Systems genomics approaches in neurologic disease

Jocelynn Renee Pearl

A dissertation
submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading committee:
Nathan D. Price, Chair
Leroy E. Hood
John A. Stamatoyannopoulos
Jay A. Shendure

Program authorized to offer degree:
Molecular and Cellular Biology

University of Washington

Abstract

Systems genomics approaches in neurologic disease

Jocelynn Renee Pearl

Chair of the Supervisory Committee:
Professor Nathan Price
Bioengineering

Neurologic disorders encompass a broad range of diseases including neurodegenerative (Huntington's disease, Parkinson's disease, Alzheimer's), neurodevelopmental (Autism, Rett syndrome), and psychiatric or mental disorders (Schizophrenia, bipolar disorder). Changes in brain gene expression accompany many of these disorders as demonstrated in studies of human post-mortem tissue. A critical objective in our understanding of gene misregulation in neurologic diseases, which range in heritability, is a comprehensive characterization of the spatial and temporal dynamics of the associated changes and how gene regulatory drivers mediate them. In this work, I explore early gene expression changes in a longitudinal study of Huntington's disease (HD) mouse models, and survey gene networks enriched for differential gene expression. I go on to investigate the contributions of sequence-specific transcription factors (TFs) to disease-specific gene expression change in HD and psychiatric disorders. I begin with a genome-scale model for TF-target gene interactions by combining publicly available DNase-seq footprinting and brain transcriptomic datasets. Using this transcriptional regulatory network (TRN), we identified TFs whose predicted target genes were overrepresented among differentially expressed genes in neurologic disorders. Following the identification of these predicted driver TFs, I applied multiple functional genomics approaches to characterize their genome-wide binding sites (ChIP-seq), survey the

impact of TF overexpression or knockdown (overexpression or CRISPR-Cas9-mediated editing), and assess the functional consequences of variation present in a motif instance (luciferase reporter assay). Together the preliminary findings from these studies further our understanding of the functional networks of genes and TFs implicated in neurologic disease and provide a methodological framework for future applications beyond the diseases covered in this thesis.

Table of Contents

1 Acknowledgements

First, I would like to thank my advisors, Lee Hood and Nathan Price. Lee and Nathan supported every direction I wanted my research to go, which was not always linear, and did not generally fit with what their labs focus on. They personally funded my experimental projects despite the fact that I essentially joined a computational study. In every interaction I had with them, they were positive and optimistic about what it was that I wanted to do. It was this confidence in my individual pursuits that helped me build independent studies and develop myself as an independent investigator.

Secondly I would like to thank Seth Ament, who was my project mentor through the majority of my work during graduate school. Part of why I joined the Institute for Systems Biology was that I was always having great scientific discussions with Seth. Seth constantly pushed me to be a better scientific writer and to be a more rigorous scientist. He was always pushing me to think harder about the problem. I benefitted immensely from his feedback on almost all of my projects. More than that, I think some really important science happened in the form of our discussions.

There have been many other mentors, collaborators, and advice-givers over the years that I would like to thank. Jeff Carroll and his lab have always been generous with their time and mouse tissues. Vanessa Wheeler and Marcy MacDonald have done more than their fair share of edits on both proposals and publications. Pete Skene helped me develop a ChIP-seq protocol at ISB, and generously advised me on science, life, and bird dog training. Chad Toledo from the Paddison Lab assisted me with human neural stem cell work and getting that working at ISB. There are so many others who helped in innumerous ways.

The MCB Department at the University of Washington has provided tremendous support through the years. I never felt alone and there was always someone there to help – thank you Michelle, Mary Ellen, Nomi, and Maia. Thank you to my committee – Jay Shendure, John Stamatoyannopoulos, Linda Buck, and Naeha Subramanian for balancing tricky schedules and pushing me to be a better scientist. I've learned a lot from all of you. John's support especially – I will always be grateful for your guidance during my first rotation and helping me write my first grant.

The Price and Hood Labs; it's been a pleasure working with such an amazing group of individuals, creative thinkers, and systems biologists. I won't name everyone here, but I appreciate all the time we spent together. I feel so lucky to have worked with Dani for almost two years. We got a lot of great science done together and had some fun doing it. Ali, you saved me when I was the only girl in the lab, and we've made some great memories (Bitterroot Science Festival). Martin thanks for always making me laugh (mostly at myself). John Earls, thanks for teaching me computational biology and always knowing exactly what I'm thinking.

Fyodor Urnov once told me that graduate school was something like a loaf of bread baking in an oven (the grad student being the bread). I don't think he was far off. His mentorship before graduate school and throughout has been invaluable.

Lastly, I must thank my 'core network' of family and friends who provide near-constant support through the good and bad days. Thank you from the bottom of my heart to Jenn, Megan, Natasha, Samantha, and Vijay. A special thanks to my "forever roommate" Laura – I got so lucky living with you the first few years of grad school. I hope I was able to offer a portion of the support you offered me on a day-to-day basis. Jennifer Cherone, despite

living across the country, you never felt far away. Our phone calls and messages made the worst weeks tolerable, and I'm glad we always felt comfortable enough to show up in lab together when one of us was on vacation. The Pam Planner really sealed the deal on me finishing things up this year.

I know I'll never have enough words or space to truly communicate how thankful I am for all that my parents, Jeff and AJ, have done for me these past 28 years. I really, really could not have done it without them. They've provided countless hours of emotional support, in person or over the phone. I aspire daily to provide others in my life with the gracious generosity and love you have shown me through everything. I dedicate this thesis to them.

This work was supported by a National Science Foundation Graduate Research Fellowship, as well as a contract from the CHDI Foundation.

2 Introduction

2.1 Summary

The scientific approach to understanding human disease has undergone several important shifts in recent years, assisted by the development of high-throughput sequencing techniques and the generation of large datasets. One of these major shifts is a movement away from single-hypothesis-driven research and reductionism towards a systems biology approach where the integration of multi-scale information and often global-genomic information allows the researcher to arrive at multiple hypotheses or network hypotheses about what drives disease. A second major shift is the incorporation of longitudinal data in order to capture subtle, early changes such as the shift from a healthy state into disease. It is these two major themes, which are incorporated into my approach towards understanding neurologic disease. Here I present an analysis of a dense time-series or longitudinal study of the Huntington's disease mutation, as well as several studies to understand gene regulatory drivers and networks contributing to gene expression changes in neurologic disorders. I employ both computational and experimental techniques for my study of transcription factor drivers and regulatory elements that I hypothesize contribute to disease. I discuss algorithmic strategies for understanding gene networks that contribute to complex gene expression changes and present experimental pipelines for validating hypotheses uncovered from *in silico* predictions. Lastly, I make an effort to frame systems-biology derived hypotheses from integrated datasets within both novel and known biology towards building a better understanding of neurologic disease mechanisms.

2.2 Background

Gene expression programs

There are approximately 20,000 protein-coding genes and a diverse set of additional non-coding RNA species encoded by the human genome (1). The regulatory state of these protein-coding and non-coding RNAs establishes specific cellular states. Misregulation of gene expression can lead to altered cellular states and cause disease. Misregulation can be caused by a diverse set of mechanisms. This work will cover genetic changes to transcription factor binding sites, the genes that encode transcription factors, gene networks, and transcription factor networks.

Gene networks

An individual gene is an incredibly small portion of the system or gene expression program of a cell. Genetic networks and pathways work concordantly to express groups of proteins that carry out cellular functions such as metabolism, signaling, and transcription. Parsing the gene expression program into coherent networks is an important part of systems genomics, with the goal of allowing us to understand groups of genes that carry out similar functions. We can then carry over this understanding into how groups of genes become misregulated in disease and how this might impact particular cellular functions.

Gene network analysis requires three components:

1. Transcriptomic data (microarray, RNA-seq)
2. Gene set annotations, such as those from KEGG or Gene Ontology
3. Analysis method or algorithm.

There are a wide variety of gene set annotations one can utilize for their analysis, in addition to generating their own *de novo* gene sets such as applied in Chapters 4 and 5. The Molecular Signatures Database, or MSigDB (http://software.broadinstitute.org/gsea/msigdb) offers an

aggregated collection of 8 major annotation sources for gene sets. As an aggregation of many different sources of gene sets, it is important to be selective in which gene sets work best for the question you are asking of your data. Reducing the number of gene sets tested will reduce the number of tests you must correct. It is also important to note that there are many different kinds of gene sets located here; some of these, such as metabolic pathways in KEGG (2), can be mapped out on a grid as part of a metabolic reaction pathway that has been biochemically defined. Other pathways are not biochemically defined but instead are based on data-driven observations such as coexpression. It is important to note the size of gene sets; many are quite large and the biological coherence of these genes can be more convoluted.

Following the selection of gene sets, one should choose a method for analysis. One straightforward statistical approach is to simply test for the enrichment of differentially expressed genes (DEGs) within each gene set. This can be accomplished, for example, using Fisher's exact test. Directionality of genes within a gene set, in this analysis, can be achieved by dividing DEGs into those with upregulated or downregulated expression as compared to controls and tested separately. In addition to this analysis, there are several publicly available algorithms which can be applied and provide different information, or 'network topology', about how gene sets are changing. One commonly applied method is Gene Set Enrichment Analysis, or GSEA, which determines whether a set of genes is statistically significant for concordant differences between two biological states or phenotypes. It is important to note that this method will achieve greater significance when the set of genes as a whole move up or down within the transcriptomic distribution.

A third approach is Differential Rank Conservation (3) or DIRAC. In contrast to GSEA, DIRAC is able to detect rank-based changes of genes within a gene set. This method thus detects gene set-level changes that might include genes within that increase and decrease in directionality.

Transcriptional Regulatory Networks

The prediction of a cell or tissue's transcriptional regulatory network (TRN) can incorporate many layers of regulation including targets of sequence-specific TFs, miRNAs, and signaling pathways that activate and inactive TFs. For the purposes of my research, I've focused on the regulatory connections of TFs and their predicted target genes, and thus proteins that act in complex with DNA. TRN maps as summarized here consist of nodes (target genes) and their directed edges (TFs). These connections (TF-gene) should be considered correct if three conditions are true:

i. **TF binding**. The TF physically interacts with the target gene by binding its regulatory DNA (promoter, enhancer, or other).

ii. **Regulation**. The TF regulates its target. Changes in TF activity within the regulatory DNA change the transcription rate of the target gene.

iii. **Direct Causation**. The physical interaction of TF binding is causal to the functional regulation of the target gene. Regulation could be affected through indirect interactions in addition to the physical interaction of binding and thus not causal. (4,5)

Various types of high-throughput data can determine the validity of TF-gene predictions, and I attempt to incorporate multiple lines of evidence in this thesis to that end. I will highlight here the data types and information used:

i. Binding locations of TFs: ChIP-seq or DNase I-seq (ENCODE)

ii. Functional regulation determined using large transcriptomic data (microarray or RNA-seq); TF knockdown or overexpression.

iii. Functional binding determined using editing of regulatory loci and functional reporter assays.

Genomic Footprinting

Enzymatic DNase I cleavage of chromatin is not uniform, and the binding of particular proteins to DNA provides 'footprints' in the cleavage pattern which can be computationally defined from the sequencing data and confer transcription factor occupancy (6). The method requires deep sequencing of high quality DNase I libraries, and has been applied to generate TF occupancy maps of yeast (7) plants (8), mouse (9,10) and human (11–13).

Numerous footprint callers have been developed since the arrival of mass amounts of DNase-seq data generated through the ENCODE consortium. One study (14) evaluated 15 different footprint-calling methods. For our purposes, we apply the Wellington (15) footprinting method here. However, more recent work in the lab relies on an ensemble approach that compares Wellington, HINT (14), and PIQ (16) calling methods. Wellington and HINT require an additional filtering step where regions of open chromatin are first identify based on the number of reads at a given location. Whereas PIQ is motif-centric and identifies all instances of a motif across the genome and then determines those most likely to be occupied by a TF based on the DNaseI cleavage pattern. Typically, footprint calling from DNase-seq data is compared to ChIP-seq data for accuracy.

Transcription factors

Transcription factors (TFs) are an integral part of regulatory networks that control gene expression and diverse cellular processes, programs, and identities. Originally, the term

'transcription factor' was coined for proteins involved in transcriptional regulation prior to our knowledge of their protein identities (thus, 'factor'). The term TF as applied in this work is synonymous here with a more definitive term in current literature: sequence-specific DNA binding protein. It should be noted that some of the literature includes other proteins such as chromatin remodelers or initiation complex proteins under the term TF.

Anderson et al (1981) (17) were the first to characterize a sequence-specific DNA binding protein, the cro repressor from λ bacteriophage. A few years later, the first human transcription factor - Sp1- was isolated and characterized from HeLa cells (18). Several methodologies enabled further characterization of this class of proteins including cloned DNA templates, DNA sensitivity assays, and protein purification and crystallization. Modern characterization benefitted greatly from deep-sequencing and proteomic strategies. Vaquerizas et al published a census of human TFs which estimates that there are approximately 1400 TFs that bind DNA in a sequence-specific manner (19). Approximately half of these are classified as zinc finger TFs, and the second largest group are the homeodomain family with approximately 250 proteins.

There are several curated databases for sequence-specific TFs, which provide motif specificities for a portion of the estimated 1400 TFs including TRANSFAC, JASPAR, and UniPROBE (20–22). The number of motif-to-TF entries varies based on the database, as well as the methodology used to define motif specificities.

Huntington's disease

Huntington's disease (HD) is an autosomal dominantly inherited disease caused by a CAG trinucleotide expansion in exon 1 of the *HTT* gene. Despite broad expression across human tissues, it causes selective neurodegeneration of the striatum with typical onset during middle age. The length of the CAG repeat is inversely correlated with the age of onset of the disease, although onset is still highly variable within a particular CAG copy number.

Despite HD being caused by a monogenic mutation, cases vs controls studies of human post-mortem brain tissue demonstrate a large portion of the transcriptome as differentially expressed. This has also been demonstrated in mouse models of the disease.

Few studies have surveyed the earliest effects of the mutation in a dense, longitudinal way. In Chapter 1 of my thesis, we have generated a dense time series study of mice with a knock-in CAG expansion in the endogenous murine *Htt* gene.

Psychiatric disorders

Psychiatric disorders are a heterogeneous group of mental disorders that collectively are the leading cause of disability worldwide. Due in part to the complexity of these disorders and the unique biological challenges that limit our understanding of their etiology and neurobiology, systems biology has risen as an essential approach to understanding how genetic variants and regulators affect gene expression and disease risk. Common, rare, inherited and *de novo* gene variants could be contributing to neurodevelopment and psychiatric disorders.

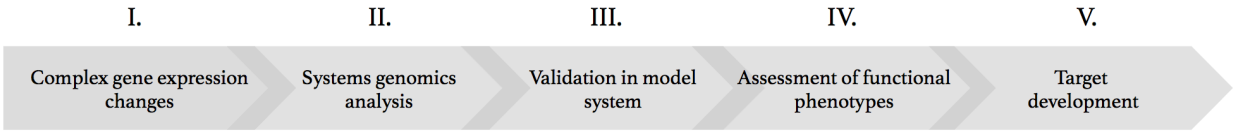
Defining features of several psychiatric disorders (23)

Name	Life prevalence	Heritability	Essential characteristics	Notable feature
Attention-deficit hyperactivity disorder (ADHD)	0.053	0.75	Persistent inattention, hyperactivity, impulsivity	Costs estimated at ~$US100 × 109 per year
Alcohol dependence (ALC)	0.178	0.57	Persistent ethanol use despite tolerance, withdrawal, dysfunction	Most expensive psychiatric disorder (total costs exceed US$225 × 109 per year)
Anorexia nervosa	0.006	0.56	Dangerously low weight from self-starvation	Notably high standardized mortality ratio
Autism spectrum disorder (ASD)	0.001	0.8	Markedly abnormal social interaction and communication beginning before age 3	Large range of function, complete daily care to exceptional occupational achievement
Bipolar disorder (BIP)	0.007	0.75	Manic-depressive illness, episodes of mania, usually with major depressive disorder	As a group, nearly as disabling as schizophrenia
Major depressive disorder (MDD)	0.13	0.37	Unipolar depression, marked and persistent dysphoria with physical and cognitive symptoms	Ranks number one in the burden of disease in the world
Schizophrenia (SCZ)	0.004	0.81	Long-standing delusions and hallucinations	Life expectancy decreased by 12–15 years

Conclusion

The importance of gene expression change in neurologic disease is clear. Recent advances in systems genomics approaches have allowed us to parse transcriptional changes into concrete hypotheses about regulatory drivers and gene networks that contribute to disease. Our understanding of the information in disease-specific genetic and transcriptomic data has been transformed by integration with expansive public datasets that describe regions of open-chromatin, spatial gene expression patterns, temporal gene expression patterns, and TF binding sites. These technological advances allow us to make systems-level hypotheses about the contributors to disease.

This work can be framed in a general pipeline with five primary phases (shown below; adapted from (24) for the development of therapies to treat neurologic disease. While my work does not include phase V – Target development – it sets the stage for assessing functional phenotypes related to gene network and gene regulatory changes.



2.3 Organization of Thesis

In this thesis, I describe the application of systems genomics approaches to understand gene networks and gene regulatory drivers in Huntington's disease and psychiatric disorders.

The first chapter describes a longitudinal analysis of early transcriptomic effects of the HD mutation. This is the first study of it's kind to sample mice weekly from 4 weeks of age to 20 weeks of age in order to understand the subtle, early effects of the HD mutation over time and across multiple genetic backgrounds. In addition to transcriptomic data generated from mouse striatal tissue, several cellular and molecular markers were also measured.

The second chapter describes gene regulatory drivers in Huntington's disease, and the application of a novel algorithm to define regulatory networks using the integration of DNase I footprinting data and large transcriptomic data. Using the *in silico* derived hypothesis that SMAD3 is a key regulator of early transcriptomic changes in HD, I went on to generate chromatin immunoprecipation sequencing data for SMAD3 from wildtype and HD mutant mouse striatal tissue, and validated predictions for SMAD3 gene targets.

The third chapter describes a parallel use of the algorithm described in Chapter 2, but as applied towards understanding gene regulatory drivers in psychiatric disorders. This study predicts regulatory loci and TF drivers of gene expression change in the brain. It goes on to validate several predictions using functional genomics approaches. We focus on a particular TF driver, POU3F2, and explore the *cis*- and *trans*-acting mechanisms of it's contributions to regulatory changes in schizophrenia and bipolar disorder.

The fourth chapter describes the application of a metabolic network analysis of transcriptomic studies of Huntington's disease. Previous literature has evaluated network changes using alternative algorithms. Here, we apply PathWave (Pinto et al) using grid-

defined KEGG pathways to understand particular metabolic pathways that are differentially expressed in both an age and mutation-severity manner.

The final chapter of this thesis discusses key conclusions from this work and future directions.

3 Longitudinal analysis of early transcriptomic effects of the Huntington's disease mutation

3.1 Abstract

Huntington's disease is a dominantly inherited neurodegenerative disease caused by the expansion of a CAG repeat in the *HTT* gene. In addition to the length of the CAG expansion, factors such as genetic background have been shown to contribute to the age at onset of neurological symptoms. A central challenge in understanding the disease progression that leads from the HD mutation to massive cell death in the striatum is the ability to characterize the subtle and early functional consequences of the CAG expansion longitudinally. We used dense time course sampling between 4 and 20 postnatal weeks to characterize early transcriptomic, molecular and cellular phenotypes in the striatum of six distinct knock-in mouse models of the HD mutation. We studied the effects of the Htt^{Q111} allele on the C57BL/6J, CD-1, FVB/NCr1, and 129S2/SvPasCrl genetic backgrounds, and of two additional alleles, Htt^{Q92} and Htt^{Q50}, on the C57BL/6J background. We describe the emergence of a transcriptomic signature in $Htt^{Q111/+}$ mice involving hundreds of differentially expressed genes and changes in diverse molecular pathways. We also show that this time course spanned the onset of mutant huntingtin nuclear localization phenotypes and somatic CAG-length instability in the striatum. Genetic background strongly influenced the magnitude and age at onset of these effects. This work provides a foundation for understanding the earliest transcriptional and molecular changes contributing to HD pathogenesis.

3.2 Introduction

Huntington's disease (HD) is a dominantly inherited neurodegenerative disorder that is characterized by progressive motor impairments, accompanied by cognitive impairment and psychiatric symptoms, leading to premature death. HD is caused by an expanded CAG trinucleotide repeat in the *HTT* gene, encoding an elongated polyglutamine (polyQ) tract in the huntingtin protein (HD Collaborative Research Group., 1993).

The age at onset of HD symptoms ranges from rare juvenile cases to more typical onset in mid- to late-adulthood. Rare, longer CAG lengths (>60) lead to juvenile onset (3–5) whereas 40-50 CAG repeats is typical of adult onset disease. Approximately 66% of variation in age at onset for adult CAG ranges (40-53 CAGs) is explained by an inverse relationship with the length of the expanded CAG tract (6). Genetic modifiers of age at onset at other genomic loci explain some of the remaining variance, with loci on chromosomes 8 and 15 reaching genome-wide significance in a recent genome-wide association study (7). Environmental and lifestyle factors may also contribute to variation in the age at onset.

Neuropathological studies of HD have revealed specific loss of GABAergic medium spiny projection neurons in the striatum (8) that occurs earlier and is more profound than cell death in other brain regions. It is thought that this loss of medium spiny neurons explains many of the symptoms of HD. Mouse models have revealed a wide range of molecular and cellular pathology in the striatum preceding cell death, including synaptic dysfunction (9), neuroinflammation (10), altered cholesterol and energy metabolism (11, 12), and changes in chromatin structure (13). Many of these phenotypes were originally discovered through gene expression profiling, which revealed thousands of differentially expressed genes in both post-mortem striatal tissue from HD patients and in mouse models of HD (14–16).

Here, we describe large-scale gene expression profiling and molecular phenotyping of striatal tissue from six distinct knock-in mouse models of HD during a time course from 4 to 20 postnatal weeks. Our study included the Htt^{Q111} allele on the C57BL/6J, CD-1, FVB/NCr1, and 129S2/SvPasCrl genetic backgrounds and two additional shorter CAG repeat alleles (Htt^{Q92} and Htt^{Q50}) on the C57BL/6J background. Some behavioral changes have been observed in Htt^{Q111} mice at 11-12 weeks of age as well as memory deficits at 16 weeks of age (17, 18). But motor phenotypes were not seen in Htt^{Q111} mice until 9 months of age (19). Overall, our longitudinal study likely encompasses some behavioral and cognitive changes, but no clear motor deficits.

The goals of our study were two-fold. First, we aimed to identify very early events in the progression of molecular and cellular phenotypes underlying HD. Identifying early events in the progression of the disease is critical, since these changes are likely to be enriched for causal mechanisms. By contrast, most previous gene expression profiling experiments and similar studies in HD mouse models have focused on later stages of the disease progression or sampled at low density time points that do not capture early inflection points (15, 16, 20).

Second, we aimed to characterize similarities and differences in the effects of *Htt* CAG expansion when occurring on different genetic backgrounds. Given the now strong evidence for genetic modifiers in the human disease, modeling the effect of genetic variation in mice is an increasingly high priority. In addition, identifying changes that are robust across multiple genetic backgrounds is a proven strategy to improve signal-to-noise and eliminate false positive results (21).

Here, we find a cascade of gene expression changes associated with *Htt* CAG expansion beginning in mice as young as 10-16-weeks old. We show that this time course spanned the onset of two known early molecular phenotypes in HD mouse models: nuclear localization of the mutant huntingtin protein and striatal CAG-length genomic instability.

Genetic background strongly influenced the rate at which molecular and transcriptomic phenotypes emerged, suggesting the presence of multiple genetic modifiers, and emphasizing the importance of studying human disease in model systems that account for genetic variation.

3.3 Results

Experimental Design

In view of the dominant nature of HD, we used heterozygous knock-in mouse lines in which expanded CAG repeats of various lengths have been inserted into one of the mouse's endogenous *Htt* alleles (Fig. 1A). The Htt^{Q111} allele (previously Hdh^{Q111}; "Q111") has been backcrossed onto four different genetic strain backgrounds: C57BL/6J ("B6"), CD-1, FVB/NCrl ("FVB"), and 129S2/SvPasCrl ("129") (22–24). On the B6 background, we also characterized knock-in mice carrying shorter alleles, Htt^{Q50} ("Q50") and Htt^{Q92} ("Q92") (16, 24, 25). Note that Q111, Q92 and Q50 are the nominal allele names, but due to intergenerational instability in Q92 and Q111 that is background strain-dependent (23, 24), mice from these lines vary in their actual CAG repeat length (Fig. 1B). Siblings of these knock-in mice that did not carry pathogenic alleles ("WT") were used as controls. 3 mice of each genotype were collected each week between 4 and 20 postnatal weeks, and an additional ten mice were collected every fourth week for the B6.Q50, B6.Q111 and CD-1.Q111 lines. 2-10 mice per week from each genotype were used for gene expression profiling and for cellular and molecular phenotyping (Fig. 1A; Supplementary Information [SI] Table S1).

We measured the body mass and brain mass of each mouse. Age and strain influenced these traits, but there were no statistically significant differences in body mass or brain mass associated with *Htt* alleles (SI Tables S2, S3).

Dynamic huntingtin nuclear localization patterns across strain and over time.

Nuclear localization of mutant huntingtin protein (mHTT) is one of the earliest known molecular phenotypes in HD knock-in mouse models (26). We measured nuclear huntingtin immunostaining phenotypes in the striata of 84 QIII mice, using a conformation-specific antibody that recognizes forms of mHTT in the nuclei of medium spiny striatal neurons (Fig. 1C,D) (26, 27). Diffuse nuclear mHTT immunostaining was visible in a fraction of striatal cells in heterozygous B6.QIII, CD-1.QIII, and FVB.QIII mice as young as 9-10 weeks old (Fig. 1D). In these strains, the intensity of nuclear staining increased over time (ANOVA, P_{age} = 1.5e-5; Fig. 1D). By 20 weeks, diffuse nuclear mHTT was prominent in striatal neurons in these strains (Fig. 1C). Punctate nuclear mHTT inclusions were visible in these strains beginning at 17 weeks. In contrast to the other three strains, diffuse nuclear huntingtin was not visible in the 129.QIII mice until they were at least 15 weeks old and was not as intense as in the other three strains (ANOVA: P_{strain} = 1.1e-3). Punctate nuclear inclusions were not observed in 129.QIII mice up to 20 weeks of age. These strain differences are consistent with the relative timing of nuclear localization and inclusion formation in previous observations in these strains (23, 26, 28).

Somatic CAG length expansion in striatum across strain and over time.

We profiled the distribution of *Htt* CAG lengths in striatal tissue from QIII mice (n=250). Previous studies showed that genetic differences between B6.QIII, 129.QIII and FVB.QIII mice influence the rate of striatal CAG length expansion (23, 29). While repeat expansion dynamics have been analyzed in CD-1.QIII mice over a broad (2-16 month) time window (22), very early repeat expansion dynamics and the relationship to molecular phenotypes are not well established. We observed age- and strain-dependent effects on the extent of striatal CAG

length instability (Fig. 1E). The "instability index" summarizes the mean change in CAG repeat length observed in each mouse's striatum. Instability index increased as a linear function of age in all four strains, but the rate of instability differed. Consistent with previous results (23, 29), CAG tracts expanded more extensively in B6.QIII and FVB.QIII mice than in 129.QIII heterozygous mice. The rates of CAG tract expansion in CD-1.QIII mice were roughly similar to those in FVB.QIII mice, but CD-1.QIII exhibited greater variation in instability indices, consistent with this being an outbred strain.

Differentially expressed genes in 4-20-week-old knock-in mouse models of the HD mutation.

We used linear modeling to characterize gene expression changes in each HD mouse model. We considered 18 conditions, defined by a mouse's genotype and age. For each of the six genetically distinct HD mouse models -- B6.QIII, CD-1.QIII, FVB.QIII, 129.QIII, B6.Q92, and B6.Q50 -- we considered three non-overlapping age windows with equal numbers of mice: "early", 4-9 postnatal weeks; "middle", 9-16 postnatal weeks; and "late", 16-20 postnatal weeks. We compared these mice to age-matched wildtype mice of the same strain. In the early age window, we detected fewer than 25 differentially expressed genes (DEGs, $q < 0.05$) in all mouse models (Fig. 2A). In the middle age window, we detected 279 DEGs in B6.QIII mice and < 25 DEGs in the other mouse models. In the late age window, we detected 1,513 DEGs in B6.QIII mice, 222 DEGs in CD-1.QIII mice, and < 25 DEGs in the other mouse models. Therefore, counts of DEGs increased progressively with age in B6.QIII and CD-1.QIII mice, while few DEGs could be detected in the other mouse models.

Replication of DEGs across conditions and in independent datasets

There was extensive overlap between the DEGs identified in different age x genotype conditions (Fig. 2B). 178 of the 222 DEGs in 16-20-week-old CD-1.Q111 mice (80%) were also differentially expressed in 16-20-week-old B6.Q111 mice (Fisher's exact test: p = 1.4e-173). 198 of the 279 DEGs in 9-16-week-old B6.Q111 mice were also differentially expressed in 16-20-week-old B6.Q111 mice (p = 1.0e-174). 77 DEGs were shared among all three conditions with > 25 DEGs.

There was also extensive overlap between the DEGs identified in our study and striatal gene expression changes in independent datasets from HD knock-in mouse models and human postmortem brain. 726 of the DEGs in 16-20-week-old B6.Q111 mice were differentially expressed in striata of 6-month-old B6.Q111 mice profiled by RNA-seq (16) (p < 2.2e-308; Fig. 2C). 761 of the DEGs in 16-20-week-old B6.Q111 mice were differentially expressed in post-mortem striatal tissue from (human) HD cases vs. controls (Hodges et al. 2006; p = 3.1e-14; Fig. 2D). Notably, there were far fewer DEGs in the < 20-week-old mice from our study than in the older mice profiled by Langfelder et al. (16) and in post-mortem human striatum. These results suggest that the DEGs identified in our study represent a robust transcriptomic signature for the early effects of the HD mutation in the striatum.

Differentially expressed genes are involved in diverse neuronal functions

We used Gene Ontology (30) and Cell-type Specific Expression Analysis (31, 32) to characterize biological functions enriched for DEGs in 16-20-week-old B6.Q111 vs. WT mice and the dominant cell types in which they are expressed. Down-regulated genes were enriched for genes that are expressed specifically in *Drd1*- and *Drd2*-positive medium spiny neurons (p = 7.4e-53, p = 9.6e-44, respectively; SI Table S4). Down-regulated genes were

enriched for 389 Gene Ontology terms (FDR < 0.05; SI Table S5). Many of the most strongly enriched pathways relate to synaptic activity, such as "dopamine receptor signaling pathway" (p=1.0e-14), "regulation of membrane potential" (p = 1.9e-19), "regulation of ion transport" (5.0e-19), "neuron projection morphogenesis" (p = 3.4e=18), and "ionotropic glutamate receptor binding" (p = 1.4e-12). Down-regulated genes were also enriched for components of intracellular signaling pathways known for their roles in coupling the activity of transmembrane receptors to downstream cellular processes. These pathways included "regulation of G-protein-coupled receptor protein signaling pathway" (p = 2.9e-22), "calmodulin binding" (p = 1.3e-25), "GTPase activator activity" (p = 1.3e-20), "response to organic cyclic compound" (p = 3.4e-14), and "response to insulin" (p = 2.7e-11).

Up-regulated genes in 16-20-week-old B6.Q111 mice were not strongly enriched for cell type-specific genes, suggesting that these expression changes are distributed across multiple cell types (SI Table S4). These genes were enriched for 188 pathways (FDR < 0.05; SI Table S6). Like down-regulated genes, up-regulated genes were enriched for neuronal functions, such as "neuron projection morphogenesis" (p = 3.2e-13). In addition, up-regulated genes were enriched for more general cellular functions, such as "translation" (p = 5.3e-13), "ubiquitin protein-ligase binding" (p = 8.2e-9), and "regulation of homeostatic process" (p = 4.7e-10). These results indicate that the *Htt* CAG expansion influences the expression of genes with diverse functions, even early in the pathogenic process.

Notably, we found few expression changes in genes related to neuroinflammation, astrogliosis, and apoptosis (SI Tables S5, S6). All of these processes are upregulated in striatal tissue from HD mouse models sampled later in life and in post-mortem tissue from HD patients (14, 16). Therefore, our results suggest that neuroinflammation-related transcriptomic changes are secondary to the "neuronal" signature detected in the young mice from our study.

Gene expression changes in 4-16-week-old HD knock-in mice

The results above suggest that robust HD-related gene expression changes occur in 16-20-week-old B6.Q111 and CD-1.Q111 mice. We next asked whether we could detect early signs of this HD molecular signature in even younger mice and in other mouse strains. Although we detected few DEGs in these conditions, we hypothesized that younger mice and/or those with lower CAG lengths might show subtle changes in expression that would be correlated with the expression changes detected at later stages of the phenotypic progression. To test this hypothesis, we focused on the expression patterns of the top 100 genes with the lowest p-values in 16-20-week-old B6.Q111 mice (Fig. 3). Many of these genes – e.g., *Pde10a, Penk, Cnr1,* and *Gpx6* – are differentially expressed in striatal tissue from HD patients (14) and in previous studies of HD knock-in mouse models (16).

We fit a linear model to test whether the fold changes of these top 100 genes predicted their fold changes in other conditions. That is, we asked to what extent the direction and magnitude of the fold changes were consistent across mouse models and time points. In this model, the correlation coefficient (*r*) describes the strength of the linear relationship between the fold changes of the top 100 genes in each pair of conditions, while the regression coefficient (b_1) describes the relative sizes of the fold changes.

In 16-20-week-old mice, the fold changes of the top 100 genes were positively correlated between B6.Q111 mice and each of the other genetic models (Fig. 4; r = 0.38 – 0.97). The fold changes varied between strains, with the largest fold changes in B6.Q111, followed by CD-1.Q111 (b_1 = 0.62, relative to 16-20-week-old B6.Q111 mice), FVB.Q111 (b_1 = 0.43), B6.Q92 (b_1 = 0.25), 129.Q111 (b_1 = 0.20) and B6.Q50 (b_1 = 0.07).

In 9-16-week-old mice, the fold changes of the top 100 genes were positively correlated in B6.Q111, CD-1.Q111, FVB.Q111, and B6.Q92 mice (r = 0.34 – 0.96) but were not correlated in 129.Q111 or B6.Q50 mice. B6.Q111 mice had the largest fold changes (b_I = 0.58), followed by CD-1.Q111 (b_I = 0.38), FVB.Q111 (b_I = 0.18), B6.Q92 (b_I = 0.06).

In 4-9-week-old mice, fold changes were more variable. Fold changes were positively correlated in B6.Q111 (r = 0.72, b_I = 0.17), FVB.Q111 (r = 0.44, b_I = 0.08), and 129.Q111 (r = 0.57, b_I = 0.15), compared to 16-20-week-old B6.Q111 mice. However, fold changes were negatively correlated in B6.Q92 mice of this age (r = -0.41, b_I = -0.11), and were uncorrelated in CD-1.Q111 and B6.Q50 mice. These results suggest that HD mutations influence many of the same genes across multiple mouse models of the HD mutation, but genetic background and CAG repeat length influence the rate at which these genes become differentially expressed.

Multiple molecular and genetic mechanisms may underlie strain differences in response to HD mutations.

Our results show that both CAG repeat length and strain genetic background modify the rate at which molecular and cellular phenotypes progress in mice with *Htt* CAG expansions. Phenotypic variation may be driven at least in part by the different inherited CAG repeat lengths in each of the four Q111 strains (Fig.1B), but is not explained by variation in *Htt* expression (SI Figure S1). As the strains also exhibit differences in the rate of somatic CAG expansion (Fig. 1E) we hypothesized that these strain differences involve a known genetic polymorphism that influences somatic instability of the CAG repeat in B6.Q111 vs. 129.Q111 mice. A previously reported genetic mapping study in C57BL/6N.Q111 x 129.Q111 mice showed that variants at the *Mlh1* locus were linked to slower *Htt* CAG somatic expansion in the striata

of 129.Q111 vs. B6N.Q111 mice (29). These variants were associated with lower expression of the *Mlh1* gene, which encodes a mismatch repair enzyme, MutL homolog 1. We examined *Mlh1* expression in mice from each of the four background strains in our study to replicate the previously reported expression polymorphism in B6N or B6J vs. 129 mice (29) and to assess differences in CD-1 and FVB mice. Wildtype 129 mice had 2.8-fold lower expression of *Mlh1*, compared to wildtype B6 mice, consistent with previous findings (Fig. 5), while B6, CD-1 and FVB mice all had similar high expression of *Mlh1*. The intermediate rates of somatic CAG expansion in FVB and CD-1 strains are consistent with high *Mlh1* expression in these strains, but suggest additional modifier genes(s) that might reduce expansion relative to B6. Nuclear mutant huntingtin immunostaining phenotypes correlate reasonably well with the rates of somatic CAG expansion across strain background, indicating that genetic modifiers of these *Htt* CAG-dependent events may be shared. Notably, however, the very different transcriptional responses to the Q111 allele in B6, FVB and CD-1 strains are unlikely to be explained by genetically encoded changes in *Mlh1* expression. Overall, these results indicate that multiple genetic factors may modify responses to the HD mutation.

3.4 Discussion

The cascade of early molecular, cellular and transcriptional changes that occurs as a result of CAG expansion in HD mouse models has not previously been studied systematically across a week-by-week timescale. We generated transcriptomic data in a dense time-series from the striatum of mice with *Htt* CAG repeat expansions in parallel with molecular and cellular phenotyping to elucidate the timing and magnitude of CAG length- and age-dependent changes across four genetic background strains. We observed robust striatal CAG

length expansion, mHTT nuclear localization, and gene expression changes in B6.Q111 mice as young as 9-16 weeks old. We find evidence that subtle changes in these same phenotypes begin to manifest even earlier in life and with shorter CAG repeat tracts. While the Q111 allele induced similar molecular changes on the other background strains studied, the rate of progression differed across these strains. Our study provides a dense and dynamic view of very early phenotypic and transcriptomic changes in the striatum of mice with *Htt* CAG repeat expansions.

