deconvoluted by our short read RNA-seq (STRT-seq) method. As we explore this method, we expect we will be able to test as little as single-cell level RNA input. Deep profiling of either cell type

specific- or single cell level- full-length cDNA repertoire will provide thorough understanding of structural variation of each transcript across different cell types within human and NHPs. This will be a great asset to our integrated analyses of RNA-seq and MS/MS proteomics.

Aim 2: Integrated analyses of multi-dimensional single cell and tissue-level genomics data. *Rationale*: Since no one method is sufficient to fully capture biological complexity, we will develop multi-level computational strategies to de-convolute identities and molecular dynamics of single cells (subaim 2.1), as well as deeply analyze and integrate multi-dimensional genomics data across multiple variables (subaim 2.2).

Subaim 2.1: Development and application of agnostic methods for analyzing single cell data. The proposed single-cell experiments will allow, for the first time, extremely high-resolution profiling of the dynamics of RNA transcription between primate neural cells, developmental periods, and species. As the mammalian brain is comprised of many distinct and specialized functional areas to achieve high resolution coverage in transcriptome profiling we will, from each of 4 regions (DFC, V1C, STR, and CBC) in all three species, select and sequence 480 single cells (96 cells X 5 sets). Preliminary analyses show the transcriptome profiles of individual single cells is extremely different from those of whole-tissue (Figure 4) (also see[30]). Our data from gene-level quantification of single cell RNA abundance shows that a minority of the total number of expressed genes are consistently expressed in multiple cells (Figure 4; dark diagonal areas). The remainder appear to behave in a binary 'off-on' transcriptional regimen suggesting that the single-cell transcriptome is variable between cells, and different cell types; a result largely consistent with the concept of transcriptional bursting[90]. This further suggests that the lower expression variability at the level of comparing whole-tissues results from the averaging of a large number of distinct cells and cell-types.

We will allocate one of the 96 wells in the single-cell profiling of each of the proposed tissues to be used as a whole-tissue quality control 'meta-cell'. To create this meta-cell data, we will obtain total RNA from a whole tissue lysis and subject this material, derived from millions of individual cells but diluted to a similar molarity as the material obtained from the individual single-cells, to the same 5' purification and amplification used for the single-cell prep. It is important to note that we will already have standard tissue-level RNA-seq for each of these meta-cell controls that will allow a identification of any technical variability introduced specifically as a result of the single-cell sample preparation (notably the 5' capture and extensive PCR amplification required for such low amounts of RNA). Further, we will use the meta-cell control to facilitate integration of the 5' single-cell sequence-read data with the standard RNA-seq read data obtained from the whole-tissues that will maximise the sensitivity and specificity of the single-cell read-alignment for inter-region, and inter-species comparisons.

Deconvolution of cell types and outlier detection of rare cell population: We will assess differences between individual cells by various unsupervised clustering methods such as principal component analysis (PCA) and hierarchical- and/or partition-based cluster analyses. The subpopulations of cells identified by such methods will provide a broad stratum of intra-tissue cellular heterogeneity. We will follow up with an analysis of the relative enrichment of each cell-specific marker genes in each subpopulation and use the expression profiles of these genes to guide the identification of an expanded set of cell-type specific markers.

Improvements in cell type and transcript discovery by targeted long-read RNA-seq: To assist in the validation of the single-cell-type results we will profile full-length cDNAs of selected cell types that were deconvoluted by our short read RNA-seq (STRT-seq) method. Deep profiling of either cell-type specific or single-cell full-length cDNA repertoire will provide thorough understanding of the structure of each transcript across different cell types within human and NHPs. This will largely overcome any issues introduced due the 5' observation bias in the single-cell analysis protocol and will therefore be a great asset to our integrated analyses of ChIP-seq, RNA-seq, and LC-MS/MS proteomics datasets. We will use the long-read RNA-seq data to search for novel transcripts in each species and cell-type and, in combination with expression information derived from single-cell RNA, whole-tissue RNA, and whole-proteome we will build the first reliable transcript models specific to brain tissues; a resource of immense value to researchers investigating gene expression in the brain.

In-situ mapping of rare or novel cell types: Genes that appear representative of particularly rare or potentially novel cell types will be prioritized, as determining the spatial distribution of such cells may be functionally relevant. Spatial distributions of these cells will be determined by *in situ* hybridization to detect the 3D gene

expression pattern of representative marker genes to create a transcriptional cell type "signature" and deduce the original anatomical localization of each cell type.

Single-cell coding and non-coding RNAs and pseudogenes: Given the heterogeneous nature of cell populations throughout the brain, we will exploit the single-cell RNA-seq data in neurons and glial cells to quantify gene expression estimates in each cell and assess consistency of expression across cells within a tissue, between regions, and within species for mRNAs, long non-coding RNAs, and pseudogenes. Pseudogenes especially, when combined with their parent-gene transcription signals, may serve as useful biomarkers to distinguish different cell types. It has also been reported that despite their low abundance, pseudogenes and ncRNAs exhibit a greater degree of cell-type specific expression than mRNAs[17]. Compared to the tissue-level, single-cell expression data are more uniquely suited to the detection of poorly expressed RNAs because it is possible to distinguish stochastic experimental noise (resulting from sample preparation and data processing) from genes that are expressed only in a small fraction of the total number of cells. Preliminary data obtained from 27 cells acutely isolated mouse embryonic neocortex (unpublished data) show that even though protein coding mRNAs are best represented and the highest-expressed, the level of pseudogenes and other ncRNAs is sufficient for a meaningful comparison (Figure 6).

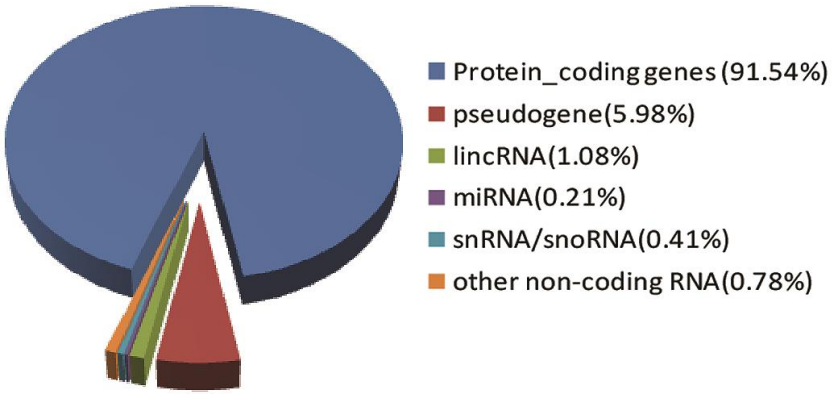

Figure 6. Genomic organization of single neocortical ventricular zone progenitor cell transcriptome based on single-cell RNA-seq. *The detected genes RNA-seq data from 27 single cells are classified into six types, of which protein-coding genes dominate the percentage distribution.*

Single-cell allele-specific expression, RNA-editing, TSS-usage, and antisense transcription. We will perform a limited but important analysis of allele-specific expression (ASE) and RNA-editing events (a.k.a. RNA-DNA differences; RDDs) using the single-cell RNA-seq data. Due to the proposed single-cell RNA prep preferentially producing sequence reads at the 5' extreme of transcripts these analyses will not be as comprehensive as those covered in much more detail in Aim 2.2, however, for a subset of the interesting cases identified as part of Aim 2.2 the extension of the analysis to single-cells will be extremely valuable.

We will exploit the 5' bias of the single-cell RNA-seq protocol to assess transcription start-sites (TSS) across cell-types, regions, development, and species. Further, our protocol includes strand information and enables an assessment of the antisense regulome. As an example, the preliminary single-cell RNA-seq data analysis found reads aligned to the antisense strand of *Nrn1* gene (see Figure 5).

Inter-region and inter-species differential expression (DEX), and comparison with whole-tissue DEX: An important concern when profiling tissue-level expression in any organ, but especially in complex organ like the brain, is the extent to which the intra-tissue cellular heterogeneity dampens the ability to detect inter-tissue or inter-species differential expression. In the brain the cell-type of greatest interest is typically a neuron. When studying differential expression between brain-regions, for example, ~60% of the signal is derived from off-target cell-types, typically astrocytes, oligodendrocytes, and microglia. This reduces the ability to detect by half differences in neurons. We will exploit the highly-specific single-cell expression data to assess the extent to which tissue-level neuronal differential expression analyses are compromised by this off-target glial signal by excluding expression estimates identified as glial (through unsupervised clustering described above) and comparing the differential expression result with that obtained from the whole tissue. This method has several potential applications more generally in that it will serve as an incredibly valuable benchmark for setting new, potentially more lenient, thresholds for calling a gene differentially-expressed in the whole-tissue analyses.

Subaim 2.2: Integrated analyses of three-species brain development data. Deep integration of different data modalities across developmental time points, tissues, cell types, and species can be achieved both computationally, through analysis workflows designed to exploit the carefully matched and curated samples, and experimentally through novel assays such as HITS-CLIP, which vastly improve our ability to resolve complex cellular regulatory mechanisms.