Our results reveal very early gene expression changes in the striatum of Q111 mice. Most of these expression changes were very small, typically less than 1.5-fold differences in expression in 16-20-week-old B6.Q111 vs. WT mice. These effect sizes in our analysis of bulk striatal tissue are consistent either with subtle changes occurring in many striatal cells or with more dramatic changes occurring in a small subset of striatal cells.

Many of the differentially expressed genes are involved in synaptic functions and are predominantly expressed in neurons rather than glia. Most of these same pathways are also differentially expressed in older mice with HD mutations and in post-mortem striatal tissue from HD cases vs. controls. Notably, however, the biological processes perturbed in young Q111 mice represent only a small fraction of processes that are perturbed later in the disease. For instance, in contrast to studies of 6- and 10-month-old B6.Q111 mice, we find few changes in genes associated with neuroinflammation and activated gliosis (16). Therefore, our results suggest that the sequence of molecular changes associated with HD mutations begins with a relatively small number of changes in neurons, before cascading onward to broader pathology in both neurons and in glia.

Our results suggest that multiple genetic modifiers influence the effects of the Q111 allele on molecular phenotypes. B6 mice consistently presented the most rapid phenotypic

progression while 129 mice were the slowest. These phenotypic differences appear to be primarily quantitative and not qualitative: i.e., we observed correlated gene expression responses in all strains, but the rate of change differed. Our study was not specifically designed to detect potential modifier effects of strain background, but rather to identify robust gene expression changes across multiple strains. Thus, the relatively low CAG repeat length in 129.Q111 mice could, at least in part, contribute to the slower phenotypic progression in this strain. However, the slow rates of the transcriptional responses in CD-1 and FVB compared to B6, despite having similar or longer CAG repeat lengths than B6, indicates the presence of modifier genes. Further, an expression QTL at the *Mlh1* locus that is linked to differences in *Htt* CAG length expansion in B6 vs. 129 mice does not appear to explain strain differences in CD-1 and FVB relative to B6, suggesting that other genetic modifiers of the transcriptional response are present in these strains. Further unbiased identification of the genetic modifiers of different phenotypes in *Htt* CAG knock-in mice, e.g. by crossing mice with HD mutations onto a mapping population such as the Collaborative Cross or Mouse Diversity outbred collection (33, 34), is a promising future direction that could reveal multiple modifier loci. Identifying these genetic modifiers could point to novel therapeutic targets.

The proximal mechanisms by which HD mutations cause early gene expression changes remain unclear. Several non-mutually-exclusive mechanisms have been proposed (35–38): (i) Both wildtype and mutant HTT form protein complexes with transcription factors, and these interactions could change the ability of these TFs to regulate their target genes; (ii) direct interaction of huntingtin protein with chromatin modifying enzymes and with genomic DNA influences gene expression through changes in chromatin or regulatory states (39–41); (iii) gene expression changes are secondary or occur indirectly as a result of another mechanism. In our study, we found that the rates of somatic *Htt* CAG length expansion,

mHTT nuclear localization, and transcriptional changes were all correlated with age. These correlations suggest that these phenotypes are connected but make it difficult to distinguish causal from correlative relationships. Unbiased genetic studies to determine whether common or different modifiers underlie these phenotypes will help to disentangle these relationships. More detailed biochemical experiments will be required to identify a subset of the observed changes that are caused by direct interactions with huntingtin protein.

Since the 1993 discovery of the HD mutation, research into the causal mechanism underlying neurodegeneration has generally focused on late stage disease. However, a wide range of pathophysiology is present in late stages of HD, much of which may be only peripherally related to pathogenesis. Our approach focuses instead on studying earlier, more subtle perturbations, enabling us to identify a smaller number of candidate mechanisms. By measuring transcriptomic, molecular, and cellular changes with many samples across a dense time series we were able to precisely time inflection points such as gene expression changes and nuclear mHTT phenotypes. Equally critical is our observation of striking differences in the magnitude and timing of these effects across strains. This study contributes to the burgeoning field of 'pre-manifest' HD research, clarifying molecular changes that occur well before the onset of diagnosable clinical symptoms with the potential of proffering new therapeutic targets

3.5 Methods

Breeding

Heterozygous males of six different strains (129S2/SvPasCrl.*Htt*$^{Q111/+}$, FVB/NCrl.*Htt*$^{Q111/+}$, CD-1.*Htt*$^{Q111/+}$, C57BL/6J.*Htt*$^{Q92/+}$, C57BL/6J.*Htt*$^{Q50/+}$, C57BL/6J.*Htt* $^{Q111/+}$) were crossed with wild type females of corresponding background strains in order to generate cohorts of age-matched heterozygous pups and WT controls for the time-course assigned tissue harvests. A tail snip was acquired at pre-weaning (2-4-weeks) to determine the CAG repeat length of each mouse.

One heterozygous animal and one age-matched wild type control animal from each strain were harvested three times weekly from 4 to 20 weeks of age. The sex of mice harvested at these times was randomized. An additional 5 female and 5 male heterozygous animals from the B6.Q50 , B6.Q111 and CD1.Q111 strains were harvested at 4, 8, 12, 16, and 20 weeks of age. Time of day of harvest was randomized within the hours of light in the mouse room, typically ≥3 hours after lights came on and ≥3 hours before lights went off.

Euthanasia was performed using live decapitation. Trunk blood was collected and plasma was frozen and stored at -80°C. The right hemisphere of each animal's brain was microdissected into thalamus, cortex, striatum, hippocampus, and remaining brain regions. These tissues were snap frozen and stored at -80°C. The left hemisphere was embedded in OCT embedding medium, snap frozen, and stored at -80°C for sectioning. Snap-frozen striata from the right hemisphere of each animal were used for isolation of RNA and DNA for transcriptional profiling and analyses of CAG instability, and sectioned tissue from the left hemisphere was used for immunostaining with mAb5374.

mAb5374 immunofluorescence

Nuclear mutant huntingtin localization was assess by immunofluorescence imaging of frozen 7 μm coronal sections with the mAb5374 antibody (Chemicon), as previously described (42). The mAb5374 antibody is a conformationally sensitive antibody that detects diffusely immunostaining nuclear mutant huntingtin and intranuclear inclusions. We used Cell Profiler image analysis software (43) to quantify the intensity of nuclear mAb5374 staining in each cell. The max intensity in each cell was normalized to intensity levels in wild type animals to calculate a normalized score.

Determination of *Htt* CAG repeat length and striatal instability

Assays for knock-in mouse genotyping and determining *Htt* CAG repeat length were as described previously (42). The *Htt* CAG repeat was amplified using a human-specific PCR assay that amplifies the repeat from the knock-in allele but does not amplify the mouse sequence. The forward primer was fluorescently labeled with 6-FAM (Applied Biosystems) and products were resolved using the ABI 3730xl DNA analyzer (Applied Biosystems) with GeneScan 500 LIZ as internal size standard (Applied Biosystems). GeneMapper v3.7 (Applied Biosystems) was used to generate CAG repeat size distribution traces. Repeat size was determined from the peak with the greatest intensity in the GeneMapper trace. An instability index was quantified from the GeneMapper CAG repeat distributions in striatal DNA as previously described (44). Briefly, the highest peak in each trace was used to determine a relative threshold of 10% and peaks falling below this threshold were excluded from analysis. Peak heights normalized to the sum of all peak heights were multiplied by the change in CAG length of each peak relative to the highest peak. These values were summed to generate an instability index, which represents the mean CAG repeat length change in the population of

cells being analyzed. Note that the tail CAG length at pre-weaning does not vary appreciably from the main CAG allele in striatum; rather the instability in striatum is reflected in a broader distribution of alleles, strongly biased to longer CAG lengths.

Microarray gene expression profiling

RNA was extracted from the striata of individual mice using Trizol reagent (DNA was isolated for instability analyses in the same procedure). We profiled gene expression on SurePrint G3 8x60k Mouse Microarrays (Agilent, Santa Clara, CA). We performed quantile normalization and log2-transformation. We used ComBat (45) to remove batch effects related to the date on which microarrays were hybridized. Linear models were fit using the limma R package (46). Age was treated as a categorical variable with three classes: "early", 4-9-weeks-old; "middle", 9-16-weeks-old; "late", 16-20-weeks-old. We considered the nominal *Htt* allele as a categorical variable. We fit a linear model to predict each gene's expression, considering effects of background strain, age, sex, and *Htt* allele. We used post-hoc contrasts to evaluate the effects of *Htt* allele in each background strain x age condition. A false discovery rate for these p-values was calculated using the Benjamini-Hochberg method (47). This dataset has been deposited in GEO, accession GSE88920. Scripts and processed data have also been deposited in a GitHub repository located at https://github.com/seth-ament/hd-time-series

3.6 Figures and Tables

Figure 1. Time course of mHTT nuclear localization and of striatal *Htt* CAG length expansion in 4- to 20-week-old HD knock-in mice. A. Striatal tissue was collected from a total of 731 4- to 20-week-old mice in order to compare mice with heterozygous mutant *Htt* alleles (Q111, Q92, Q50) to mice with wild-type *Htt* alleles (WT) on four genetic backgrounds. B. Inherited CAG repeat length of *Htt* alleles in Htt^{Q111} mice varies with genetic background. The CAG repeat length in each mouse with a nominal Htt^{Q111} allele was measured in a tail snip sample at weaning. C. Nuclear mutant huntingtin immunofluorescence using mAb5374 (red) and DAPI (blue) counterstain in 20-week-old B6.Q111, CD1.Q111, FVB.Q111, 129.Q111 mice, along with negative control 20 week-old B6.WT. All images were acquired at the same sensitivity and displayed with equal adjustment of the levels to enhance the red signal relative to background and nuclear staining. To illustrate the pattern of the weak staining in 129.Q111 the cell in the inset was further enhanced; arrows indicate positive cells. The white scale bar represents 10 μm. D. Quantitation of mHtt immunofluorescence intensity in 9- to 20-week-old Q111 mice from each background (n=84 total mice, n=20-25 mice per strain). Each point indicates the immunofluorescence intensity in a single cell. E. Age- and background strain-dependent expansion of *Htt* CAG tracts in the striata of 4- to 20-week-old Q111 mice. Each point represents the "instability index" for a single mouse, a quantitative metric representing the mean CAG length change in the population, relative to the modal allele.

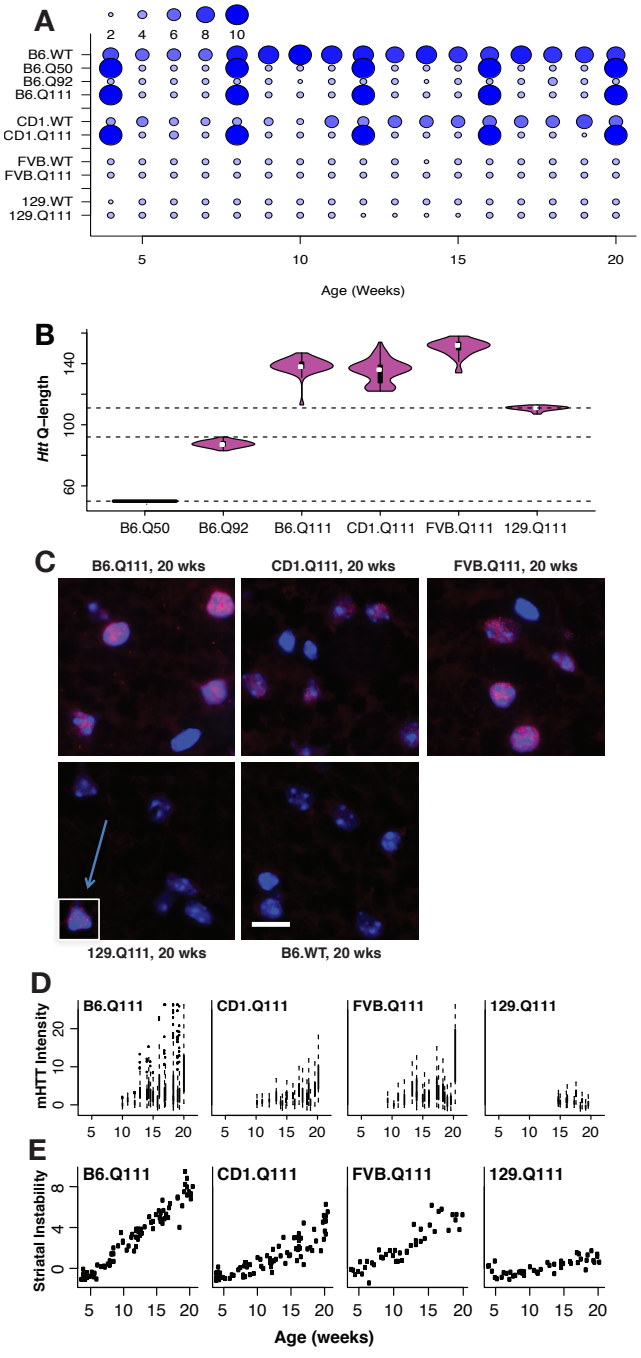

A
2
4
6
8
10
B6.WT
B6.Q50
B6.Q92
B6.Q111
CD1.WT
CD1.Q111
FVB.WT
FVB.Q111
129.WT
129.Q111
5
10
15
20
Age (Weeks)
B
Htt Q–length
140
100
60
B6.Q50
B6.Q92
B6.Q111
CD1.Q111
FVB.Q111
129.Q111
C
B6.Q111, 20 wks
CD1.Q111, 20 wks
FVB.Q111, 20 wks
129.Q111, 20 wks
B6.WT, 20 wks
D
mHTT Intensity
0
10
20
B6.Q111
CD1.Q111
FVB.Q111
129.Q111
5
10
15
20
E
Striatal Instability
0
4
8
B6.Q111
CD1.Q111
FVB.Q111
129.Q111
5
10
15
20
Age (weeks)

Figure 2. Age- and strain-dependent effects of *Htt* alleles on gene expression. A. Counts of up- and down-regulated genes (FDR q < 0.05) in each genotype at three time windows: "early" (4-8-weeks-old), "middle" (9-15-weeks-old), and "late" (16-20-weeks-old). B. Overlap between differentially expressed genes (DEGs) detected in the three conditions with >25 DEGs. C. Overlap between DEGs in 16-20-week-old B6.Q111 mice (this study) and DEGs in 6-month-old B6.Q111 mice from Langfelder et al. (2016). D. Overlap between DEGs in 16-20-week-old B6.Q111 mice (this study) and DEGs in post-mortem caudate nucleus from HD cases vs. controls (Hodges et al. 2006).

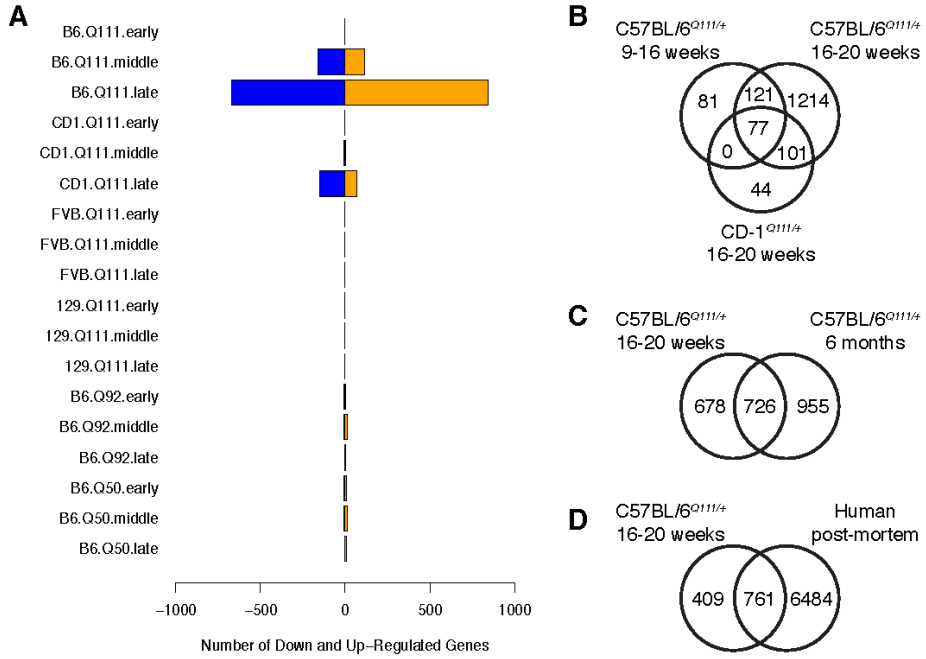


Figure 3. Expression patterns of the top 100 differentially expressed genes. The top 100 genes were defined by their p-values in 16-20-week-old B6.Q111 mice. The heatmap indicates the -$\log_{10}$(p-values) of these genes in 18 conditions defined by genotype and age window.

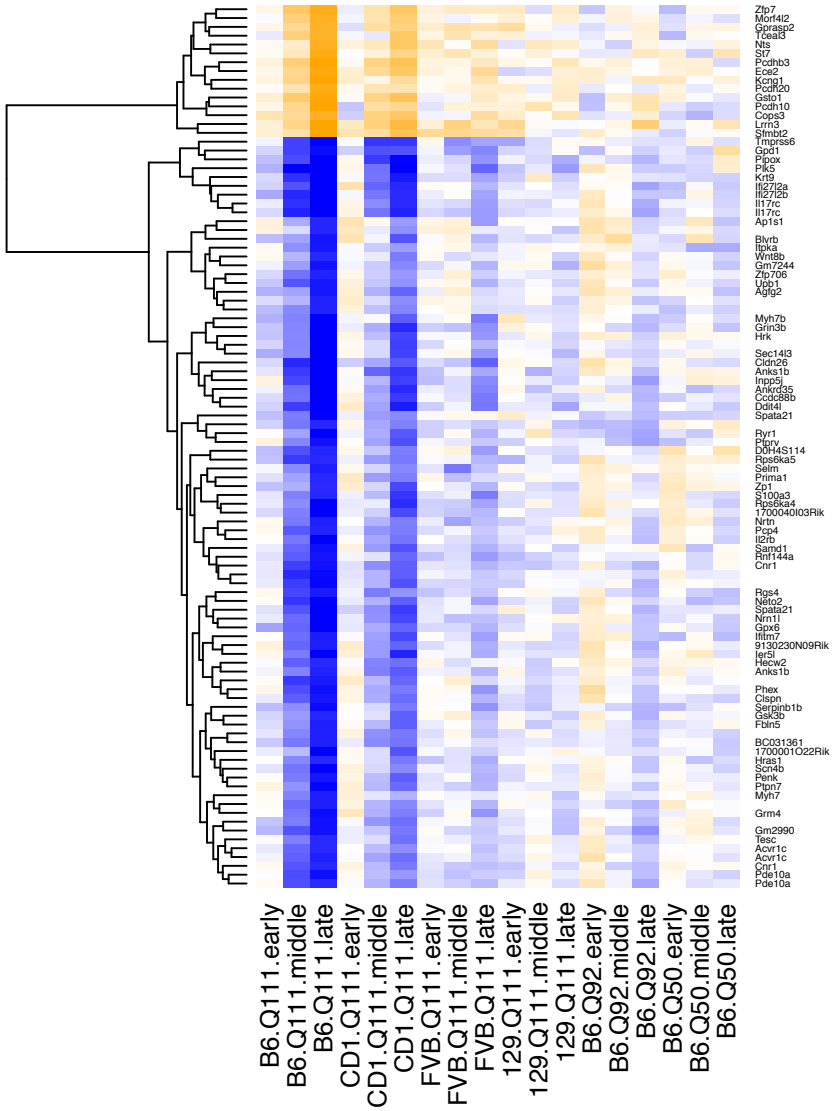


Figure 4. Age, CAG length and genetic background influences the magnitude but not the direction of gene expression changes in HD knock-in mice. Each scatterplot compares the fold changes of the top 100 genes in 16-20-week-old B6.Q111 mice (x-axis) to the fold changes of these genes in another condition, defined by age and genotype (y-axis). Each point on the scatterplot indicates the fold change estimate of a single gene.

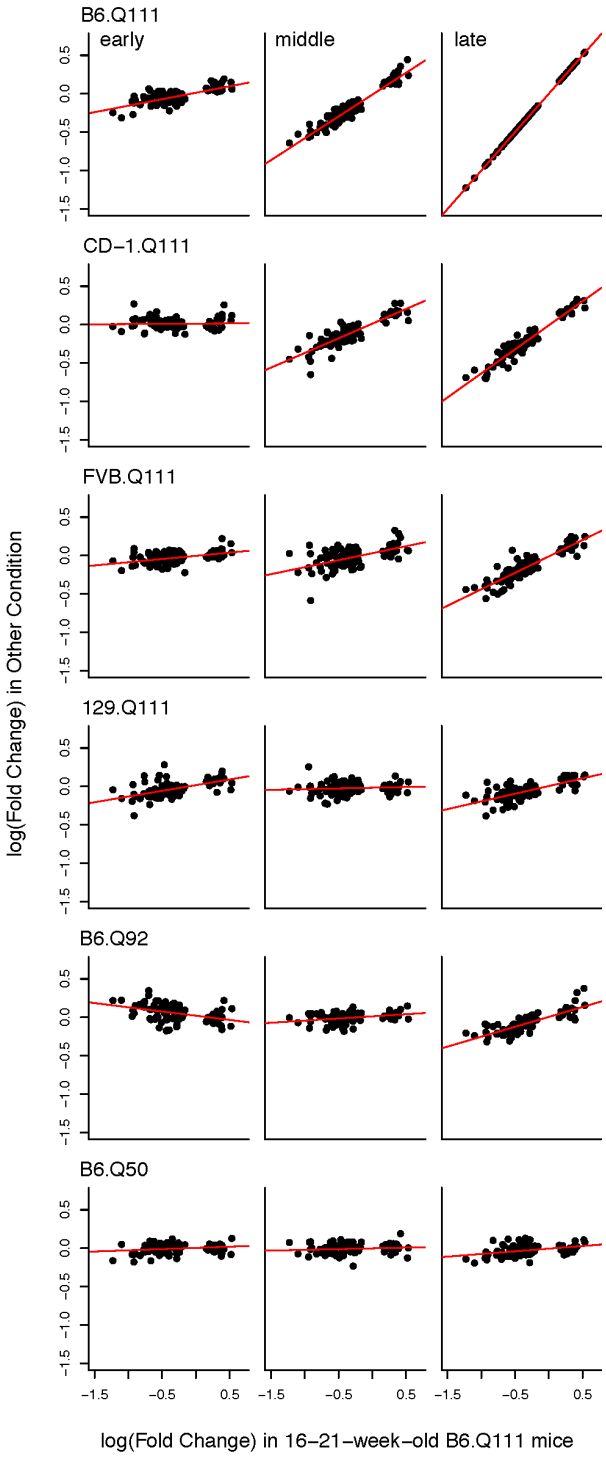

B6.Q111
early
middle
late
CD−1.Q111
FVB.Q111
129.Q111
B6.Q92
B6.Q50
log(Fold Change) in Other Condition
log(Fold Change) in 16−21−week−old B6.Q111 mice
−1.5
−1.0
−0.5
0.0
0.5

Figure 5. Effects of genetic background and of HD mutations on *Mlh1* expression. Box plots represent mean expression of *Mlh1* in each strain and genotype group.

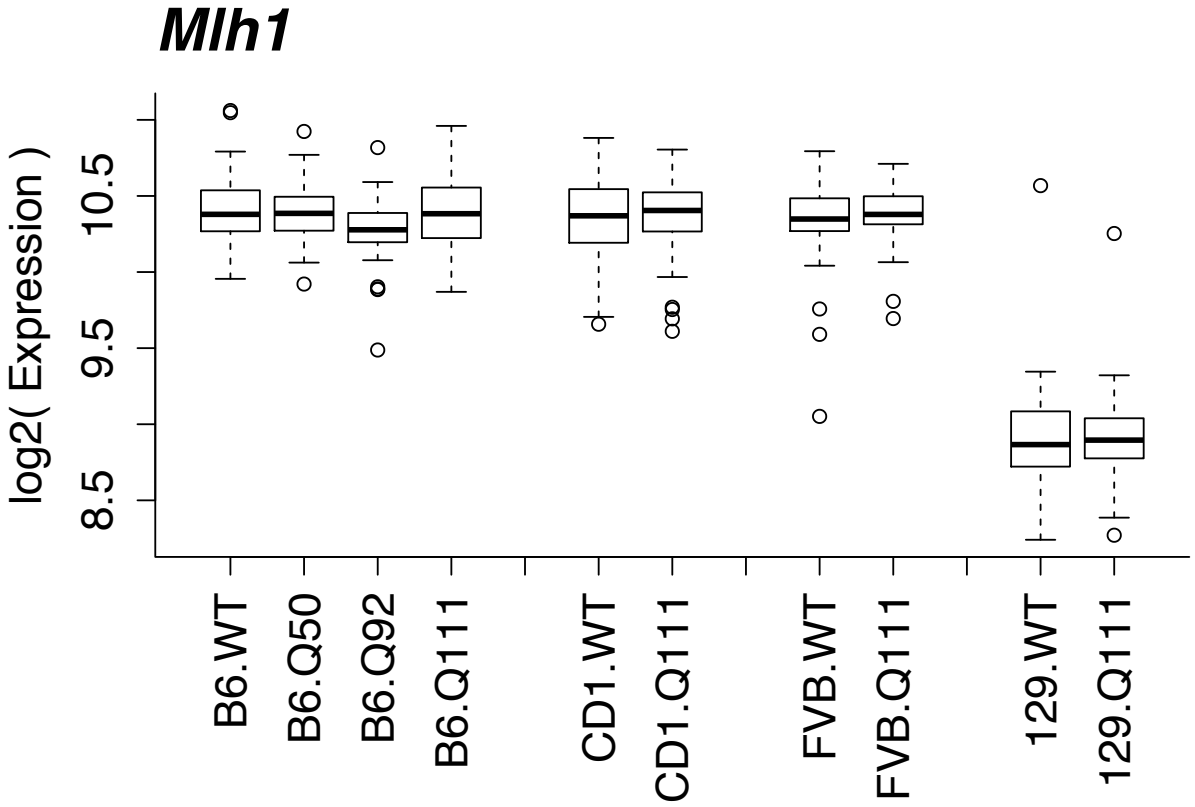


Table S1. Sample sizes (number of mice) for each group used in microarray gene expression profiling.

	4-9-weeks-old	10-16-weeks-old	16-21-weeks-old
B6.WT	41	53	43
B6.Q50	31	26	29
B6.Q92	17	16	17
B6.Q111	32	25	29
CD1.WT	23	33	32
CD1.Q111	32	25	29
FVB.WT	17	17	16
FVB.Q111	17	18	16
129.WT	17	17	16
129.Q111	18	14	15

Table S2. No association between Htt mutations and body mass. Each mouse used in the gene expression profiling study was weighed. A linear model was fit to assess effects of age, sex, strain, Htt CAG length, and strain x CAG length interactions, using the formula Linear model: body weight ~ age + sex + strain + cag.actual + strain*cag.actual, using female mice and 129 strain background as baseline comparators (note that "cag.actual" is the CAG repeat length in weaning tail DNA".

	Estimate	p-value
(Intercept)	-9.10526	2.00E-16
age	0.730815	2.00E-16
sexM	5.290804	2.00E-16
strainB6	1.2182	0.00293
strainCD1	15.49909	2.00E-16
strainFVB	7.909084	2.00E-16
cag.actual	-0.00346	0.62023
strainB6:cag.actual	-0.00099	0.89944
strainCD1:cag.actual	0.0168	0.03718
strainFVB:cag.actual	0.003544	0.67825

Table S3. No association between Htt mutations and brain mass. The brain of each mouse used in the gene expression profiling study was weighed. A linear model was fit to assess effects of age, sex, strain, Htt CAG length, and strain x CAG length interactions, using the formula brain weight ~ age + sex + strain + cag.actual + strain*cag.actual, using female mice and 129 strain background as baseline comparators (note that "cag.actual" is the CAG repeat length in weaning tail DNA".

Estimate	Estimate	p-value
(Intercept)	-3.17E-02	2.00E-16
age	1.36E-03	1.70E-10
sexM	1.49E-03	0.483
strainB6	2.77E-03	0.438
strainCD1	8.04E-02	2.00E-16
strainFVB	3.34E-02	2.56E-13
cag.actual	-1.62E-05	0.792
strainB6:cag.actual	-6.46E-05	0.346
strainCD1:cag.actual	6.86E-06	0.922
strainFVB:cag.actual	-1.48E-05	0.843

Table S4. Enrichment of up- and down-regulated genes in the striatum of 16-20-week-old B6.Q111 mice for cell type-specific genes. Cell type specific genes were derived from translational profiling of 35 different cell types in the brain using the pSI R package (31). Table describes the brain region and cell type marker used in each translational profiling experiment, as well as p-values for the enrichment of genes specific to each each cell type (specificity p-value < 0.01) among differentially expressed genes (Fisher's exact test).

Region	Marker	Up-Regulated Genes	Down-Regulated Genes
Striatum	Drd1	1	7.4E-53
Striatum	Drd2	1	9.6E-44
Striatum	(whole tissue)	1	7.5E-29
Basal Forebrain	(whole tissue)	0.27	0.014
Spinal Cord	Chat	1	0.028
Basal Forebrain	Chat	0.17	0.029
Cerebellum	Pcp2	0.98	0.044
Striatum	Chat	0.42	0.12
Cerebellum	(whole tissue)	0.63	0.26
Cerebellum	Neurod1	0.92	0.29
Cortex	Aldh1L1	1	0.32
Brainstem	Chat	0.94	0.35
Cerebellum	Septin4	1	0.36
Cortex	Pnoc	0.76	0.36
Habenula	Chat	0.57	0.5
Cortex	Cort	0.68	0.55
Cerebellum	Grm2	0.74	0.59

Retina	Cones	0.94	0.65
Cerebellum	Lypd6	0.99	0.67
Cortex	(whole tissue)	0.065	0.71
Cortex	Etv1_ts88	0.95	0.74
Retina	Rods	0.99	0.87
Hypothalamus	(whole tissue)	1	0.87
Habenula	(whole tissue)	0.65	0.87
Cortex	Ntsr	0.13	0.89
Cortex	Glt25d2	0.0017	0.89
Spinal Cord	(whole tissue)	0.96	0.89
Cerebellum	Grp	0.87	0.94
Cerebellum	Aldh1l1	1	0.97
Hypothalamus	Hcrt	0.014	0.98
Cortex	Pdgfrajd340	0.8	0.99
Brainstem	(whole tissue)	1	1
Brainstem	Slc6a4	0.68	1
Cerebellum	Cnp	1	1
Cortex	Cnp	1	1

Table S5. Gene Ontology terms enriched among genes down-regulated in the striatum of 16-20-week-old B6.Q111 mice.

GO term	# DEGs	odds ratio	p-value	FDR
purine_ribonucleotide_catabolic_process	64	6.74	1.8E-29	8.3E-26
regulation_of_GTP_catabolic_process	61	7.12	2.3E-29	1.1E-25
regulation_of_ion_transmembrane_transport	57	7.08	1.9E-27	8.6E-24
GTP_metabolic_process	58	6.81	3.9E-27	1.8E-23
positive_regulation_of_GTPase_activity	51	7.84	1.4E-26	6.3E-23
neuronal_cell_body	60	5.55	7.6E-24	3.5E-20
calmodulin_binding	37	9.73	1.3E-22	5.8E-19
regulation_of_G-protein_coupled_receptor_protein_signaling_pathway	29	14.58	2.9E-22	1.4E-18
cognition	40	8.14	8.5E-22	3.9E-18
regulation_of_GTPase_activity	44	7.13	9.1E-22	4.2E-18
regulation_of_Ras_protein_signal_transduction	46	6.72	9.9E-22	4.6E-18
regulation_of_transmembrane_transporter_activity	33	10.11	7.9E-21	3.7E-17
GTPase_activator_activity	40	7.47	1.3E-20	6.1E-17
actin_filament-based_process	57	4.94	1.5E-20	6.8E-17
dendritic_spine	29	11.31	1.2E-19	5.4E-16
regulation_of_membrane_potential	47	5.67	1.9E-19	8.6E-16
regulation_of_Ras_GTPase_activity	35	7.98	4.6E-19	2.1E-15
regulation_of_ion_transport	50	5.17	5.0E-19	2.3E-15
Ras_protein_signal_transduction	49	5.04	2.8E-18	1.3E-14
neuron_projection_morphogenesis	49	5.01	3.4E-18	1.6E-14
intrinsic_component_of_plasma_membrane	40	6.15	6.2E-18	2.9E-14
axon	49	4.90	8.1E-18	3.7E-14
single-organism_behavior	43	5.60	9.3E-18	4.3E-14
regulation_of_small_GTPase_mediated_signal_transduction	47	4.92	3.2E-17	1.5E-13
positive_regulation_of_Ras_GTPase_activity	23	10.77	1.6E-15	7.3E-12
striated_muscle_tissue_development	40	5.13	1.7E-15	8.0E-12
response_to_hormone	48	4.29	1.8E-15	8.5E-12
cell_morphogenesis_involved_in_neuron_differentiation	43	4.61	5.2E-15	2.4E-11
postsynaptic_density	26	8.11	9.8E-15	4.5E-11
dopamine_receptor_signaling_pathway	12	47.55	1.0E-14	4.6E-11
glutamate_receptor_signaling_pathway	17	17.25	1.1E-14	4.9E-11
muscle_tissue_development	42	4.55	1.5E-14	7.0E-11
dendritic_shaft	18	14.80	1.7E-14	7.6E-11

membrane_raft	33	5.83	1.7E-14	7.7E-11
positive_regulation_of_hydrolase_activity	42	4.49	2.4E-14	1.1E-10
response_to_organic_cyclic_compound	46	4.09	3.4E-14	1.5E-10
calcium_ion_transport	36	5.14	3.9E-14	1.8E-10
transmission_of_nerve_impulse	24	8.43	4.7E-14	2.1E-10
regulation_of_growth	48	3.85	8.2E-14	3.8E-10
response_to_alkaloid	17	14.66	1.0E-13	4.6E-10
neuron-neuron_synaptic_transmission	21	9.90	1.1E-13	5.0E-10
muscle_structure_development	47	3.81	2.0E-13	9.2E-10
muscle_system_process	33	5.26	2.3E-13	1.1E-09
cellular_response_to_organonitrogen_compound	29	6.05	2.6E-13	1.2E-09
monovalent_inorganic_cation_transport	39	4.42	3.0E-13	1.4E-09
negative_regulation_of_cellular_protein_metabolic_process	42	4.13	3.0E-13	1.4E-09
contractile_fiber	28	6.08	5.9E-13	2.7E-09
actin_binding	38	4.39	7.1E-13	3.3E-09
regulation_of_ion_transmembrane_transporter_activity	20	9.51	8.0E-13	3.6E-09
muscle_contraction	29	5.70	9.9E-13	4.5E-09
positive_regulation_of_Rac_GTPase_activity	14	18.52	1.0E-12	4.8E-09
Rho_protein_signal_transduction	27	6.17	1.1E-12	5.1E-09
regulation_of_synaptic_transmission	28	5.86	1.3E-12	6.1E-09
ionotropic_glutamate_receptor_binding	13	21.61	1.4E-12	6.3E-09
regulation_of_cell_morphogenesis	34	4.74	1.5E-12	6.9E-09
cellular_response_to_hormone_stimulus	35	4.59	1.7E-12	7.7E-09
protein_serine/threonine_kinase_activity	44	3.74	2.0E-12	9.2E-09
learning_or_memory	24	6.89	2.3E-12	1.0E-08
divalent_inorganic_cation_transport	38	4.19	2.5E-12	1.2E-08
axon_ensheathment	18	10.47	2.6E-12	1.2E-08
regulation_of_synaptic_plasticity	18	10.47	2.6E-12	1.2E-08
axonogenesis	32	4.90	3.0E-12	1.3E-08
regulation_of_protein_kinase_activity	44	3.68	3.2E-12	1.4E-08
regulation_of_Rho_GTPase_activity	17	11.13	4.5E-12	2.0E-08
cation_channel_activity	28	5.50	5.2E-12	2.4E-08
cellular_response_to_insulin_stimulus	20	8.43	5.7E-12	2.6E-08
transmembrane_transporter_complex	32	4.76	5.9E-12	2.7E-08
neuronal_action_potential	23	6.90	6.2E-12	2.8E-08
regulation_of_neuron_differentiation	37	4.14	6.8E-12	3.1E-08
actin_cytoskeleton	32	4.69	8.4E-12	3.8E-08
actin_cytoskeleton_organization	38	4.01	8.4E-12	3.8E-08

negative_regulation_of_transport	36	4.20	8.5E-12	3.8E-08
postsynaptic_membrane	29	5.10	1.2E-11	5.4E-08
learning	20	7.91	1.6E-11	7.2E-08
ion_channel_activity	36	4.05	2.2E-11	1.0E-07
regulation_of_protein_localization	44	3.44	2.4E-11	1.1E-07
ephrin_receptor_binding	13	16.35	2.5E-11	1.1E-07
cardiac_chamber_development	22	6.75	2.6E-11	1.2E-07
action_potential	23	6.38	2.6E-11	1.2E-07
response_to_insulin	21	7.16	2.7E-11	1.2E-07
regulation_of_cell_morphogenesis_involved_in_differentiation	28	5.04	3.4E-11	1.5E-07
positive_regulation_of_cell_differentiation	42	3.51	3.7E-11	1.7E-07
regulation_of_cell_projection_organization	38	3.79	4.0E-11	1.8E-07
regulation_of_synaptic_transmission,_glutamatergic	10	30.27	4.2E-11	1.9E-07
developmental_growth	34	4.15	4.2E-11	1.9E-07
locomotory_behavior	22	6.56	4.3E-11	2.0E-07
GTPase_regulator_activity	21	6.89	5.1E-11	2.3E-07
cellular_response_to_peptide	23	6.12	5.6E-11	2.5E-07
cell_leading_edge	23	6.12	5.6E-11	2.5E-07
actin_filament-based_movement	17	9.25	5.8E-11	2.6E-07
response_to_inorganic_substance	29	4.74	6.1E-11	2.8E-07
costamere	12	17.66	6.8E-11	3.1E-07
muscle_cell_differentiation	32	4.27	7.8E-11	3.6E-07
positive_regulation_of_protein_kinase_activity	32	4.26	8.3E-11	3.8E-07
actin_cytoskeleton_reorganization	15	11.07	8.5E-11	3.9E-07
drug_binding	20	7.09	9.1E-11	4.1E-07
chemotaxis	37	3.72	1.1E-10	5.1E-07
dendrite	36	3.79	1.2E-10	5.4E-07
SH3_domain_binding	20	6.95	1.3E-10	5.7E-07
behavioral_response_to_cocaine	9	35.57	1.4E-10	6.3E-07
adenylate_cyclase-modulating_G-protein_coupled_receptor_signaling_pathway	15	10.61	1.4E-10	6.5E-07
positive_regulation_of_protein_phosphorylation	40	3.47	1.5E-10	6.7E-07
muscle_organ_development	33	4.04	1.5E-10	6.8E-07
cell_growth	34	3.94	1.5E-10	7.0E-07
striated_muscle_cell_differentiation	27	4.84	1.7E-10	7.7E-07
potassium_ion_transmembrane_transporter_activity	20	6.81	1.7E-10	7.8E-07
response_to_ammonium_ion	12	15.85	1.9E-10	8.7E-07
metal_ion_transmembrane_transporter_activity	34	3.89	2.0E-10	9.2E-07
response_to_amphetamine	9	33.03	2.2E-10	1.0E-06

cell_body	24	5.37	2.5E-10	1.1E-06
channel_activity	36	3.68	2.6E-10	1.2E-06
cell-substrate_adhesion	24	5.32	2.9E-10	1.3E-06
regulation_of_actin_filament-based_process	28	4.54	3.1E-10	1.4E-06
voltage-gated_potassium_channel_activity	16	8.70	4.7E-10	2.1E-06
regulation_of_cation_channel_activity	12	14.38	5.0E-10	2.2E-06
peptidyl-serine_phosphorylation	24	5.15	5.5E-10	2.5E-06
actin_filament_bundle_assembly	17	7.85	5.6E-10	2.5E-06
sarcomere	21	5.91	6.5E-10	2.9E-06
cardiac_muscle_tissue_development	23	5.30	7.3E-10	3.3E-06
protein_heterodimerization_activity	39	3.32	7.6E-10	3.4E-06
adherens_junction	37	3.44	8.3E-10	3.8E-06
leukocyte_differentiation	37	3.43	9.1E-10	4.1E-06
negative_regulation_of_protein_complex_assembly	15	9.00	1.1E-09	4.8E-06
cellular_extravasation	9	25.68	1.2E-09	5.4E-06
negative_regulation_of_phosphorylation	29	4.11	1.3E-09	5.7E-06
regulation_of_Rho_protein_signal_transduction	22	5.38	1.3E-09	5.9E-06
regulation_of_neuron_projection_development	25	4.70	1.4E-09	6.1E-06
regulation_of_protein_transport	34	3.59	1.4E-09	6.3E-06
response_to_peptide_hormone	23	5.05	1.7E-09	7.6E-06
G-protein_coupled_receptor_binding	22	5.28	1.8E-09	8.1E-06
cellular_response_to_organic_cyclic_compound	24	4.77	2.2E-09	9.8E-06
potassium_ion_transmembrane_transport	21	5.47	2.4E-09	1.1E-05
voltage-gated_ion_channel_activity	23	4.95	2.5E-09	1.1E-05
GTPase_binding	25	4.54	2.6E-09	1.2E-05
myofibril	22	5.16	2.7E-09	1.2E-05
positive_regulation_of_transferase_activity	34	3.49	2.7E-09	1.2E-05
negative_regulation_of_transmembrane_transport	9	22.01	3.5E-09	1.6E-05
voltage-gated_potassium_channel_complex	14	9.03	3.7E-09	1.6E-05
glutamate_receptor_activity	10	16.59	4.4E-09	2.0E-05
memory	14	8.80	4.9E-09	2.2E-05
ruffle_membrane	13	9.85	5.2E-09	2.3E-05
axon_development	30	3.73	5.5E-09	2.5E-05
transmembrane_receptor_protein_tyrosine_kinase_signaling_pathway	29	3.83	5.6E-09	2.5E-05
negative_regulation_of_protein_kinase_activity	20	5.45	5.8E-09	2.6E-05
peptidyl-threonine_phosphorylation	15	7.82	5.9E-09	2.6E-05
brain_development	37	3.18	6.0E-09	2.7E-05
positive_regulation_of_cell_development	23	4.68	6.6E-09	2.9E-05

monovalent_inorganic_cation_transmembrane_transporter_activity	27	4.03	6.7E-09	3.0E-05
positive_regulation_of_positive_chemotaxis	8	27.37	6.9E-09	3.1E-05
skeletal_muscle_contraction	10	15.58	7.2E-09	3.2E-05
dendrite_development	19	5.65	8.1E-09	3.6E-05
regulation_of_cellular_extravasation	7	39.89	9.4E-09	4.2E-05
cyclic_nucleotide_metabolic_process	20	5.25	1.0E-08	4.6E-05
activation_of_MAPK_activity	14	8.20	1.1E-08	4.8E-05
dendrite_morphogenesis	16	6.77	1.2E-08	5.3E-05
actin_filament_organization	26	4.04	1.2E-08	5.3E-05
inflammatory_response	33	3.34	1.2E-08	5.3E-05
developmental_growth_involved_in_morphogenesis	20	5.20	1.2E-08	5.3E-05
regulation_of_behavior	20	5.17	1.3E-08	5.8E-05
cell_junction_assembly	17	6.23	1.3E-08	5.8E-05
negative_regulation_of_locomotion	22	4.67	1.4E-08	6.2E-05
positive_regulation_of_cell_adhesion	18	5.78	1.4E-08	6.4E-05
enzyme_activator_activity	24	4.28	1.5E-08	6.9E-05
inorganic_cation_transmembrane_transporter_activity	35	3.17	1.6E-08	7.1E-05
small_GTPase_binding	22	4.61	1.7E-08	7.6E-05
regulation_of_cell-substrate_adhesion	16	6.45	2.2E-08	9.7E-05
cell-cell_junction_organization	18	5.60	2.2E-08	9.8E-05
regulation_of_establishment_of_protein_localization	34	3.19	2.2E-08	9.9E-05
activin_binding	7	32.63	2.5E-08	1.1E-04
ion_channel_complex	22	4.50	2.5E-08	1.1E-04
endomembrane_system_organization	23	4.32	2.6E-08	1.1E-04
myotube_differentiation	14	7.60	2.6E-08	1.2E-04
erythrocyte_differentiation	14	7.60	2.6E-08	1.2E-04
apical_part_of_cell	17	5.89	2.7E-08	1.2E-04
B_cell_activation	19	5.20	2.7E-08	1.2E-04
neuromuscular_process_controlling_balance	13	8.37	2.9E-08	1.3E-04
activation_of_protein_kinase_B_activity	8	21.60	2.9E-08	1.3E-04
focal_adhesion	32	3.27	3.3E-08	1.5E-04
cell-substrate_junction_assembly	13	8.27	3.3E-08	1.5E-04
regulation_of_neurological_system_process	13	8.27	3.3E-08	1.5E-04
positive_regulation_of_cell-matrix_adhesion	10	12.85	3.4E-08	1.5E-04
regulation_of_cellular_component_size	23	4.21	3.9E-08	1.7E-04
membrane_hyperpolarization	8	20.53	4.0E-08	1.8E-04
regulation_of_actin_cytoskeleton_organization	23	4.20	4.2E-08	1.9E-04
sodium_ion_transport	20	4.79	4.2E-08	1.9E-04

epithelial_cell_differentiation	34	3.10	4.2E-08	1.9E-04
regulation_of_sodium_ion_transport	11	10.48	4.4E-08	2.0E-04
regulation_of_cellular_response_to_insulin_stimulus	11	10.48	4.4E-08	2.0E-04
protein_dephosphorylation	20	4.75	4.8E-08	2.1E-04
response_to_wounding	36	2.96	5.2E-08	2.3E-04
cell-substrate_junction	32	3.19	5.4E-08	2.4E-04
positive_regulation_of_organelle_organization	27	3.61	5.5E-08	2.4E-04
small_GTPase_regulator_activity	16	5.94	6.2E-08	2.8E-04
G-protein_coupled_receptor_signaling_pathway,_coupled_to_cyclic_nucleotide_second_messenger	14	7.01	6.4E-08	2.8E-04
potassium_ion_transport	22	4.25	6.5E-08	2.9E-04
homeostasis_of_number_of_cells	22	4.25	6.5E-08	2.9E-04
insulin_receptor_signaling_pathway	14	6.94	7.2E-08	3.2E-04
negative_regulation_of_ion_transport	11	9.93	7.2E-08	3.2E-04
potassium_channel_activity	16	5.86	7.4E-08	3.3E-04
in_utero_embryonic_development	32	3.14	7.5E-08	3.3E-04
developmental_cell_growth	16	5.81	8.1E-08	3.6E-04
striated_muscle_cell_development	17	5.42	8.3E-08	3.7E-04
regulation_of_sodium_ion_transmembrane_transport	9	14.01	8.8E-08	3.9E-04
cellular_response_to_peptide_hormone_stimulus	16	5.77	8.9E-08	3.9E-04
positive_regulation_of_blood_pressure	10	11.42	8.9E-08	3.9E-04
negative_regulation_of_apoptotic_process	35	2.93	9.8E-08	4.3E-04
cell_maturation	14	6.68	1.1E-07	4.8E-04
positive_regulation_of_transporter_activity	11	9.43	1.1E-07	5.0E-04
skeletal_muscle_fiber_development	11	9.43	1.1E-07	5.0E-04
scaffold_protein_binding	11	9.43	1.1E-07	5.0E-04
positive_regulation_of_peptidyl-serine_phosphorylation	12	8.24	1.1E-07	5.0E-04
protein_localization_to_plasma_membrane	16	5.66	1.2E-07	5.1E-04
neuron_projection_extension	13	7.28	1.3E-07	5.5E-04
positive_regulation_of_heart_contraction	7	23.92	1.3E-07	5.6E-04
regulation_of_sodium_ion_transmembrane_transporter_activity	7	23.92	1.3E-07	5.6E-04
phosphatidylinositol-4,5-bisphosphate_binding	10	10.94	1.3E-07	5.6E-04
negative_regulation_of_cytoskeleton_organization	14	6.56	1.3E-07	5.9E-04
pre-B_cell_differentiation	5	85.48	1.4E-07	6.1E-04
protein_N-terminus_binding	16	5.54	1.5E-07	6.5E-04
positive_regulation_of_kinase_activity	28	3.33	1.6E-07	6.8E-04
forebrain_development	27	3.41	1.6E-07	7.1E-04
regulation_of_intracellular_protein_transport	21	4.18	1.7E-07	7.3E-04
phosphoprotein_phosphatase_activity	20	4.35	1.8E-07	8.0E-04

monoamine_transport	12	7.82	1.9E-07	8.3E-04
telencephalon_development	20	4.31	2.0E-07	9.0E-04
termination_of_G-protein_coupled_receptor_signaling_pathway	10	10.28	2.1E-07	9.3E-04
regulation_of_epidermal_growth_factor_receptor_signaling_pathway	8	15.79	2.1E-07	9.4E-04
regulation_of_glucose_transport	11	8.70	2.3E-07	1.0E-03
cellular_response_to_reactive_oxygen_species	11	8.70	2.3E-07	1.0E-03
protein_kinase_C_binding	11	8.70	2.3E-07	1.0E-03
gliogenesis	20	4.26	2.5E-07	1.1E-03
positive_regulation_of_synaptic_transmission	9	12.16	2.5E-07	1.1E-03
heart_development	32	2.96	2.6E-07	1.1E-03
neuronal_postsynaptic_density	8	15.21	2.7E-07	1.2E-03
myeloid_cell_homeostasis	16	5.26	2.8E-07	1.2E-03
regulation_of_blood_pressure	15	5.65	2.9E-07	1.3E-03
regulation_of_MAP_kinase_activity	18	4.65	2.9E-07	1.3E-03
negative_regulation_of_transferase_activity	20	4.20	2.9E-07	1.3E-03
fibroblast_migration	9	11.85	3.0E-07	1.3E-03
positive_regulation_of_cell_projection_organization	13	6.69	3.0E-07	1.3E-03
actin-mediated_cell_contraction	11	8.44	3.1E-07	1.3E-03
sodium_ion_transmembrane_transport	15	5.61	3.1E-07	1.4E-03
protein_targeting_to_membrane	10	9.70	3.4E-07	1.5E-03
regulation_of_receptor_activity	10	9.70	3.4E-07	1.5E-03
positive_regulation_of_glycoprotein_metabolic_process	6	30.76	3.4E-07	1.5E-03
glutamate_receptor_binding	8	14.67	3.4E-07	1.5E-03
G1/S_transition_of_mitotic_cell_cycle	17	4.85	3.5E-07	1.6E-03
phosphatase_activity	22	3.82	3.6E-07	1.6E-03
regulation_of_cell_size	13	6.56	3.7E-07	1.6E-03
divalent_metal_ion_transport	26	3.34	3.9E-07	1.7E-03
glial_cell_development	10	9.52	3.9E-07	1.7E-03
Schwann_cell_differentiation	8	14.16	4.3E-07	1.9E-03
cellular_chemical_homeostasis	32	2.88	4.5E-07	1.9E-03
long-term_synaptic_potentiation	10	9.34	4.6E-07	2.0E-03
myeloid_cell_differentiation	25	3.40	4.7E-07	2.0E-03
mitochondrial_transport	12	7.10	4.8E-07	2.1E-03
regulation_of_MAPK_cascade	32	2.87	5.0E-07	2.2E-03
regulation_of_proton_transport	6	27.93	5.2E-07	2.3E-03
regulation_of_cytokine_production	29	3.04	5.6E-07	2.4E-03
cytokine_production	32	2.85	5.7E-07	2.5E-03
negative_regulation_of_protein_complex_disassembly	11	7.86	5.8E-07	2.5E-03

cellular_homeostasis	31	2.90	5.8E-07	2.5E-03
cAMP_biosynthetic_process	13	6.26	6.1E-07	2.6E-03
activation_of_Ras_GTPase_activity	7	17.95	6.1E-07	2.6E-03
catecholamine_secretion	9	10.75	6.1E-07	2.7E-03
cell-cell_junction	27	3.16	6.4E-07	2.8E-03
fat_cell_differentiation	15	5.26	6.6E-07	2.9E-03
glucose_import	11	7.65	7.4E-07	3.2E-03
axon_extension	12	6.79	7.4E-07	3.2E-03
positive_regulation_of_protein_binding	10	8.71	8.1E-07	3.5E-03
regulation_of_calcium_ion_transmembrane_transport	10	8.71	8.1E-07	3.5E-03
protein_kinase_C-activating_G-protein_coupled_receptor_signaling_pathway	8	12.83	8.2E-07	3.5E-03
positive_regulation_of_sodium_ion_transport	8	12.83	8.2E-07	3.5E-03
phosphatidylinositol-3-phosphate_binding	8	12.83	8.2E-07	3.5E-03
growth_cone	15	5.15	8.3E-07	3.6E-03
adenylate_cyclase-activating_G-protein_coupled_receptor_signaling_pathway	9	10.27	8.6E-07	3.7E-03
protein_localization_to_membrane	20	3.90	8.8E-07	3.8E-03
ion_channel_binding	13	6.03	8.8E-07	3.8E-03
heart_morphogenesis	20	3.89	9.2E-07	4.0E-03
glial_cell_differentiation	17	4.48	9.9E-07	4.3E-03
olfactory_lobe_development	7	16.32	1.0E-06	4.5E-03
apical_plasma_membrane	22	3.57	1.0E-06	4.5E-03
regulation_of_developmental_growth	16	4.72	1.1E-06	4.7E-03
positive_regulation_of_MAPK_cascade	22	3.56	1.1E-06	4.8E-03
blood_circulation	27	3.06	1.1E-06	4.9E-03
kinase_binding	21	3.67	1.2E-06	5.2E-03
regulation_of_neuronal_synaptic_plasticity	8	12.07	1.2E-06	5.2E-03
protein_phosphatase_inhibitor_activity	8	12.07	1.2E-06	5.2E-03
negative_regulation_of_cell_differentiation	27	3.04	1.3E-06	5.5E-03
regulation_of_cytoskeleton_organization	25	3.21	1.3E-06	5.5E-03
regulation_of_cell_adhesion	23	3.40	1.3E-06	5.6E-03
regulation_of_protein_tyrosine_kinase_activity	7	15.61	1.3E-06	5.7E-03
myelination	11	7.16	1.3E-06	5.7E-03
positive_regulation_of_intracellular_signal_transduction	29	2.89	1.4E-06	5.9E-03
protein_complex_scaffold	10	8.16	1.4E-06	6.0E-03
cardiocyte_differentiation	14	5.30	1.4E-06	6.0E-03
acid-amino_acid_ligase_activity	23	3.37	1.5E-06	6.4E-03
dephosphorylation	26	3.09	1.5E-06	6.5E-03
atrioventricular_valve_development	6	21.95	1.5E-06	6.7E-03

positive_regulation_of_protein_serine/threonine_kinase_activity	17	4.32	1.5E-06	6.7E-03
Ras_GTPase_binding	10	8.03	1.6E-06	6.8E-03
regulation_of_heart_rate	9	9.43	1.6E-06	6.9E-03
regulation_of_homeostatic_process	24	3.25	1.6E-06	7.0E-03
extrinsic_component_of_membrane	17	4.30	1.6E-06	7.1E-03
potassium_channel_regulator_activity	7	14.96	1.7E-06	7.3E-03
long-term_memory	8	11.40	1.8E-06	7.6E-03
regulation_of_actin_filament_depolymerization	8	11.40	1.8E-06	7.6E-03
ligase_activity	31	2.73	1.9E-06	8.4E-03
phosphoric_ester_hydrolase_activity	29	2.83	2.0E-06	8.7E-03
actin_filament_binding	14	5.11	2.1E-06	8.9E-03
heart_contraction	16	4.46	2.1E-06	9.1E-03
calmodulin-dependent_protein_kinase_activity	7	14.36	2.1E-06	9.1E-03
epidermal_growth_factor_receptor_signaling_pathway	11	6.73	2.3E-06	9.8E-03
peripheral_nervous_system_development	10	7.67	2.3E-06	9.8E-03
nucleic_acid_binding_transcription_factor_activity	30	2.74	2.6E-06	1.1E-02
regulation_of_glycoprotein_biosynthetic_process	7	13.81	2.6E-06	1.1E-02
negative_regulation_of_protein_phosphorylation	19	3.76	2.7E-06	1.2E-02
nuclear_hormone_receptor_binding	13	5.40	2.7E-06	1.2E-02
lymphocyte_differentiation	20	3.60	2.7E-06	1.2E-02
positive_regulation_of_protein_export_from_nucleus	6	19.21	2.9E-06	1.2E-02
positive_regulation_of_renal_sodium_excretion	5	31.99	3.0E-06	1.3E-02
regulation_of_chemotaxis	14	4.94	3.0E-06	1.3E-02
glycogen_metabolic_process	11	6.50	3.1E-06	1.3E-02
cyclic_purine_nucleotide_metabolic_process	13	5.31	3.2E-06	1.4E-02
cell-cell_adherens_junction	9	8.56	3.3E-06	1.4E-02
direct_ligand_regulated_sequence-specific_DNA_binding_transcription_factor_activity	7	13.30	3.3E-06	1.4E-02
regulation_of_protein_depolymerization	11	6.43	3.4E-06	1.5E-02
sarcolemma	14	4.87	3.5E-06	1.5E-02
regulation_of_leukocyte_migration	12	5.77	3.5E-06	1.5E-02
regulation_of_axonogenesis	12	5.77	3.5E-06	1.5E-02
negative_regulation_of_protein_polymerization	8	10.27	3.5E-06	1.5E-02
filamentous_actin	8	10.27	3.5E-06	1.5E-02
cellular_ion_homeostasis	29	2.74	3.7E-06	1.6E-02
muscle_cell_development	17	4.02	3.8E-06	1.6E-02
proton_transport	11	6.35	3.8E-06	1.6E-02
regulation_of_protein_complex_disassembly	11	6.35	3.8E-06	1.6E-02
protein_complex_binding	30	2.68	3.8E-06	1.6E-02

regulation_of_fatty_acid_oxidation	6	18.08	3.8E-06	1.6E-02
cyclic_nucleotide_biosynthetic_process	14	4.81	4.0E-06	1.7E-02
cardiac_septum_development	10	7.14	4.1E-06	1.7E-02
response_to_morphine	5	28.42	4.5E-06	1.9E-02
positive_regulation_of_insulin_receptor_signaling_pathway	5	28.42	4.5E-06	1.9E-02
peroxidase_activity	7	12.37	4.9E-06	2.1E-02
transforming_growth_factor_beta-activated_receptor_activity	6	17.08	5.0E-06	2.1E-02
positive_regulation_of_programmed_cell_death	15	4.39	5.1E-06	2.2E-02
polysaccharide_metabolic_process	13	5.07	5.1E-06	2.2E-02
Rho_GTPase_binding	8	9.55	5.6E-06	2.4E-02
regulation_of_heart_contraction	13	4.99	6.0E-06	2.6E-02
cellular_carbohydrate_metabolic_process	21	3.28	6.1E-06	2.6E-02
cell_cortex	18	3.69	6.2E-06	2.6E-02
lyase_activity	17	3.86	6.2E-06	2.7E-02
lipid_localization	19	3.53	6.3E-06	2.7E-02
glucose_transport	12	5.42	6.3E-06	2.7E-02
voltage-gated_cation_channel_activity	12	5.42	6.3E-06	2.7E-02
vocalization_behavior	6	16.18	6.5E-06	2.8E-02
positive_regulation_of_glucose_import	8	9.33	6.5E-06	2.8E-02
positive_regulation_of_myelination	5	25.59	6.7E-06	2.9E-02
cardiac_ventricle_development	12	5.37	6.9E-06	2.9E-02
phosphatidylinositol_phosphate_binding	10	6.67	7.0E-06	3.0E-02
motor_neuron_axon_guidance	7	11.58	7.2E-06	3.1E-02
ventricular_cardiac_muscle_tissue_morphogenesis	7	11.58	7.2E-06	3.1E-02
UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate-D-alanyl-D-alanine_ligase_activity	21	3.24	7.3E-06	3.1E-02
ribosomal_S6-glutamic_acid_ligase_activity	21	3.24	7.3E-06	3.1E-02
coenzyme_F420-0_gamma-glutamyl_ligase_activity	21	3.24	7.3E-06	3.1E-02
coenzyme_F420-2_alpha-glutamyl_ligase_activity	21	3.24	7.3E-06	3.1E-02
protein-glycine_ligase_activity	21	3.24	7.3E-06	3.1E-02
protein-glycine_ligase_activity,_initiating	21	3.24	7.3E-06	3.1E-02
protein-glycine_ligase_activity,_elongating	21	3.24	7.3E-06	3.1E-02
tubulin-glycine_ligase_activity	21	3.24	7.3E-06	3.1E-02
protein-glutamic_acid_ligase_activity	21	3.24	7.3E-06	3.1E-02
tubulin-glutamic_acid_ligase_activity	21	3.24	7.3E-06	3.1E-02
membrane_depolarization	11	5.89	7.3E-06	3.1E-02
carbohydrate_transport	14	4.53	7.5E-06	3.2E-02
hydrogen_peroxide_catabolic_process	6	15.37	8.3E-06	3.5E-02
heart_valve_development	7	11.21	8.6E-06	3.7E-02

associative_learning	7	11.21	8.6E-06	3.7E-02
retinoic_acid_receptor_signaling_pathway	7	11.21	8.6E-06	3.7E-02
delayed_rectifier_potassium_channel_activity	8	8.92	8.7E-06	3.7E-02
cellular_cation_homeostasis	28	2.66	8.9E-06	3.8E-02
stress_fiber_assembly	10	6.42	9.5E-06	4.0E-02
DNA_damage_response,_signal_transduction_by_p53_class_mediator_resulting_in_cell_cycle_arrest	5	23.26	9.6E-06	4.1E-02
exploration_behavior	5	23.26	9.6E-06	4.1E-02
activin_receptor_signaling_pathway	7	10.88	1.0E-05	4.3E-02
regulation_of_long-term_neuronal_synaptic_plasticity	7	10.88	1.0E-05	4.3E-02
positive_regulation_of_developmental_growth	7	10.88	1.0E-05	4.3E-02
regulation_of_myelination	6	14.64	1.0E-05	4.4E-02
protein_complex_disassembly	13	4.71	1.1E-05	4.5E-02
regulation_of_cell_migration	25	2.81	1.1E-05	4.8E-02
protein_tyrosine_phosphatase_activity	13	4.68	1.1E-05	4.8E-02

Table S6. Gene Ontology terms enriched among gene up-regulated in the striatum of 16-20-week-old B6.Q111 mice.

GO term	# DEGs	odds ratio	p-value	FDR
cell-substrate_junction	53	4.42	2.7E-17	1.3E-13
axon_development	47	4.87	5.3E-17	2.4E-13
regulation_of_cell_projection_organization	53	4.34	5.4E-17	2.5E-13
adherens_junction	52	3.97	3.0E-15	1.4E-11
regulation_of_neuron_differentiation	47	4.27	4.8E-15	2.2E-11
focal_adhesion	48	4.05	1.6E-14	7.6E-11
negative_regulation_of_phosphorylation	40	4.66	3.6E-14	1.7E-10
monovalent_inorganic_cation_transport	44	3.99	2.9E-13	1.3E-09
neuron_projection_morphogenesis	47	3.77	3.2E-13	1.5E-09
translation	49	3.59	5.3E-13	2.4E-09
cell_morphogenesis_involved_in_neuron_differentiation	45	3.82	6.2E-13	2.8E-09
centrosome	42	3.87	2.4E-12	1.1E-08
positive_regulation_of_neuron_projection_development	22	7.48	5.6E-12	2.6E-08
regulation_of_cell_morphogenesis_involved_in_differentiation	32	4.64	1.3E-11	6.0E-08
regulation_of_release_of_sequestered_calcium_ion_into_cytosol_by_sarcoplasmic _reticulum	10	37.08	1.8E-11	8.1E-08
negative_regulation_of_cellular_protein_metabolic_process	44	3.43	2.9E-11	1.3E-07
negative_regulation_of_peptidyl-serine_phosphorylation	10	29.15	9.1E-11	4.2E-07
regulation_of_cell_morphogenesis	35	3.87	1.5E-10	6.8E-07
axonogenesis	33	4.01	2.1E-10	9.5E-07
regulation_of_cell_size	18	7.61	3.5E-10	1.6E-06
regulation_of_neuron_projection_development	29	4.39	3.5E-10	1.6E-06
regulation_of_homeostatic_process	34	3.77	4.7E-10	2.2E-06
regulation_of_cellular_component_size	29	4.31	5.2E-10	2.4E-06
developmental_cell_growth	20	5.95	1.8E-09	8.1E-06
vacuole	41	3.08	2.4E-09	1.1E-05
microtubule-based_process	40	3.13	2.5E-09	1.2E-05

sarcoplasmic_reticulum_calcium_ion_transport	10	17.75	3.5E-09	1.6E-05
growth_cone	20	5.66	3.8E-09	1.7E-05
negative_regulation_of_intracellular_signal_transduction	28	4.00	4.6E-09	2.1E-05
release_of_sequestered_calcium_ion_into_cytosol	16	7.28	5.6E-09	2.6E-05
regulation_of_ion_homeostasis	22	4.88	7.7E-09	3.5E-05
ubiquitin_protein_ligase_binding	26	4.15	8.2E-09	3.8E-05
peptidyl-serine_phosphorylation	25	4.26	9.3E-09	4.3E-05
lipid_binding	42	2.88	9.4E-09	4.3E-05
heat_shock_protein_binding	13	9.16	1.4E-08	6.3E-05
cytosolic_calcium_ion_transport	18	5.81	1.5E-08	7.1E-05
carbohydrate_catabolic_process	19	5.45	1.6E-08	7.3E-05
regulation_of_ion_transport	38	3.00	1.8E-08	8.0E-05
peptide_hormone_processing	7	35.62	2.5E-08	1.2E-04
regulation_of_axonogenesis	16	6.36	3.1E-08	1.4E-04
protein_kinase_inhibitor_activity	12	9.61	3.1E-08	1.4E-04
autophagy	20	4.91	3.3E-08	1.5E-04
ATP_hydrolysis_coupled_proton_transport	9	16.69	3.3E-08	1.5E-04
translational_initiation	17	5.85	3.5E-08	1.6E-04
neuron_projection_extension	15	6.82	3.6E-08	1.7E-04
negative_regulation_of_locomotion	24	4.07	4.0E-08	1.8E-04
negative_regulation_of_transcription,_DNA-templated	44	2.65	4.3E-08	2.0E-04
actin_cytoskeleton_organization	36	2.98	4.4E-08	2.0E-04
response_to_lipid	34	3.10	4.5E-08	2.1E-04
phosphoric_ester_hydrolase_activity	37	2.92	4.9E-08	2.2E-04
protein_kinase_complex	14	7.25	5.2E-08	2.4E-04
neuronal_cell_body	40	2.77	5.6E-08	2.5E-04
glycerolipid_metabolic_process	28	3.52	5.6E-08	2.6E-04
phosphoprotein_binding	15	6.46	6.8E-08	3.1E-04
sequestering_of_calcium_ion	16	5.95	7.0E-08	3.2E-04
regulation_of_sequestering_of_calcium_ion	16	5.95	7.0E-08	3.2E-04
chaperone-mediated_protein_folding	10	11.66	9.4E-08	4.3E-04
positive_regulation_of_cell_projection_organization	15	6.26	9.8E-08	4.5E-04

filopodium_membrane	8	18.11	1.2E-07	5.4E-04
chemotaxis	36	2.85	1.3E-07	5.7E-04
negative_regulation_of_kinase_activity	20	4.48	1.3E-07	5.8E-04
pyrimidine_nucleotide_biosynthetic_process	9	13.59	1.4E-07	6.3E-04
protein_ubiquitination	35	2.89	1.4E-07	6.4E-04
regulation_of_ion_transmembrane_transport	33	3.00	1.4E-07	6.5E-04
regulation_of_endocytosis	19	4.64	1.6E-07	7.4E-04
cell_aging	13	7.18	1.7E-07	7.7E-04
ATPase_activity,_coupled	21	4.16	2.0E-07	8.9E-04
regulation_of_cell_growth	18	4.82	2.0E-07	9.1E-04
response_to_starvation	16	5.46	2.0E-07	9.3E-04
protein_folding	19	4.55	2.1E-07	9.4E-04
perinuclear_region_of_cytoplasm	42	2.54	2.3E-07	1.0E-03
regulation_of_calcium_ion_transport_into_cytosol	12	7.78	2.3E-07	1.1E-03
membrane_coat	14	6.29	2.5E-07	1.1E-03
regulation_of_ubiquitin-protein_transferase_activity	7	21.93	2.7E-07	1.2E-03
RNA_splicing	30	3.09	2.8E-07	1.3E-03
AMP-activated_protein_kinase_complex	6	34.87	2.9E-07	1.3E-03
negative_regulation_of_cell_development	17	4.97	2.9E-07	1.3E-03
microtubule	29	3.15	3.0E-07	1.4E-03
glycolytic_process	13	6.64	3.8E-07	1.7E-03
negative_regulation_of_DNA_damage_response,_signal_transduction_by_p53_clas s_mediator	7	20.36	4.0E-07	1.8E-03
protein_complex_binding	37	2.66	4.0E-07	1.8E-03
structural_constituent_of_muscle	8	14.81	4.0E-07	1.8E-03
developmental_growth_involved_in_morphogenesis	20	4.12	4.4E-07	2.0E-03
phospholipid_biosynthetic_process	18	4.47	5.5E-07	2.5E-03
actin_filament-based_process	39	2.55	5.8E-07	2.6E-03
ribonucleoprotein_complex_assembly	13	6.25	7.0E-07	3.2E-03
regulation_of_striated_muscle_contraction	12	6.90	7.3E-07	3.3E-03
regulation_of_calcium_ion_transmembrane_transport	11	7.74	7.7E-07	3.5E-03
ATPase_activity	30	2.92	7.9E-07	3.6E-03

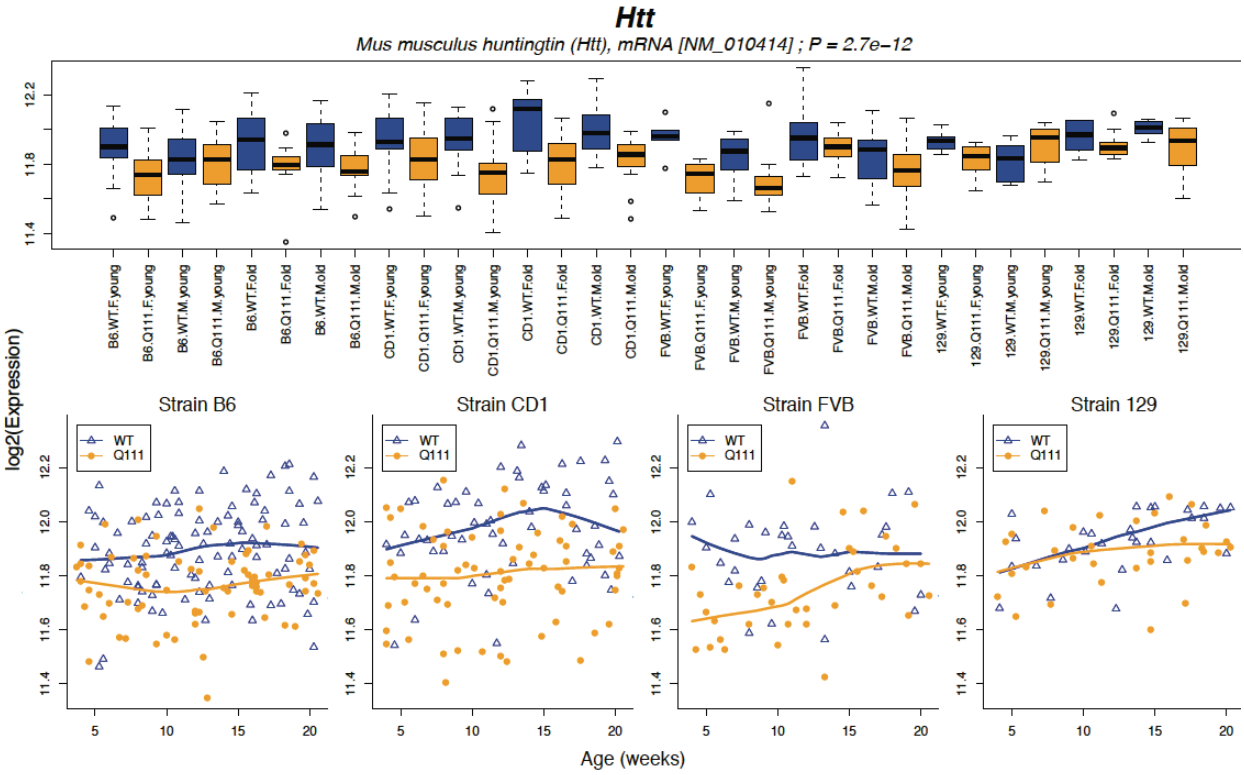
cranial_nerve_morphogenesis	7	17.82	8.1E-07	3.7E-03
axon_guidance	18	4.34	8.1E-07	3.7E-03
lamellipodium	18	4.34	8.1E-07	3.7E-03
protein_kinase_binding	35	2.66	8.2E-07	3.7E-03
cell_cortex	22	3.64	8.3E-07	3.8E-03
RNA_polymerase_II_repressing_transcription_factor_binding	9	10.48	8.8E-07	4.0E-03
proton_transport	13	6.11	8.8E-07	4.0E-03
striated_muscle_contraction	14	5.56	9.4E-07	4.3E-03
negative_regulation_of_protein_kinase_activity	19	4.08	9.7E-07	4.4E-03
regulation_of_calcium_ion_transport	19	4.08	9.7E-07	4.4E-03
mRNA_processing	33	2.73	9.9E-07	4.5E-03
aging	16	4.78	1.0E-06	4.6E-03
glycerophospholipid_metabolic_process	19	4.03	1.1E-06	5.0E-03
brain_morphogenesis	8	12.54	1.2E-06	5.2E-03
regulation_of_heart_rate	10	8.50	1.2E-06	5.2E-03
cardiac_muscle_contraction	14	5.45	1.2E-06	5.2E-03
cytosolic_calcium_ion_homeostasis	18	4.21	1.2E-06	5.4E-03
axon_extension	13	5.90	1.2E-06	5.6E-03
protein_K11-linked_ubiquitination	9	9.92	1.3E-06	5.9E-03
RNA_splicing,_via_transesterification_reactions	18	4.16	1.4E-06	6.3E-03
regulation_of_muscle_contraction	16	4.64	1.4E-06	6.5E-03
positive_regulation_of_ubiquitin-protein_transferase_activity	7	15.84	1.5E-06	6.9E-03
negative_regulation_of_protein_phosphorylation	22	3.50	1.5E-06	6.9E-03
hydrogen_ion_transmembrane_transport	12	6.36	1.6E-06	7.0E-03
unfolded_protein_binding	12	6.36	1.6E-06	7.0E-03
postsynaptic_membrane	24	3.27	1.6E-06	7.1E-03
oxidoreductase_activity,_acting_on_paired_donors,_with_incorporation_or_reduction_of_molecular_oxygen,_2-oxoglutarate_as_one_donor,_and_incorporation_of_one_atom_each_of_oxygen_into_both_donors	11	7.13	1.6E-06	7.1E-03
kinase_inhibitor_activity	10	8.16	1.6E-06	7.2E-03
negative_chemotaxis	8	11.64	1.8E-06	8.3E-03

regulation_of_alternative_mRNA_splicing,_via_spliceosome	9	9.41	1.9E-06	8.6E-03
ubiquitin_ligase_complex	20	3.71	2.0E-06	8.8E-03
long_term_synaptic_depression	7	15.00	2.0E-06	9.2E-03
protein_kinase_A_catalytic_subunit_binding	7	15.00	2.0E-06	9.2E-03
hippocampus_development	10	7.85	2.2E-06	9.8E-03
regulation_of_cell-substrate_adhesion	15	4.76	2.3E-06	1.0E-02
negative_regulation_of_transferase_activity	21	3.51	2.4E-06	1.1E-02
epithelial_cell_differentiation	35	2.52	2.5E-06	1.1E-02
calcium_ion_transport_into_cytosol	14	5.06	2.5E-06	1.1E-02
lipid_biosynthetic_process	32	2.64	2.6E-06	1.2E-02
spliceosomal_snRNP_assembly	7	14.25	2.7E-06	1.2E-02
actin-mediated_cell_contraction	11	6.70	2.7E-06	1.2E-02
positive_regulation_of_neurogenesis	15	4.68	2.7E-06	1.2E-02
nuclear_speck	15	4.68	2.7E-06	1.2E-02
response_to_hormone	36	2.46	3.0E-06	1.4E-02
GTP_binding	30	2.72	3.0E-06	1.4E-02
regulation_of_sodium_ion_transport	10	7.42	3.4E-06	1.5E-02
cation-transporting_ATPase_activity	11	6.51	3.5E-06	1.6E-02
protein_homotrimerization	8	10.51	3.5E-06	1.6E-02
positive_regulation_of_protein_ubiquitination	11	6.41	3.9E-06	1.8E-02
guanyl_nucleotide_binding	31	2.63	4.0E-06	1.8E-02
nuclease_activity	19	3.67	4.0E-06	1.8E-02
chaperone_mediated_protein_folding_requiring_cofactor	6	18.78	4.0E-06	1.8E-02
ribosome_biogenesis	16	4.25	4.1E-06	1.8E-02
regulation_of_vesicle-mediated_transport	25	2.99	4.2E-06	1.9E-02
regulation_of_developmental_growth	17	4.00	4.4E-06	2.0E-02
ER_to_Golgi_vesicle-mediated_transport	10	7.16	4.5E-06	2.0E-02
mRNA_splice_site_selection	7	12.96	4.5E-06	2.0E-02
retinal_ganglion_cell_axon_guidance	7	12.96	4.5E-06	2.0E-02
peptide_biosynthetic_process	7	12.96	4.5E-06	2.0E-02
regulation_of_synaptic_transmission	21	3.36	4.6E-06	2.1E-02
positive_regulation_of_cell_development	21	3.35	4.8E-06	2.2E-02

regulation_of_growth	38	2.34	4.9E-06	2.2E-02
positive_regulation_of_ion_transport	15	4.44	4.9E-06	2.2E-02
phospholipid_metabolic_process	22	3.23	5.0E-06	2.2E-02
vesicle_localization	17	3.95	5.1E-06	2.3E-02
pyruvate_metabolic_process	10	7.03	5.2E-06	2.3E-02
regulation_of_mitosis	10	7.03	5.2E-06	2.3E-02
neuron_projection_terminus	13	5.11	5.3E-06	2.4E-02
negative_regulation_of_cell_projection_organization	9	8.15	5.3E-06	2.4E-02
hydrolase_activity,_acting_on_acid_anhydrides,_catalyzing_transmembrane_mov ement_of_substances	12	5.57	5.4E-06	2.4E-02
ATPase_activity,_coupled_to_movement_of_substances	12	5.57	5.4E-06	2.4E-02
calcium_ion_transmembrane_transport	19	3.59	5.4E-06	2.4E-02
monosaccharide_binding	6	17.44	5.6E-06	2.5E-02
protein_complex_disassembly	15	4.38	5.8E-06	2.6E-02
cytoplasmic_vesicle_membrane	15	4.38	5.8E-06	2.6E-02
phosphatidylinositol_biosynthetic_process	7	12.39	5.8E-06	2.6E-02
early_endosome_to_late_endosome_transport	7	12.39	5.8E-06	2.6E-02
isomerase_activity	17	3.91	5.8E-06	2.6E-02
monocarboxylic_acid_metabolic_process	31	2.58	5.8E-06	2.6E-02
monocarboxylic_acid_transport	14	4.65	6.2E-06	2.8E-02
muscle_tissue_development	31	2.57	6.3E-06	2.8E-02
cAMP_binding	8	9.59	6.3E-06	2.8E-02
protein_serine/threonine_kinase_activity	36	2.36	7.0E-06	3.1E-02
purine_nucleoside_monophosphate_catabolic_process	26	2.82	7.3E-06	3.2E-02
gliogenesis	20	3.37	7.3E-06	3.3E-02
long-chain_fatty_acid_transport	7	11.88	7.3E-06	3.3E-02
response_to_organic_cyclic_compound	35	2.39	7.5E-06	3.4E-02
cytoplasmic_membrane-bounded_vesicle	35	2.39	7.5E-06	3.4E-02
B_cell_homeostasis	8	9.31	7.6E-06	3.4E-02
ephrin_receptor_activity	6	16.28	7.7E-06	3.4E-02
regulation_of_cardiac_muscle_contraction	9	7.65	8.4E-06	3.7E-02
positive_regulation_of_cell_differentiation	36	2.34	8.6E-06	3.8E-02

blood_circulation	29	2.62	8.8E-06	3.9E-02
cell_growth	29	2.61	9.1E-06	4.0E-02
regulation_of_cellular_response_to_growth_factor_stimulus	16	3.97	9.1E-06	4.1E-02
muscle_system_process	24	2.92	9.4E-06	4.2E-02
protein_binding_transcription_factor_activity	34	2.39	9.9E-06	4.4E-02
dioxygenase_activity	13	4.78	1.0E-05	4.4E-02
transcription_from_RNA_polymerase_I_promoter	6	15.26	1.0E-05	4.6E-02

Fig S1. *Htt* gene expression across strains.