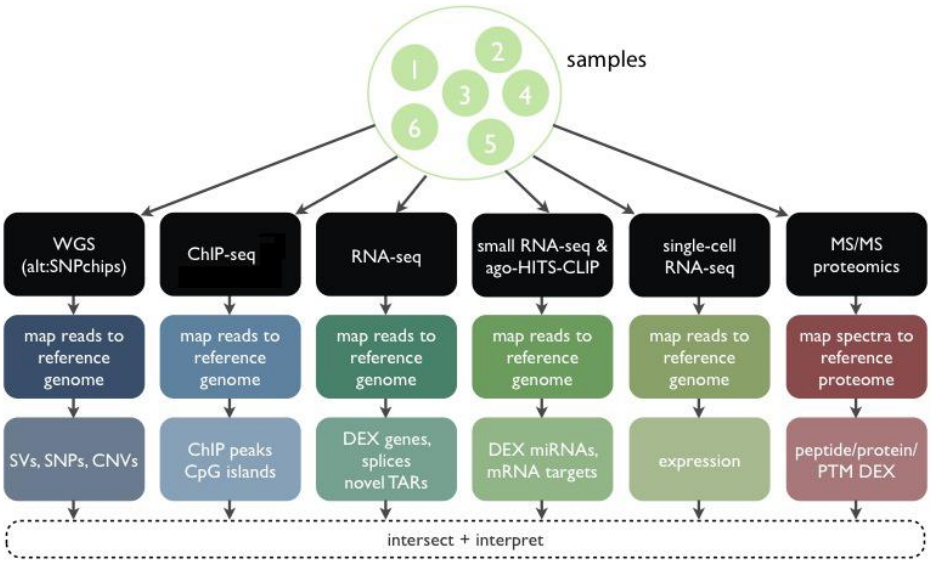

Figure 7. Standard multi-omic analysis paradigm. *In multi-level genomic analyses, each data modality tends to be treated independently until downstream integration, which is usually a form.*

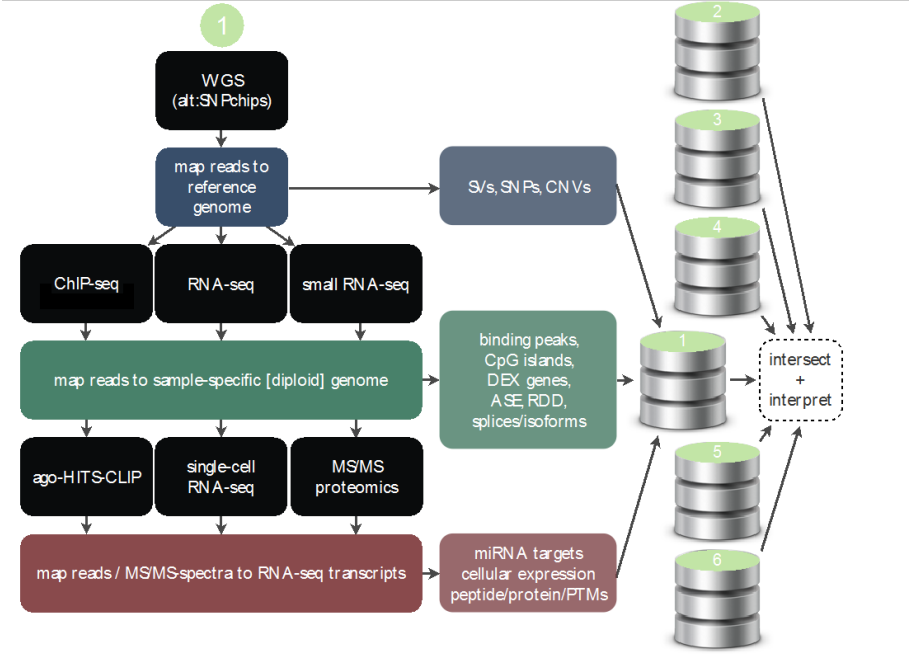

Figure 8. Proposed analysis paradigm. *Reflecting the logic provided by the central dogma of molecular biology, we propose to develop a completely novel integrated data analysis platform where actionable sample- or subject-specific information regarding genomic variants and isoform usage are fully integrated in downstream analysis steps.*

Development of an integrated platform for computational analysis of all single-cell and tissue-level data: In every mammalian species, the brain is the most complex and 'personal' organ. A wide variety of cell-types co-exist to ensure the proper functioning of the brain and neighboring neurons, exhibit different behavior and perform different neurological functions. Despite often-small differences in absolute gene expression between neurons or brain regions, the brain appears to exploit a highly complex choreography of splicing and epigenetics in order to retain molecular diversity. Therefore to fully elucidate these neuronal molecular networks, we propose to develop a computational analysis framework that is similar to, but more powerful, than those that are starting to be applied in the analyses of personal genomics.

In multi-level genomic studies of a given system, the normal analysis approach is to integrate the various levels of information in the form of a network analysis. Several high-profile analysis efforts have pioneered the integration of ChIP-seq, RNA-seq, and proteomic datasets derived from comparable biological sources to produce interesting representations of the biosynthetic landscape of cellular systems. While this has repeatedly been shown to produce useful knowledge, such an approach does not make full use of the available data as each genomic experiment is essentially analyzed independently up until the point at which the summary data are integrated into the network analysis (Figure 7). We propose to develop a completely novel depth-first approach to integrate and analyze the multi-dimensional omics data produced in Aim 1 of this proposal (Figure 8). For each brain (a.k.a. 'subject') we will first map and analyze the whole-genome sequence data to obtain subject-specific information regarding variants, structural variation, and copy-number variation. Using this genomic information we will automate the creation of subject-specific diploid reference genomes that we will use to map the sequence reads produced from our RNA-seq, small-RNA-seq, and ChIP-seq experiments. The main advantage of such an approach is that it reduces the reliance of all of these analyses on a single generic reference genome (e.g. hg19) which can result in artificially fewer reads mapping to regions with higher-rates of mutation[91]. Another large advantage is that it allows the automatic integration of allele-specific binding and [mRNA/miRNA] expression as well as RNA-editing. We will use transcripts assembled using the RNA-seq reads to map reads obtained from the HITS-CLIP, single-cell-RNA-seq, and mass-spectrometry proteomics assays.

Proof of concept: Integrated analysis of RNA-seq and proteomic datasets. Protein expression is often of most interest and relevance in molecular-biological research although, due mainly to technological limitations, expression at the RNA-level continues to be used as a proxy for genome-wide gene-expression estimates. For example, protein structure and/or abundance is likely to be much more relevant to disease mechanism than RNA structure and/or abundance. However, for many genes the relationship between mRNA and protein expression has been found to be poor[92-95] and these poor correlations are believed to result from alternative splicing, mRNA stability, micro-RNA induced mRNA degradation, post-translational modifications, or protein-turnover. The successful identification of any such event is often an important finding for exactly this reason.

Here we propose to develop and refine a novel method that exploits the single-nucleotide, unbiased nature of RNA-seq to predict a reference proteome for the subject and tissue of interest in an effort to improve the resolution and throughput of LC-MS/MS proteomic analyses (Figure 9a). RNA-seq can potentially predict the entire proteome without requiring any a priori information regarding the transcriptome sequence, and we have already shown that this approach is capable of predicting the proteome with very high resolution[7,96]. We have shown through preliminary investigations that the number of identified proteins can be greatly increased (by up to 150% compared to SWISS-Prot) through the use of a tissue-specific reference proteome derived directly from an RNA-seq experiment (Figure 9b). We have also found that the coverage of the detected proteins, in terms of the number of distinct peptides observed for each, is consistently greater when using a specific proteome reference compared to a generic reference such as SWISS-Prot, leading to more accurate protein-quantitation and increased likelihood of validating variants, ASE, or RNA-editing at the whole-protein level. Therefore we will expand on this very promising preliminary investigation by constructing species- and tissue-specific transcriptomes which we will use to map LC-MS/MS spectra, affording a much greater specificity of protein-level analyses than would be available via conventional methods.

Integration of transcriptome analyses in tissues and single-cells: Using the data for each sample output from the integrated analysis workflow, we will perform various leading-edge transcriptome analyses across tissues, subjects, and species. We will individually assess exon-splicing and alternative transcript usage, pseudogene activity, nc RNAs, ASE, and RNA-editing, all of which are described in detail in the following sub-sections. Finally we will perform a final integration of these data by through a comprehensive network analysis of whole-tissue

and single-cell multi-dimensional genomic data. Through such an approach we expect to elucidate complex signaling and evolutionary mechanisms that would be impossible to resolve without such a tightly integrated computational analysis of these high quality matched brain tissues.

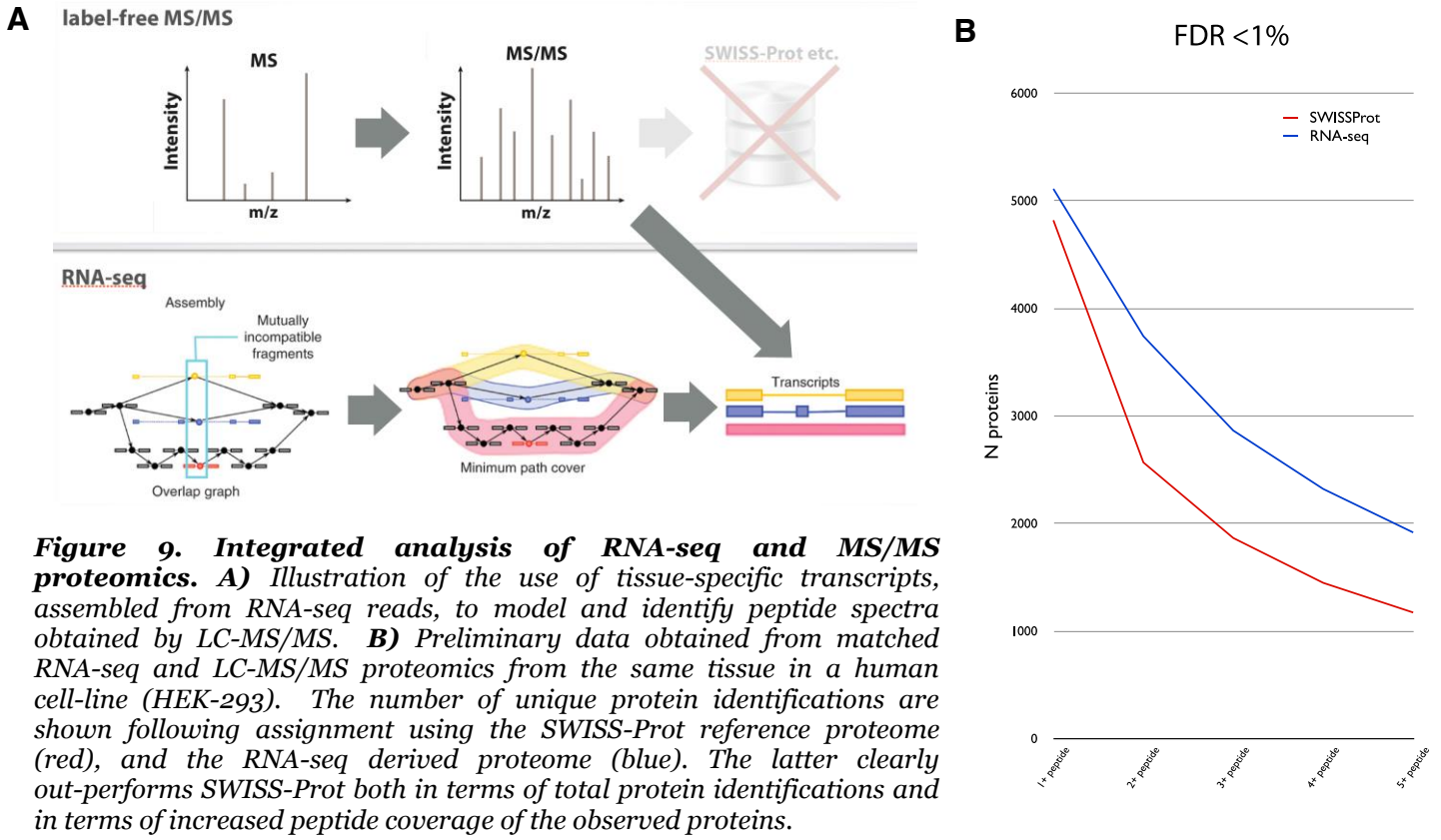

Figure 9. Integrated analysis of RNA-seq and MS/MS proteomics. ***A)*** *Illustration of the use of tissue-specific transcripts, assembled from RNA-seq reads, to model and identify peptide spectra obtained by LC-MS/MS.* ***B)*** *Preliminary data obtained from matched RNA-seq and LC-MS/MS proteomics from the same tissue in a human cell-line (HEK-293). The number of unique protein identifications are shown following assignment using the SWISS-Prot reference proteome (red), and the RNA-seq derived proteome (blue). The latter clearly out-performs SWISS-Prot both in terms of total protein identifications and in terms of increased peptide coverage of the observed proteins.*

Individual/group difference in splicing: Alternative processing of primary RNAs generates transcripts that may differ in localization, stability, and function. We will characterize the patterns of alternative transcript usage (e.g. promoter usage, exon-skipping, intron-retention, alternative transcription termination site) across different conditions (species and brain regions) in the proposed analysis. We will assess the complexity of alternative transcript usage in each condition, as well as compare quantitative changes between conditions. Differential alternative transcripts will be detected and such signatures can be used as biomarkers and or to prioritize targets for further investigation.

We will employ computational methods that we have previously developed, as well as develop new analytical procedures for such purposes. We have developed a tool, IQSeq[97], that can be used to quantify transcript abundance using RNA-seq. We have also developed a statistical method that can be used to assess the significance of quantitative differences in alternative transcript usage between conditions[98] and unpublished work).

Transcribed pseudogenes: Although pseudogenes have long been considered as nonfunctional genomic loci, recent studies have shown that, in some cases, pseudogenes are not only transcribed, but perform crucial regulatory roles via their mature RNA products[99-102]. In our previous work, we have shown that pseudogene transcription exhibits tissue specificity, and many pseudogenes are specifically expressed in brain[103]. These observations imply that pseudogenes could have unique biological activities in brain.

We propose to identify the transcriptional activity for each pseudogene annotation using the following protocol. First, we will remove pseudogene regions with mappability lower than 1. Second, we will discard the pseudogene regions shorter than 100 nucleotides after the mappability filtering. Only RNA-seq reads mapped to the remaining unique regions will be used to compute a normalized expression value (RPKM). Next, given previously published results on human pseudogenes with small-scale validation[102,103], which imply that ~15% of human pseudogenes are transcribed, we can set an RPKM threshold for human analysis such that it gives an

approximate agreement with the previous validation. With the assumption that the transcription of protein coding genes in human, chimpanzee, and macaque samples have similar distributions, we may apply quantile normalization on the gene transcription data for chimpanzee and macaque samples, using human as a reference. We will then apply this normalization to the pseudogene transcription data, and consistently apply the human threshold across the three species.

Assessment of non-coding transcription: We will utilize a statistical approach that compares the levels of expression in the known exon regions to threshold the RNA-seq signal and identify the intergenic and intronic regions that show significant expression. Next, we will utilize the methods we developed (e.g., incRNA[104]) to further classify and characterize these regions. Specifically, we will use the known coding sequences, UTRs, and non-coding RNAs to train a random forest algorithm and apply the trained algorithm to classify the novel transcript regions to one of the classes. Next we will assign targets to the classified regions by comparing them both with the annotated *cis*-regulatory elements (e.g. enhancers) and with proximal genes. We will also utilize statistical methods to identify antisense transcripts that have roles in regulating the overlapping transcript. We will compare the identified antisense transcription between different regions of the brain and identify the differentially regulated transcripts.

ASE and RNA editing: Gerstein lab has developed a software pipeline, AlleleSeq[105], for the processing and analysis of RNA-seq or ChIP-seq datasets in order to determine sites of ASE. AlleleSeq first uses genome variants (SNPs, indels, and deletions) to construct a 'personalized' diploid genome sequence, specific to the individual organism from which the samples were obtained. The pipeline then aligns RNA-seq reads, in parallel, to each of the diploid reference sets and examines reads overlapping heterozygous SNPs to statistically determine regions that exhibit ASE. Using generated WGS data (subaims 1.1-2) in the three species, we propose to perform an extremely broad analysis of ASE in the brain. Similar to the detection of ASE, it is also possible to study the phenomena of RNA editing. RNA editing occurs where RNAs undergo post-transcriptional modification. RNA editing events can be detected by comparing nucleotide changes in RNA sequences that are not present in the corresponding genomic DNA[106]. We propose to modify the AlleleSeq pipeline to incorporate automatic detection and reporting of RNA editing events to enable broad comparisons across the diverse sample-sets already generated and those proposed for this project.

In addition, we intend to apply and extend the AlleleSeq pipeline to the application of ASE of miRNAs and RBPs. ASE of miRNA expression is in principal straightforward to measure for those miRNAs that contain heterozygous SNPs either within the mature miRNA or the miRNA hairpin precursor. However, due to the short lengths (21-22nts) of miRNAs, care needs to be taken to correctly count RNA-Seq sequencing reads originating from each haplotype – to avoid miscounting reads that are not proportional to the expression of miRNA from each allele. One way this can be done is by incorporating custom barcodes into the sequencing library (to better measure library complexity for miRNAs) or integrating small variations in the sequence obtained. We also plan on investigating how ASE of miRNAs affects and correlates with ASE of their target mRNAs determined from matched mRNA-Seq data. We will also extend the AlleleSeq pipeline to RBPs using sequencing data from CLIP-seq. CLIP-seq data can be analyzed for allele-specific binding (ASB) in the same way as ChIP-seq data, making it possible to detect if there is preferential binding of a RBP to the mRNA (or any other ncRNA) expressed from one allele. As with ChIP-seq, the presence of a heterozygous SNP is required to differentiate between the sequences from the maternal and paternal haplotypes. One key difference is that, unlike for DNA (ASB from ChIP-seq data) where epigenetic effects are important to binding of a RBP to RNAs from either allele. This can either be due to the differences in ASE of the RNA itself or to differential binding (the actual affinity of the RBP to the target RNA molecule) of the RBP to the different RNA molecules expressed from each allele. Care must be taken to construct maternal and paternal sets of transcript annotation, since different RBP can bind either to the mature spliced mRNA or to the unspliced precursor. We also intend to correlate the ASB of RBPs from CLIP-seq with ASE from RNA-seq for the RBP-targeted mRNAs.

Extending AlleleSeq to integrate transcriptome output with proteomics: Further to the species- and tissue-specific transcriptomes, we will modify the AlleleSeq pipeline outlined above to produce subject-specific transcriptomes that allow whole-proteome level analysis and validation of any ASE, RDD, or variants detected via non-synonymous coding sequence variation obtained from the RNA- and DNA-level analyses. Moreover, peptide-level abundance differences between competing alleles, where available, will be assessed to provide a

quantitative, protein-level validation of the ASE detected at the RNA-level. Such an approach affords a very rare opportunity to perform an unbiased, whole-proteome, validation of putative novel ASE/RDD events. To the best of our knowledge such a methodical, large-scale, and integrated analysis of genomic variants, RNA expression, and protein abundance has never before been proposed and the software tools and protocols developed as part of the proposed work would likely be of large benefit to the wider community.