3.7 Notes and Acknowledgements

Note: This chapter was published in the February 2017 edition of *Human Molecular Genetics* as

> **High resolution time-course mapping of early transcriptomic, molecular and cellular phenotypes in Huntington's disease CAG knock-in mice across multiple genetic backgrounds.** Seth A. Ament[*] (1,8), Jocelynn R. Pearl[*] (1,7), Andrea Grindeland[*] (2), Jason St. Claire (3), John C. Earls (1,9), Marina Kovalenko (3), Tammy Gillis (3), Jayalakshmi Mysore (3), James F. Gusella (3), Jong-Min Lee (3), Seung Kwak (4), David Howland (4), Minyoung Lee (1), David Baxter (1), Kelsey Scherler (1), Kai Wang (1), Donald Geman (6), Jeffrey B. Carroll (5), Marcy E. MacDonald (3), George Carlson (2), Vanessa C. Wheeler (3), Nathan D. Price (1), and Leroy E. Hood (1). Hum Mol Genet (2017) 26 (5): 913-922.

[*] *co-first authors*

ACKNOWLEDGEMENTS

We thank Nathan Goodman for his role in initiating this project and for performing microarray data pre-processing. This work was supported by the University of Luxembourg-Institute for Systems Biology Strategic Partnership, a contract from the CHDI Foundation, a National Science Foundation Graduate Student Research Fellowship to JRP, and NIH grant NS049206 to VCW.

4 Gene regulatory drivers in Huntington's disease

4.1 Abstract

Transcriptional changes occur presymptomatically and throughout Huntington's Disease (HD), motivating the study of transcriptional regulatory networks (TRNs) in HD. We reconstructed a genome-scale model for the target genes of 718 TFs in the mouse striatum by integrating a model of the genomic binding sites with transcriptome profiling of striatal tissue from HD mouse models. We identified 48 differentially expressed TF-target gene modules associated with age- and *Htt* allele-dependent gene expression changes in the mouse striatum, and replicated many of these associations in independent transcriptomic and proteomic datasets. Strikingly, many of these predicted target genes were also differentially expressed in striatal tissue from human disease. We experimentally validated a key model prediction that SMAD3 regulates HD-related gene expression changes using chromatin immunoprecipitation and deep sequencing (ChIP-seq) of mouse striatum. We found *Htt* allele-dependent changes in the genomic occupancy of SMAD3 and confirmed our model's prediction that many SMAD3 target genes are down-regulated early in HD. Importantly, our study provides a mouse and human striatal-specific TRN and predicts transcription factor drivers of striatal gene expression changes in HD.

4.2 Introduction

Massive changes in gene expression accompany many human diseases, yet we still know relatively little about how specific transcription factors (TFs) mediate these changes. Comprehensive characterization of disease-related transcriptional regulatory networks (TRNs) can clarify potential disease mechanisms and prioritize targets for novel therapeutics. A variety of approaches have been developed to reconstruct interactions between TFs and their target genes, including models focused on reconstructing the physical locations of transcription factor binding (Neph *et al*, 2012; Gerstein *et al*, 2012), as well as computational algorithms utilizing gene co-expression to infer regulatory relationships (Marbach *et al*, 2012; Margolin *et al*, 2006; Bonneau *et al*, 2006; Friedman *et al*, 2000; Huynh-Thu *et al*, 2010; Reiss *et al*, 2015). These approaches have yielded insights into the regulation of a range of biological systems, yet accurate, genome-scale models of mammalian TRNs remain elusive.

Several lines of evidence point to a specific role for transcriptional regulatory changes in Huntington's disease (HD). HD is a fatal neurodegenerative disease caused by dominant inheritance of a polyglutamine (polyQ)-coding expanded trinucleotide (CAG) repeat in the *HTT* gene (MacDonald *et al*, 1993). Widespread transcriptional changes have been detected in post-mortem brain tissue from HD cases vs. controls (Hodges *et al*, 2006), and transcriptional changes are among the earliest detectable phenotypes in HD mouse models (Luthi-Carter *et al*, 2000; Seredenina & Luthi-Carter, 2012). These transcriptional changes are particularly prominent in the striatum, the most profoundly impacted brain region in HD (Tabrizi *et al*, 2013; Vonsattel *et al*, 1985) . Replicable gene expression changes in the striatum of HD patients and HD mouse models include down-regulation of genes related to synaptic function in medium spiny neurons accompanied by up-regulation of genes related to neuroinflammation (Seredenina & Luthi-Carter, 2012).

Some of these transcriptional changes may be directly related to the functions of the HTT protein. Both wildtype and mutant HTT (mHTT) protein have been shown to associate with genomic DNA, and mHTT also interacts with histone modifying enzymes and is associated with changes in chromatin states (Thomas *et al*; Benn *et al*; Seong *et al*, 2010). Wildtype HTT protein has been shown to regulate the activity of some TFs (Zuccato *et al*, 2007). Also, high concentrations of nuclear mHTT aggregates sequester TF and co-factor proteins and interfere with genomic target finding, though it is unknown if this occurs at physiological concentrations of mHTT (Wheeler *et al*, 2000; Shirasaki *et al*, 2012; Li *et al*, 2016). Roles for several TFs in HD have been characterized (Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes, 2003; Dickey *et al*, 2015; Arlotta *et al*, 2008; Tang *et al*, 2012), but we lack a global model for the relationships between HD-related changes in the activity of specific TFs and the downstream pathological processes that they regulate.

The availability of large transcriptomics datasets related to HD is now making it possible to begin comprehensive network analysis of the disease, particularly in mouse models. Langfelder et al. (Langfelder *et al*, 2016) generated RNA-seq from the striatum of 144 knock-in mice heterozygous for HD mutations and 64 wildtype littermate controls, and they used gene co-expression networks to identify modules of co-expressed genes with altered expression in HD. However, their analyses did not attempt to identify any of the TFs responsible for these gene expression changes.

Here, we investigated the roles of core TFs that are predicted to drive the gene expression changes in Huntington's disease, using a comprehensive network biology approach. We used a machine learning strategy to reconstruct a genome-scale model for TF-target gene interactions in the mouse striatum, combining publicly available DNase-seq with

brain transcriptomics data HD mouse models. We identified 48 core TFs whose predicted target genes were overrepresented among differentially expressed genes in at least five of fifteen conditions defined by a mouse's age and *Htt* allele, and we replicated the predicted core TFs and differential gene expression associations in multiple datasets from HD mouse models and from HD cases and controls. Based on the gene expression signature of SMAD3 and its predicted target genes, we hypothesized that SMAD3 is a core regulator of early gene expression changes in HD. Using chromatin immunoprecipitation and deep sequencing (ChIP-seq), we demonstrate *Htt*-allele-dependent changes in SMAD3 occupancy and down-regulation of SMAD3 target genes in mouse brain tissue. In conclusion, the results from our TRN analysis and ChIP-seq studies of HD reveal new insights into predicted transcription factor drivers of complex gene expression changes in this neurodegenerative disease.

4.3 Results

A genome-scale transcriptional regulatory network model of the mouse striatum.

We reconstructed a model of TF-target gene interactions in the mouse striatum by integrating information about transcription factor binding sites (TFBSs) with evidence from gene co-expression in the mouse striatum (Fig. 1a).

We predicted the binding sites for 871 TFs in the mouse genome using digital genomic footprinting. We identified footprints in DNase-seq data from 23 mouse tissues (Yue *et al*, 2014), using Wellington (Piper *et al*, 2013). Footprints are defined as short genomic regions with reduced accessibility to the DNase-I enzyme in at least one tissue. Our goal in combining DNase-seq data from multiple tissues was to reconstruct a single TFBS model that could make useful predictions about TF target genes, even in conditions for which DNase-seq data were not available. We identified 3,242,454 DNase-I footprints. Genomic footprints are often

indicative of occupancy by a DNA-binding protein. We scanned these footprints for 2,547 sequence motifs from TRANSFAC (Matys *et al*, 2006), JASPAR (Mathelier *et al*, 2014), UniProbe (Hume *et al*, 2015), and high-throughput SELEX (Jolma *et al*, 2013) to predict binding sites for specific TFs (TFBSs), and we compared these TFBSs to the locations of transcription start sites. We considered a TF to be a potential regulator of a gene if it had at least one binding site within 5kb of that gene's TSS. We showed previously that a 5kb region upstream and downstream of the TSS maximizes target gene prediction from digital genomic footprinting of the human genome (Plaisier *et al*, 2016).

To assess the accuracy of this TFBS model, we compared our TFBS predictions to ChIP-seq experiments from ENCODE (Yue *et al*, 2014) and ChEA (Lachmann *et al*, 2010) (SI Fig. 2). For 50 of 52 TFs, there was significant overlap between the sets of target genes predicted by our TFBS model vs. ChIP-seq (FDR < 1%). Our TFBS model had a median 78% recall of target genes identified by ChIP-seq, and a median 22% precision. That is, our model identified the majority of true-positive target genes but also made a large number of false-positive predictions. Low precision is expected in this model, since TFs typically occupy only a subset of their binding sites in a given tissue. Nonetheless, low precision indicates a need for additional filtering steps to identify target genes that are relevant in a specific context.

We sought to identify TF-target gene interactions that are active in the mouse striatum, by evaluating gene co-expression patterns in RNA-seq transcriptome profiles from the striatum of 208 mice (Langfelder *et al*, 2016). The general idea is that active regulation of a target gene by a TF is likely to be associated with strong TF-gene co-expression, and TFBSs allow us to identify direct regulatory interactions. This step also removes TFs with low expression: of the 871 TFs with TFBS predictions we retained as potential regulators the 718 TFs that were expressed in the striatum. We fit a regression model to predict the expression of

each gene based on the combined expression patterns of TFs with one or more TFBSs ±5kb of that gene's transcription start site. We used LASSO regularization to select the subset of TFs whose expression patterns together predicted the expression of the target gene. This approach extends several previous regression methods for TRN reconstruction (Friedman *et al*, 2010; Tibshirani, 1996; Bonneau *et al*, 2006; Haury *et al*, 2012) by introducing TFBS-based constraints. In preliminary work, we considered a range of LASSO and elastic net (alpha = 0.2, 0.4, 0.6, 0.8, 1.0) regularization penalties and evaluated performance in five-fold cross-validation (see Methods). We selected LASSO based on the highest correlation between prediction accuracy in training vs. test sets.

We validated the predictive accuracy of our TRN model by comparing predicted vs. observed expression levels of each gene. Our model explained >50% of expression variation for 13,009 genes in training data (Fig. 1b). Prediction accuracy in five-fold cross-validation was nearly identical to prediction accuracy in training data. That is, genes whose expression was accurately predicted in the training data were also accurately predicted in the test sets (r=0.94; Fig. 1b). Genes whose expression was not accurately predicted generally had low expression in the striatum (Supplementary Information [SI] Fig. 1). We removed poorly predicted genes, based on their training set accuracy before moving to the test set. The final TRN model contains 13,009 target genes regulated by 718 TFs via 176,518 interactions (SI Dataset 1). Our model predicts a median of 14 TFs regulating each target gene and a median of 147 target genes per TF (Fig. 1c,d). 15 TFs were predicted to regulate >1,000 target genes (SI Figure 3). Importantly, TF-target gene interactions retained in our striatum-specific TRN model were enriched for genomic footprints in the adult (p = 1.4e-82) and fetal (p = 2.1e-88) brain, supporting the idea that these TF-target gene interactions reflect TF binding sites in the brain.

We defined as "TF-target gene modules" the sets of genes predicted to be direct targets of each of the 718 TFs. 135 of these 718 TF-target gene modules were enriched for a functional category from Gene Ontology (Ashburner *et al*, 2000) (FDR < 5%, adjusting for 4,624 GO terms). 337 of the 718 TF modules were enriched ($p < 0.01$) for genes expressed specifically in a major neuronal or non-neuronal striatal cell type (Doyle *et al*, 2008; Zhang *et al*, 2014; Dougherty *et al*, 2010), including known cell type-specific activities for both neuronal (e.g., *Npas1-3*) and glia-specific TFs (e.g., *Olig1, Olig2*) (SI Fig. 4). These results suggest that many TRN modules reflect the activities of TFs on biological processes within specific cell types.

Prediction of core TFs associated with transcriptional changes in HD mouse models.

We next sought to identify TFs that are core regulators of transcriptional changes in HD. Of the 208 mice in the RNA-seq dataset used for network reconstruction, 144 were heterozygous for a human *HTT* allele knocked into the endogenous *Htt* locus (Wheeler *et al*, 1999), and the remaining 64 mice were C57BL/6J littermate controls. Six distinct *HTT* alleles differing in the length of the poly-Q repeat were knocked in. In humans, the shortest of these alleles -- Htt^{Q20} -- is non-pathogenic, and the remaining alleles -- Htt^{Q80}, Htt^{Q92}, Htt^{Q111}, Htt^{Q140}, and Htt^{Q175} -- are associated with progressively earlier onset of symptoms. We used RNA-seq data from four male and four female mice of each genotype at each of three time points: 2-months-old, 6-months-old, and 10-months-old. These mouse models undergo subtle age- and allele-dependent changes in behavior, and all of the ages profiled precede detectable neuronal cell death (Alexandrov *et al*, 2016; Rothe *et al*, 2015; Carty *et al*, 2015).

We evaluated gene expression differences between $Htt^{Q20/+}$ mice and mice with each of the five pathogenic *HTT* alleles at each time point, a total of 15 comparisons. The extent of gene expression changes increased in an age- and Q-length-dependent fashion, with extensive

overlap between the DEGs identified in each condition (Fig. 2). 8,985 genes showed some evidence of differential expression (DEGs; $p < 0.01$) in at least one of the 15 conditions, of which 5,132 were significant at a stringent False Discovery Rate < 1%. These results suggest that robust and replicable gene expression changes occur in the striatum of these HD mouse models at ages well before the onset of neuronal cell death or other overt pathology.

The predicted target genes of 209 TFs were overrepresented for DEGs in at least one of the 15 conditions (3 ages x 5 mouse models; Fisher's exact test, $p < 1e\text{-}6$; SI Dataset 2). Repeating this analysis in 1,000 permuted data sets indicated that enrichments at this level of significance never occurred in more than four conditions. We therefore focused on a core set of 48 TFs whose predicted target genes were overrepresented for DEGs in five or more conditions. Notably, 44 of these 48 TFs were differentially expressed (FDR < 0.01) in at least one of the 15 conditions (SI Fig. 4). We refer to these 48 TFs as core TFs.

Replication of core TFs in independent datasets.

We sought to replicate these associations by testing for enrichment of TF-target gene modules for differentially expressed genes in independent HD-related datasets. First, we conducted a meta-analysis of differentially expressed TF-target gene modules in four independent microarray gene expression profiling studies of striatal tissue from HD mouse models (Becanovic *et al*, 2010; Giles *et al*, 2012; Fossale *et al*, 2011; Kuhn *et al*, 2007). Targets of 46 of the 48 core TFs were enriched for DEGs (meta-analysis p-value < 0.01) in the microarray data. The overlap between TFs whose target genes were differentially expressed in HD vs. control mice in microarray datasets and the core TFs from our primary dataset was significantly greater than expected by chance (Fisher's exact test: $p = 5.7e\text{-}32$). These results suggest that transcriptional changes in most of the core TF-target gene modules were preserved across multiple datasets and mouse models of HD.

Next, we asked whether the target genes of core TFs were also differentially abundant at the protein level. We studied quantitative proteomics data from the striatum of 64 6-month-old HD knock-in mice (Langfelder *et al*, 2016). These were a subset of the mice profiled with RNA-seq in our primary dataset. Targets of 22 of the 48 core TFs were enriched for differentially abundant proteins (Fisher's exact test, $p < 0.01$). The overlap between TFs whose target genes were differentially abundant between HD vs. wildtype mice and the core regulator TFs was significantly greater than expected by chance (Fisher's exact test: $p = 5.7e\text{-}20$).

Third, we asked whether target genes of these same TFs are differentially expressed in late-stage human disease. We reconstructed a TRN model for the human striatum integrating a map of TFBSs (Plaisier *et al*, 2016) based on digital genomic footprinting of 41 human cell types(Neph *et al*, 2012) with microarray gene expression profiles of post-mortem striatal tissue from 36 HD cases and 30 controls (Hodges *et al*, 2006). As in our TRN model for the mouse striatum, we fit a LASSO regression model to predict the expression of each gene in human striatum from the expression levels of TFs with predicted TFBSs within 5kb of its transcription start sites (SI Fig. 6). We studied the enrichments of TF-target gene modules from this human striatum TRN model for differentially expressed genes.

We compared HD-related TF-target gene modules identified in mouse and human striatum, focusing on 616 TFs with one-to-one orthology and ≥10 predicted target genes in both the mouse and human striatum TRN models. We conducted a meta-analysis of two independent datasets from the dorsal striatum of HD cases vs. controls (Hodges *et al*, 2006; Durrenberger *et al*, 2015) to identify TF-target gene modules enriched for DEGs. Targets of 13 of the 48 core TFs from mouse striatum were over-represented among differentially expressed genes in HD cases vs. controls. This overlap was not statistically greater than expected by

chance (odds ratio = 1.79; p = 0.05). However, when we considered the broader set of 209 TF-target gene modules that were enriched for differentially expressed genes in any of the 15 conditions from the primary RNA-seq dataset, we found significant overlap for TF-target gene modules that were down-regulated both in HD and in HD mouse models (28 shared TF-target gene modules; odds ratio = 3.6, p = 5.0e-5; SI Fig. 6d) and for TF-target gene modules that were up-regulated both in HD and in HD mouse models (26 shared TF-target gene modules; odds ratio = 1.8, p = 0.02; SI Fig. 6e). The striatum is heavily degraded in late-stage HD, with many dead neurons and extensive astrogliosis. Nonetheless, these results suggest that some transcriptional programs are shared between the earliest stages of molecular progression (assayed in mouse models) and late stages of human disease.

Notably, targets of 13 of the 48 core regulator TFs were enriched for differentially expressed genes in all four datasets: *Gli3*, *Irf2*, *Klf16*, *Npas2*, *Pax6*, *Rarb*, *Rfx2*, *Rxrg*, *Smad3*, *Tcf12*, *Tef*, *Ubp1*, and *Vezf1*. These 13 TFs may be especially interesting for follow-up studies.

Biological associations of core TFs.

We evaluated relationships among the 48 core TFs based on clustering and network topology. Plotting TF-to-TF regulatory interactions among the 48 core TFs (Fig. 4) revealed two distinct TF-to-TF sub-networks, characterized by numerous positive interactions within sub-networks and by fewer, mostly inhibitory interactions between sub-networks. The target genes of TFs in the first sub-network were predominantly down-regulated in HD, while the target genes of TFs in the second module were predominantly up-regulated. Hierarchical clustering of the 48 core TFs based on the expression patterns of their predicted target genes revealed similar groupings of TFs whose target genes were predominantly down- vs. up-regulated (Fig. 5).

We studied the predicted target genes of each core TF to characterize possible roles for these TFs in HD. Down-regulated TF-target gene modules were overrepresented for genes specifically expressed in Drd1+ and Drd2+ medium spiny neurons (Fig. 5). Functional enrichments within these modules were mostly related to synaptic function, including metal ion transmembrane transporters (targets of *Npas2*, p = 2.3e-4), voltage-gated ion channels (targets of *Mafa*, p = 8.1e-4), and protein localization to cell surface (targets of *Rxrg*, p = 1.7e-4). These network changes may be linked to synapse loss in medium spiny neurons, which is known to occur in knock-in mouse models of HD (Deng *et al*, 2013).

Some up-regulated TF-target gene modules were overrepresented for genes specifically expressed in oligodendrocytes or astrocytes, while others were overrepresented for genes specifically expressed in neurons (Fig. 5). Functional enrichments within these modules included Gene Ontology terms related to apoptosis ("positive regulation of extrinsic apoptotic signaling pathway via death domain receptors", targets of *Wt1*, p = 1.8e-4) and DNA repair (targets of *Runx2*, "single-strand selective uracil DNA N-glycosylase activity", p = 2.0e-4). Therefore, core TFs whose target genes were predominantly up-regulated may contribute to a variety of pathological processes both in neurons and in glia. The number of oligodendrocytes is basally increased in HD mutation carriers, while activated gliosis is thought to begin later in disease progression (Vonsattel *et al*, 1985).

An open question in the field is whether the same sequence of pathogenic events underlies disease progression in juvenile-onset HD due to *HTT* alleles with very long poly-Q tracts vs. adult-onset HD due to *HTT* alleles with relatively short poly-Q tracts. This question is of practical relevance for modeling HD in mice, since mouse models with very long *HTT* alleles are often used in research due to their faster rates of phenotypic progression within a two-year lifespan. To address this question, we evaluated overlap between TF-target gene

modules activated at the earliest time points in mice with each of the five pathogenic *Htt* alleles in our dataset. In the mice with the longest *HTT* alleles -- Htt^{Q175} and Htt^{Q140} -- the target genes of core TFs first became enriched for differentially expressed genes in two-month-old mice. In mice with relatively short *HTT* alleles – Htt^{Q111}, Htt^{Q92} and Htt^{Q80} -- target genes of core TFs became enriched for differentially expressed genes beginning in six-month-old mice. We found that eight modules – the predicted target genes of IRF2, MAFA, KLF16, LMO2, NPAS2, RUNX2, RXRG, and VEZF1 – were significantly enriched for DEGs in at least three of these five conditions (two-month-old $Htt^{Q175/+}$, two-month-old $Htt^{Q140/+}$, six-month-old $Htt^{Q111/}$, six-month-old $Htt^{Q92/+}$, and six-month-old $Htt^{Q80/+}$). A limitation of this analysis is that all of the alleles used in this study are associated with juvenile onset disease, and the extent to which these results extend to adult-onset alleles remains to be detemined. Nonetheless, these results suggest that many aspects of the trajectory of transcriptional changes are shared across the *HTT* Q-lengths that have been studied. Notably, all of the TFs whose target genes were enriched for differentially expressed genes at the very earliest timepoints were enriched primarily for genes that were down-regulated in HD. Strong enrichments of TF-target gene modules for up-regulated genes occurred only at slightly later time points.

Genome-wide characterization of SMAD3 binding sites in the mouse striatum supports a role in early gene dysregulation in HD.

SMAD3 was one of 13 core TFs whose predicted target genes were overrepresented among differentially expressed genes across all four independent datasets. Progressive down-regulation of *Smad3* mRNA (Fig. 6a) and of predicted SMAD3 target genes (Fig. 5) occurred in an age- and *Htt*-allele-dependent fashion, beginning at or before six postnatal months.

We characterized the binding sites of SMAD3 in the striatum of four-month-old $Htt^{Q111/+}$ mice and wild-type littermate controls by chromatin immunoprecipitation and deep sequencing (ChIP-seq, n=2 pooled samples per group, with each pool containing DNA from three mice). Peak-calling revealed 57,772 SMAD3 peaks (MACS2.1, FDR < 0.01 and >10 reads in at least two of the four samples; Dataset 3). 34,633 of the 57,772 SMAD3 peaks (59.9%) were located within 10kb of transcription start sites (TSSs), including at least one peak within 10kb of the TSSs for 11,727 genes (Fig. 6b). The summits of SMAD3 peaks were enriched for the SMAD2:SMAD3:SMAD4 motif (p-value = 7.2e-85; Fig. 6c). Importantly, the TSSs for 753 of the 938 computationally predicted SMAD3 target genes in our TRN model were located within 10kb of at least one ChIP-based SMAD3 binding site. This overlap was significantly greater than expected by chance (odds ratio = 4.33, p-value = 2.8e-84).

We characterized the relationship between SMAD3 occupancy and transcriptional activation by measuring the genomic occupancy of RNA Polymerase II (RNAPII) in the striatum of $Htt^{Q111/+}$ and wildtype mice. RNAPII occupancy is a marker of active transcription and of active transcription start sites. Occupancy of SMAD3 and of RNAPII were positively correlated, across all genomic regions (r = 0.70) and specifically within SMAD3 peaks (r = 0.71).

Similarly, we characterized the relationship between SMAD3 occupancy and chromatin accessibility, using publicly available DNase-seq of midbrain tissue from wildtype mice. 22,650 of the 57,772 SMAD3 peaks (39.2%) overlapped a DNase hypersensitive site in the midbrain. Occupancy of SMAD3 was positively correlated with DNase-I hypersensitivity across all genomic regions (r = 0.33) and specifically within SMAD3 peaks (r = 0.25).

We ranked genes from highest to lowest SMAD3 regulatory potential based on the number of SMAD3 peaks within 10kb of their transcriptional start sites. We focused on the top 837 genes with SMAD3 peak counts > 2 standard deviations above the mean. These top 837

SMAD3 target genes were enriched (FDR < 0.01) for 24 non-overlapping clusters of Gene Ontology terms (SI Table 1). These enriched GO terms prominently featured pathways related to gene regulation ("mRNA processing", p = 4.2e-9; "histone modification", p = 1.7e-7; "transcriptional repressor complex", p = 3.7e-5), as well as functions more specifically related to brain function ("neuromuscular process controlling balance", p = 1.2e-7; "brain development", p = 1.27e-6; "neuronal cell body", p = 2.5e-5).

We performed quantitative and qualitative analyses to compare SMAD3 occupancy in $Htt^{Q111/+}$ vs. wildtype mice. 51,721 of the 57,772 SMAD3 peaks (89.5%) were identified in both $Htt^{Q111/+}$ and wildtype mice. 5,419 peaks (9.4%) were identified only in wildtype mice, while only 632 peaks (1.1%) were identified only in $Htt^{Q111/+}$ mice (Fig. 6d). Quantitative analyses of differential binding with edgeR revealed four peaks whose occupancy was significantly different (FDR < 0.05) between $Htt^{Q111/+}$ and wildtype mice. All four of these peaks were more weakly occupied in $Htt^{Q111/+}$ mice. 138 peaks had nominally significant differences in occupancy between genotypes (p < 0.01). 133 of these 138 peaks (96.4%) were more weakly occupied in $Htt^{Q111/+}$ mice (Fig. 6e). These results suggest that SMAD3 occupancy is decreased at a subset of its binding sites in four-month-old $Htt^{Q111/+}$ mice.

Finally, we tested whether the top 837 SMAD3 target genes from ChIP-seq were differentially expressed in HD knock-in mice. The top 837 SMAD3 target genes from ChIP-seq were significantly overrepresented among genes that became down-regulated in the striatum of HD knock-in mice (223 down-regulated SMAD3 target genes; odds ratio = 2.0, p-value = 3.4e-15; Fig. 6f). By contrast, SMAD3 target genes were not overrepresented among genes that became up-regulated in the striatum of HD mouse models (143 up-regulated SMAD3 target genes, odds ratio = 0.92, p = 0.40). These results are consistent with our computational model,

in which SMAD3 target genes were primarily down-regulated in HD knock-in mice. Therefore, SMAD3 binding is associated with down regulation in HD mouse models.

4.4 Discussion

Here, we identified putative core TFs regulating gene expression changes in Huntington's disease by reconstructing genome-scale transcriptional regulatory network models for the mouse and human striatum. Identifying core TFs in HD provides insights into the mechanisms of this devastating, incurable disease. This method to reconstruct models of mammalian transcriptional regulatory networks can be readily applied to find regulators underlying any trait of interest.

Our model extends prior knowledge about the TFs involved in HD. A role in HD for *Rarb* is supported by ChIP-seq and transcriptome profiling of striatal tissue from $Rarb^{-/-}$ mice (Niewiadomska-Cimicka *et al*, 2016). A role in HD for *Foxo1* is supported by experimental evidence that FOXO signaling influences the vulnerability of striatal neurons to mutant Htt (Parker *et al*, 2012). A role in HD for *Relb* is supported by experimental evidence that NF-kB signaling mediates aberrant neuroinflammatory responses in HD and HD mouse models (Hsiao *et al*, 2013). Notably, microglia counts in 10-12 month $Htt^{Q111/+}$ mice indicate that these cells are not proliferating, suggesting that the transcriptional changes observed in our study represent a proinflammatory state, rather than microgliosis *per se*. Other predicted core TFs, including *Klf16* and *Rxrg*, have previously been noted among the most consistently differentially expressed genes in mouse models of HD (Seredenina & Luthi-Carter, 2012). In some cases, known functions for core TFs suggest hypotheses about their roles in HD. For instance, *Npas2* is a component of the molecular clock, so its dysfunction could contribute to

circadian disturbances in HD (Morton *et al*, 2005). Notably, the predicted target genes for several TFs whose functions in HD have been studied by other investigators -- e.g., *Rest* (Zuccato *et al*, 2003), *Srebf2* (Valenza *et al*, 2005), and *Foxp1* (Tang *et al*, 2012) – were overrepresented for differentially expressed genes in our model, but only at later time points or more weakly than our top 48 core regulator TFs.

Our results suggest that HD involves parallel changes in distinct down- vs. up-regulated TF sub-networks. Targets of TFs in the down-regulated sub-network are enriched for synaptic genes and appear to be primarily neuronal. Targets of TFs in the up-regulated sub-network are enriched for stress response pathways (e.g., DNA damage repair, apoptosis). These up-regulated networks appear to involve processes occurring in both neurons and glia. Several previous studies provide independent support for synaptic changes in medium spiny neurons and of activated gliosis in HD pathogenesis (Deng *et al*, 2013; Singhrao *et al*, 1999; Hsiao *et al*, 2013).

Replication across four independent datasets revealed 13 TFs whose target genes were most consistently enriched among differentially expressed genes. We propose that these TFs should be prioritized for follow-up experiments, both to validate predicted target genes and to evaluate specific biological functions for each TF. For instance, it will be interesting to determine which (if any) of the core TFs have direct protein-protein interactions with the HTT protein and to test our model's predictions about TF perturbations with specific aspects of HD pathology. The target genes for most of these 13 TFs were enriched for genes that were down-regulated in HD and for neuron-specific genes, consistent with the idea that pathological changes originate in medium spiny neurons.

Our ChIP-seq data confirm an association between SMAD3 binding sites and genes that are down-regulated in HD. SMAD3 is best known for its role in mediating Transforming

Growth Factor-Beta (TGF-β) signaling (Kandasamy *et al*, 2011). Several studies have described dysregulation of the TGF-β signaling pathwway in the early stages of HD (Ring *et al*, 2015; Battaglia *et al*, 2011). A recent study went on to characterize SMAD transcription factors within the TGF- β pathway in mouse and human cell models of HD (Bowles *et al,* 2017), and found a SMAD binding site present in the *HTT* promoter. SMAD3 in particular was found to regulate *Htt* expression in mouse striatal lines. These findings suggest an intriguing possibility that agonists of TGF-β signaling could have therapeutic benefit in HD patients. Consistent with this possibility, TGF-β treatment has recently been shown to reduce apoptotic cell death in neural stem cells with expanded *HTT* polyQ tracts (Ring *et al*, 2015).

Our method to reconstruct TRNs by integrating information about TF occupancy with gene co-expression is likely to be broadly applicable, providing a strategy to optimize both mechanistic and quantitative accuracy. TRN reconstruction methods based purely on gene co-expression struggle to distinguish direct vs. indirect interactions. Physical models of TF occupancy provide poor quantitative predictions because many TF binding sites are non-functional or do not regulate the nearest gene. Our study demonstrates that integrated TRN modeling can be utilized effectively to study neurodegenerative diseases such as HD, combining data from the ENCODE project with disease specific transcriptome profiling.

4.5 Methods

Referenced datasets. We obtained RNA-seq and microarray gene expression profiling data from the following GEO Datasets (http://www.ncbi.nlm.nih.gov/geo/): GSE65776 (Langfelder *et al*, 2016), GSE18551 (Becanovic *et al*, 2010), GSE32417 (Giles *et al*, 2012), GSE9038 (Fossale *et al*, 2011), GSE9857 (Kuhn *et al*, 2007), GSE26927 (Durrenberger *et al*, 2015), GSE3790 (Hodges *et al*,

2006). We obtained proteomics data from the PRIDE archive (https://www.ebi.ac.uk/pride/archive/), accession PXD003442 (Langfelder *et al*, 2016). For RNA-seq data (GSE65776), we downloaded read counts and FPKM estimates, mapped to ENSEMBL gene models. For Affymetrix microarrays (GSE18551, GSE32417, GSE9038, GSE9857, GSE26927, and GSE3790) we downloaded raw image files and used the affy package in R to perform within-sample RMA normalization and between-sample quantile normalization. For proteomics data, we downloaded MaxQuant protein quantities.

Genomic footprinting. DNase-I digestion of genomic DNA followed by deep sequencing (DNase-seq) enables the identification of genomic footprints across the complete genome. We predicted genome-wide transcription factor binding sites (TFBSs) in the mouse and human genomes based on instances of TF sequence motifs in digital genomic footprints from the ENCODE project. Short regions of genomic DNA occupied by DNA-binding proteins produce characterizes characteristic "footprints" with altered sensitivity to the DNase-I enzyme. DNase-I digestion of genomic DNA followed by deep sequencing (DNase-seq) enables the identification of genomic footprints across the complete genome.