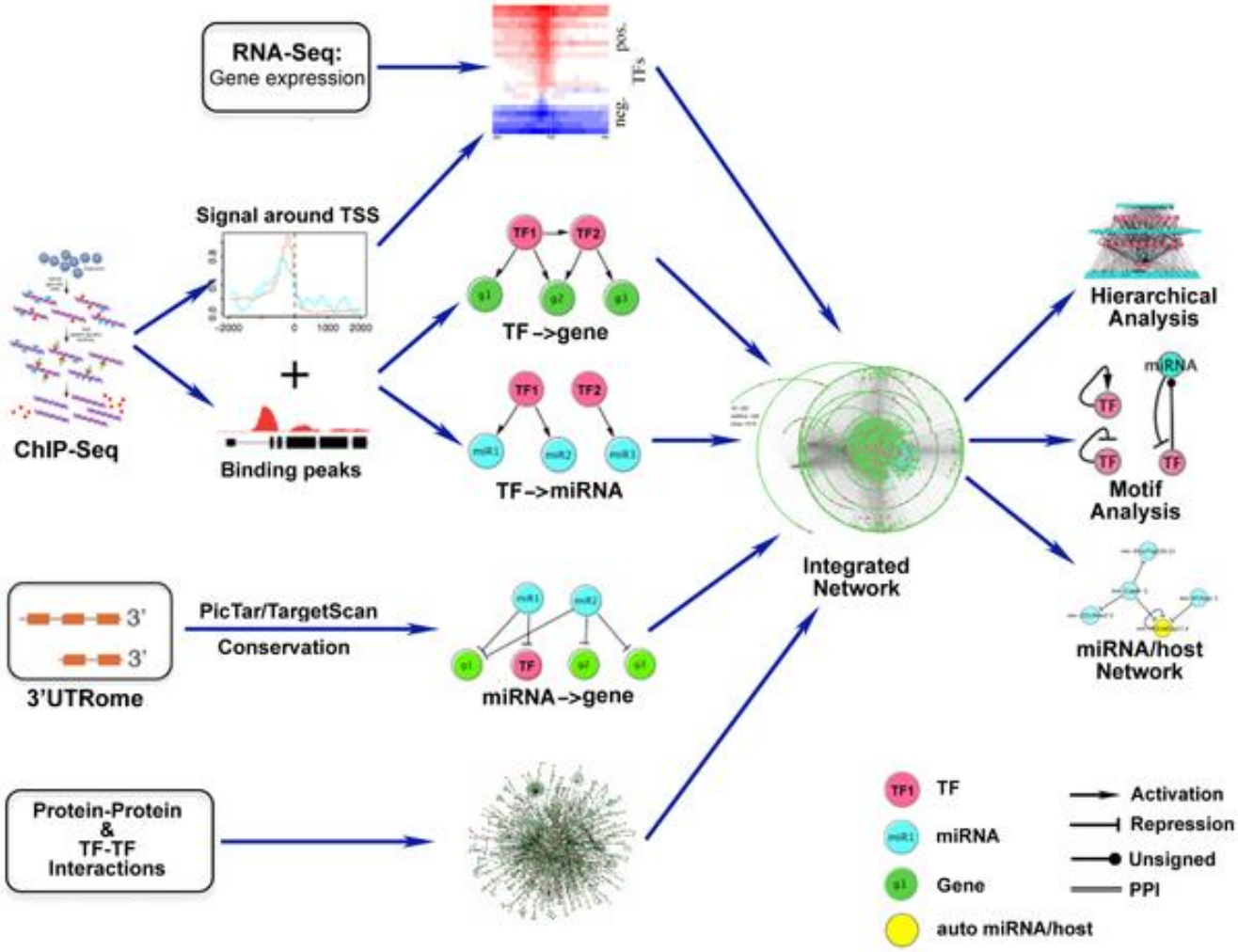

Figure 10. Schematic diagram of the construction and analysis of the integrative regulatory network. *ChIP-seq data were used to determine target genes and miRNAs of transcription factors. miRNA target genes were predicted using PicTar or TargetScan algorithms together with conservation information. The three types of regulations form the basic network. The sign of each regulatory interaction was determined based on the correlation between TF binding and gene expression, Extra edges of protein-protein or TF-TF combinatorial interactions were incorporated. We studied the topological structure of the integrated network, including hierarchical organization and motif enrichment.*

Construction and integrated analysis of regulatory networks: We constructed a probabilistic framework called TIP (Target Identification from Profile) for identifying regulatory targets (genes and miRNAs) of TFs based on ChIP-seq profiles[107]. To explore how biological regulation is carried out in multiple levels in addition to transcriptional regulation, we developed a computational framework to integrate various regulatory relationships such as transcriptional regulation, post-transcriptional regulation mediated by miRNAs, and post-translational regulation mediated by protein-protein interactions[107] (see Figure 10). These wiring diagrams can be served as maps for interpreting personal genome sequences and understanding basic principles of human biology and disease. In previous studies[18,108], we performed a variety of analyses on the topology of the regulatory networks. We found that TFs are organized in a hierarchical fashion in which information is transmitted via a cascade of regulation from upstream master regulators to downstream workhorses. We further examined the so-called network motifs in these integrated regulatory networks. We found co-regulations between TFs give rise to many enriched network motifs (for example, noise-buffering feed-forward loops), and we found several novel

composite motifs that include both TFs and miRNAs regulation. We plan to extend the motifs analysis in Gerstein et al.,[18] to become a machine-learning method that use the occurrence of motifs as features to predict genomics properties of TFs. A machine-learning approach goes beyond simple enrichment analysis and will make better use the high-dimensional network properties in understanding the functions of various TFs.

A wiring diagram offers a static picture on how components could interact. With the availability of the comprehensive battery of experimental data generated as part of this proposal, we plan to construct and analyze cell-type-specific, tissue-specific, and species-specific networks in order to shed light on the dynamics of such wiring diagram. In the past, we developed a computational approach for the statistical analysis of network dynamics (SANDY) that combine well-known global topological measures, local motifs and newly derived statistics[109]. SANDY was applied for microarray data in yeast to explore various environmental specific sub-networks. We plan to generalize SANDY to investigate tissue-specific sub-networks in multi-cellular organisms. In this specific case, dynamics across different types of neurons and evolutionary dynamics across three species will be explored.

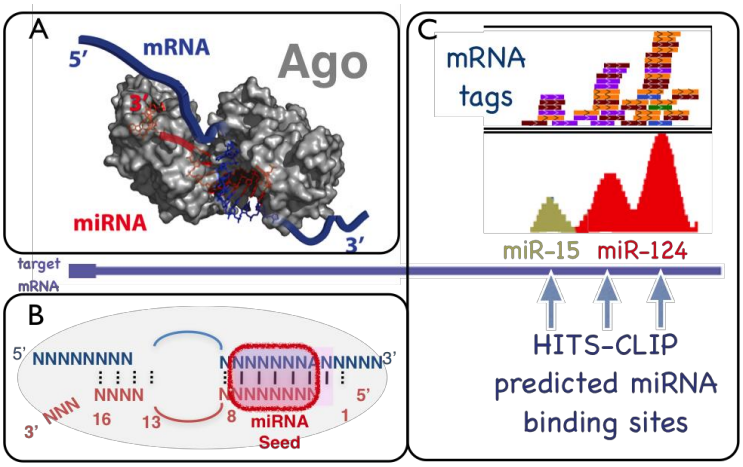

Figure 11. Ago HITS-CLIP. *(A) Ago binds miRNAs intimately, but also is in sufficient proximity to mRNA to crosslink to both RNA species. (B) miRNA seeds need to match only 6 nt of mRNA to function, making bioinformatic prediction of targets very difficult. (C) Ago mRNA tags, when aligned (each color is from a biologic replicate in mouse brain), form clusters that accurately predict a seed sequence in the center. These Ago-mRNA footprints identify genome-wide miRNA binding sites on target mRNAs.*

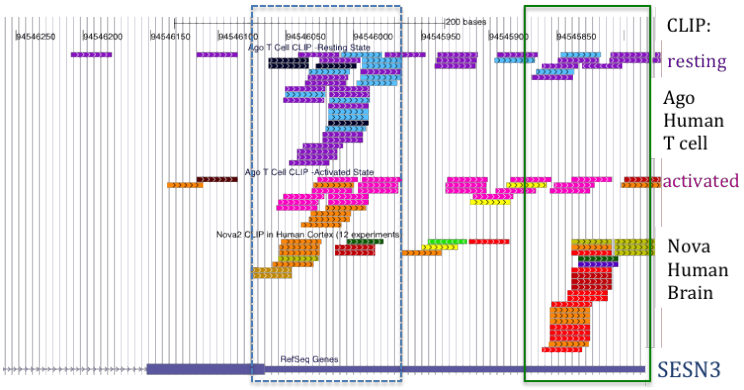

Figure 12. *Mouse Ago/ Nova HITS-CLIP maps overlayed on Itgb1 3' UTR (compare **Figure 11**), suggesting various potential agonistic or antagonistic interactions. Each color represents a CLIP tag from a different animal.*

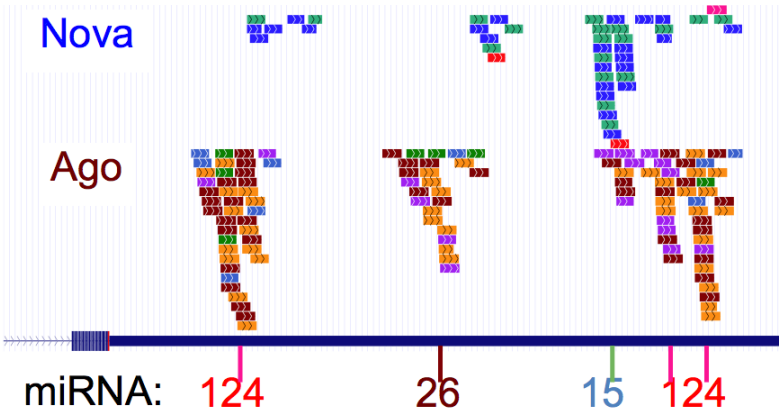

Figure 13. Nova and Ago HITS-CLIP in human cells. *Ago CLIP results are shown for resting vs. activated (1 hr) human T cells; Nova is a composite of frontal cortex CLIP from 12 samples. Dashed blue box indicates a site where Ago binding is decreased after T cell activation, and might overlap a Nova-regulated site in human fetal brain. Green box indicates a possible site where Nova regulates alternative polyA (an end of one 3' UTR is shown. Each color represents a different biologic sample.*

Development of an integrated analysis of RNA regulation: HITS-CLIP has become accepted as the preferred method to identify RNA-protein interaction sites *in vivo*[110,111]. We extended HITS-CLIP to study Ago-miRNA regulation[112], in which HITS-CLIP defined precise, genome-wide footprints where Ago-miRNA complexes crosslinked to brain mRNA (Figure 11). In preliminary data, we have succeeded in developing HITS-CLIP and mapping RNA tags in fresh human T cells, and in frozen samples of human brain (Figure 12). Darnell lab has developed a computational expertise in analysis of HITS-CLIP[112], in which we allied data with a Bayesian network, generating a comprehensive map of Nova splicing targets and revealing unsuspected functional interactions with the splicing factor RbFox1/2[113]. We also developed improvements in data analysis, allowing us to map Nova and Ago interaction sites with single nucleotide resolution[114]. We will apply these analysis tools to decipher differential regulation of RNA-protein interaction among the in each tissue condition (brain regions across developmental periods in 3 species) and in each cell type differentiated from iPS cell. Comparing the number of CLIP tags on a transcript between such samples or conditions requires normalization to transcript abundance. One of the advantages of the current joint project is normalization of that for CLIP, we will have deep RNA-seq available from paired brain samples. This data will provide the added advantage of assessing RNA variation (in levels, splicing, 3' UTR usage) under different conditions that can be directly compared with CLIP maps. For cell-specific studies this will be particularly important, as cell-cell variation may be anticipated to be extensive.

Integration between the tissue and iPSC HITS-CLIP experiments will be afforded by orthogonal studies being done in mouse single cell types (not part of this proposal; being done in the context of NIH NS34389; see Figure 13), using Cre-lox mediated generation of individual epitope tagged RBPs for single cell-type HITS-CLIP analysis in mouse brain. Such data will provide a bridge for bioinformatic overlay between the single cell-type advantage of iPS cell HITS-CLIP and the more complex but physiologically relevant analysis of human brain tissue.

Aim 3: Elucidation of common and cell type specific regulatory and molecular networks compromised in autism spectrum disorders (ASD).

Rationale: Recent genomics research into ASD has largely been restricted to the evaluation of coding mutations via WES. Although this strategy has identified high confidence *de novo* mutations clearly associated with risk, the majority of affected individuals, including in the Simons Simplex Collection (SSC) which we have extensively studied, are still without a known genetic contribution. Interestingly, recent ENCODE and GWAS studies show that many common disease-associated variations localize to cis-regulatory loci, implicating these elements in human disease[115]. Moreover, there are multiple lines of evidence suggesting that non-coding variation contributes to ASD risk via influencing the expression of protein-coding risk genes[116]. For example, the identification of associated copy number variations (CNV) strongly suggests that gene dosage is playing an important role in ASD[117]. In addition, a substantial portion of those genes most strongly associated with ASD in WES studies, have known functions as transcription factors and chromatin modifiers. Whole genome sequencing (WGS) which would allow for the analysis of the contribution of non-coding mutations is still cost prohibitive and limited by poor annotations of non-coding elements active in regions of developing human brain. In this Aim, we will overcome these obstacles by the targeted sequencing of cis-regulatory elements identified in Aims 1 and 2 in 250 ASD quartets from the SSC. High priority (see below) mutations will be subjected to functional analysis under Aim 4. This approach will provide an important complement to current WES efforts.

Subaim 3.1: Targeted re-sequencing of cis-regulatory elements in ASD. We will design a custom targeting library with non-coding RNAs selected from Aims 1 and 2, using NimbelGen's SeqCap Choice XL Library platform. This library will provide for high-performance capture of up to a total of 50 Mb of the non-coding genome. This should be sufficient to include the majority of target loci based on the existing human RNA-seq and ChIP-seq dataset. However, if regions identified under prior aims exceed this total, we will prioritize based on the following criteria: (1) loci associated with previously implicated ASD risk genes, (2) loci developmentally regulated in the neocortex, a brain region implicated in ASD pathology, and (3) human-specific loci. The workflow for Aim 3 is shown in Figure 14.

Sample selection: We will select 250 SSC quartets (including a proband, an unaffected sibling, and two unaffectged parents). We will restrict to Caucasian ancestry to simplify the interpretation of allele frequencies and select families that have not been found to carry either a *de novo* CNV or *de novo* loss of function mutations [64, 117]. We will divide the sample into two halves: 125 families will be chosen based on on criteria likely to increase the likelihood of identifying highly penetrant *de novo* variants, specifically: female proband, low proband IQ, multiple unaffected siblings, and advanced paternal age. The other 125 families will be selected at random to be representative of the SSC.

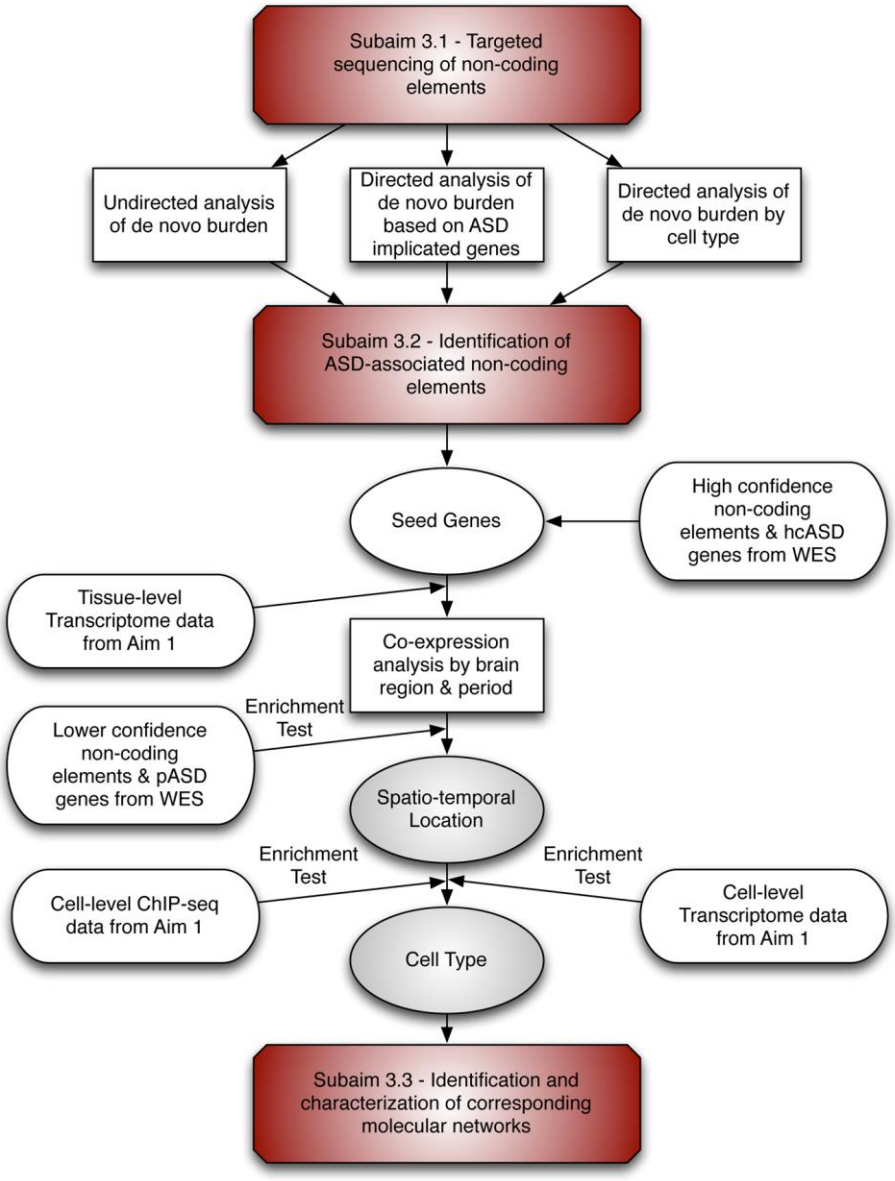

Figure 14. Aim 3 workflow.

Sequencing methods: We will hybridize whole blood derived DNA with the custom Nimblegen Libraries. The captured DNA will then be sequenced using Illumina HiSeq2000 instruments to generate 75 bp paired-end reads. Capture and sequencing are performed in pools of six barcoded samples to minimize costs while yielding consistently high coverage. All proband and sibling pairs are processed in the same batch.

Data quality, analysis and management: We will monitor data quality for all samples to ensure >30x mean coverage of non-duplicate reads and <1% error rate per base. All data is initially processed with Illumina's CASAVA pipeline. The data is aligned using BWA/Bowtie and stored in BAM format files; SNVs are predicted using GATK, and indels are predicted using Dindel. All variants are stored in Variant Calling Format (VCF) files. *De novo* variants are identified using the methods outlined in our preliminary data. The raw data, FASTQ files are stored on the high-speed parallel storage system of the Yale BulldogN high-performance cluster (HPC). We will purchase dedicated storage and computing nodes to facilitate rapid alignment and variant prediction.

Confirmation: We will confirm all *de novo* SNV and indel predictions using PCR and Sanger sequencing. We have achieved a 96% confirmation rate for *de novo* SNVs and a 94% confirmation rate for *de novo* indels[64].

Annotation: The absence of a triplet code in non-coding regions makes interpretation of the functional consequences of *de novo* variation harder to predict. Therefore, to allow functional interpretation of the *de novo* variants, we will use data from Aims 1 and 2 and publically available tools or databases, such as the Genomic Regions Enrichment of Annotations Tool (GREAT)[37] or Vista Browser[118].

Subaim 3.2: Characterization of mutation burden and identification of ASD-associated regulatory regions. We will use three complementary strategies to evaluated for an ASD-associated subset of *de novo* variants and then to prioritize regulatory features for functional analysis in Aim 4. As an initial test, we will evaluate whether there is a statistically significant increase overall in non-coding de novo mutations in cases vs controls. If present, this finding can be leveraged to establish the number of *de novo* mutations required in a given non-coding region to reach a genome-wide significance threshold[64]. However, the directed analyses described below provide an important alternative strategy that is not dependent on identifying an overall increased burden in cases vs. controls.

Undirected analysis of de novo burden: Genomics can generate data-driven hypotheses of ASD causation without imposing prior assumptions regarding biological mechanisms. For example, WES studies have uncovered the surprising contribution of chromatin regulation to ASD[64-67]. We will evaluate all non-coding de novo, mutations in cases and control in search of an increased rate in affected individuals, and then leverage this finding to identify specific regulatory loci for functional analysis. Importantly, as each regulatory locus is small (e.g. 300-1,000bp for an enhancer) there is limited power, compared to genic analyses, to observe multiple *de novo* mutations in a specific element. Therefore, we will group regulatory loci by target gene (determined by Aim1/2 data, GREAT or Vista Browser) to achieve detection power comparable to gene discovery efforts in ASD by WES[88].

Directed analysis of de novo burden based on genes already implicated in ASD: Regardless of the outcome of the un-directed analysis, we will test the hypothesis that *de novo* mutations in the regulatory machinery proximal to known ASD genes contributes to ASD risk. To do so, we will restrict our analysis to the burden of *de novo* non-coding mutations in cases vs. controls within the regulatory loci of 1) nine high-confidence ASD genes (with multiple recurrent *de novo* LoF mutations in independent probands) and 2) 116 probable ASD genes (with a single *de novo* LoF mutation)[64].

Directed analysis of de novo burden by cell and tissue type: We will group regulatory loci by cell and tissue type based on chromatin state marks and corresponding *cis*-regulatory loci identified in Aim 1, and assess the burden of non-coding variation by cell or tissue type. Enrichment of variation in particular cell types would implicate a subset of cell types in pathogenesis as well as indicate the relevant set of non-coding loci.