For the human TFBS model, we used a previously described database (Plaisier *et al*, 2016)_of footprints from DNase-seq of 41 cell types (Neph *et al*, 2012). For the mouse TFBS model, we downloaded digital genomic footprinting data (deep DNase-seq) for 23 mouse tissues and cell types (Yue *et al*, 2014) from the UCSC ENCODE portal on October 29, 2013: ftp://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/. We detected footprints in each sample with Wellington (Piper *et al*, 2013), using a significance threshold, $p < 1e\text{-}10$. Using FIMO (Grant *et al*, 2011), we scanned the mouse genome (mm9) for instances of 2,547 motifs from TRANSFAC_(Matys *et al*, 2006), JASPAR (Mathelier *et al*, 2014), UniPROBE (Hume *et al*,

2015), and high-throughput SELEX (Jolma *et al*, 2013). We intersected footprints from all tissues with motif instances to generate a genome-wide map of predicted TFBSs. A motif can be recognized by multiple TFs with similar DNA-binding domains. We assigned motifs to TF families using annotations from the TFClass database (Wingender *et al*, 2013). In total, our model included motifs recognized by 871 TFs.

Regression-based transcriptional regulatory network models. We fit a regression model to predict the expression of each gene in mouse or human striatum, based on the expression patterns of TFs that had predicted binding sites within 5kb of that gene's transcription start sites. We applied LASSO regularization to penalize regression coefficients and remove TFs with weak effects, using the glmnet package in R. These methods were optimized across several large transcriptomics datasets, prior to their application to the Huntsington's disease data. To reconstruct the TRN model for mouse striatum, we used RNA-seq data from the striatum of 208 mice (Langfelder *et al*, 2016). Prior to network reconstruction, we evaluated within and between group variance and detected outlier samples using hierarchical clustering and multidimensional scaling. No major differences in variance were identified between groups, and no outlier samples were detected or removed.

We considered a variety of model parameterization during the initial model formulation. We considered elastic net regression and ridge regression as alternatives to LASSO regression. We selected LASSO based on the least falloff in performance from the training data to test sets in five-fold cross-validation. We note that when multiple TFs have correlated expression, the LASSO will generally retain only one for the final model. This feature of the LASSO has been considered advantageous, since it can eliminate indirect interactions. However, there is virtually no doubt that the TFs selected by our model

underestimate the true number of TF-target gene interactions. We would only pick up dominant effects where a linear model works reasonably well. Our primary interest is ultimately in using this approach to find a relatively small number of targets based on multiple lines of evidence. We are less concerned with finding everything than in trying to make sure what we do find is as highly enriched for true positives as possible.

We also considered a variety of strategies to select an appropriate penalty parameter. For instance, we could apply an independent penalty parameter for each gene, or we could use a uniform penalty parameter across all genes. We found that optimal performance was obtained in both training data and in five-fold cross-validation when we applied a uniform penalty parameter across all genes. We assigned this penalty parameter by evaluating performance in cross-validation across a range of possible parameters for a random subset of 100 genes. For each gene, we identified the most stringent penalty such that the unfitted variance was < 1 standard error greater than the minimum unfitted variance across all the penalty parameters considered. We selected the median penalty defined by this procedure across the 100 randomly selected gene.

Not all genes' expression can be accurately predicted based on the expression of TFs. To select genes for the final model, we evaluated the variance explained by the model in a training set consisting of 80% of the data. We selected those genes for which the model explained >50% of expression variance in the training set and carried these genes forward to a test set, consisting of the remaining 20% of genes. We found that training set performance accurately predicted test performance ($r = 0.94$). We therefore fit a final model for genes whose expression could be accurately predicted in the training set. The result of these procedures is a tissue-specific TRN model, predicting the TFs that regulate each gene in the

striatum and assigning a positive or negative weight for each TF's effect on that gene's expression in the striatum.

Enrichments of TF-target gene modules in ChIP-seq data. We downloaded ChIP-seq data from the ENCODE website (encodeproject.org, accessed August 20, 2015) for 33 mouse transcription factors included in our TRN model. We identified genes whose transcription start sites were located within 5kb of a narrowPeak in each ChIP experiment. We also downloaded a table of ChIP-to-gene annotations for 19 additional mouse TFs from the ChEA website (http://amp.pharm.mssm.edu/lib/chea.jsp, accessed August 6, 2015). We tested for enrichments of the target genes identified by ChIP for each of these 52 TFs to predicted TFBSs from our model.

Enrichments of TF-target gene modules for Gene Ontology terms. We downloaded Gene Ontology (GO) annotations for mouse genes from GO.db on November 4, 2015, using the topGO R package. We extracted the genes annotated to each GO term and its children, and we used Fisher's exact tests to characterize enrichments of TF-target gene modules for the 4,624 GO terms that contain between 10 and 500 genes.

Enrichments of TF-target gene modules for cell type-specific genes. We characterized sets of genes expressed in each striatal cell type using gene expression profiles from purified cell types (Doyle *et al*, 2008; Zhang *et al*, 2014) and the pSI R package for Cell-type Specific Expression Analysis (Dougherty *et al*, 2010). We used Fisher's exact tests to characterize enrichments of TF-target gene modules for genes expressed specifically in each cell type.

Enrichments of TF-target gene modules for differentially expressed genes. We identified genes that were differentially expressed in HD vs. control samples. In the primary dataset, we compared mice with the non-pathogenic Q20 allele and mice with each of the other five alleles, separately for 2-, 6-, and 10-month-old mice. We used the edgeR R package to fit generalized linear models and test for significance of each contrast. We used Fisher's exact tests to characterize enrichments of down-regulated genes and up-regulated genes in each condition (significance threshold for differentially expressed genes, $p < 0.01$) for the target genes of each TF. We considered enrichments to be statistically significant at a raw p-value threshold < 1e-6, or an adjusted p-value < 0.02 after accounting for 19,170 tests (639 TFs x 5 *Htt* alleles x 3 time points x 2 tests / condition).

To identify top TFs, accounting for non-independence among genes and conditions, we calculated an empirical false discovery rate for these enrichments. We repeated the edgeR and enrichment analyses 1,000 times with permuted sample labels. We found that no module had a p-value < 1e-6 in more than four conditions in any of the permuted datasets. Therefore, we focused on TFs whose target genes were overrepresented for differentially expressed genes in five or more conditions.

We performed similar analyses to characterize TF-target gene modules enriched for genes that were differentially expressed in replication samples. We used the limma R package to calculate differentially expressed genes in each of the four microarray studies from mouse striatum (Giles *et al*, 2012; Kuhn *et al*, 2007; Fossale *et al*, 2011; Becanovic *et al*, 2010). We calculated enrichments of the DEGs from each study for TF-target gene modules. We then combined the enrichment p-values across the four studies using Fisher's method to produce a meta-analysis p-value for the association of each TF-target gene module in HD mouse models.

We used quantitative proteomics data from 6-month old $Htt^{Q20/+}$, $Htt^{Q80/+}$, $Htt^{Q92/+}$, $Htt^{Q111/+}$, $Htt^{Q140/+}$ and $Htt^{Q175/+}$ mice (n = 8 per group) (Langfelder *et al*, 2016). We characterized proteins whose abundance was correlated with *Htt* CAG length in the striatum of 6-month-old mice, using MaxQuant protein quantities. We then calculated enrichments of CAG-length correlated proteins (Pearson correlation, $p < 0.01$) for each TF-target gene module with Fisher's exact test, separately for proteins whose abundance was positively or negatively correlated with CAG length.

We used the limma R package to fit a linear model to characterize differentially expressed genes in each of two microarray datasets (Hodges *et al*, 2006; Durrenberger *et al*, 2015) profiling dorsal striatum of HD cases vs. controls, treating sex as a covariate. We calculated enrichments of the DEGs from each study for TF-target gene modules. We then combined the enrichment p-values across the two studies using Fisher's method to produce a meta-analysis p-value for the association of each TF-target gene module with HD.

Mouse Breeding, Genotyping, and microdissection. The B6.$Htt^{Q111/+\ mice}$ (Strain 003456; JAX) used for the ChIP-seq study have a targeted mutation replacing a portion of mouse *Htt* (formerly *Hdh*) exon 1 with the corresponding portion of human *HTT* (formerly *IT15*) exon 1, including an expanded CAG tract (originally 109 repeats). Mice used in the present study were on the C57BL/6J inbred strain background. The targeted *Htt* allele was placed from the CD-1 background onto the C57BL/6J genetic background by selective backcrossing for more than 10 generations to the C57BL/6J strain at Jackson laboratories. Cohorts of heterozygote and wild-type littermate mice were generated by crossing B6.$Htt^{Q111/+}$ and B6.$Htt^{+/+}$ mice. Male mice were sacrificed at 122 ± 2 days of age (or 16 weeks) via a sodium phenobarbital based euthanasia solution (Fatal Plus, Henry Schein). Both hemispheres of each animal's brain was

microdissected on ice into striatum, cortex, and remaining brain regions. These tissues were snap frozen and stored in -80°C. Experiments were approved by an institutional review board in accordance with NIH animal care guidelines.

High resolution X-ChIP-seq. We prepared duplicate ChIP samples for each antibody from four-month-old *Htt*$^{Q111/+}$ and from age-matched wildtype mice. For each ChIP preparation, chromatin DNA was prepared using the combined striatal tissue from both hemispheres of three mice. Preliminary experiments suggested that this was the minimal amount of material required to provide enough material for multiple IPs. Striata were transferred to a glass dounce on ice and homogenized in cold PBS with protease inhibitors. High-resolution X-ChIP-seq was performed as described (Skene *et al*, 2010), with slight modifications. IPs were performed using Abcam Anti-SMAD3 antibody ab28379 [ChIP grade] or Anti-RNA polymerase II CTD repeat YSPTSPS antibody [8WG16] [ChIP Grade] ab817. Sequencing libraries were prepared from the isolated ChIP DNA and from input DNA controls as previously described (Orsi *et al*, 2015). Libraries were sequenced on an Illumina HiSeq 2500 sequencer to a depth of ~17-25 million paired-end 25 bp reads per sample. Sequence reads have been deposited in GEO, accession GSE88775.

ChIP-seq analysis. Sequencing reads were aligned to the mouse genome (mm9) using bowtie2 (Langmead & Salzberg, 2012). Peak-calling on each sample was performed with MACS v2.1 (Zhang *et al*, 2008), scaling each library to the size of the input DNA sequence library to improve comparability between samples. We retained peak regions with a significant MACS p-value (FDR < 0.01 and a read count ≥10 in at least two of the individual ChIP samples). Enrichment of the SMAD3 motif (JASPAR CORE MA0513.1) was performed with CentriMo

(Bailey & Machanick, 2012), using the 250bp regions around peak summits obtained by running MACS on the combined reads from all the samples. Peaks were mapped to genes using the chipenrich R package (Welch *et al*, 2014), and genes were ranked by the number of peaks within 10kb of each gene's transcription start sites. Gene Ontology enrichment analysis of the top SMAD3 target genes (peak counts >2 s.d. above the mean), was performed using Fisher's exact test, using the same set of GO terms used to analyze the computationally derived TF-target gene modules. Statistical analysis of differential occupancy in $Htt^{Q111/+}$ vs. wildtype mice was performed with edgeR (Robinson *et al*, 2010).

4.6 Figures and Tables

Figure 1. Reconstruction and validation of a transcriptional regulatory network (TRN) model of the mouse striatum. a. Schematic for reconstruction of tissue-specific TRN models by combining information about TF binding sites with evidence from co-expression. b. Training (black) and test set (blue) prediction accuracy for genes in the mouse striatum TRN model. Genes are ordered on the x-axis according to their training set prediction accuracy (r^2, predicted vs. actual expression). c. Distribution for the number of predicted regulators per target gene. d. Distribution for the number of predicted target genes per TF. e. Enrichments of TF-target gene interactions in the mouse striatum TRN for TFBSs supported by DNase footprints identified in 23 tissues.

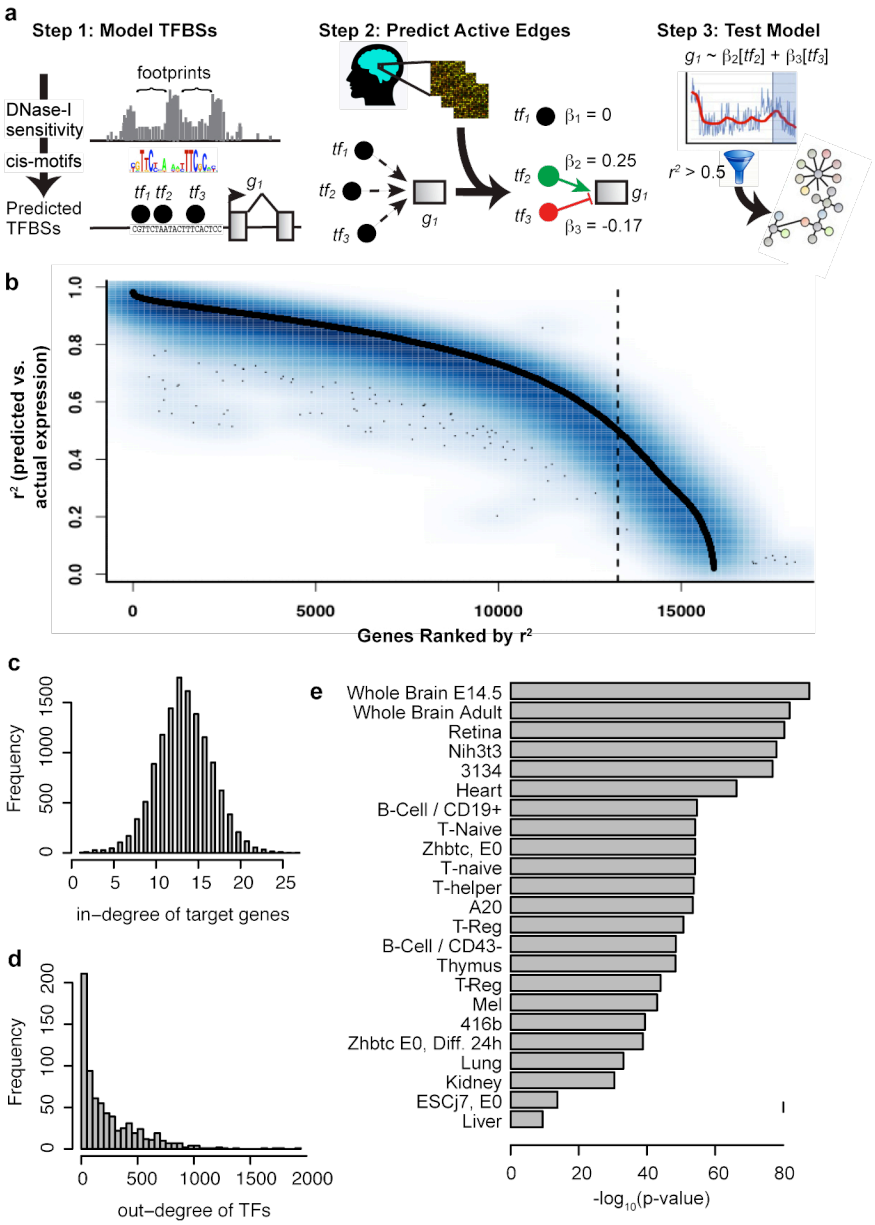

a
Step 1: Model TFBSs
footprints
DNase-I sensitivity
cis-motifs
Predicted TFBSs
Step 2: Predict Active Edges
Step 3: Test Model
$g_1 \sim \beta_2[tf_2] + \beta_3[tf_3]$
$r^2 > 0.5$
$\beta_1 = 0$
$\beta_2 = 0.25$
$\beta_3 = -0.17$
b
r^2 (predicted vs. actual expression)
Genes Ranked by r^2
c
Frequency
in−degree of target genes
d
Frequency
out−degree of TFs
e
Whole Brain E14.5
Whole Brain Adult
Retina
Nih3t3
3134
Heart
B-Cell / CD19+
T-Naive
Zhbtc, E0
T-naive
T-helper
A20
T-Reg
B-Cell / CD43-
Thymus
T-Reg
Mel
416b
Zhbtc E0, Diff. 24h
Lung
Kidney
ESCj7, E0
Liver
$-\log_{10}$(p-value)

Figure 2. Robust changes in striatal gene expression in two-, six-, and ten-month-old HD knock-in mice. Counts of differentially expressed genes ($p < 0.01$) in each mouse model at each time point.

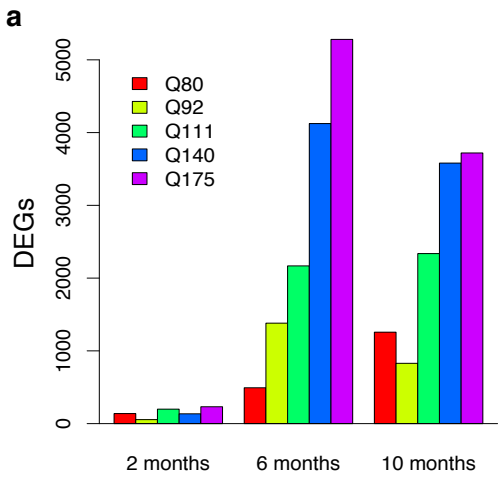
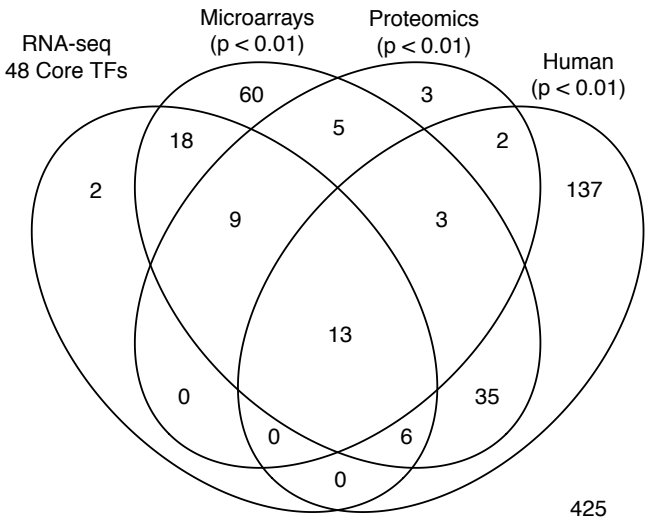


Figure 3. Replication of core TFs in independent datasets. a. Venn Diagram showing overlap between core regulator TF-target gene modules identified in the primary RNA-seq dataset, compared to TF-target gene modules enriched for differentially expressed genes in three independent datasets. b. $-\log_{10}$(p-values) for the strength of enrichment of each of the core regulator TF-target gene modules for differentially expressed genes in each of the four datasets.

A


RNA-seq
48 Core TFs
Microarrays
(p < 0.01)
Proteomics
(p < 0.01)
Human
(p < 0.01)
60
3
18
5
2
2
137
9
3
13
0
35
0
6
0
425

B

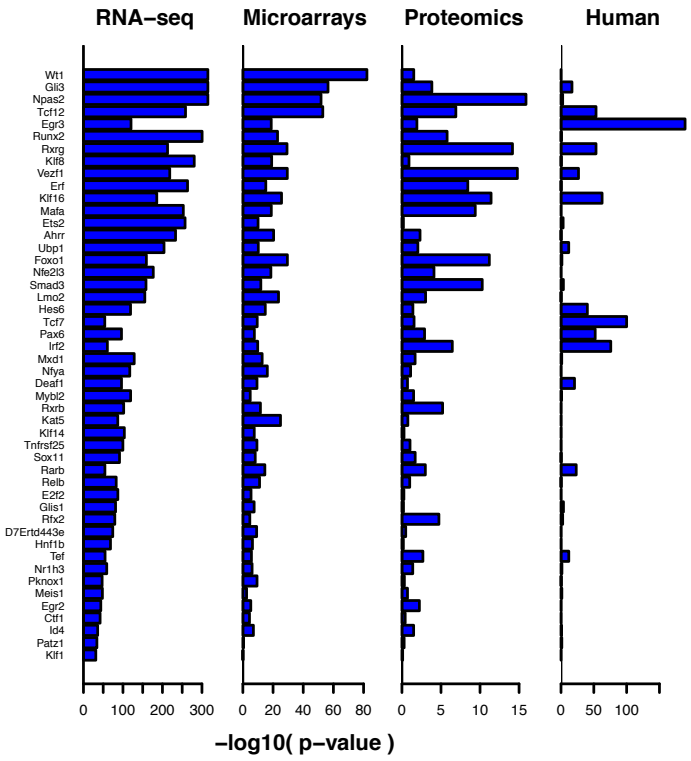

RNA−seq
Microarrays
Proteomics
Human
Wt1
Gli3
Npas2
Tcf12
Egr3
Runx2
Rxrg
Klf8
Vezf1
Erf
Klf16
Mafa
Ets2
Ahrr
Ubp1
Foxo1
Nfe2l3
Smad3
Lmo2
Hes6
Tcf7
Pax6
Irf2
Mxd1
Nfya
Deaf1
Mybl2
Rxrb
Kat5
Klf14
Tnfrsf25
Sox11
Rarb
Relb
E2f2
Glis1
Rfx2
D7Ertd443e
Hnf1b
Tef
Nr1h3
Pknox1
Meis1
Egr2
Ctf1
Id4
Patz1
Klf1
0 100 200 300
0 20 40 60 80
0 5 10 15
0 50 100
−log10(p−value)

Figure 4. Predicted TF-to-TF interactions among 48 putative core regulators of transcriptional changes in mouse models of Huntington's disease. Nodes and edges indicate direct regulatory interactions between TFs predicted by the mouse striatum TRN model. Solid black arrows and dotted red arrows indicate positive vs. inhibitory regulation, respectively, and the width of the line is proportional to the predicted effect size. Blue and orange shading of nodes indicates that the TF's target genes are overrepresented for down-regulated vs. up-regulated genes in HD mouse models. If a TF's target genes are enriched in both directions, the stronger enrichment is shown. Each panel indicates the network state in a specific condition. a. two-month-old $Htt^{Q92/+}$ mice. b. six-month-old $Htt^{Q92/+}$ mice. c. two-month-old $Htt^{Q175/+}$ mice. d. six-month-old $Htt^{Q175/+}$ mice.

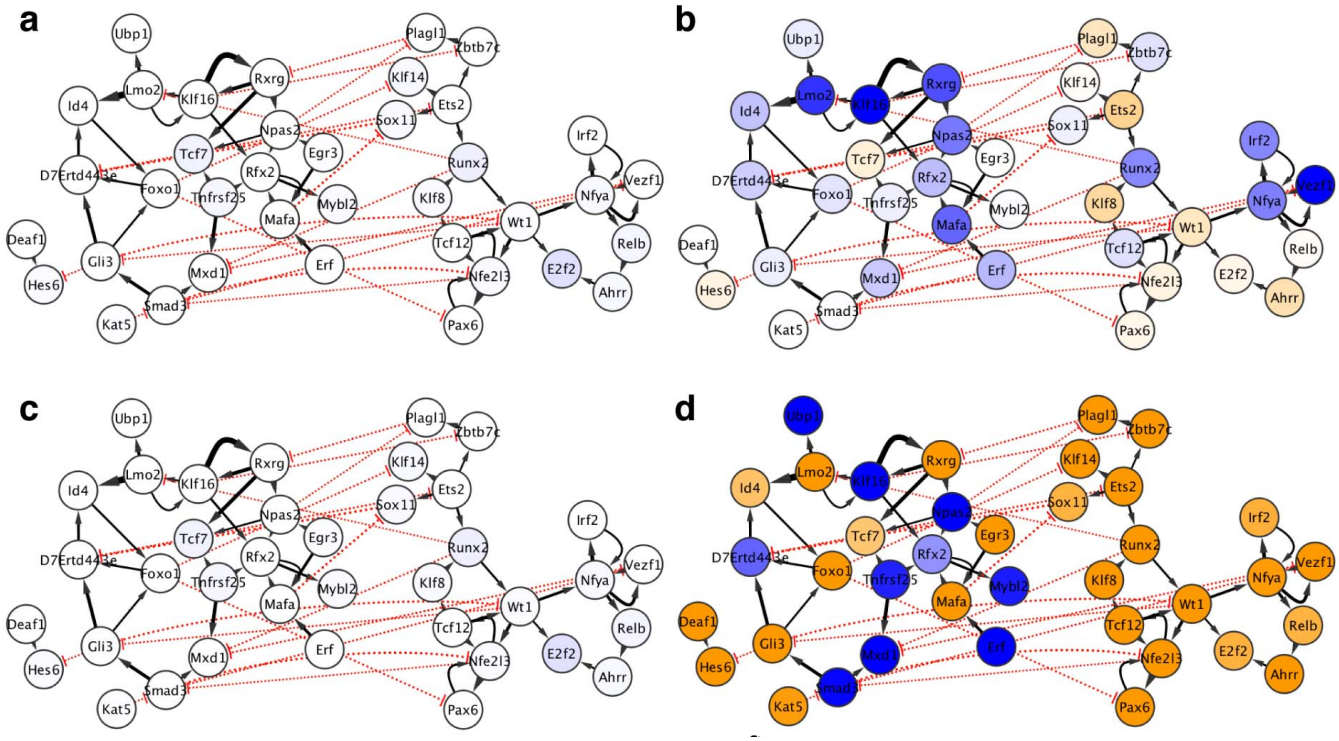


Figure 5. Enrichments of the 48 core TFs for differentially expressed genes in each condition and for cell type-specific genes. a. Enrichments of each TF's target genes for down- and up-regulated genes for each HTT allele at each time point. b. Enrichments of each TF's target genes for genes expressed specifically in one of seven major cell types in the mouse striatum.

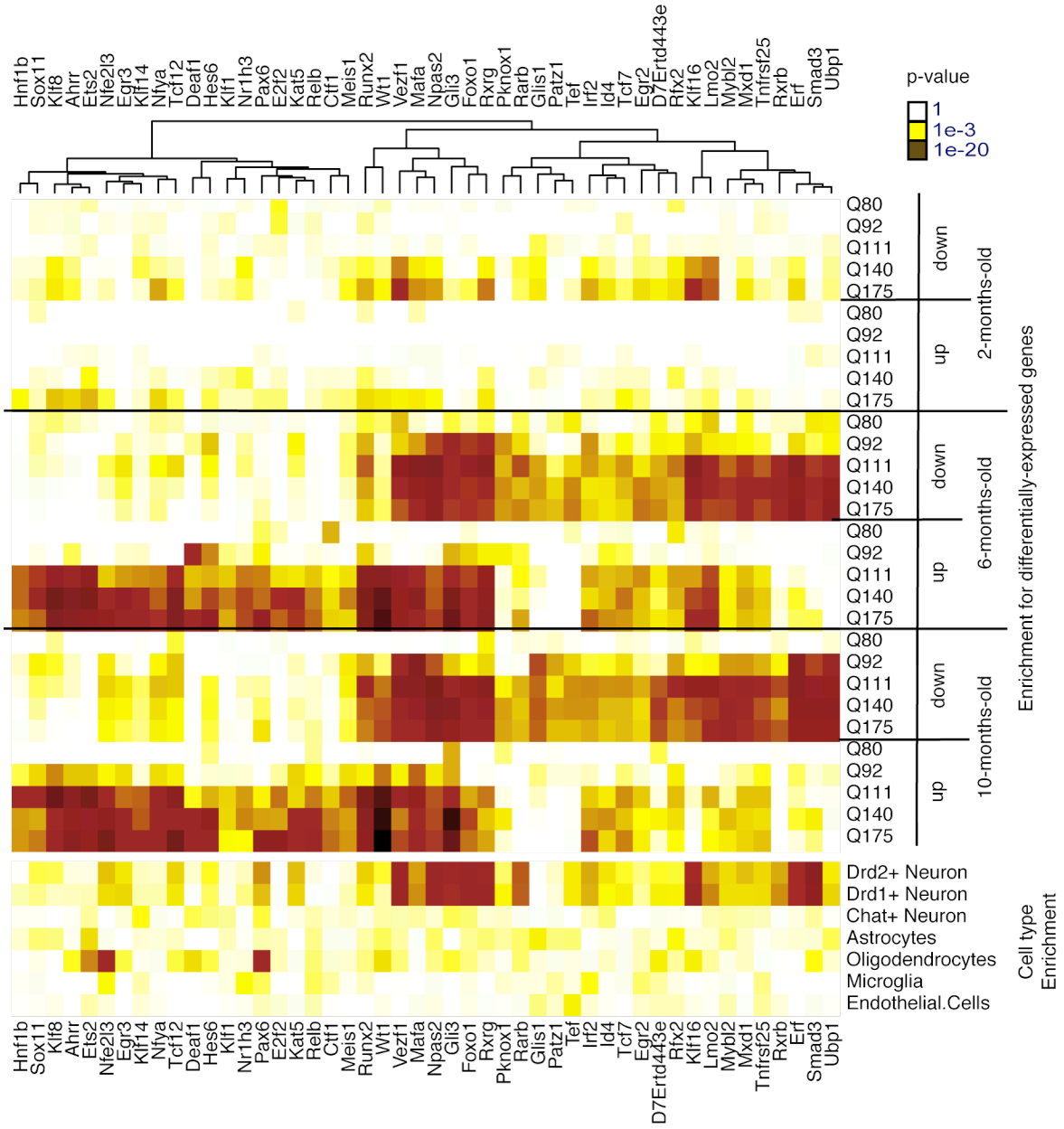
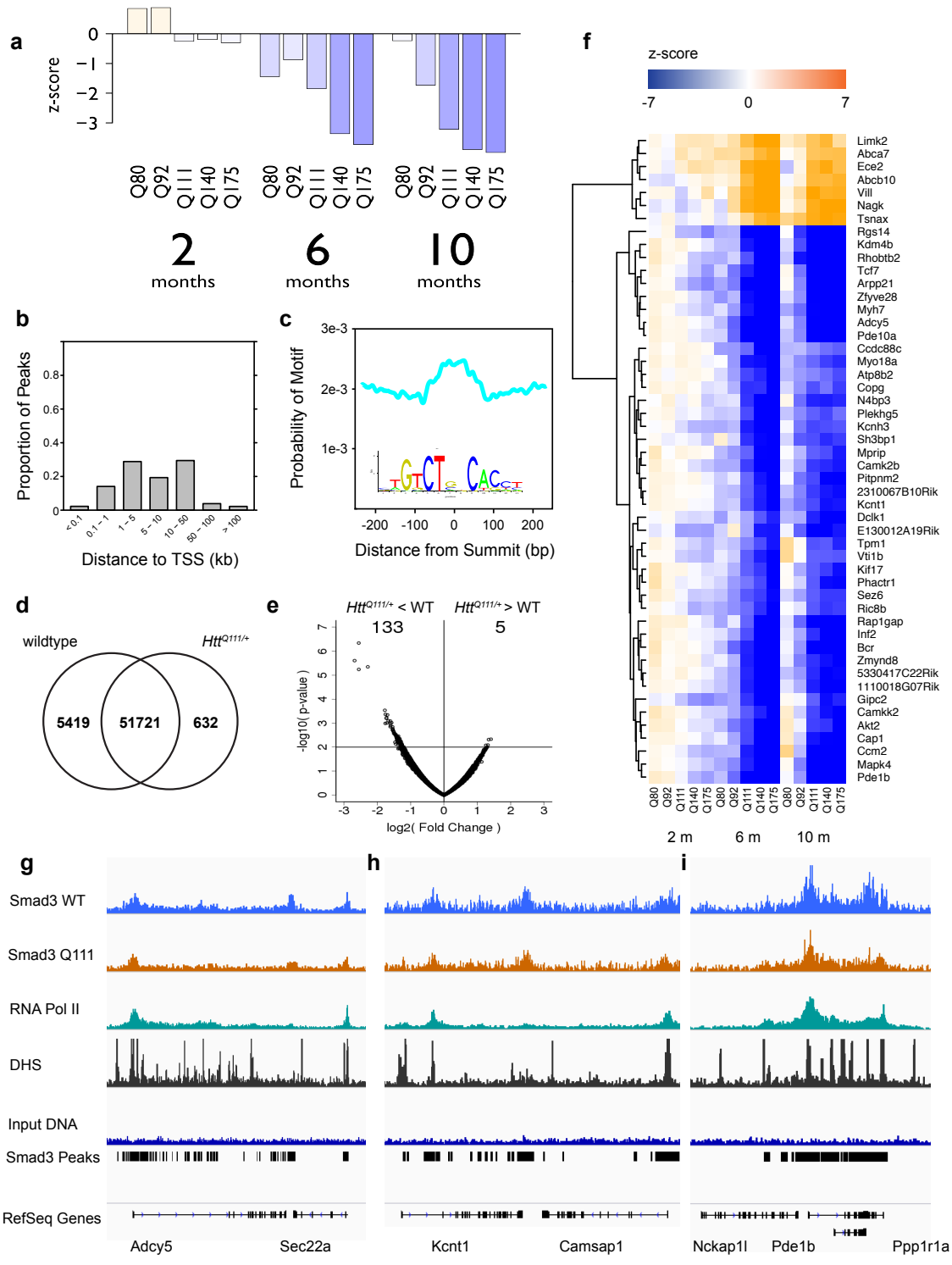
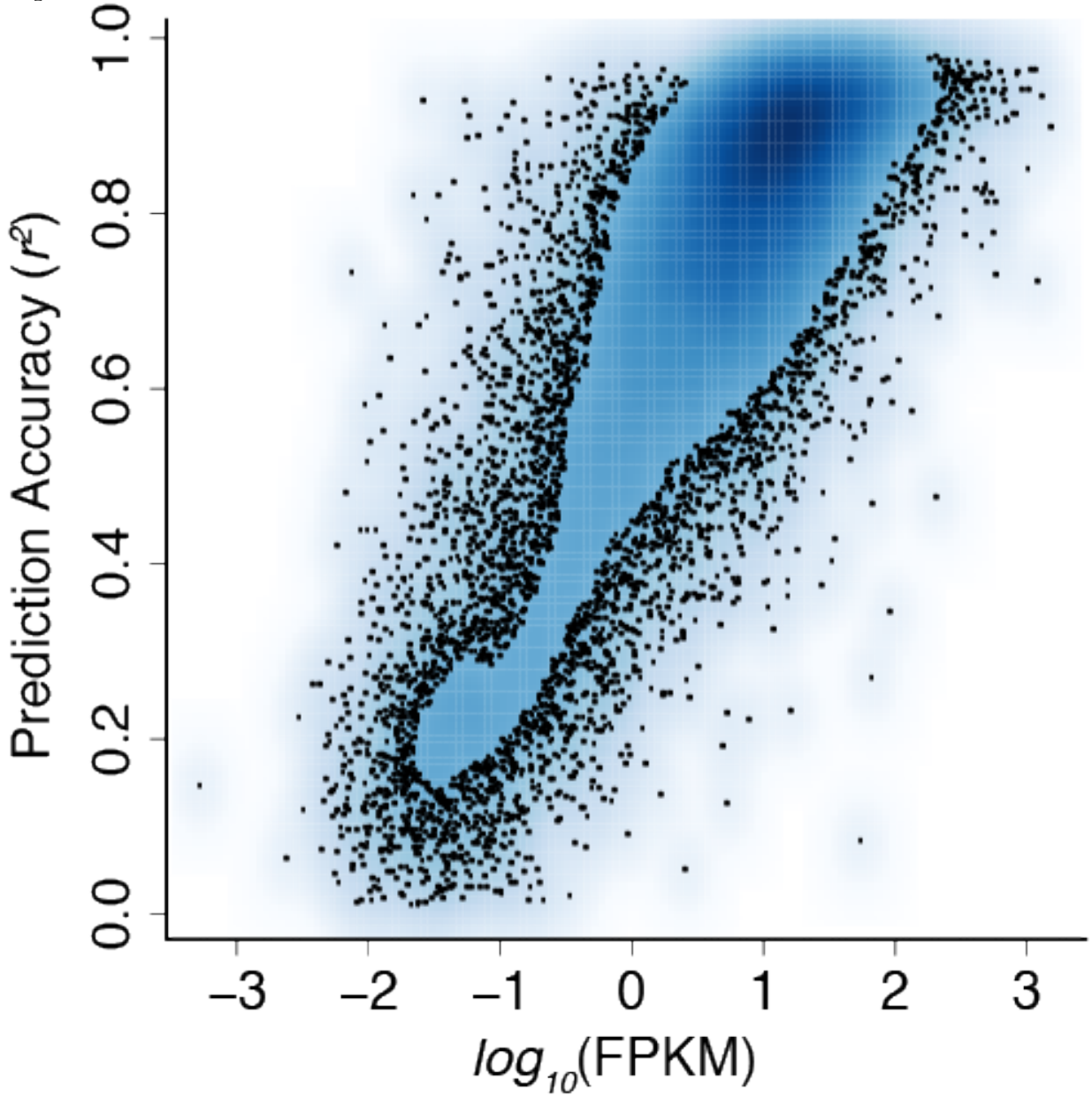


Figure 6. SMAD3 expression, genomic occupancy, and target gene expression in the striatum of HD mouse models. a. Progressive age- and *Htt*-allele-dependent changes in the expression of SMAD3 in mouse striatum. Bars indicate z-scores for the expression level in heterozygous mice with each pathogenic *Htt* allele compared to age-matched $Htt^{Q20/+}$ mice. b. Distribution of the distances of 57,772 SMAD3 peaks identified by ChIP-seq to the nearest transcription start site (TSS). c. The summits of SMAD3 peaks are enriched for the sequence motif recognized by SMAD3 (JASPAR CORE MA0513.1, shown in inset). d. Overlap between peaks identified in $Htt^{Q111/+}$ vs. wildtype mice. e. SMAD3 occupancy is decreased at a subset of peaks in $Htt^{Q111/+}$ vs. wildtype mice. x-axis and y-axis represent the log2(fold change) and –log10(p-value), respectively, for each peak region. f. Age- and *Htt*-allele-dependent expression patterns of the top 50 most strongly differentially expressed SMAD3 target genes. g, h, i. Genomic occupancy of SMAD3 and RNA polymerase II and accessibility of genomic DNA to DNase-I near *Adcy5* (g), *Kcnt1* (h), and *Pde1b* (i).