Subaim 3.3: Identification of corresponding regulatory and molecular networks and their role in ASD. Aim 1 will generate complementary maps of gene expression and non-coding *cis*-regulatory loci. Furthermore, Aim 3.1 and 3.2 will identify ASD-associated non-coding regions of the human genome. The goal of this subaim is therefore to integrate this information to achieve: 1) finer cellular resolution of ASD pathology, 2) increased understanding of the contribution of coding and non-coding loci to ASD, and 3) an actionable understanding of ASD biology. The biological relevance of gene co-expression networks is well supported[43,119], and since highly correlated gene expression reflects shared function and/or regulation, we anticipate that ASD-associated genes (preliminary data) and ASD-associated *cis*-regulatory elements (Aim 3.2) will coalesce in the same co-expression networks – either because these regulatory regions target ASD genes or are targets of other ASD genes (i.e., TFs like *TBR1* – preliminary data). Therefore, combining these two sets of data should strengthen our network-based co-expression analysis. Hence, similar to the approach outlined in the preliminary data, we will use high confidence (hc) ASD genes as seeds in a bottom-up co-expression analysis but will also use genes corresponding to non-coding regions implicated in ASD (Subaim 3.2) with similar confidence as seeds. We will generate co-expression networks by calculating pairwise Pearson correlation coefficients between each seed gene and all genes measured in the Aim 1 human RNA-seq data.

By restricting the Aim 1 RNA-seq data to smaller developmental subsets, brain regions or cell types, as shown in the preliminary analyses, we can achieve greater spatio-temporal resolution. Thus, using a similar sliding window analysis, we will construct co-expression networks and assess enrichment of other ASD risk genes, with the hypothesis that greater enrichment reflects more relevance to ASD etiology. We will identify overlap of these networks with probable ASD genes (preliminary data) and genes corresponding to non-coding regions with lower confidence association to ASD (Subaim 3.2). By permuting control networks, we can assess the significance of enrichment, identifying networks most important to ASD, and when and where these genes converge to influence ASD. In addition, using *in silico* analyses based on known position weight matrices (PWMs)[120-122], we will also determine the TFs that likely bind to ASD-associated *cis*-regulatory regions of genes present within the co-expression networks, and then similarly test for enrichment. We expect that co-expression networks enriched for ASD risk genes will also be enriched for these TFs, thus explaining phenotypic convergence of coding and non-coding mutations in ASD.

After identifying co-expression networks that meet these criteria, we will utilize the data generated in Aim 1 from single cell and iPS cell-derived RNA-seq and ChIP-seq experiments to identify enrichment of genes with expression patterns or chromatin state signatures characteristic of particular cell types. As an alternative approach, we can also assess correlations between network genes across transcriptome data from a particular cell type. We will calculate the connectivity (sum of pairwise correlations) of each co-expression network by cell type – the cell type with the highest network connectivity being the most relevant as more strongly correlated genes are expected to be more functionally related. The gene networks identified through this analysis will be highly informative for the underlying biology by telling us not only *when* and *where*, but also in *which* cell types specific subsets of genes contribute to ASD and other phenotypes. Regulatory regions mutated in ASD and highlighted in co-expression networks will be prioritized for follow-up in Aim 4.

Potential pitfalls: *1) Data comparability*: To ensure that differences in *de novo* mutation rate between probands and unaffected siblings are due to biological effects rather than sequencing bias we will: 1) sequence all proband and sibling pairs on the same Illumina flowcell to minimize variability; 2) concentrate the analysis on bases in which all four family members (both parents, proband and unaffected sibling) have adequate read coverage to allow highly accurate variant detection as demonstrated in our whole-exome research[64] and 3) compare the rate of variants with low expectation of carrying risk (e.g. variants in poorly conserved repetitive regions) between probands and unaffected siblings to demonstrate the absence of bias.
2) Ability to analyze data: We have expertise in dealing with large datasets, as exemplified by our recent publications in which we analyzed WES data from 928 individuals[64] and expression data from 1,340 samples of the developing human brain[43]. 3) ASD genes identified by coding and by non-coding variants may not converge in the same networks, regions of the brain, or periods of development. For example, genes identified by non-coding variants may be temporally disconnected from other ASD genes if they are downstream in a prolonged regulatory interaction. If our initial approach is unsuccessful, we will modify the approach to account for such dynamics or analyze these subsets of genes separately.

Alternate strategies: *1) WGS* would increase the completeness of our analysis of the non-coding elements, however there are several compelling reasons to opt for targeted sequencing: 1.1) WGS remains at least 4-fold more expensive for sequencing alone without considering the necessary investment in computing infrastructure to analyze the data; 1.2) The non-coding elements are harder to interpret than the exome, however they are orders of magnitude easier to interpret than genomic regions without known functional roles. By concentrating on the 1.5% of the regulatory machinery that this best annotated we maximize our ability to ask biologically relevant questions.
2) Case-control targeted sequencing: While a case control strategy would allow us to double the number of cases (500 instead of 250), ultimately this would decrease our power to detect meaningful variants. This is best illustrated by the recent work looking at exome analysis in ASD. While a *de novo* strategy was able to detect genome-wide significant association for *SCN2A* in 200 quartets[64], a larger study of 1,039 cases and 870 controls lacked the power to identify any novel ASD-associated genes[123].
3) Molecular inversion probes (MIPs): While MIPs are a highly cost-effective method of analyzing small numbers of genetic loci in large numbers of samples, it is less cost-effective and more labor intensive for a project of this scale. Our analysis shows the inflection point to be at a target of 6.5Mbp (with MIPs preferred below this figure and array-based capture above); thus non-coding element capture is the better choice.

Aim 4: Modeling and functional characterization of human-specific and disease-associated *cis*-regulatory elements.
Rationale: Many sequence analyses have identified non-coding regulatory regions of the human genome that appear unique to our species due to features such as accelerated nucleotide substitution[124-126] and human-specific deletions[37]. However, there has been very little success linking these regions, especially those with putative cis-regulatory function, with human-specific features of brain development, largely due to the difficulty of investigating functionality in these non-coding regions. Similarly, the role of disease variants in regulatory regions has remained difficult to investigate, in contrast to mutations in coding regions, where the deleterious impact can be partially inferred by their effect on the coding sequence..

In this aim, we will combine BAC recombineering and transgenesis to introduce a region of a human chromosome containing a gene and the surrounding regulatory elements into a mouse and functionally characterize putative human-specific (subaim 4.1) or ASD-related (subaim 4.2) regulatory loci, prioritized in Aim1 and Aim3, respectively, through site-specific mutagenesis. This approach has several advantages over the standard transgenic reporter assays. 1) BACs have been previously shown, most prominently by the GENSAT project[127], to faithfully recapitulates the endogenous expression pattern. 2) As recently demonstrated by the Sestan lab[128] the same technology can be used to identify a cis-regulatory element within a BAC that is necessary for the proper spatio-temporal expression of a gene within the BAC by a “negative” selection of candidate *cis*-regulatory sequences. 3) Finally, the entire human gene and protein will be expressed under its own regulatory control, which allow us to decipher possible functional consequences of misexpression of the human protein in the context of neurodevelopment.

Subaim 4.1: Putative human-specific regulatory element identification and functional characterization. First, we will prioritize genes with human-specific expression by the pattern of differential expression between brain regions and cells in Aim 1. We will give priority to genes exhibiting human-specific expression in neurons of the developing prefrontal cortex (e.g., DFC) due to the importance of DFC in human cognition and behavior. Next, we will consider genes for which a BAC containing the entire gene locus is available. Finally, we will only consider those BACs that also have human-specific ChIP-seq peaks based histone marks (H3K4me3, H3K27ac, and H3K27me3) and CTCF, which identify a large fraction of active enhancers, promoters, and repressors, as well as insulators. To identify putative upstream TFs, we will select for further analysis TFs with an available PWM[120-122] that are within the top 50 most correlated genes in the developing human brain transcriptome[43] based on gene co-expression analyses outlined in Aim3. This list will serve as the starting point for ranking the genes with human-specific expression for further functional studies. We will score genes by the following criteria: 1) co-expression correlation of TF (maximum if more than one TF); 2) the gene's developmental and cellular expression profile in the human brain; 3) proximity and number of previously identified human-specific accelerated or lost elements; and 4) quality of the NHP sequence containing the gene.

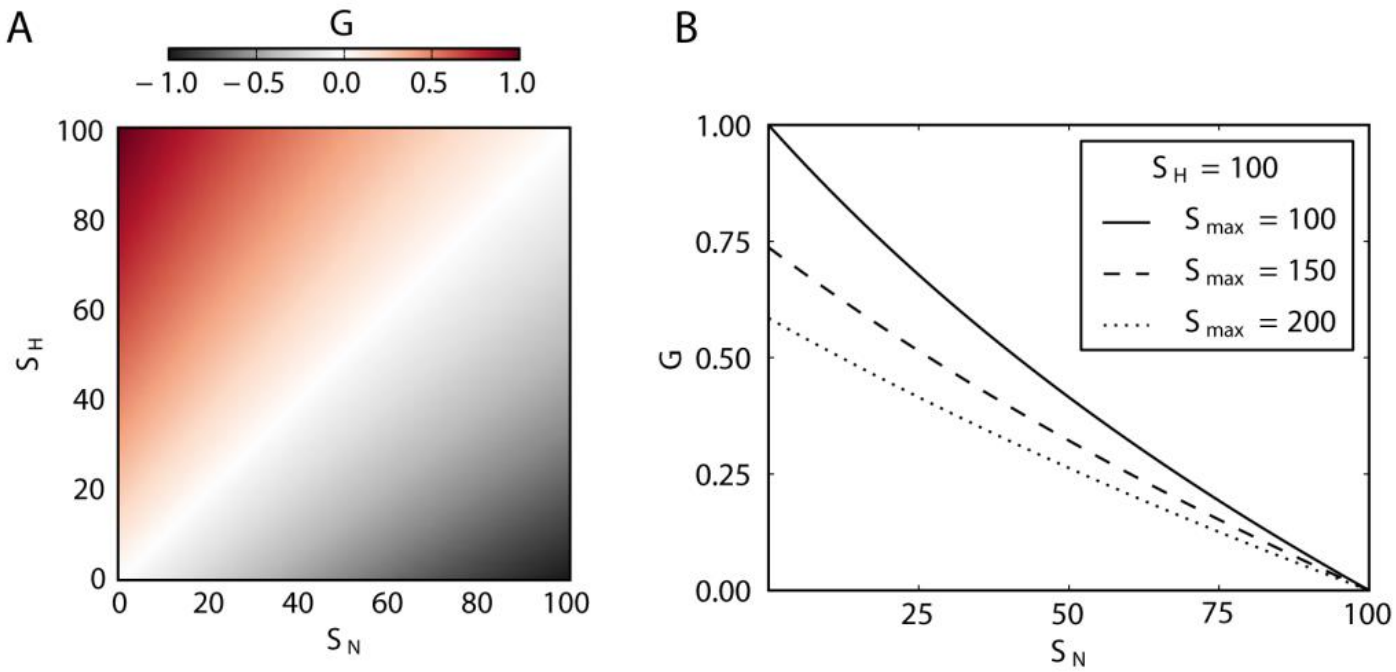

Figure 15. Transcription factor (TF) site gain metric. *(**A**) The value of G increases as the score of the human binding site score (SH) increases and the NHP binding site score (SN) decreases. (**B**) As the distance between the maximum possible score (S_{max}) and SH increases, G decreases.* $G = \mathbf{log}_2\left(\frac{S_{max}+S_H}{S_{max}+S_N}\right)$