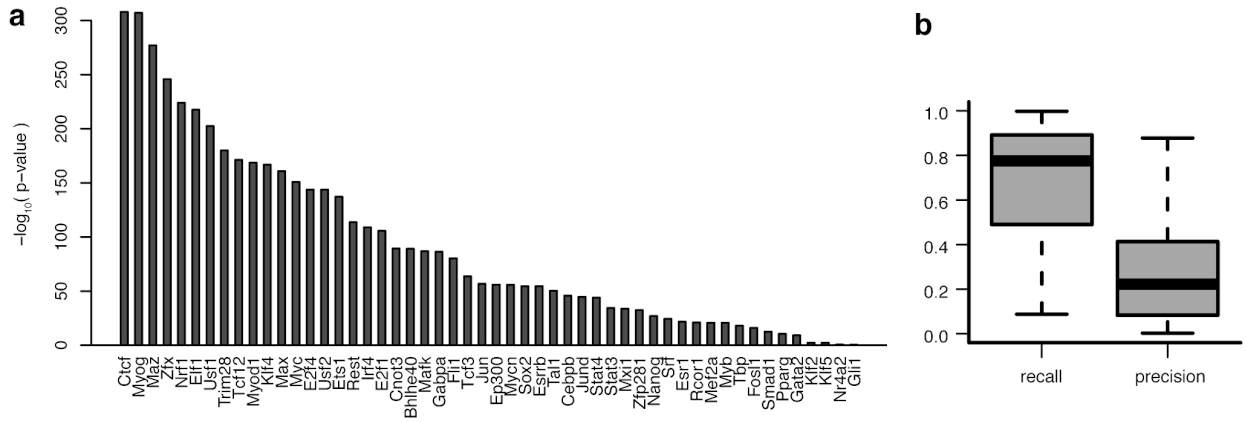

a
z-score
0
−1
−2
−3
Q80 Q92 Q111 Q140 Q175 Q80 Q92 Q111 Q140 Q175 Q80 Q92 Q111 Q140 Q175
2 months
6 months
10 months
b
Proportion of Peaks
0.8
0.6
0.4
0.2
0
<0.1
0.1 – 1
1 – 5
5 – 10
10 – 50
50 – 100
>100
Distance to TSS (kb)
c
Probability of Motif
3e-3
2e-3
1e-3
-200 -100 0 100 200
Distance from Summit (bp)
d
wildtype
Htt^Q111/+
5419
51721
632
e
Htt^Q111/+ < WT
133
Htt^Q111/+ > WT
5
-log10(p-value)
log2(Fold Change)
f
z-score
-7
0
7
Limk2
Abca7
Ece2
Abcb10
Vill
Nagk
Tsnax
Rgs14
Kdm4b
Rhobtb2
Tcf7
Arpp21
Zfyve28
Myh7
Adcy5
Pde10a
Ccdc88c
Myo18a
Atp8b2
Copg
N4bp3
Plekhg5
Kcnh3
Sh3bp1
Mprip
Camk2b
Pitpnm2
2310067B10Rik
Kcnt1
Dclk1
E130012A19Rik
Tpm1
Vti1b
Kif17
Phactr1
Sez6
Ric8b
Rap1gap
Inf2
Bcr
Zmynd8
5330417C22Rik
1110018G07Rik
Gipc2
Camkk2
Akt2
Cap1
Ccm2
Mapk4
Pde1b
Q80 Q92 Q111 Q140 Q175 Q80 Q92 Q111 Q140 Q175 Q80 Q92 Q111 Q140 Q175
2 m
6 m
10 m
g
h
i
Smad3 WT
Smad3 Q111
RNA Pol II
DHS
Input DNA
Smad3 Peaks
RefSeq Genes
Adcy5
Sec22a
Kcnt1
Camsap1
Nckap1l
Pde1b
Ppp1r1a

4.7 Supplementary Information

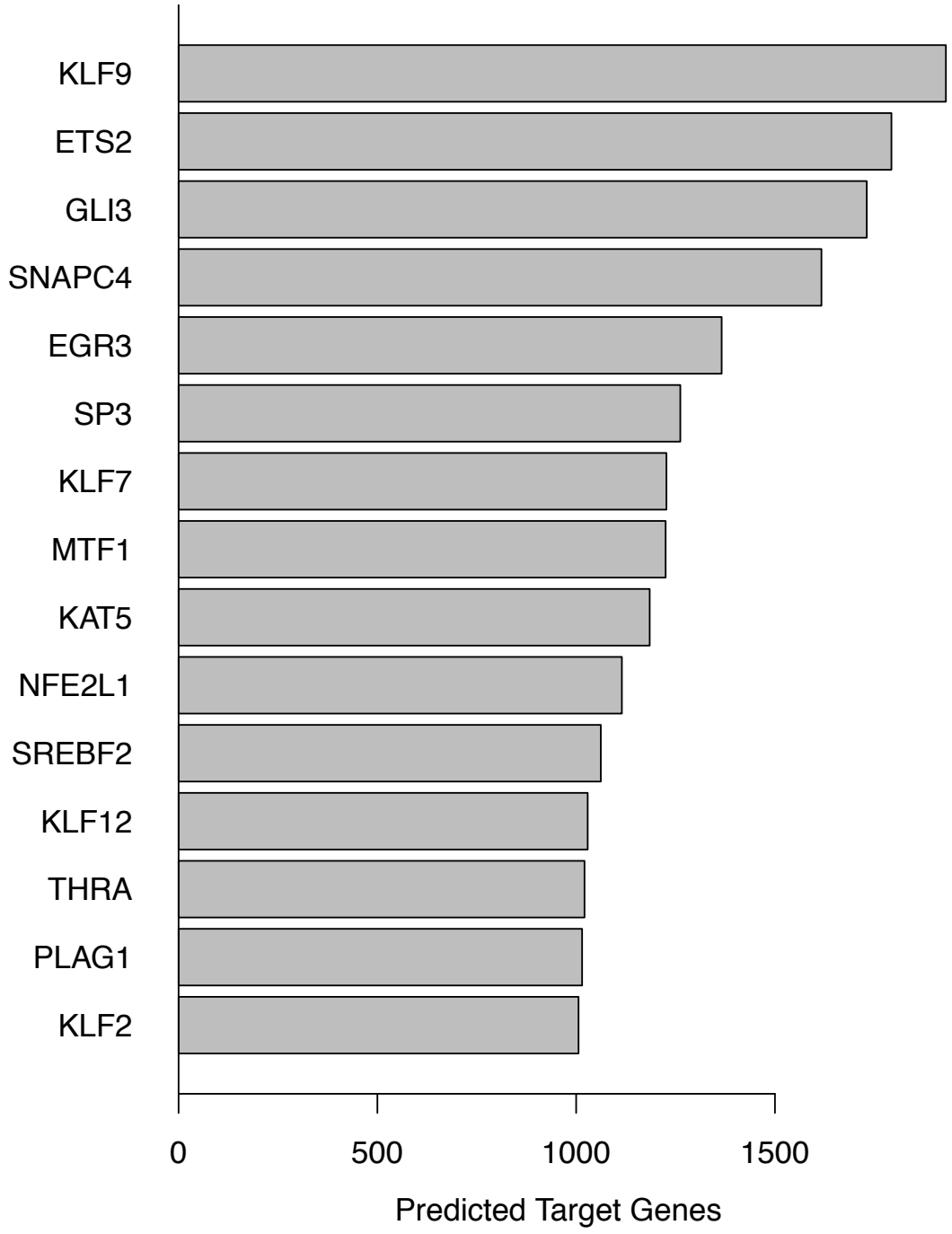
SI Figure 1. Association between TRN prediction accuracy and expression level. Each point on the scatterplot represents the mean expression level of a gene in the striatum (x-axis; fragments per kilobase million, FPKM) and the prediction accuracy for that gene in the transcriptional regulatory network model (r^2, predicted vs. observed expression across all samples).



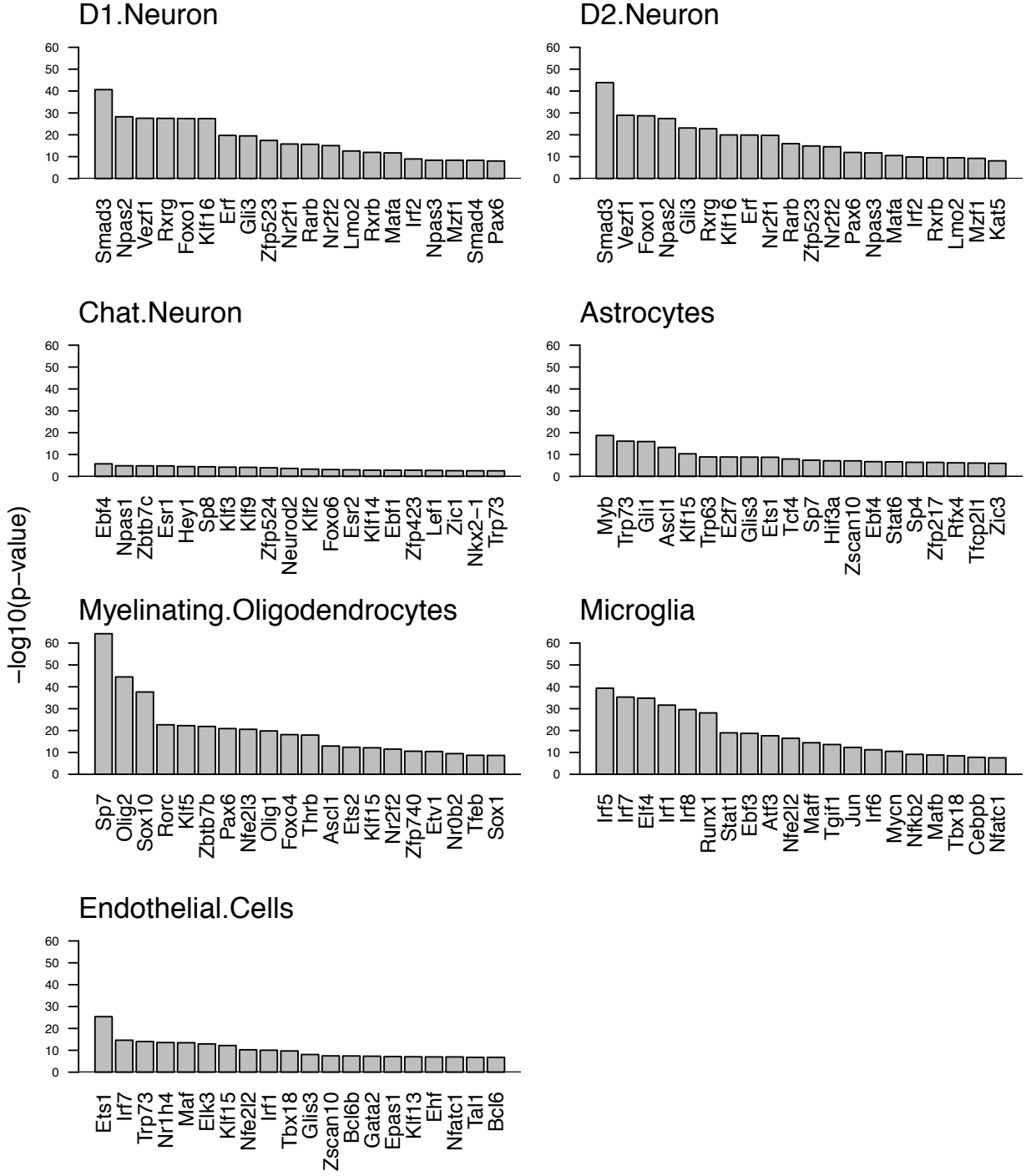
SI Figure 2. Comparison of TF binding site predictions to ChIP-seq data. For each of 52 TFs, we compared the sets of genes adjacent to predicted TF binding sites in our model to the sets of genes adjacent to observed binding sites from ChIP-seq studies. **a.** $-\log_{10}$(p-values) for overlap between modeled vs. observed gene sets (Fisher's exact test). **b.** Distribution of recall (sensitivity) and precision (positive predictive accuracy) of the TFBS model for identifying the target genes of each TF identified by ChIP-seq.



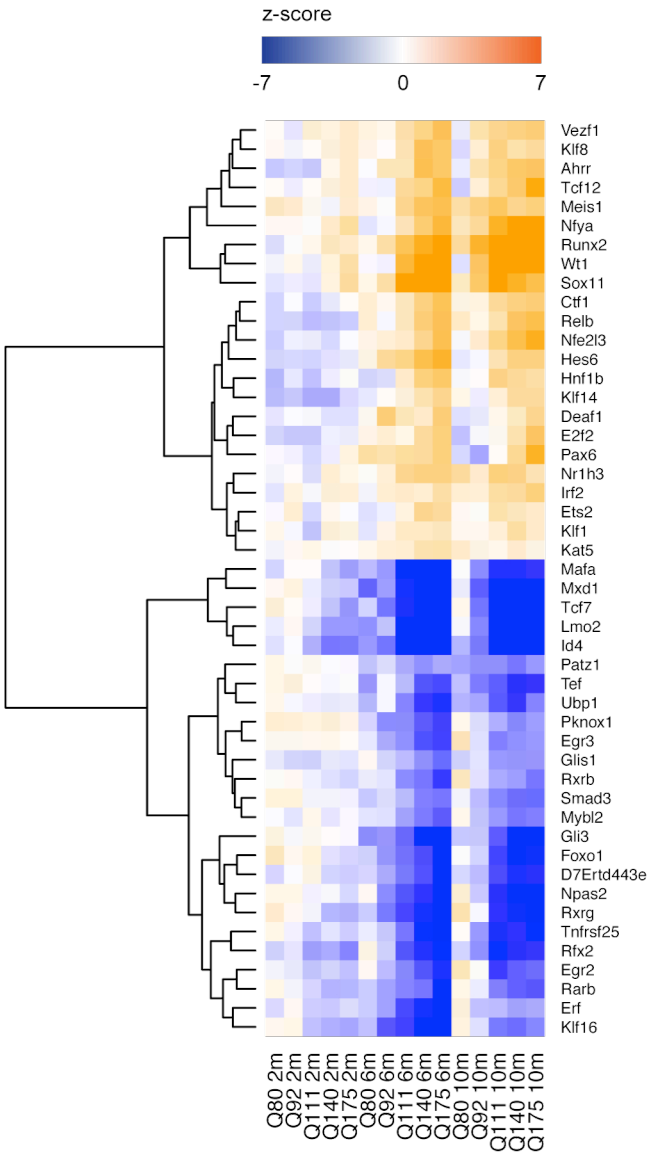
SI Figure 3. TFs with >1,000 predicted target genes. Bars indicate the number of predicted target genes for each of the 15 TFs with >1,000 predicted target genes in the TRN model for the mouse striatum.



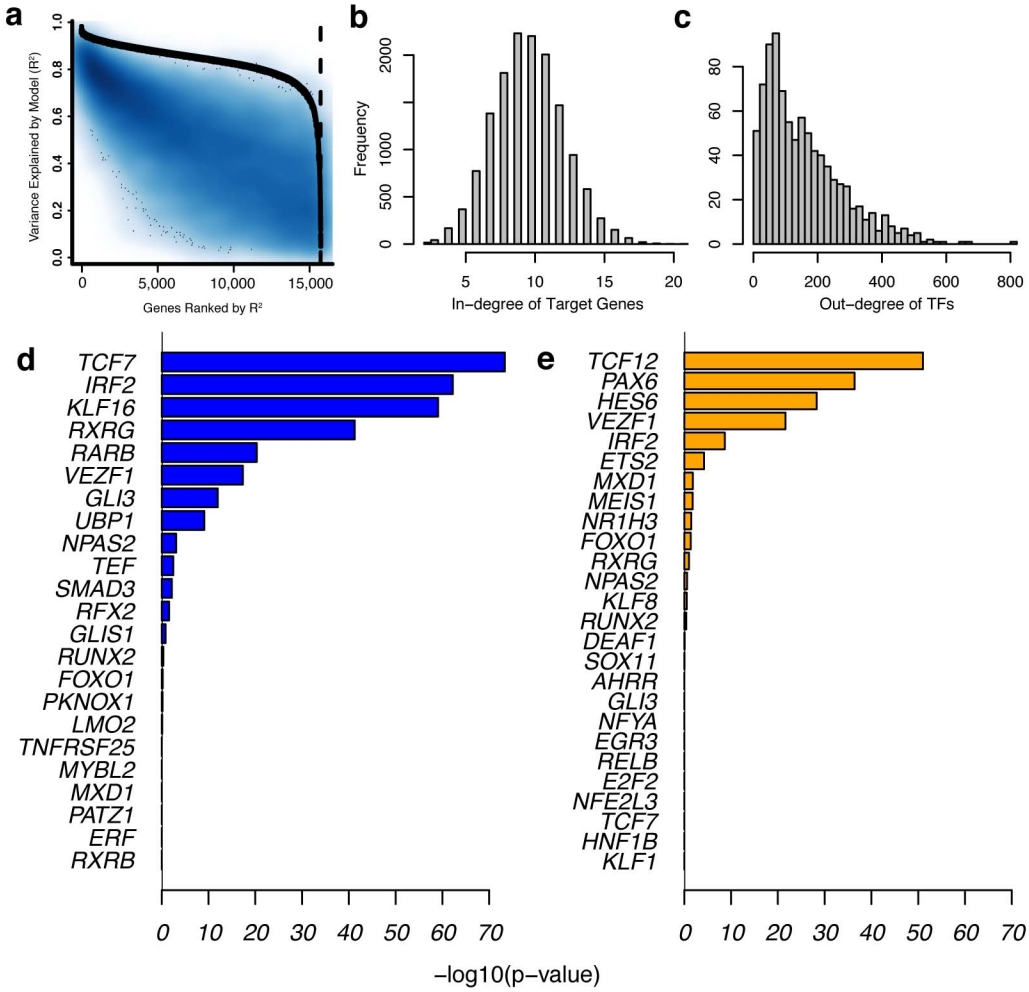
SI Figure 4. Enrichments of TF modules within each striatal cell type. Enrichments of the predicted target genes of each TF for genes expressed specifically in one of seven major cell types in the mouse striatum. The top 20 TF modules are shown for each cell type, ranked by the –log10(p-value) for the strength of enrichment in a one-sided Fisher's exact test.



SI Figure 5. Core regulator TFs are differentially expressed in the striatum of HD CAG knock-in mice. z-scores indicate significance and direction of expression changes in each condition, relative to age-matched *Htt*$^{Q20/+}$ mice.



SI Figure 6. Reconstruction of a TRN model of the human striatum. a. Training (black) and test set (blue) prediction accuracy for genes in the human striatum TRN model. b. Distribution for the number of predicted regulators per target gene. c. Distribution for the number of predicted target genes per TF. d. Enrichment of down-regulated core regulator TFs identified in mouse striatum for down-regulated genes in HD cases vs. controls. e. Enrichment of up-regulated core regulator TFs identified in mouse striatum for up-regulated genes in HD cases vs. controls.



SI Table 1. GO enrichments of top 837 SMAD3 target genes.

Term	Set Size	SMAD3 Targets	Odds Ratio	P-Value	FDR
Actin Filament-Based Process	442	41	3.6	4.2E-11	1.9E-07
mRNA Processing	344	32	3.7	4.2E-09	9.6E-06
Acting Binding	332	30	3.6	2.1E-08	2.4E-05
Neuromuscular Process Controlling Balance	59	12	8.5	1.2E-07	9.4E-05
Histone Modification	293	26	3.6	1.7E-07	1.1E-04
Brain Development	432	32	2.8	1.3E-06	4.9E-04
Chromatin Binding	387	29	2.8	2.7E-06	9.3E-04
Actin Filament-Based Movement	65	11	6.8	2.8E-06	9.3E-04
Regulation of Cell Projection Organization	380	28	2.8	4.5E-06	1.3E-03
Lamellipodium	114	14	4.8	5.5E-06	1.4E-03
Protein Serine/Threonine Kinase Activity	409	29	2.6	9.9E-06	2.3E-03
Protein Deacetylation	49	9	7.5	1.1E-05	2.4E-03
Centrosome	350	25	2.8	1.4E-05	2.6E-03
Purine Ribonucleotide Catabolic Process	382	27	2.7	1.6E-05	3.0E-03
Kinase Binding	199	18	3.4	2.3E-05	4.0E-03
Phosphoric Ester Hydrolase Activity	374	26	2.6	2.5E-05	4.0E-03
Neuronal Cell Body	411	28	2.5	2.5E-05	4.0E-03
Protein Kinase Binding	369	26	2.6	2.8E-05	4.2E-03
Cellular Protein Catabolic Process	415	28	2.5	3.0E-05	4.2E-03
Transcriptional Repressor Complex	70	10	5.6	3.7E-05	4.4E-03
Respiratory System Development	138	14	3.8	5.5E-05	5.3E-03
Kinesin Binding	29	6	9.5	1.1E-04	9.3E-03
Endocytosis	334	23	2.6	1.1E-04	9.3E-03
Negative Regulation of ERBB Signaling Pathway	10	4	22.1	1.4E-04	9.7E-03

4.8 Notes

This work was performed in collaboration with Seth A. Ament, Robert M. Bragg, Peter J. Skene, Jeffrey Cantle, Sydney R. Coffey, Dani E. Bergey, Christopher L. Plaisier, Vanessa C. Wheeler, Marcy E. MacDonald, Nitin S. Baliga, Jim Rosinski, Leroy E. Hood, Jeffrey B. Carroll, and Nathan D. Price.

Software and Primary Data Resources. Code for analysis of gene expression, transcriptional regulatory networks, and ChIPseq data for this manuscript are publicly available in the github repository located at https://github.com/seth-ament/hd-trn. Bedgraph files and raw sequencing data for SMAD3 and RNA Pol2 ChIP-seq can be accessed at the GEO repository #GSE88775 prior to publication at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=oryzgqeerzmvdaf&acc=GSE88775.

Acknowledgements

This work was supported by a contract from the CHDI Foundation (N.D.P., Principal Investigator and J.B.C, Principal Investigator). J.R.P. is supported by a National Science Foundation Graduate Research Fellowship. J.B.C. is supported by Huntington Society of Canada New Pathways Program.

Author Contributions

S.A., J.P., J.C., and N.P. designed research. S.A. performed the computational analysis and built the TRN. R.B. and S.C. conducted the mouse work. J.P. conducted the ChIP-seq experiments. S.A. and J.P. wrote the paper.

5 Gene regulatory drivers in psychiatric disease

5.1 Abstract

Genetic and genomic studies suggest an important role for transcriptional regulatory changes in brain diseases, but roles for specific transcription factors remain poorly understood. By integrating human brain-specific DNase I footprinting and TF-gene co-expression we reconstructed a transcriptional regulatory network (TRN) model for the human brain, predicting the brain-specific binding sites and target genes for 741 transcription factors. Using our TRN model, we predicted TFs that regulate prefrontal cortex gene expression changes in psychiatric and neurodegenerative diseases, as well as functional, disease-associated SNPs that disrupt transcription factor binding sites. Our results suggest that disease-related transcriptomic and genetic changes converge on small sets of disease-specific regulators, with distinct networks underlying neurodegenerative vs. psychiatric diseases. Through *in vitro* studies in human neural stem cells, we validated network predictions that link the transcription factor *POU3F2* to schizophrenia and bipolar disorder via both cis- and trans-acting mechanisms. We show that *POU3F2* has anti-proliferative effects, involving transcriptional repression of a cluster of cell cycle genes that is highly expressed in early proliferative neural stem cells and in adult astrocytes, and which is up-regulated in the prefrontal cortex of individuals with schizophrenia and bipolar disorder.

5.2 Introduction

Convergent evidence suggests that altered transcriptional regulation is a prominent mechanism of common human diseases, including psychiatric and neurodegenerative disorders. Many disease states are accompanied by characteristic, tissue-specific changes in gene expression. Neurodegenerative diseases such as Alzheimer's disease and Huntington's disease involve progressive changes in the expression of thousands of genes in vulnerable brain regions[1,2]. Psychiatric disorders such as schizophrenia, bipolar disorder, major depression, and autism also involve brain gene expression changes, with replicable changes in the expression of hundreds of genes observed in neocortical regions that influence cognition and emotional control[3–7].

While non-genetic factors could contribute to brain gene expression changes in psychiatric and neurodegenerative disorders -- including effects of medications, lifestyle factors, and differences in cell type distributions -- multiple lines of evidence point to a significant role for genetic factors. Hundreds of genetic haplotypes associated with risk for psychiatric and neurodegenerative disorders co-localize with gene expression quantitative trait loci (eQTLs)[8,9]. Genetic variants associated with risk for common diseases are enriched in promoters and enhancers[10], and genetic liability due to risk variants in regulatory regions marked by DNase I hypersensitivity may be as high as 80%[11]. Thus, it has been proposed that the causal variants at many risk loci alter the expression of nearby target genes via mechanisms such as changes in transcription factor binding sites.

Genetically encoded changes in transcription factors (TFs) and other transcriptional regulatory proteins may also contribute to disease risk. Both common and rare genetic variation associated with psychiatric disorders are enriched for genes in transcriptional regulation and chromatin-remodeling pathways[12,13]. 29 TFs are located at genome-wide

significant risk loci from large-scale GWAS of schizophrenia or bipolar disorder[14–17]. Additional TFs have been implicated in risk for these diseases due to their disruption by rare variants[18].

We hypothesized that psychiatric and neurodegenerative disorders involve dysregulation within networks of TFs, TF binding sites, and TF target genes in the brain. Disease related changes in transcriptional networks could occur either at the transcriptional level (i.e., disease-associated changes in gene expression) or at the genetic level (i.e., disease-associated variation in the DNA sequence). Previous studies have used gene co-expression networks to identify disease-perturbed networks in the brain[1,7,19,20]. It has also been observed that disease-associated changes in gene expression may be enriched for genes that are also associated with genetic risk for that disease[1,7]. Other studies have utilized epigenomics techniques and expression quantitative trait loci (eQTLs) to annotate non-coding genetic variation associated with risk for psychiatric and neurodegenerative disorders[4,9], and to fine-map transcriptional regulatory mechanisms at specific disease risk loci[21]. These studies lay the groundwork for an analysis of disease-associated changes linking TFs, TF binding sites, and TF target genes. However, genome-wide studies have not previously integrated all of these levels of TRN organization to understand the transcriptional regulatory architecture of brain diseases.

Here, we reconstructed a transcriptional regulatory network (TRN) model that predicts the genomic binding sites and target genes of 741 transcription factors in the human brain by integrating tissue-specific DNase-seq footprinting with TF-gene co-expression. We used our TRN model to predict key regulators of transcriptomic changes in psychiatric and neurodegenerative diseases, as well as disease-associated SNPs that disrupt transcription factor binding sites. We integrate these results to characterize *cis*- and *trans*-acting

mechanisms by which key regulator TFs influence brain gene expression changes in schizophrenia and bipolar disorder (Fig. S1).

5.3 Results

Reconstruction of a transcriptional regulatory network model for the human brain.

To predict the genome-wide binding sites for transcription factors in the human brain, we performed digital genomic footprinting analysis with 15 DNase-seq experiments from human brain regions and cell types generated by the ENCODE project (Table S1). DNase I cleavage patterns predict occupied binding sites for TFs and other DNA binding proteins[22,23]. We downloaded raw sequencing reads from each DNase-seq experiment (www.encodeproject.org), aligned the reads to the human genome (hg38) with SNAP[24], identified DNase I hypersensitive regions with F-seq[25], and located footprints of DNA binding proteins with Wellington[26]. We intersected DNase I footprints with DNA sequence motifs from JASPAR[27], HOCOMOCO[28], and SwissRegulon[29] to predict binding sites for specific TFs, focusing on 741 TFs that are expressed in the human brain[30] and have known sequence specificity. We used the union of TF binding sites predicted across all 15 DNase-seq experiments to identify 2,637,487 DNase hypersensitive sites (DHSs), containing 1,121,670 footprints (Supplementary Dataset 1).

To predict TF-target gene interactions, we integrated our model of brain TFBSs with evidence of co-expression between TF-gene pairs (Fig. 1a). The model was trained using gene expression data from five of the six brains in the Allen Human Brain Atlas[30], leaving the final brain as a test set. Each brain in this dataset was microdissected into 400-900 tissue samples, with each sample representing the expression levels of transcripts in ~10^4 cells and mapped

onto a reference atlas with 862 brain structures[31]. Thus, the Allen Human Brain Atlas defines a cellular resolution map of gene expression in the human brain.

We applied two algorithms to predict TF-target gene interactions: (i) LASSO regression, in which combinations of TFs are used as predictors of the target gene's expression[32–35], and (ii) Pearson correlation between TF-gene pairs. TFBSs were used as constraints, so that only those TFs whose TFBSs were enriched +/-10 kb from a gene's transcription start site were considered as candidate regulators. We constructed separate models for each of the five brains. We then created a final, consensus model, which retained TF-target gene interactions that were supported by both LASSO and Pearson correlation in at least two of the five brains. In addition, we removed genes for which the LASSO regression model explained <50% of the variance in gene expression. The resulting transcriptional regulatory network (TRN) model includes 741 TFs, 11,093 target genes, and 201,218 TF-target gene interactions (Supplementary Dataset 2).

We evaluated the statistical robustness of our model using the Allen Brain Atlas samples held out from TRN reconstruction (893 microdissected brain tissue samples from a single, microdissected brain). 96% of predicted TF-target gene pairs had correlated expression patterns in the test set ($p < 0.05$), with 150,677 of 201,218 TF-gene pairs (75%) correlated at $|r| > 0.25$. In addition, we used the LASSO regression models from the training sets to predict gene expression in the test set samples. Prediction accuracy was strongly correlated between training and test sets ($r = 0.75$), with 8,369 of the 11,093 genes in our TRN model achieving a test set prediction accuracy greater than the inclusion threshold from training sets ($r^2 > 0.5$; Fig. 1b). These results suggest that the statistical features used to predict TF-target gene interactions are robust in independent test sets.

Many TFs regulate cell-type specific processes in the brain[36]. Therefore, we asked whether the sets of predicted target genes for each TF were enriched in specific brain cell types and whether these cell-type enrichments were concordant with the expression patterns of the TFs themselves. We used gene expression profiles from purified neurons, astrocytes, oligodendrocytes, microglia, or endothelial cells[37] and the pSI algorithm[38] to test for cell-type specific expression of each of the 741 TFs in our TRN model and of each TF's predicted target genes. We identified 135 cell-type specific TFs and 147 cell-type enriched TF-target gene modules (Fig. 1c, Table S2). Among 47 TFs that could be assigned to a specific cell-type by both methods, we observed 91% concordance in cell-type assignments (inter-rater reliability: Cohen's kappa = 0.89, p << 2e-16). Consistent with evidence from the literature, our model assigned the cell-type markers *OLIG2* to oligodendrocytes[39] and *NEUROD2* and *NEUROD6* to neurons[40,41], among many such examples. These results support the biological interpretability of our TF-target gene predictions and suggest that a subset of TRN modules reflect cell-type specific processes. Non-cell-type-specific modules may regulate a wide variety of biological functions that are important in multiple cell types.

Key regulators of disease-related transcriptional changes in prefrontal cortex

Next, we sought to identify TFs that are key regulators of gene expression changes in brain diseases. We focused on gene expression changes in the prefrontal cortex, a neocortical region that is involved in both emotional control and cognition[42]. Altered prefrontal cortex structure and function are implicated in a wide range of neurodevelopmental, psychiatric, and neurodegenerative disorders[43–48].

We studied post-mortem prefrontal cortex gene expression profiles from 1,372 human subjects, with two to three independent cohorts from each of five common brain diseases: schizophrenia[4,19,49], bipolar disorder[49,50], major depression[49,51], autism spectrum disorder[7,52],

and Alzheimer's disease[1,53], as well as non-diseased controls (Table S3). We identified differentially expressed genes in the cases vs. controls from each dataset and tested for over-representation of each TF's target genes among these differentially expressed genes (Fisher's exact test). We considered a TF to be a key regulator in a disease if its predicted target genes were over-represented among the differentially expressed genes in that disease, with a meta-analytic q-value < 0.05 across all cohorts and a p-value < 0.05 in at least two independent cohorts. We identified 69 key regulators for schizophrenia, 13 for bipolar disorder, none for major depression, 58 for autism spectrum disorder, and 78 for Alzheimer's disease (Figure 2a, Tables S4-S7).

We found significant overlap between regulators of disorders with psychotic features (schizophrenia vs. bipolar disorder, 13 shared TFs, 1 expected by chance, p = 1.3e-14; Fig. 2a) and between disorders with a strong neurodevelopmental component (schizophrenia vs. autism, 22 shared TFs, 5 expected by chance, p = 3.1e-9). Intriguingly, while key regulator TFs for schizophrenia were shared with both bipolar disorder and autism spectrum disorder, there was very little overlap between bipolar disorder and autism (3 shared TFs, 1 expected by chance, p = 0.07), raising the possibility that distinct transcriptional networks may underlie the psychotic and neurodevelopment dimensions of schizophrenia. Also, there was relatively little overlap between regulators of bipolar disorder or schizophrenia, compared to key regulators of Alzheimer's disease (BD vs. AD, 1 shared regulator, 1 expected by chance, p = 0.76; SCZ vs. AD, 13 shared regulators, 7 expected by chance, p = 0.02). Taken together, these results suggest that shared key regulator TFs connect diseases with overlapping clinical features.

TFs that are predicted to be key regulators of gene expression changes in a brain disease often co-localized with genetic risk for the same disease. We searched the GWAS catalog[54] for genetic associations at loci containing TF genes, and we identified 13 instances in

which a key regulator TF from our TRN model has been reported to be associated with genetic risk for the same disease (Fig. 2b; Table S8). This is twice as many instances of overlap between GWAS and TRN regulators as expected by chance (odds ratio = 2.05; p-value = 2.6e-2). TFs associated with both genetic and transcriptional changes in the same disease include key regulators for schizophrenia (*SREBF1, NPAS3, POU3F2, RFX2, KLF13, FOXN2, FOXO3*), bipolar disorder (*POU3F2*, *NPAS3*), and Alzheimer's disease (*MEF2C, GLIS3, TFEB, NR3C2*). These results highlight the convergence of genetic and transcriptional changes on shared transcriptional regulatory networks in the brain.

Disease-associated genetic variation influencing TF binding sites

The results above highlight putative *trans*-acting effects in which genetic changes at a small number of TF loci are linked to changes in the expression of many downstream target genes. We next sought to elucidate *cis*-acting network perturbations: genetic variants that alter a transcription factor binding site, perturbing a single edge in the network. eQTL analyses suggest that such effects may be present at thousands of genes, explaining hundreds of GWAS risk loci[4,9,55,56].

To identify TFBS-disrupting SNPs, we intersected our model of human brain TFBSs with genetic variants from Kaviar[57], focusing on SNPs that overlap TFBSs +/- 10 kb of the transcription start site for one of that TF's predicted target genes. We identified 52,705 putative TFBS-disrupting SNPs (Supplementary Dataset 3). These SNPs are predicted to modify 67,152 of the 201,218 TF-target gene interactions in our TRN model (the number of modulated TF-target gene interactions is greater than the number of SNPs, since some SNPs overlap the binding sites for more than one TF and some TFs regulate more than one adjacent gene). These results support the idea that there are widespread effects of non-coding SNPs on gene regulation[58].

To explore the potential impact of TFBS-disrupting variants on disease risk, we used our model to predict functional variants and target genes at schizophrenia risk loci. We found 17 schizophrenia-associated TFBS-disrupting SNPs ($p < 1e\text{-}4$), located at 11 of the 108 genome-wide significant risk loci from the landmark Psychiatric Genomic Consortium GWAS of schizophrenia[14] (Table S9). Fifteen of these 17 schizophrenia-associated TFBS-disrupting SNPs are in strong linkage disequilibrium [LD] ($r^2 > 0.9$) with an eQTL targeting the same gene. Notably, 11 of the 17 schizophrenia-associated TFBS-disrupting SNPs are predicted to disrupt a binding site for one of the schizophrenia key regulator TFs, including binding sites for KLF15 (two binding sites), NR1H2, NR1H3, NR2F6, POU3F2 (two binding sites), PRRX1, RFX4, and SP3 (two binding sites) (Table 1; Table S9). This is significantly more overlap than expected by chance (odds ratio = 3.3, p-value = 0.015). Thus, TFBS-disrupting SNPs predict the functional SNPs on some eQTL haplotypes. The enrichment for binding sites of key regulator TFs again highlights the convergence of genetics and gene expression on shared transcriptional networks.

TFBS-disrupting variants suggested a transcriptional regulatory mechanism for the schizophrenia risk locus at 17p11.2. The extent of LD at this locus spans eight genes: *ATPAF2, DRG2, GID4, LRRC48, MYO15A, RAI1, SREBF1,* and *TOM1L2* (Fig. S2). Our model predicts three disease-associated, TFBS-disrupting SNPs at this locus -- rs7359509, rs4925119, and rs6502618. -- located 1, 4, and 6 kb upstream of *SREBF1*. Each of these SNPs disrupts a putative binding site for a TF that targets *SREBF1* in our TRN model, with eQTLs providing independent support for these SNP-gene associations (eQTL p-value = 9.4e-91). Both rs7359509 and rs4925119 disrupt binding sites for the key regulator TF *KLF15*, while rs6502618 disrupts a binding site for the key regulator TF *RFX4*. These results implicate *SREBF1* as the causal gene at this locus, and propose altered *KLF15* and *RFX4* regulation of *SREBF1* transcription as a mechanism.

Notably, *SREBF1* itself is a key regulator TF of schizophrenia in our TRN model, suggesting dysregulation of *SREBF1* target genes in schizophrenia. *SREBF1* target genes in our TRN model were over-represented for oligodendrocyte-specific genes (p-value = 3.2e-8). Also, *SREBF1* target genes were enriched for the Gene Ontology terms "lipid binding" (p-value = 6.6e-5) and "phospholipid metabolism" (p-value = 2.1e-3), consistent with the role of *SREBF1* in lipid biosynthesis[59]. We speculate that genetic perturbation of the (*RFX4-KLF15*)-*SREBF1*-(downstream target genes) circuit could contribute to deficits in white matter tracts and myelination, which have been described in the cortex of schizophrenia patients[60].

Modulation of a POU3F2 binding site by a schizophrenia-associated SNP in the *VRK2* promoter

We selected the transcription factor *POU3F2* (also known as *BRN2*) for experimental validation for the following four reasons: (i) *POU3F2* is a predicted key regulator of schizophrenia (p-value = 9.7e-6) and bipolar disorder (p-value = 5.2e-6) with predicted target genes over-represented among up-regulated genes in prefrontal cortex tissue from affected individuals with both diseases; (ii) The *POU3F2* genomic locus is associated with risk for bipolar disorder[15,17]; (iii) Predicted binding sites for *POU3F2* are disrupted by disease-associated SNPs at two genome-wide significant risk loci for schizophrenia (the *VRK2* and *SRPK2* loci); and (iv) *POU3F2* has well known roles in neural progenitors and neuronal differentiation[61], suggesting a tractable, disease-relevant cell type in which to characterize its functions.