Ranking human-specific regulatory sites: Using the list of genes from above, we will scan the sequence within and flanking the gene for PWMs of the top-correlated TFs. Multiple alignments will define human-specific changes (SNPs, human regions with accelerated nucleotide substitutions (HARs), or human-specific deletions) within these putative binding sites. Human-specific mutations are ranked by the impact on the binding site prediction. The impact of the human-specific mutation is calculated (Figure 15) where SH is the PWM score for the human sequence, SN is the average PWM matrix score for the NHP sequences and Smax is the maximum score possible to the TF. This metric preferentially scores sites that have a large difference between the human and NHP binding site predictions and favors sites where the human score is close to the maximum possible scores (Figure 15). The binding site with the highest gain metric (G) will be prioritized for functional characterization. Also, we will only consider sites that appear fixed in humans. Ranked sites are filtered to remove variants that appear in command dbSNP or Aim 3 (ASD), as well as those that are variable in our NHP and human sequencing data.

BAC prioritization: We will use the following criteria to prioritize human BACs for functional analyses: 1) BACs containing an entire gene exhibiting human-specific gene and protein expression pattern in the developing and/or adult prefrontal neocortex; 2) BACs with an overlapping human-specific sequence and 3) human-specific ChIP-seq signature. We will then generate transgenic mice using prioritized BACs and functionally characterize them using methods described below (see also[128]). We plan to functionally characterize at least 30 putative human-specific regulatory elements over 5 years using this approach. Based on the preliminary analyses and criteria 1 and 2 outlined above we have identified hundreds of candidate human BACs. We will further prioritize these candidates once we generate NHP ChIP-seq, RNA-seq and proteomic data in Aim 1.

Subaim 4.2: Modeling the effects of ASD-associated regulatory mutations on gene regulation in BAC transgenic mice. We will evaluate the effects of ASD-associated mutations in regulatory elements identified and prioritized in Aim 3 using BAC transgenesis. We expect to identify and prioritize at least 10 ASD-associated regulatory elements. First, we will establish in a transgenic mouse whether a control human BAC clone containing the same regulatory sequences found in normal populations and in unaffected family members is sufficient to drive the expression of the harboring gene in a manner similar to the human pattern. We will then

engineer these BAC clones to create ASD-associated mutations by homologous recombination and create transgenic mice harboring these mutations, as we have previously described[128]. We will examine these mice at different time points during embryonic and/or postnatal development, depending on when the control human BAC gene is expressed, to determine whether regulatory mutations lead to the loss or gain (including neomorphic) of gene expression. Furthermore, we will also identify TFs binding to the ASD-mutated using the approach described in subaim 4.1 and functionally characterized as described below. Depending on the functional outcome of a regulatory mutation, we will follow up with studies of loss or gain-function experiments involving genes controlled by the element or upstream TFs whose binding site is affected by the mutation.

Generation of transgenic mice carrying human BACs: Because BACs can include up to 300 kb of genomic DNA, it is very likely that there are BAC clones containing entire coding and regulatory regions. To increase the efficiency of transgenesis, and to preserve the intactness of integrated BACs (obtained from CHORI), a mixture of the transposon-containing BAC and transposase mRNA is injected into the pronucleus and cytoplasm of fertilized mouse eggs as previously described[129,130]. Injected eggs are transferred to the ampulla of foster females. BAC integration into the genome is checked by PCR and amplicon sequencing using human-specific primer sets. To test whether BAC integration into the genome is sufficient to drive the human expression pattern in the mouse brain, the spatio-temporal expression patterns of the human BAC gene and the endogenous mouse ortholog is confirmed by *in situ* hybridization using human-, and mouse-specific DIG-RNA probes. Those that can successfully drive the human expression pattern, and in particular those that differ from endogenous mouse orthologous transcript will be selected for sequence mutagenesis.

Human BAC-recombineering to mutate putative TF binding site sequence: After confirming the human pattern of expression after BAC integration described above, we will interrogate the putative regulatory elements. For this purpose, predicted TF binding sites are replaced by chimpanzee sequence or the ASD-associated variant using a homologous recombination-based method[127,131], and then the changes in expression pattern will be assessed in BAC transgenic mice. Targeted analyses of TFs binding to the affected cis-regulatory loci, genes controlled by them, and their roles in brain development will complement the above experiments exploiting well established methods in use by our teams, and rapidly developing genomic technologies.

Potential pitfalls and alternatives: The proposed techniques are well established in the Sestan lab, which has its own transgenic facility. It is possible that targeted human TFs are unique to primates or humans[132-134], have unique functional properties due to amino acid changes that lead to human-specific transcriptional activities[135,136], or have expression patterns different from mouse orthologs[42,43,48]. In such cases, human TFs would be knocked-down in human cells instead of mice, using RNAi in iPS cells expressing human BACs. The main issue, however, is the number of regulatory elements to be analyzed. Based on our preliminary analyses of human and NHP data and previous studies of human-specific genomic changes, we already have a large list of potential candidates. We will expand but also prioritize the list of potential candidates based on experiments outlined in Aim 1. It is less certain how many regulatory loci will be linked to ASD in Aim 3, but a number of sequence or CNVs within non-coding regions are already implicated by GWAS or CNV studies in ASD[117,137,138]. Our ChIP-seq experiments in Aim 1 will help determine which of these are active in the developing human brain; we could then functionally characterize these using the above described BAC approach. Finally, we can also rely on reverse experiments in which we can prioritize putative cis-regulatory elements based on their proximity to hcASD genes identified in WES studies[64,139,140].

Aim 5: Creation of an infrastructure to facilitate research and teaching activities. Research efforts described in the above Aims will be coupled with robust data and resource sharing opportunities (subaim 5.1) as well as training and outreach programs in basic and translational neurogenomics (subaim 5.2).

Subaim 5.1. Data and resource sharing opportunities. We will create an infrastructure to facilitate internal and external research activities, by disseminating our bio-specimens (i.e., tissues, single cells preparations, iPS cells), tools, methods, protocols, and data to the community as well as by soliciting outside ideas through our **www.neuroCEGS.org** website. Additional information is provided in Management and Training Core and Resource Sharing Plan.

Subaim 5.2. Training program and opportunities. We will also establish training and outreach programs in basic and translational neurogenomics. Through this program we will create new opportunities to considerably expand the pool of underrepresented minorities in the contemporary fields of genomics, computational biology and neurobiology. Our CEGS training plan is described in Management and Training Core and our CEGS Diversity Action Plan is described in the R25 proposal.

1 Barak, T. *et al.* Recessive LAMC3 mutations cause malformations of occipital cortical development. *Nat Genet* **43**, 590-594 (2011).

2 Liao, B. Y. & Zhang, J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**, 6987-6992 (2008).

3 Varki, N. M. *et al.* Biomedical differences between human and nonhuman hominids: potential roles for uniquely human aspects of sialic acid biology. *Annual review of pathology* **6**, 365-393 (2011).

4 Soon, W. W. *et al.* High-throughput sequencing for biology and medicine. *Mol Syst Biol* **9** (2013).

5 Mortazavi, A. *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628 (2008).

6 Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349 (2008).

7 Wang, Z. *et al.* RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).

8 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).

9 Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods* **4**, 651-657 (2007).

10 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-U552 (2007).

11 Johnson, D. S. *et al.* Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502 (2007).

12 Cox, J. & Mann, M. Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology. *Annu Rev Biochem* **80**, 273-299 (2011).

13 Craft, G. E. *et al.* Recent advances in quantitative neuroproteomics. *Methods* **61**, 186-218 (2013).

14 Mann, M. *et al.* The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. *Mol Cell* **49**, 583-590 (2013).

15 Geiger, T. *et al.* Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Mol Cell Proteomics* **11** (2012).

16 Feingold, E. A. *et al.* The ENCODE (ENCyclopedia of DNA elements) Project. *Science* **306**, 636-640 (2004).

17 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

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

19 Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190-1195 (2012).

20 Williamson, I. *et al.* Enhancers: From Developmental Genetics to the Genetics of Common Human Disease. *Dev Cell* **21**, 17-19 (2011).

21 Evrony, G. D. *et al.* Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. *Cell* **151**, 483-496 (2012).

22 O'Huallachain, M. *et al.* Extensive genetic variation in somatic human tissues. *Proc Natl Acad Sci U S A* **109**, 18018-18023 (2012).

23 Frumkin, D. *et al.* Genomic variability within an organism exposes its cell lineage tree. *PLoS Comp Biol***1**, 382-394 (2005).

24 Gage, F. H. & Muotri, A. R. What Makes Each Brain Unique. *Sci Am* **306**, 26-31 (2012).

25 Garcia-Perez, J. L. *et al.* LINE-1 retrotransposition in human embryonic stem cells. *Hum Mol Genet* **16**, 1569-1577 (2007).
26 Abyzov, A. *et al.* Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* **492**, 438-+ (2012).
27 Okaty, B. W. *et al.* Cell type-specific transcriptomics in the brain. *J Neurosci* **31**, 6939-6943 (2011).
28 Hashimshony, T. *et al.* CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports* **2**, 666-673 (2012).
29 Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**, 1160-1167 (2011).
30 Ramskold, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777-782 (2012).
31 Sasagawa, Y. *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA-Seq reveals non-genetic gene expression heterogeneity. *Genome Biol* **14**, R31 (2013).
32 Pan, X. *et al.* Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc Natl Acad Sci U S A* **110**, 594-599 (2013).
33 Eberwine, J. & Bartfai, T. Single cell transcriptomics of hypothalamic warm sensitive neurons that control core body temperature and fever response Signaling asymmetry and an extension of chemical neuroanatomy. *Pharmacol Therapuet* **129**, 241-259 (2011).
34 Spaethling, J. M. & Eberwine, J. H. Single-cell transcriptomics for drug target discovery. *Curr Opin Pharmacol* (2013).
35 Qiu, S. *et al.* Single-neuron RNA-Seq: technical feasibility and reproducibility. *Frontiers in genetics* **3**, 124 (2012).
36 Azevedo, F. A. C. *et al.* Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-Up Primate Brain. *J Comp Neurol* **513**, 532-541 (2009).
37 McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216-219 (2011).
38 Charrier, C. *et al.* Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**, 923-935 (2012).
39 Dennis, M. Y. *et al.* Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**, 912-922 (2012).
40 King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).
41 Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391-399 (2012).
42 Johnson, M. B. *et al.* Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62**, 494-509 (2009).
43 Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489 (2011).
44 Lipovich, L. *et al.* Developmental Changes in the Transcriptome of Human Cerebral Cortex Tissue: Long Noncoding RNA Transcripts. *Cereb Cortex* (2013).
45 Shulha, H. P. *et al.* Coordinated cell type-specific epigenetic remodeling in prefrontal cortex begins before birth and continues into early adulthood. *PLoS genetics* **9**, e1003433 (2013).
46 Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* **23**, 555-567 (2013).

47 Visel, A. *et al.* A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**, 895-908 (2013).

48 Zeng, H. *et al.* Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* **149**, 483-496 (2012).

49 Altevogt, B. M. *et al.* Research agenda. Guiding limited use of chimpanzees in research. *Science* **335**, 41-42 (2012).

50 Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**, 206-216 (2007).

51 Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* **134**, 25-36 (2008).

52 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858 (2009).

53 Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature* **468**, 911-920 (2010).

54 Noonan, J. P. & McCallion, A. S. Genomics of Long-Range Regulatory Elements. *Annu Rev Genom Hum G* **11**, 1-24 (2010).

55 Bilguvar, K. *et al.* Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467**, 207-210 (2010).

56 Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106**, 19096-19101 (2009).

57 Johnston, J. J. *et al.* Massively Parallel Sequencing of Exons on the X Chromosome Identifies RBM10 as the Gene that Causes a Syndromic Form of Cleft Palate. *Am J Hum Genet* **86**, 743-748 (2010).

58 Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-U41 (2010).

59 Pierce, S. B. *et al.* Mutations in the DBP-Deficiency Protein HSD17B4 Cause Ovarian Dysgenesis, Hearing Loss, and Ataxia of Perrault Syndrome. *Am J Hum Genet* **87**, 282-288 (2010).

60 Berman, B. P. *et al.* Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol* **5**, R61 (2004).

61 Hong, J. W. *et al.* Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314-1314 (2008).

62 Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Gen Mol Biol* **4**, Article17 (2005).

63 Kwan, K. Y. *et al.* Species-dependent posttranscriptional regulation of NOS1 by FMRP in the developing cerebral cortex. *Cell* **149**, 899-911 (2012).

64 Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012).

65 Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012).

66 O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* **43**, 585-589 (2011).

67 O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).

68 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012).

69 Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-299 (2012).
70 Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-872 (2007).
71 Clancy, B. *et al.* Web-based method for translating neurodevelopment from laboratory species to humans. *Neuroinformatics* **5**, 79-94 (2007).
72 Okita, K. *et al.* A more efficient method to generate integration-free human iPS cells. *Nature methods* **8**, 409-412 (2011).
73 Espuny-Camacho, I. *et al.* Pyramidal neurons derived from human pluripotent stem cells integrate efficiently into mouse brain circuits in vivo. *Neuron* **77**, 440-456 (2013).
74 Shi, Y. *et al.* Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. *Nature Neurosci* **15**, 477-486, S471 (2012).
75 Ingolia, N. T. *et al.* Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223 (2009).
76 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003).
77 Patel, V. J. *et al.* A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *Journal of proteome research* **8**, 3752-3759 (2009).
78 Mann, M. Comparative analysis to guide quality improvements in proteomics. *Nature methods* **6**, 717-719 (2009).
79 Bantscheff, M. *et al.* Quantitative mass spectrometry in proteomics: a critical review. *Analyt Bioanalytical Chem* **389**, 1017-1031 (2007).
80 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337-342 (2011).
81 Newman, J. R. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840-846 (2006).
82 Pan, S. *et al.* Mass spectrometry based targeted protein quantification: methods and applications. *Journal of proteome research* **8**, 787-797 (2009).
83 Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312**, 212-217 (2006).
84 Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature methods* **9**, 555-566 (2012).
85 Van Gelder, R. N. *et al.* Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc Natl Acad Sci U S A* **87**, 1663-1667 (1990).
86 Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A* **89**, 3010-3014 (1992).
87 Kamme, F. *et al.* Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *J Neurosci* **23**, 3607-3615 (2003).
88 Kurimoto, K. *et al.* An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucl Acid Res* **34**, e42 (2006).
89 Matz, M. *et al.* Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucl Acid Res* **27**, 1558-1560 (1999).
90 Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236-240 (2013).
91 Dewey, F. E. *et al.* Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS genetics* **7**, e1002280 (2011).

92 Greenbaum, D. *et al.* Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* **4**, 117 (2003).
93 Griffin, T. J. *et al.* Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Mol Cell Proteomics* **1**, 323-333 (2002).
94 Bitton, D. A. *et al.* Exon level integration of proteomics and microarray data. *BMC bioinformatics* **9**, 118 (2008).
95 Bitton, D. A. *et al.* An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. *PloS one* **5**, e8949 (2010).
96 Zimmer, J. S. *et al.* Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spec Rev***25**, 450-482 (2006).
97 Du, J. *et al.* IQSeq: integrated isoform quantification analysis based on next-generation sequencing. *PloS one* **7**, e29175 (2012).
98 Cotney, J. *et al.* Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res* **22**, 1069-1080 (2012).
99 Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-1038 (2010).
100 Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534-538 (2008).
101 Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539-543 (2008).
102 Zheng, D. *et al.* Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res* **17**, 839-851 (2007).
103 Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol* **13**, R51 (2012).
104 Lu, Z. J. *et al.* Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res* **21**, 276-285 (2011).
105 Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).
106 Park, E. *et al.* RNA editing in the human ENCODE RNA-seq data. *Genome Res* **22**, 1626-1633 (2012).
107 Cheng, C. *et al.* TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* **27**, 3221-3227 (2011).
108 Cheng, C. *et al.* Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comp Biol***7**, e1002190 (2011).
109 Luscombe, N. M. *et al.* Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308-312 (2004).
110 Chen, M. & Manley, J. L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Bio* **10**, 741-754 (2009).
111 Sharp, P. A. The Centrality of RNA. *Cell* **136**, 577-580 (2009).
112 Chi, S. W. *et al.* Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**, 479-486 (2009).
113 Zhang, C. L. *et al.* Integrative Modeling Defines the Nova Splicing-Regulatory Network and Its Combinatorial Controls. *Science* **329**, 439-443 (2010).
114 Zhang, C. L. & Darnell, R. B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol* **29**, 607-U686 (2011).

115 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).
116 Judson, M. C. *et al.* A new synaptic player leading to autism risk: Met receptor tyrosine kinase. *Journal of neurodevelopmental disorders* **3**, 282-292 (2011).
117 Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011).
118 Visel, A. *et al.* VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucl Acid Res* **35**, D88-92 (2007).
119 Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**, 218-223 (2009).
120 Ben-Gal, I. *et al.* Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* **21**, 2657-2666 (2005).
121 Jolma, A. *et al.* DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327-339 (2013).
122 Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucl Acid Res* **34**, D108-110 (2006).
123 Lui, J. H. *et al.* Development and evolution of the human neocortex. *Cell* **146**, 18-36 (2011).
124 Bird, C. P. *et al.* Fast-evolving noncoding sequences in the human genome. *Genome Biol* **8** (2007).
125 Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167-172 (2006).
126 Prabhakar, S. *et al.* Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786-786 (2006).
127 Gong, S. C. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917-925 (2003).
128 Shim, S. B. *et al.* Cis-regulatory control of corticospinal system development and evolution. *Nature* **486**, 74-U177 (2012).
129 Suster, M. L. *et al.* Transposon-mediated BAC transgenesis in zebrafish and mice. *Bmc Genomics* **10** (2009).
130 Rostovskaya, M. *et al.* Transposon mediated BAC transgenesis via pronuclear injection of mouse zygotes. *Genesis* **51**, 135-141 (2013).
131 Lee, E. C. *et al.* A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. *Genomics* **73**, 56-65 (2001).
132 Nowick, K. *et al.* Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc Natl Acad Sci U S A* **106**, 22358-22363 (2009).
133 O'Bleness, M. *et al.* Evolution of genetic and genomic features unique to the human lineage. *Nat Rev Genet* **13**, 853-866 (2012).
134 Vaquerizas, J. M. *et al.* A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252-263 (2009).
135 Enard, W. *et al.* A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell* **137**, 961-971 (2009).
136 Konopka, G. *et al.* Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462**, 213-217 (2009).
137 Weiss, L. A. *et al.* A genome-wide linkage and association scan reveals novel loci for autism. *Nature* **461**, 802-808 (2009).

138 Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886-897 (2011).
139 Eichler, E. E. & Zimmerman, A. W. A hot spot of genetic instability in autism. *New England J Med* **358**, 737-739 (2008).
140 Lim, E. T. *et al.* Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77**, 235-242 (2013).