Our TRN model implicates POU3F2 in schizophrenia and bipolar disorder both through *cis*-acting effects of disease-associated variants that disrupt POU3F2 binding sites and through *trans*-acting effects on prefrontal cortex gene expression. To validate a *cis*-acting effect of POU3F2, we characterized the regulatory impact of rs13384219 (chr2:58134458 A/G).

This variant is located on one of several haplotypes at this locus that are associated with risk for schizophrenia (Fig. 3a; lead SNP at this locus, p = 1e-11; rs13384219, p = 2.1e-5), and it is located 336 bp upstream of the transcription start site for *VRK2*. Our TRN model predicted a functional effect of this SNP on *POU3F2* binding for two reasons. First, the SNP overlaps a DNase I footprint spanning a sequence motif recognized by POU-domain transcription factors (Fig. 3b). Second, *POU3F2* expression was strongly correlated with *VRK2* expression in the Allen Human Brain Atlas (r = 0.67; Fig. 3c).

We used luciferase assays to test the activity of a 436 bp fragment of the *VRK2* promoter containing either the A or G allele of rs13384219. The risk-associated G allele decreased the activity of the *VRK2* promoter in HEK293 cells (p = 4e-24; Fig. 3d). We repeated this experiment in combination with POU3F2 over-expression to test the effect of POU3F2 on the activity of the *VRK2* promoter fragment. *POU3F2* overexpression caused a dose-dependent decrease in the activity of the *VRK2* promoter fragment with the rs13384219 A allele. By contrast, POU3F2 over-expression caused a dose-dependent increase in the activity of the *VRK2* promoter fragment with the rs13384219 G allele ($P_{SNP \times POU3F2}$ = 8e-4). These results validate our model's predictions that POU3F2 regulates the *VRK2* promoter and that rs13384219 modifies this effect.

POU3F2 regulates a schizophrenia- and bipolar disorder-related gene network in neural stem cells

To validate trans-acting effects of POU3F2 and gain insight into its potential role in psychiatric disease, we overexpressed it in primary human neural stem cells (phNSCs; NSCs). NSCs have low POU3F2 expression, and *in vivo* induction of POU3F2 naturally occurs in early neural precursors, just subsequent to the stage of neurodevelopment at which our phNSCs were derived[61]. We collected RNA for microarray analysis at 3 and 10 days following infection and puromycin selection of cells transduced with a POU3F2 lentiviral expression vector. K-

means clustering of these microarray data revealed nine gene clusters, two of which were differentially expressed upon POU3F2 overexpression specifically at the day 3 time point in all 4 replicate samples: cluster C6 showing repression and cluster C1 up-regulation (Fig. 4a).

We compared genes in clusters C6 and C1 to our TRN model to evaluate the overlap with our *in silico* network predictions. Genes in C6 were enriched for genes whose promoters and proximal enhancers (+/-10 kb) harbor multiple brain-occupied POU3F2 binding sites from our model (p = 3.7e-3), as well as for computationally-predicted *POU3F2* target genes in the adult brain (p = 8.5e-4, odds ratio = 1.44, 105 shared genes). Importantly, genes in C6 were over-represented among genes that were up-regulated in prefrontal cortex from individuals with schizophrenia (p = 2.4e-4) and bipolar disorder (p = 6.7e-5).

Genes in C1 were also enriched for genes whose promoters and proximal enhancers harbor multiple brain-occupied POU3F2 binding sites from our model (p = 4.7e-3). However, the genes in C1 did not overlap computationally-predicted POU3F2 target genes or genes up-regulated in schizophrenia or bipolar disorder, suggesting that these POU3F2 target genes are either non-targeted in the adult brain or are missed by our computational model. Therefore, POU3F2 overexpression in phNSCs validated our network prediction that POU3F2 is a key regulator of gene expression changes in prefrontal cortex from individuals with schizophrenia and bipolar disorder and identified additional target genes that may be specific to neurodevelopment.

To gain perspective on how these expression changes relate to *in vivo* development, we explored the expression patterns of genes from clusters C6 and C1 in several other neurodevelopmental gene expression datasets: (i) RNA-seq of dorsolateral prefrontal cortex development from BrainSpan[62]; (ii) gene expression microarrays of laser-captured regions of the developing macaque cortex[63]; (iii) single-cell RNA-seq of the developing human cortex[64];

and (iv) RNA-seq of *in vitro* differentiation of human induced pluripotent stem cells toward neurons[65].

Genes in C6, down-regulated upon POU3F2 overexpression, were enriched for genes involved in the cell cycle (Gene Ontology, p=5.8e-75). Genes in this cluster overlapped significantly with clusters obtained from *in vivo* neocortical data, showing high expression during the first trimester of *in utero* development and low expression thereafter, in both human and non-human primate (Fig. 4b, Fig. S4). TRN analysis of the genes in both cluster C6 and the overlapping human DFC cluster indicated the involvement of several TFs known to control cell cycle, including the top hit E2F1 (p=1.3e-32), which shows high correlation with the DFC cluster itself (r=0.65, p=2.5e-5). Single-cell data from the developing human neocortex indicate that these genes are expressed specifically in dividing cells of the prenatal neocortex (Fig. 4c). Remarkably, a subset of the genes in the C6 cluster, including GLI3, are also highly expressed in astrocytes of the adult cortex (Fig. S5).

Genes in C1, up-regulated upon POU3F2 overexpression, were enriched for genes involved in transcription (Gene Ontology, p=4.1e-13). These genes overlapped a cluster of genes whose neocortical expression peaks in the middle of the second trimester and which correlates with POU3F2 expression itself (Fig. 4b, r=0.87). Single-cell data suggest that these genes are expressed in cells of the developing neocortex that are not dividing and that these putative targets of POU3F2 transcriptional activation increase further as neurons mature (Fig. 4c).

These observations suggest that POU3F2 overexpression in primary human phNSCs *in vitro* recapitulates an *in vivo* cellular transition out of an early highly proliferative state and into a subsequent POU3F2-driven state during the second trimester of human neocortical development. Similar cluster overlapping with data from the CORTECON project[60] indicates

that the dynamics of these gene clusters are robust in distinct *in vitro* neural differentiation protocols (Fig. 4d).

POU3F2 represses cellular proliferation in primary human neural stem cells

To directly test the effects of POU3F2 on the transition from proliferative to non-proliferative cell states, we quantified cell proliferation, cell size, and nucleus size in phNSCs and in cells differentiated for two weeks toward neurons or astrocytes. We overexpressed POU3F2 as in the gene expression profiling experiments above, and we knocked it out via lentiviral delivery of a CRISPR/Cas9 sgRNA. We observed a significant decrease in proliferation in POU3F2 overexpressing phNSCs compared to control phNSCs (~2 fold decrease, p=0.0043, Fig. 5a). There was a trend toward increased proliferation in POU3F2 knockout phNSCs, but this was not significant (p=0.1523). In addition, we observed a significant increase in cell size in POU3F2-overexpressing cells compared to controls, both in phNSCs (13 fold increase, p<0.0001) and in cells differentiated toward cortical neurons for two weeks (~6 fold increase, p<0.001) (Fig. 5b,c)). We also observed a significant increase in nucleus size in POU3F2 overexpressing cells in the NSC state (~1.5 fold increase, p=0.01). These findings suggest that POU3F2 overexpression inhibits cellular proliferation, possibly via its repression of cell cycle genes and of genes expressed in proliferative phNSCs.

5.4 Discussion

In this study, we have presented a genome-scale transcriptional regulatory network model for the human brain. Our computational analyses show that gene expression changes and putative cis- and trans-acting genetic changes associated with psychiatric and neurodegenerative disorders involve shared key regulator TFs. We validated these

mechanisms for *POU3F2*, providing convergent evidence for a role of this TF in bipolar disorder and schizophrenia.

Our results support the view that common human diseases involve changes within polygenic biological networks. Recently, Boyle et al. (2017)[66] invoked the small-world property of biological networks – nearly all genes in a cell form a single interconnected network of direct and indirect connections -- to explain how virtually any gene expressed in a disease-relevant cell type could contribute to genetic risk for a disease. By contrast, our results emphasize that an understanding of network structure can reveal a relatively small number of core genes.

Our results suggest that schizophrenia and bipolar disorder are associated with transcriptional regulatory changes in multiple brain cell types. Most research on schizophrenia and bipolar disorder has focused on neuronal mechanisms. Our results suggest that neuronal gene expression changes in these diseases are mediated by TFs such as *MEF2A*, a key regulator TF whose target genes were enriched in neurons (Table S3) and which has previously been shown to influence synaptic function[67]. However, our network models also reveal disease-perturbed networks in glial cells. The oligodendrocyte-enriched *SREBF1* network was implicated by both genetic and gene expression changes.

Strikingly, several key regulator TFs for schizophrenia and bipolar disorder are well known for their roles in neural stem cells, including *HES1*, *MEIS2*, *NKX2-1*, *NPAS3*, RFX2, *RFX4*, *SOX2*, and *SOX9*, among others[36]. Of these, *POU3F2*[15,17], *SOX2*[14,21], *NPAS3*[18,68], and *RFX4*[69] are also associated with genetic risk for schizophrenia or bipolar disorder, suggesting an etiological role. Interestingly, we identified these NSC-related key regulator TFs based on differential expression in the adult brain. Most of these TFs are also highly expressed in adult astrocytes, and their target genes in our TRN model were enriched for astrocyte-specific genes

(Table S3). These pleiotropic effects make it difficult to discern the time point(s) and cell types in which these TFs influence psychiatric disorders. A neurodevelopmental hypothesis is compelling, since adverse events during fetal development are among the strongest non-genetic risk factors for schizophrenia and mood disorders and many schizophrenia risk genes are most highly enriched during fetal brain development[70,71]. However, changes in adult astrocytes likely also contribute to psychiatric disorders. Several of these NSC/astrocyte TFs have been implicated in the malignancy of astrocytic tumors[72] and in astrocyte differentiation[73], suggesting that they may play a role in astrocyte proliferation.

Convergent lines of evidence led us to focus on POU3F2 as a central regulator of gene expression changes in schizophrenia and bipolar disorder. We validated key network predictions that POU3F2 target genes are over-represented among differentially expressed genes in prefrontal cortex from SCZ and BD vs. controls and that a risk-associated SNP near *VRK2* influences gene expression through an interaction with *POU3F2*. We show that POU3F2 represses cell cycle genes and regulates the proliferation of neural stem cells. These anti-proliferative effects are consistent with two recent reports linking *POU3F2* to neural proliferation phenotypes in stem cell models of autism[74,75]. This manuscript is being submitted in parallel with a second paper demonstrating that POU3F2 regulates SCZ-related gene expression changes in prefrontal cortex. The two studies converged on this conclusion independently, despite utilizing different transcriptomics datasets and network reconstruction approaches. We exchanged gene lists and found that there is significant overlap between the lists of computationally predicted POU3F2 target genes described in the two studies (p-value = 4.5e-7, odds ratio=2.4). Therefore, our shared conclusions represent a true convergence of our approaches on the same SCZ-related gene network working from independent starting data. This convergence across studies emphasizes the reproducibility of results in systems biology

and the importance of POU3F2 as a key regulator of bipolar disorder. Identifying risk genes for bipolar disorder has been particularly difficult, with only ~10 replicable GWAS risk loci. The convergent evidence that we present adds POU3F2 to a very short list of well-supported bipolar disorder risk genes.

Our results suggest a promising path forward to the identification of additional genes and gene networks contributing to psychiatric disorders. Several additional TFs are promising candidates for experimental validation, including an assessment of combinatorial effects. Understanding how genetic polymorphisms associated with disease risk alter the function of these networks in the developing and adult brain will provide valuable insights for both basic neurodevelopmental biology and disease mechanisms.

Our TRN model is a broadly applicable resource for future genetic and genomic studies of the human brain and of brain diseases. We have made available 741 gene sets defined as the target genes for each TF in our TRN model. These gene sets can now be used to identify key regulators of gene expression changes in any brain-related transcriptomics experiment. In addition, we have made available our models of TFBSs and of TFBS-disrupting SNPs. We anticipate these resources will be broadly useful for functional analysis and fine-mapping of brain-related GWAS. As such, this study presents a roadmap for understanding brain gene regulation in human health and disease.

5.6 Methods

DNase-seq of human brain tissue.

We used publicly available data from 15 DNase-seq experiments with human brain tissue, generated by the ENCODE project (Table S1). FASTQ files were downloaded from

https://www.encodeproject.org in January, 2016. Note that ENCODE makes a distinction between sample and experiment in that one experiment can contain more than one sample. An experiment can contain both single or paired-end reads, with varying depth of sequencing and varying read length in a single experiment.

https://www.encodeproject.org/search/?type=Experiment&assay_slims=DNA+accessibility&assay_title=DNase-seq&award.project=ENCODE&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&organ_slims=brain

Alignment

For each DNase-seq experiment, we started with the FASTQ files available on https://www.encodeproject.org/. Each FASTQ file (or paired-end files) was aligned to GRCh38 using the SNAP algorithm[24]. SNAP creates a large hash table of indices prior to alignment with a default seed length of 20. Because most of the FASTQ files were short (< 50 bases), we reduced the seed size to a length of 16. The number of footprints found from an experiment is strongly correlated to the depth of sequencing. As the length of reads was variable, the number of mismatches allowed was adjusted accordingly as a parameter of SNAP.

Identifying regions of open chromatin

To identify regions of open chromatin from the aligned BAM files, we used F-seq[25]. Our choice of F-seq was based on work from Koohy *et al.*[77], who compared four different approaches (F-seq, Hotspot, MACS and ZINBA) on DHS data. We utilized the parameters they found to produce the best results in their benchmark comparison to ENCODE ChIP-seq data, using a minimum length of 400 bases.

Running Wellington

Using the output bed files from F-seq, we identified putative footprints using Wellington[26]. Wellington was run with standard settings and -fdrlimit set to -1. Instructions for running Wellington can be found here: http://pythonhosted.org/pyDNase/tutorial.html.

FIMO Database

All potential motif matches in GRCh38 were identified using FIMO[78] and parsed using customized bash scripts. As an input to FIMO, we used JASPAR CORE Vertebrate 2016 collection (http://jaspar.genereg.net/html/DOWNLOAD/), and additional motifs from the HOMOCOMO and Swiss Regulon databases that match TF families not included in JASPAR CORE. Motif point-weight matrices were downloaded with the MEME suite[79]. The footprints identified from Wellington were intersected with the FIMO catalog. To maximize coverage, and because of the potential imprecise nature of footprints, if any part of a known motif overlapped with a single base of the footprint, an entry was created. We assigned motifs to their cognate TFs, as well as to additional TFs in the same DNA-binding domain family from the TFClass database[80].

Transcriptional regulatory network model for the human brain

We predicted TF-target gene interactions by integrating brain-specific DNase-seq genomic footprints with human brain gene expression profiles from the Allen Brain Atlas (http://human.brain-map.org/static/download). The Allen Brain Atlas consists of 2,748 microarray gene expression profiles of microdissected tissue from six human brains. We

applied two gene expression-based methods to predict active TF-target gene interactions among these candidate regulators: (1) Pearson correlation and (2) LASSO regression (see below for details). We constructed separate models using the gene expression data from each of five brains in the ABA, leaving the data from the last of the six Allen Brain Atlas brains as a test set. Finally, we created a consensus model, retaining TF-target gene interactions that were selected by both the Pearson and LASSO methods in at least two of the five brains.

Selection of candidate regulators based on genomic footprinting. We counted the number of footprints for each TF in proximal and distal regions around each gene's canonical transcription start site (TSS; ENSEMBL gene models), with proximal regions defined as +/- 10kb from the TSS and distal regions defined as +/- 1Mb from the target gene's TSS and >10kb from any gene's TSS. TFs with low-complexity sequence motifs generally have large numbers of motifs. To reduce the bias of our models toward TFs with low-complexity motifs, we quantile normalized the footprint counts for each TF across all genes. We selected as candidate regulators those TFs that had normalized footprint counts in the upper quartile, in either the proximal or distal region around each gene's TSS.

Methods to predict active TF-target gene interactions.

1. Pearson correlation. We calculated the Pearson correlation (r) between each TF-gene pair, across the samples from each of the six ABA brains, separately. TF-gene pairs with $|r| > 0.25$ were considered correlated.

2. LASSO regression. We fit a LASSO regression model to predict the expression of each target gene based on a linear combination of the expression patterns of its candidate regulators. These LASSO regression models were fit using the glmnet R package[81]. We considered candidate regulators with proximal binding sites and with distal binding sites, sequentially. First, we constructed a model considering only TFs whose binding sites were

enriched in proximal regions (+/-10kb from the TSS). We then used the predictions from this model as an offset in a second model in which we considered TFs whose binding sites were enriched only in distal regions (>10kb and <1Mb from the TSS).

Over-representation of TF target genes among differentially expressed genes, cell-type specific genes, functional categories, and curated gene sets. We tested for over-representation using one-tailed hypergeometric distributions. We calculated a False Discovery Rate for each test using the Benjamini-Hochberg method.

Prediction of disease-associated TFBS-disrupting variants

Using bedtools, we intersected TFBSs with genetic variants from the Kaviar database (http://db.systemsbiology.net/kaviar/), restricting our analysis to variants in dbSNP. We included all variants that overlapped a predicted TFBS by at least 1 bp. Next we intersected this table of putative TFBS-disrupting variants with summary statistics from the Psychiatric Genomics Consortium GWAS of schizophrenia[14] (https://www.med.unc.edu/pgc/results-and-downloads). We focused on 108 LD-independent, genome-wide significant loci. We selected TFBS-disrupting variants that were associated with schizophrenia risk at p-value < 1e-4 and located within the extent of LD at each genome-wide significant locus.

Neural Stem Cells

NSC lines were grown in neural expansion media (NEM) supplemented with EGF and FGF-2 (20 ng/ml) (PeproTech) on laminin (Sigma L2020) -coated polystyrene plates and passaged as previously described[82]. Neural expansion media (NEM): NeuroCult NS-A Basal Medium

(Stem Cell Technologies), DMEM/F12, Antibiotic-Antimycotic, GlutaMax, B-27 supplement, N-2 supplement, 1:1000 EGF, and 1:1000 FGF.

Lentivirus Infections

VSV-G pseudotyped, self-inactivating lentivirus was prepared by transfecting 293T cells with 1.5 μg pVSV-G, 3 μg psPAX-2, and 6 μg pRRL lentiviral vectors - either lentiCRISPR or lenti-SFFV-*POU3F2*-Myc-DDK. Lentiviral supernatants were collected 48 hours later and transferred to human neural stem cells dishes (MOI = ~ 1). Positively transduced cells were selected with 0.6 ug/mL puromycin for 3 days. Gene editing was evaluated using the Surveyor® Assay (Integrated DNA Technologies). Overexpression was evaluated using qRT-PCR and Western blotting.

Generation and Analysis of CRISPR-Cas9 mutants

POU3F2 was edited in neural stem cells using a custom guide site targeting the following sequence: ATCGTGCACGCCGAGCCGCCCGG. Genomic DNA was extracted from transduced and puromycin-selected cells (*POU3F2* targeted and non-targeting guide site control) using Thermo genomic DNA purification kit (K0512) and amplified using AccuPrime *Taq* DNA polymerase. The following primers were used to amplify a 431 bp region framing the target site for analysis of editing using the Surveyor® Mutation Detection Kit (IDT #706025).

POU3F2_For: AGAGCGAGAAGGAGGGAGAG

POU3F2_Rev: GTGATCCACTGGTGAGCGTG

Four microliters of PCR product was used for TOPO cloning for sequencing (Thermo #K457501) and transformed into TOP10 cells. 34 colonies were picked, grown up, and plasmid

DNA extracted (Qiagen Miniprep Kit). Inserts were sequenced at GeneWiz, and percentage gene editing in cell population was determined from alignment to reference sequence. Results from these validation experiments are shown in Fig. S5 and Table S10, indicating ~50% editing efficiency.

Western Blot Analysis

We used Western blots and qPCR to ePCR to evaluate POU3F2 overexpression. Cells were harvested following lentiviral infection and selection, washed with PBS, and protein extracted using RIPA buffer on ice for 20 minutes. 30 ul of Nupage LDS buffer was added to 100 ul of sample (in RIPA) and boiled at 80 degrees C for 10 min. 12 microliters was loaded onto 4-12% Bis-Tris gel and run for 1 hour at 180 volts. iBlot transfer system was used to transfer to PVDF membrane, according to manufacturer's instructions. Membrane was cut between 51 and 39 kDa bands indicated by See Blue Plus 2 pre-stained ladder. The following commercial antibodies were used: POU3F2 (Abcam ab137469, 1:1000), FLAG (Sigma #F1804, 1:5000), and GAPDH (Abcam ab37168, 1:5000). A FluorChem was used to image the blots using SuperSignal West Dura chemiluminescent substrate (Thermo #34075). Results are shown in Fig. S6.

qPCR

Total RNA was extracted using Qiagen miRNeasy kit. cDNA was reverse transcribed using the VILO mastermix kit (Thermo #11755050). A multiplex qPCR assay was run in triplicate for each sample, from duplicate RNA extractions of each modified or control NSC cell line. Multiplex reactions contained SsoFast Universal Probes mix, POU3F2 assay (FAM; IDT custom primer/probe assay) and ACTB assay (HEX; IDT #Hs.PT.56a.40703009). Samples were

run on BioRad CFX96 Real-Time system and quantified using the CFX software. Results are shown in Fig. S7.

Neural stem cell differentiation

Protocols for NSC differentiation toward astrocytes and neurons were modified from previously reported methods[83]. Following cell count, 15,000 cells were seeded into each well on 4-well chamber slides (Thermo #154526PK). Cells were treated with either neuronal-differentiation media (NeuroCult NS-A Basal Medium Stem Cell Technologies, 2% B-27 Serum-Free Supplement, 1% GlutaMax, 1% Antibiotic-Antimycotic) or astrocyte differentiation media (DMEM L-glutamine, high glucose, 1% N-2 supplement, 1% GlutaMax, 1% FBS). Control cells were maintained in NEM. Media was changed every 3-4 days over an 18-day differentiation period. After one week, 0.05 mM dibutyryl cAMP was added to neuronal differentiation media for the remaining differentiation.

Immunofluorescence assays

Cells were fixed for 10 minutes in warm 4% paraformaldehyde. Immediately following fixation, standard IF protocol (Cell Signaling Technologies) was conducted with citrate buffer antigen retrieval. Both primary and secondary antibodies were allowed to incubate at 4 degrees overnight. To detect NSC, neuronal, and glial markers we used Nestin (R&D anti-hNestin purified mouse monoclonal IgG #5568P), b-III-tubulin (CST beta3-Tubulin Rabbit mAb #5568P), and GFAP antibodies (CST GFAP (GA5) mouse mAb #3670S), respectively. Secondary antibodies were purchased from Cell Signaling Technologies (Anti-Mouse IgG - Alexa Fluor 647 conjugate #4410, Anti-Rabbit IgG - Alexa Fluor 488 conjugate #4412, and Anti-Rabbit IgG - Alexa Fluor 555 conjugate #4413).

Quantification of IF images for cell proliferation and size analysis

Cells were stained for DAPI (CST 4063S), stem cell marker Nestin (R&D MAB1259), and neuronal marker b-III-tubulin (CST 5568P). Image J software (FIJI) was used for all analyses. Images for quantifying cell size were taken on a Leica DM IRBE with a 40X NA 1.25 Oil Immersion objective at University of Washington's W.M. Keck Microscopy Center. A total area of 0.553 mm^2 per cell line per condition was visualized by combining 18 individual images within the imaging program MetaMorph. Cell size was quantified by tracing and recording the area of a randomly selected subset of cells after immunofluorescence staining using the marker that best captured the entire cell (Nestin for NSC state, b-III-tubulin for neuronal differentiation). Cell size was not quantified for astrocytic differentiation due to poor resolution of cell boundaries. Cell proliferation was quantified by using DAPI to count the number of nuclei within an image of each well that was 2.5% of the total well (4.6 mm^2 sampled of 1.7 cm^2 total well area). Images for quantifying cell size were taken on a Leica DM IRBE with a 10X objective at University of Washington's W.M. Keck Microscopy Center. We calculated the fold change in number of cells by extrapolating the number of nuclei in the imaged area to the number of cells total in each well. Finally, given the seed density, the fold increase was quantified.

Microarrays

Total RNA was extracted using the miRNeasy extraction kit (Qiagen) at three days and ten days post-puromycin selection of lentiviral transduced hNSCs. Two individual RNA extractions were performed from each plate at each time point, and each viral construct was transduced into two plates of cells. Gene expression was quantified using SurePrint G3

Human Gene Expression 8x60K v2 Microarray (#G4851B). Samples overexpressing POU3F2 were compared to samples that were transduced with a control vector that did not contain the *POU3F2* coding sequence.

Gene Expression Analysis.

In Figure 4, we used the intersectoR() function in the projectoR package in the R statistical language (https://github.com/genesofeve/projectoR) to explore gene membership overlaps between the gene clusters defined in the POU3F2 over-expression experiments we performed here and other key public data sets relevant to brain development and neuropsychiatric disease. This uses a hypergeometric test, phyper(), to determine the statistical significance of the number of genes shared across clusters in the different experiments. Correlation between individual genes' expression and cluster averages were determined using the cor.test() function. Enrichment of disease gene DEG lists and genes annotated to particular GO terms was calculated using the geneSetTest() function in the limma package within Bioconductor.

Luciferase Reporter Assay

VRK2 promoter sequences were amplified from the genomic DNA of an individual whose genome was heterozygous at this position based on whole-genome sequencing. Amplified DNA was cloned into the pGL4.10[*luc2*] reporter plasmid (Promega #E6651) upstream of the *luc* transcription start site. Reporter constructs were co-transfected with a pRL-CMV Renilla vector (ratio of 100 ng luciferase: 5 ng renilla control plasmid) into HEK293 cells using Lipofectamine 2000. After 48 hrs, cells were harvested and luciferase and renilla activity was assayed on a Synergy H4 Plate Reader using the Dual-Luciferase® Reporter Assay (Promega #E1910). All reported values were normalized to renilla co-transfection controls. All

experiments were performed with 3 biological replicates per condition in a 24-well plate format. Results are representative of at least 2 independent experiments. Barplots are presented as mean +/- s.d.

5.7 Figures and Tables

Figure 1. Reconstruction of a transcriptional regulatory network (TRN) model for the human brain. (a) The genomic binding sites and target genes for 741 TFs were predicted by integration of digital genomic footprinting with TF-gene coexpression and LASSO regression. In the LASSO regression model, the expression of each gene is predicted based on the expression levels of transcription factors whose binding sites are enriched +/-10kb from the transcription start site. (b) LASSO regression models for each gene were evaluated by comparing predicted vs. observed expression levels (Pearson's r^2) in training and in test sets from the Allen Human Brain Atlas. (c) Predicted regulator interactions among the 741 TFs, and enrichment of each TF's target genes in one of the major brain cell types (FDR < 0.05). Node size indicates out-degree in the TF-to-TF network.

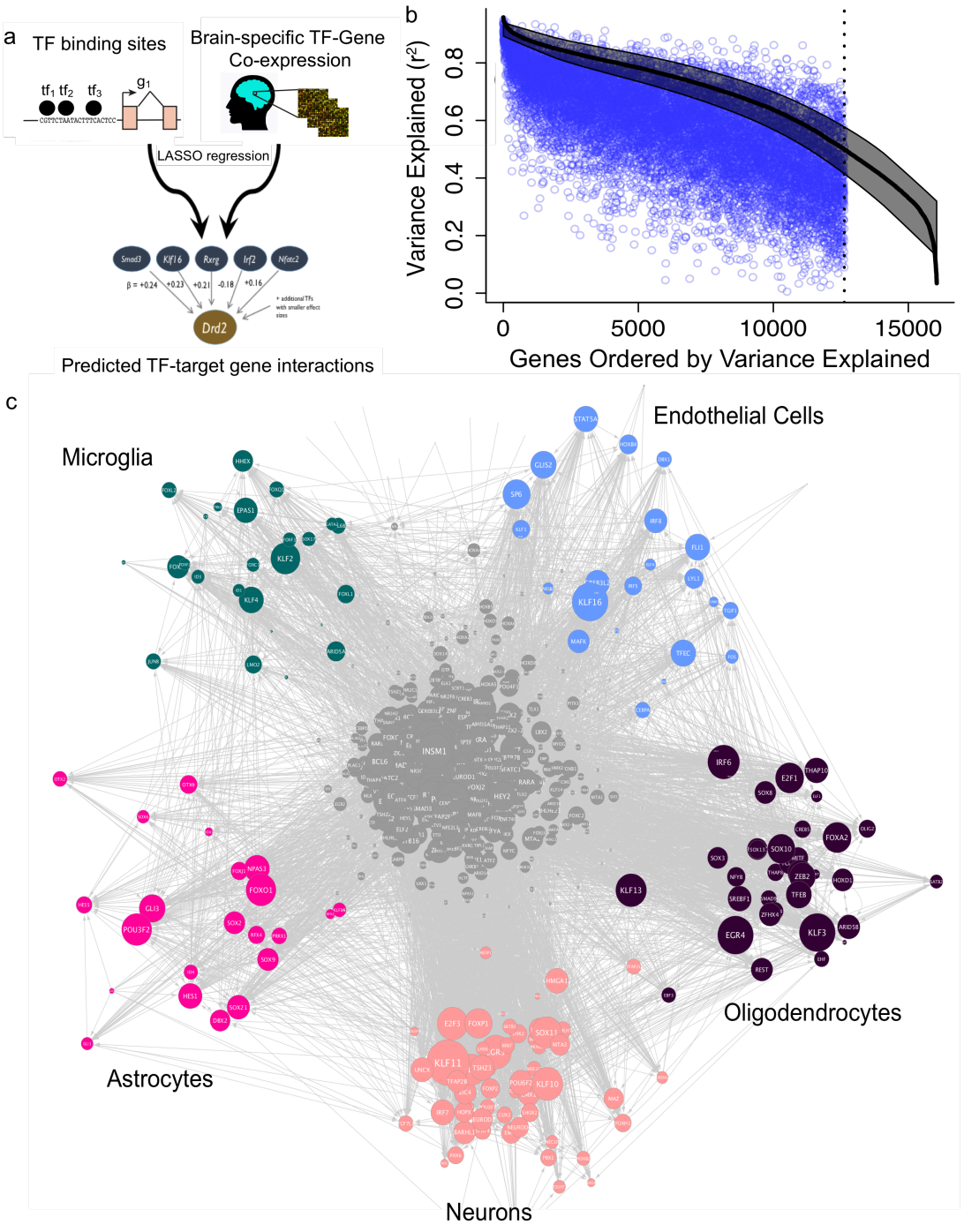

a
TF binding sites
Brain-specific TF-Gene Co-expression
LASSO regression
Predicted TF-target gene interactions
b
Variance Explained (r²)
Genes Ordered by Variance Explained
c
Microglia
Endothelial Cells
Astrocytes
Oligodendrocytes
Neurons

Figure 2. Key regulators of disease-related transcriptional changes in prefrontal cortex. (a) Venn diagram describing overlap of key regulators in schizophrenia (SCZ), bipolar disorder (BD), autism spectrum disorder (ASD), and Alzheimer's disease (AD). (b) –log10(p-values) for genetic associations of TF loci and enrichment of TF's target genes among disease-specific differentially expressed genes for 13 TFs with both a genetic association and a target gene enrichment in one of the four diseases. GWAS p-values are based on studies in the GWAS catalog (Table S8).

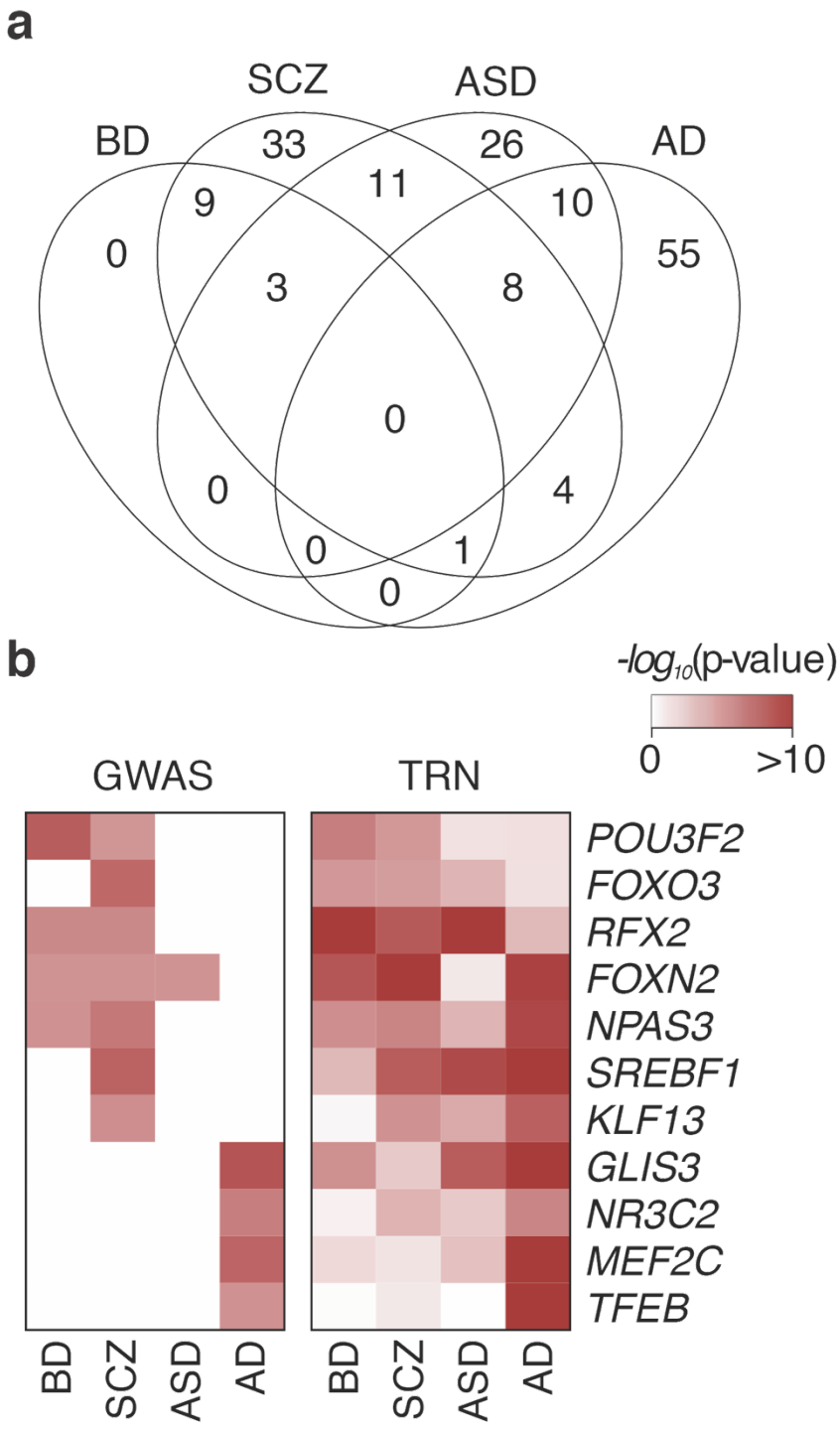

a
SCZ
ASD
BD
AD
33
26
11
9
10
0
55
3
8
0
0
4
0
1
0
b
-log10(p-value)
0
>10
GWAS
TRN
POU3F2
FOXO3
RFX2
FOXN2
NPAS3
SREBF1
KLF13
GLIS3
NR3C2
MEF2C
TFEB
BD
SCZ
ASD
AD
BD
SCZ
ASD
AD

Figure 3. Modulation of a POU3F2 binding site by a schizophrenia-associated SNP in the *VRK2* promoter (a) Region plot of the *VRK2* locus from GWAS of schizophrenia[14]. TRN analysis revealed a single risk-associated SNP at this locus ($p < 1e\text{-}4$) that overlaps a binding site for a TF that is a predicted regulator of VRK2, rs13384219. **(b)** rs13384219 disrupts a key residue in a sequence motif recognized by POU-domain TFs (the Pou2f2 motif is shown). **(c)** *POU3F2* expression is positively correlated with *VRK2* expression in the Allen Human Brain Atlas ($r = 0.67$) **(d)** Dual luciferase reporter assay comparing the activity of a *VRK2* promoter fragment containing either the A or G allele of rs13384219 in HEK293 cell. POU3F2 was overexpressed in combination with transfection of each luciferase construct to assess dose-dependent effects of POU3F2 and interactions with rs13384219 genotype.

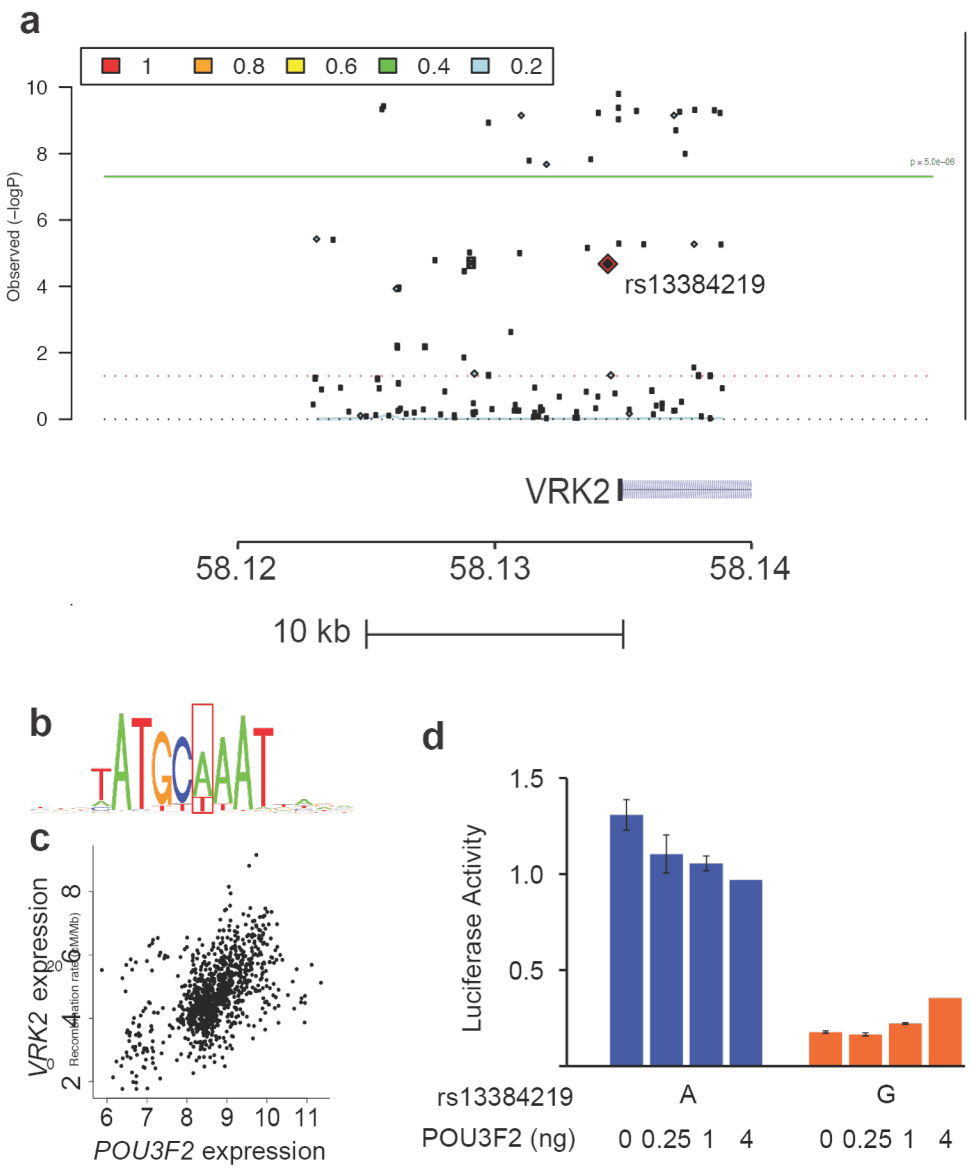
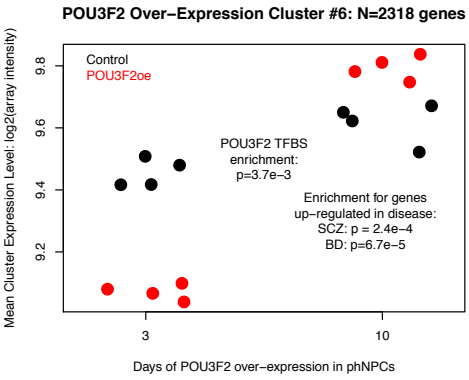
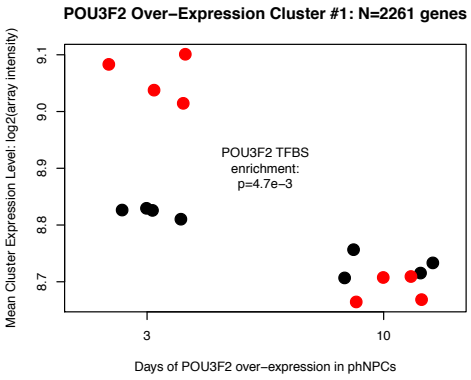
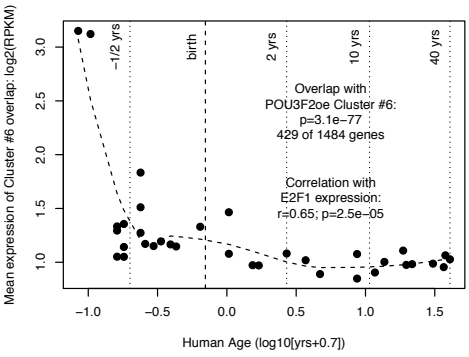
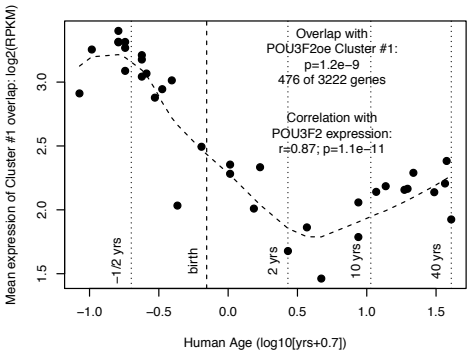

a
1
0.8
0.6
0.4
0.2
Observed (−logP)
rs13384219
VRK2
58.12
58.13
58.14
10 kb
b
c
VRK2 expression
POU3F2 expression
d
Luciferase Activity
rs13384219
A
G
POU3F2 (ng)
0 0.25 1 4
0 0.25 1 4

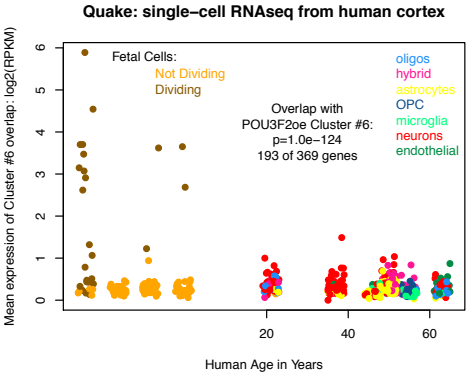
Figure 4. POU3F2 regulates a schizophrenia- and bipolar disorder-related gene network in neural stem cells. We generated gene expression profiles from primary human neural stem cells (phNSCs) after 3 and 10 days of transgenic overexpression of POU3F2 and from control phNSCs. We used K-means clustering to identify POU3F2-responsive gene clusters, then we compared these gene clusters to reference datasets from cortical development. **(a)** Mean cluster expression level (log2 array intensity) from genes in clusters C6 and C1 in POU3F2-overexpressing phNSCs and controls. **(b)** Mean expression (log2 RPKM) of genes in overlapping clusters from RNA-seq of developing dorsolateral prefrontal cortex from BrainSpan[62]. **(c)** Mean expression (log2 RPKM) of genes in overlapping clusters from single-cell RNA-seq of fetal and adult human cortex[64]; single-cell transcriptomes were assigned to brain cell types based on cell type markers and clustering, as described by the study's authors. **(d)** Mean expression (log2 RPKM) of genes in overlapping clusters from *in vitro* differentiation of human ES cells into forebrain neurons.[65]

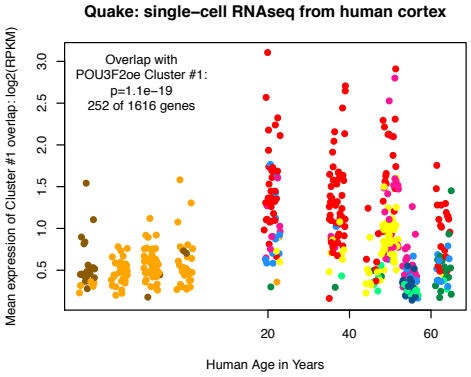

POU3F2 Over−Expression Cluster #6: N=2318 genes
Control
POU3F2oe
POU3F2 TFBS
enrichment:
p=3.7e−3
Enrichment for genes
up−regulated in disease:
SCZ: p = 2.4e−4
BD: p=6.7e−5
Mean Cluster Expression Level: log2(array intensity)
Days of POU3F2 over−expression in phNPCs

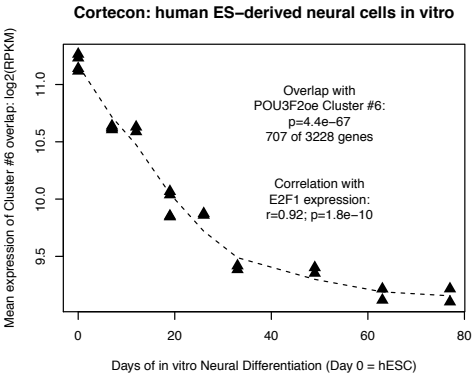

POU3F2 Over−Expression Cluster #1: N=2261 genes
POU3F2 TFBS
enrichment:
p=4.7e−3
Mean Cluster Expression Level: log2(array intensity)
Days of POU3F2 over−expression in phNPCs


BrainSpan: human DLPFC bulk tissue
−1/2 yrs
birth
2 yrs
10 yrs
40 yrs
Overlap with
POU3F2oe Cluster #6:
p=3.1e−77
429 of 1484 genes
Correlation with
E2F1 expression:
r=0.65; p=2.5e−05
Mean expression of Cluster #6 overlap: log2(RPKM)
Human Age (log10[yrs+0.7])


BrainSpan: human DLPFC bulk tissue
Overlap with
POU3F2oe Cluster #1:
p=1.2e−9
476 of 3222 genes
Correlation with
POU3F2 expression:
r=0.87; p=1.1e−11
−1/2 yrs
birth
2 yrs
10 yrs
40 yrs
Mean expression of Cluster #1 overlap: log2(RPKM)
Human Age (log10[yrs+0.7])


Quake: single−cell RNAseq from human cortex
Fetal Cells:
Not Dividing
Dividing
oligos
hybrid
astrocytes
OPC
microglia
neurons
endothelial
Overlap with
POU3F2oe Cluster #6:
p=1.0e−124
193 of 369 genes
Mean expression of Cluster #6 overlap: log2(RPKM)
Human Age in Years


Quake: single−cell RNAseq from human cortex
Overlap with
POU3F2oe Cluster #1:
p=1.1e−19
252 of 1616 genes
Mean expression of Cluster #1 overlap: log2(RPKM)
Human Age in Years


Cortecon: human ES−derived neural cells in vitro
Overlap with
POU3F2oe Cluster #6:
p=4.4e−67
707 of 3228 genes
Correlation with
E2F1 expression:
r=0.92; p=1.8e−10
Mean expression of Cluster #6 overlap: log2(RPKM)
Days of in vitro Neural Differentiation (Day 0 = hESC)

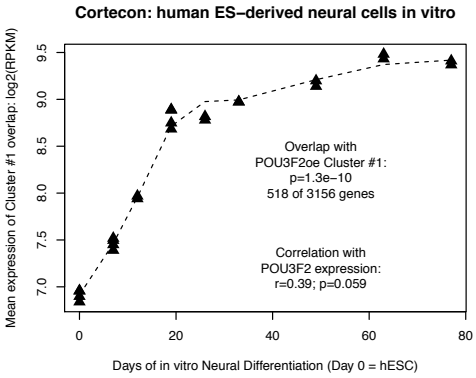

Cortecon: human ES−derived neural cells in vitro
Overlap with
POU3F2oe Cluster #1:
p=1.3e−10
518 of 3156 genes
Correlation with
POU3F2 expression:
r=0.39; p=0.059
Mean expression of Cluster #1 overlap: log2(RPKM)
Days of in vitro Neural Differentiation (Day 0 = hESC)

Figure 5. POU3F2 represses cellular proliferation in primary human neural stem cells. We perturbed POU3F2 expression in primary human neural stem cells (phNSCs) by knocking it out with a CRISPR/Cas9 sgRNA or with transgenic over-expression, then assessed effects on cellular prolifersation. (a) POU3F2 overexpression decreased cellular proliferation in phNSCs to levels comparable to proliferation in cells differentiated toward neurons or astrocytes. POU3F2 sgRNA did not significantly influence proliferation. (b) POU3F2 overexpression increases cell size, both in phNSCs and in cells differentiated toward neurons. (c) Immunofluorescence images of phNSCs and of differentiated neurons, showing effects of POU3F2 overexpression on cell size and morphology.

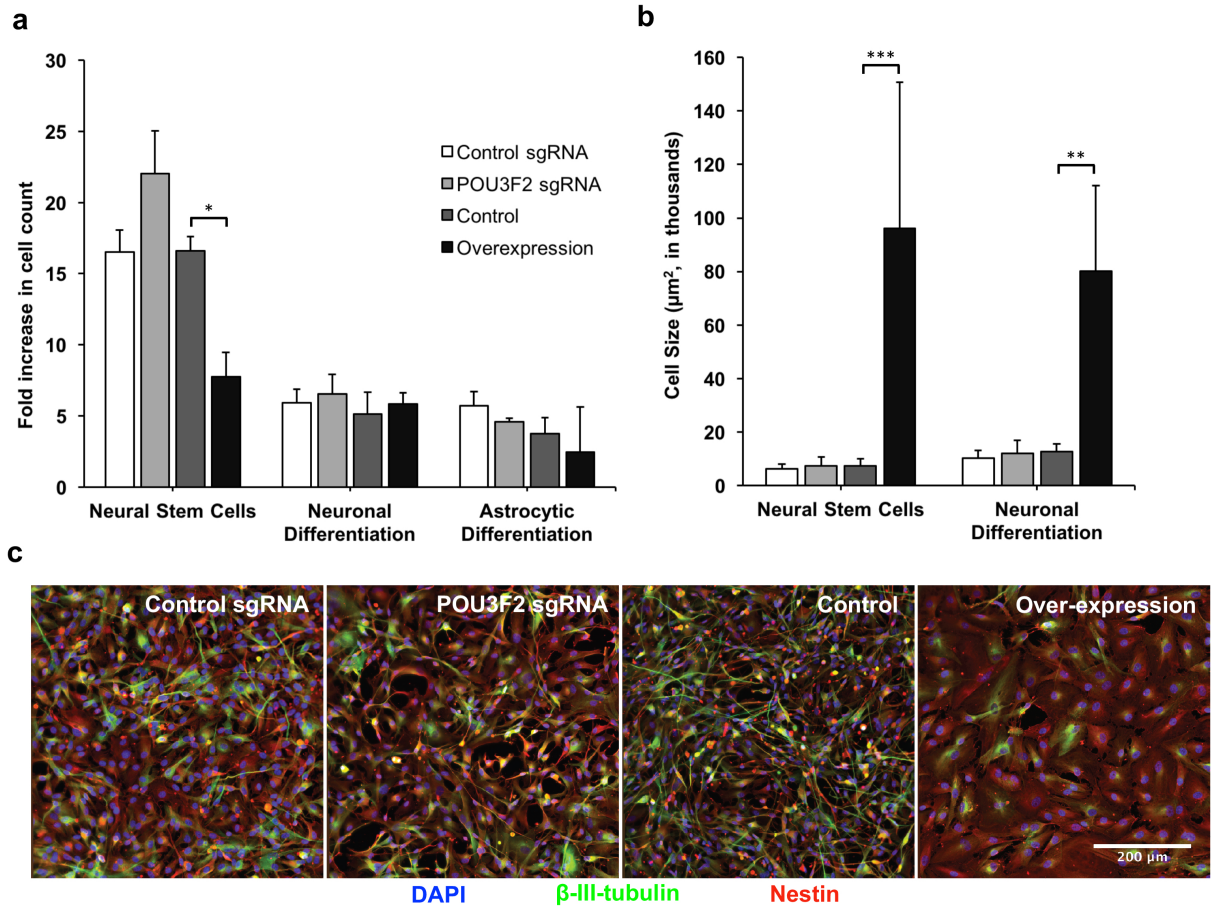


Table 1. Binding sites for key regulator TFs are disrupted by SNPs associated with risk for schizophrenia. Schizophrenia-associated SNPs ($p < 1e-4$) that disrupted TF binding sites in the human brain were identified by scanning SNPs within 108 genome-wide significant schizophrenia risk loci[14] for TFBSs from our TRN model. These loci contained 17 risk-associated, TFBS-disrupting SNPs, including 11 that involve key regulator TFs, shown in Table 1. For each SNP, the table shows the risk locus, as described by Ripke et al. (2014)[14]; the location of the risk-associated, TFBS-disrupting SNP; the p-value for the association of this SNP with schizophrenia, from Ripke et al. (2014)[14]; the key regulator TF whose binding site is predicted to be disrupted by this SNP; the predicted target gene, based on TRN analysis; and supporting evidence from eQTLs, compiled from HaploReg[84] and the CommonMind Consortium[4]. Details of eQTL studies are shown in Table S9).

SCZ Risk Locus	SNP	Postion	SCZ pval	TF	Target Gene	Distance to TSS	TF-gene Correlation (r)	eQTL pvalue
chr2:57943593-58502192	rs13384219	chr2:58134458	2.1E-05	*POU3F2*	*VRK2*	-336	0.67	2.0E-03
chr2:198148577-198835577	rs12618612	chr2:198171058	3.7E-08	*PRRX1*	*ANKRD44*	-4839	0.46	7.2E-18
chr3:135807405-136615405	rs9845788	chr3:135914715	6.0E-07	*SP3*	*MSL2*	-1371	0.49	
chr5:137598121-137948092	rs154069	chr5:137876920	2.3E-07	*NR1H3* *NR2F6*	*ETF1*	-2077	0.64 0.62	1.7E-10
chr5:137598121-137948092	rs11746692	chr5:137946355	4.6E-07	*NR1H2*	*CTNNA1*	-313	0.33	1.1E-04
chr7:104598064-105063064	rs62484724	chr7:105049431	2.1E-05	*POU3F2*	*SRPK2*	9668	0.64	
chr10:104423800-105165583	rs79780963	chr10:104952499	7.5E-16	*SP3*	*NT5C2*	-564	0.46	8.9E-57
chr17:17722402-18030202	rs7359509	chr17:17741875	1.4E-06	*KLF15*	*SREBF1*	1546	0.67	9.4E-91
chr17:17722402-18030202	rs4925119	chr17:17742904	5.9E-06	*KLF15*	*SREBF1*	2575	0.67	9.4E-91
chr17:17722402-18030202	rs6502618	chr17:17746741	2.1E-06	*RFX4*	*SREBF1*	6411	0.50	9.4E-91

5.8 Supplementary Figures and Tables

Figure S1. Graphical Abstract

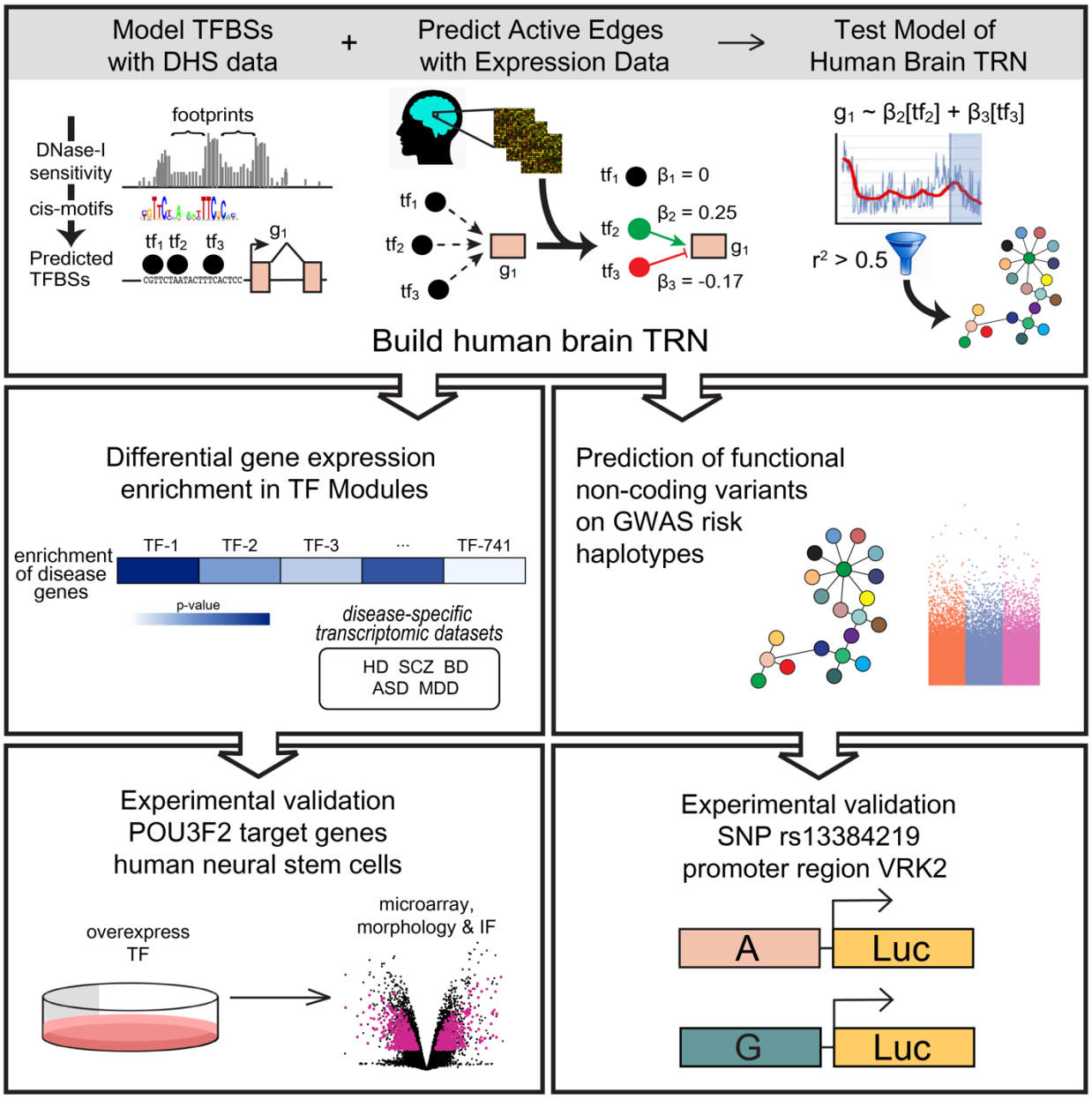


Figure S2. Genetic associations with risk for schizophrenia at the *SREBF1* locus. Region plot was generated with data from Ripke et al. (2014)[14], using Ricopili (https://data.broadinstitute.org/mpg/ricopili/).

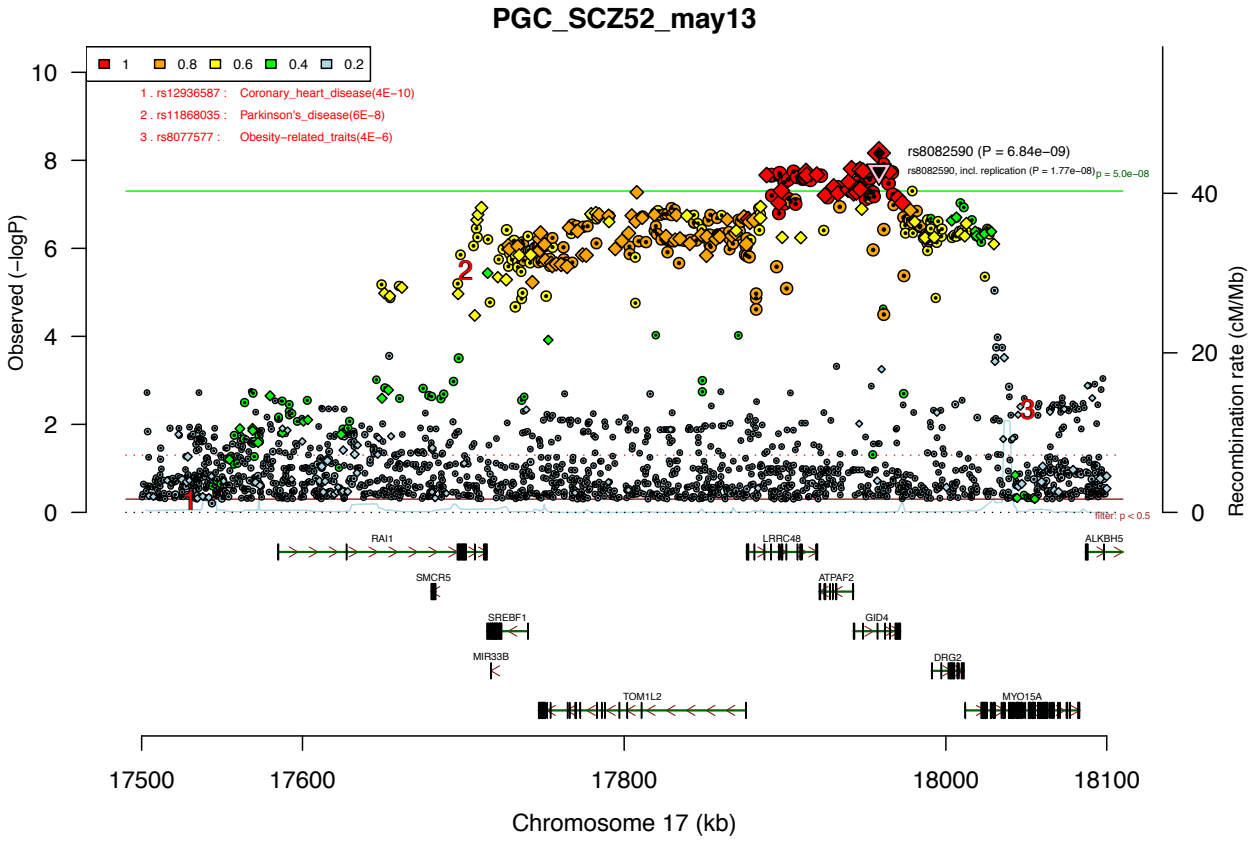


Figure S3. Mean expression of a C6-overlapping cluster in the developing cortex of the non-human primate. We applied k-means clustering to cortical data from the NIH Blueprint Non-Human Primate Brain Atlas[63] (http://www.blueprintnhpatlas.org), identifying a single gene cluster that overlaps cluster C6 from over-expression of POU3F2 in primary human neural stem cells. Figure S3 shows mean expression levels of the genes in this cluster in laser microdissected non-human primate brain tissue samples. Genes in this cluster had very high expression levels in the VZ and SVZ, and low levels in the developing cortical plate. In contrast to the BrainSpan RNA-seq of developing human brain, the non-human primate dataset captures a larger number of early fetal time points with a gradual decline in the expression of the C6-overlapping genes.

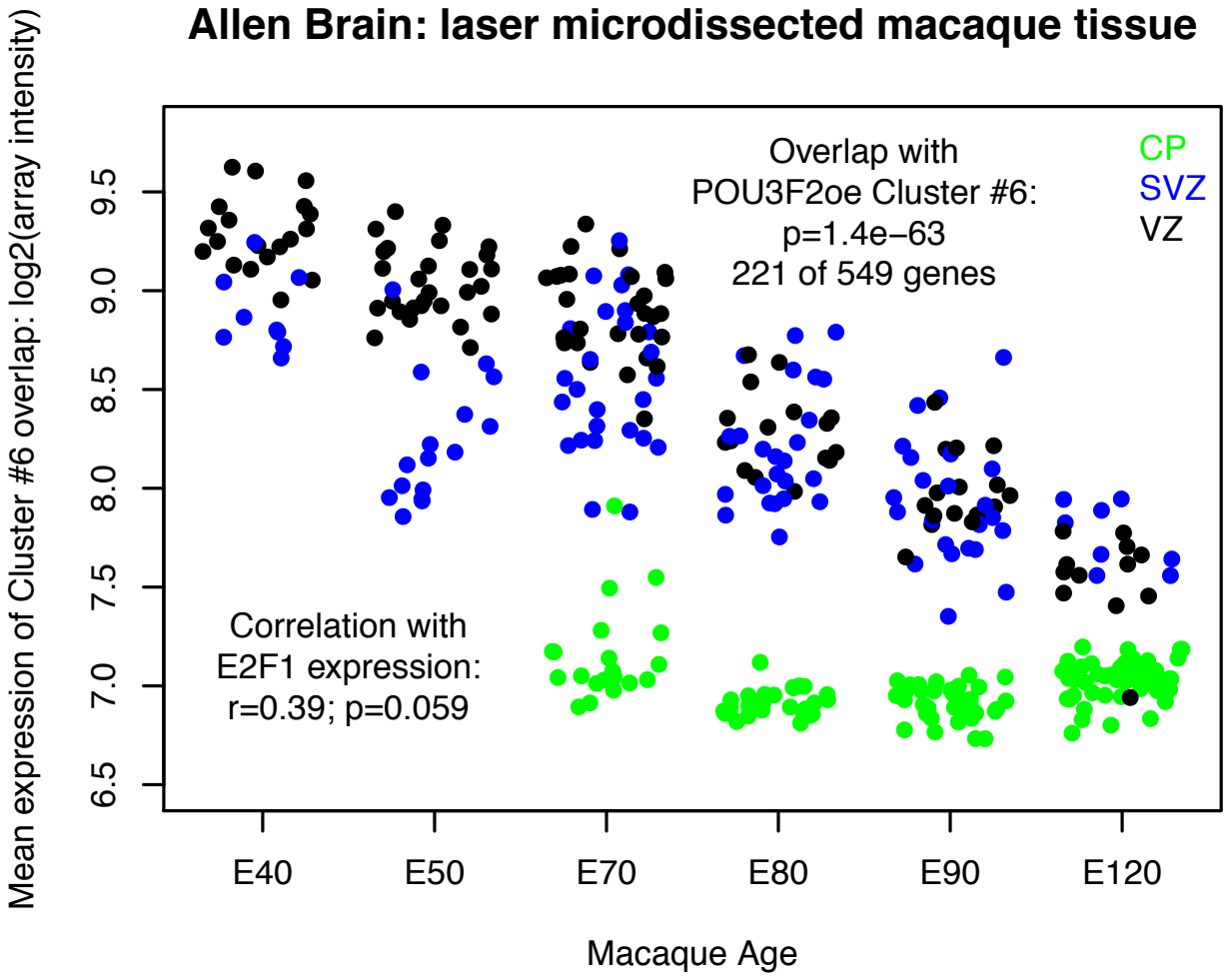
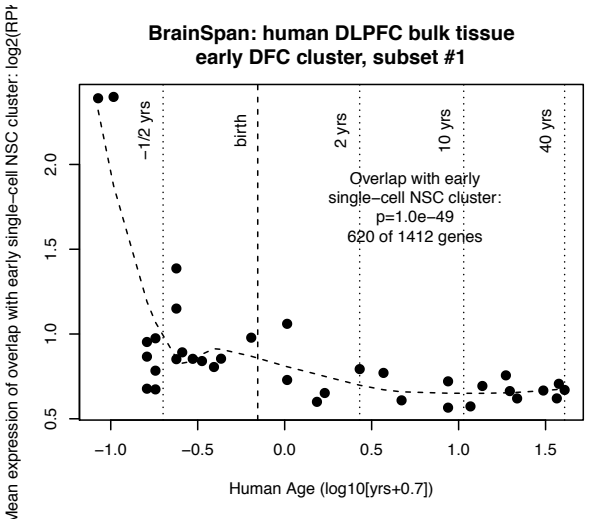
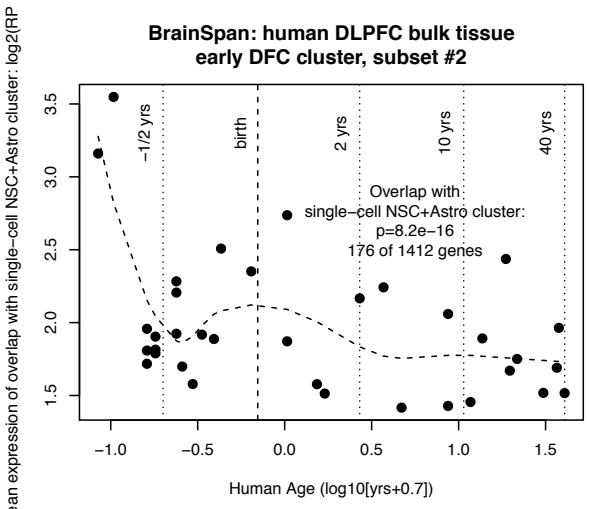
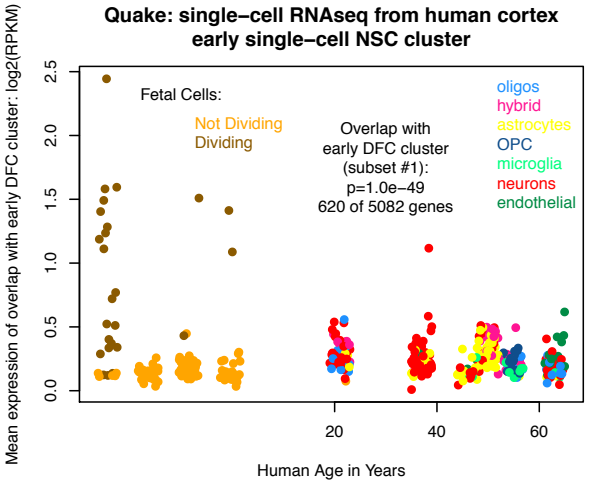
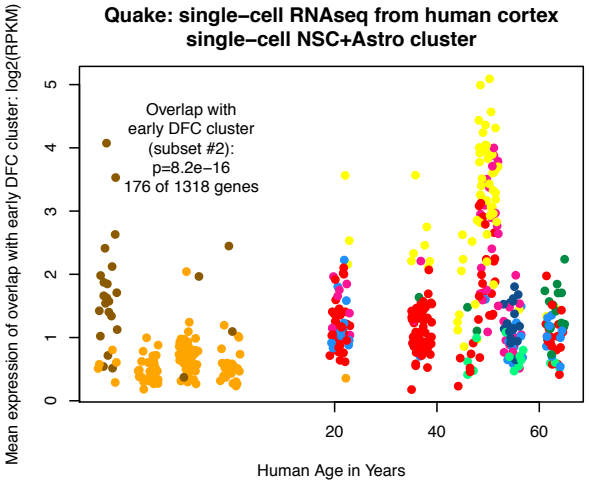


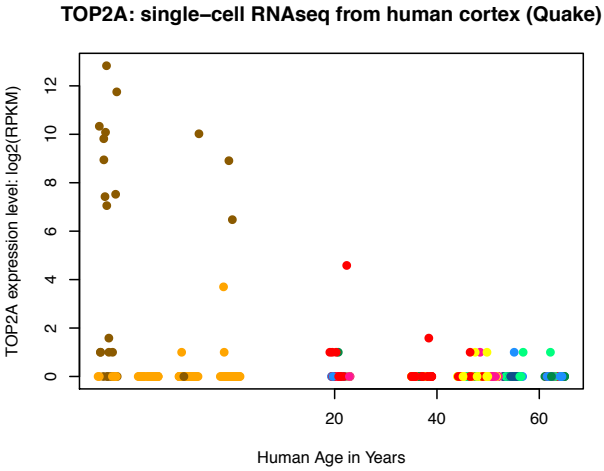
Figure S4. A subset of genes in the C6 cluster are highly expressed in adult astrocytes. Subclustering of the C6-overlapping gene clusters in *in vivo* datasets revealed that subsets of the genes in cluster C6 are expressed both in neural stem cells and in adult astrocytes. The preservation of this gene cluster in adult astrocytes may have allowed us to detect stem cell-enriched POU3F2 target genes while studying adult brain tissue.


BrainSpan: human DLPFC bulk tissue
early DFC cluster, subset #1
Mean expression of overlap with early single−cell NSC cluster: log2(RPK
−1/2 yrs
birth
2 yrs
10 yrs
40 yrs
Overlap with early
single−cell NSC cluster:
p=1.0e−49
620 of 1412 genes
Human Age (log10[yrs+0.7])


BrainSpan: human DLPFC bulk tissue
early DFC cluster, subset #2
Mean expression of overlap with single−cell NSC+Astro cluster: log2(RP
−1/2 yrs
birth
2 yrs
10 yrs
40 yrs
Overlap with
single−cell NSC+Astro cluster:
p=8.2e−16
176 of 1412 genes
Human Age (log10[yrs+0.7])


Quake: single−cell RNAseq from human cortex
early single−cell NSC cluster
Mean expression of overlap with early DFC cluster: log2(RPKM)
Fetal Cells:
Not Dividing
Dividing
Overlap with
early DFC cluster
(subset #1):
p=1.0e−49
620 of 5082 genes
oligos
hybrid
astrocytes
OPC
microglia
neurons
endothelial
Human Age in Years


Quake: single−cell RNAseq from human cortex
single−cell NSC+Astro cluster
Mean expression of overlap with early DFC cluster: log2(RPKM)
Overlap with
early DFC cluster
(subset #2):
p=8.2e−16
176 of 1318 genes
Human Age in Years


TOP2A: single−cell RNAseq from human cortex (Quake)
TOP2A expression level: log2(RPKM)
Human Age in Years

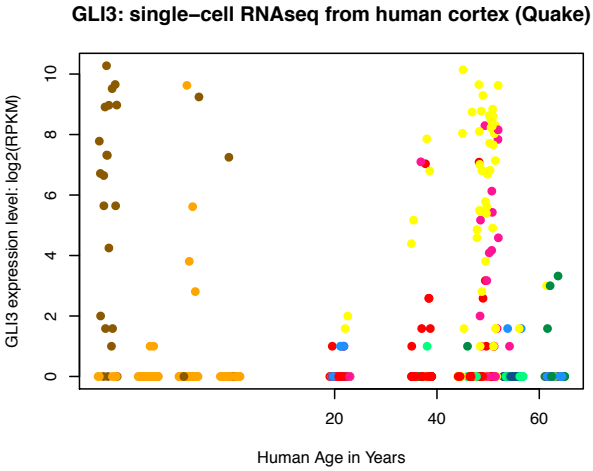

GLI3: single−cell RNAseq from human cortex (Quake)
GLI3 expression level: log2(RPKM)
Human Age in Years

Figure S5. Surveyor Nuclease Assay validation of CRISPR-Cas9 POU3F2 sgRNA from genomic DNA of transduced primary human neural stem cells.

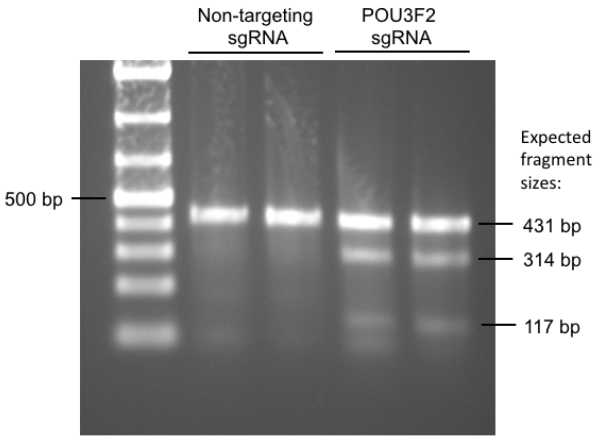


Figure S5. Western blot results from overexpression of FLAG-tagged POU3F2 in phNSCs

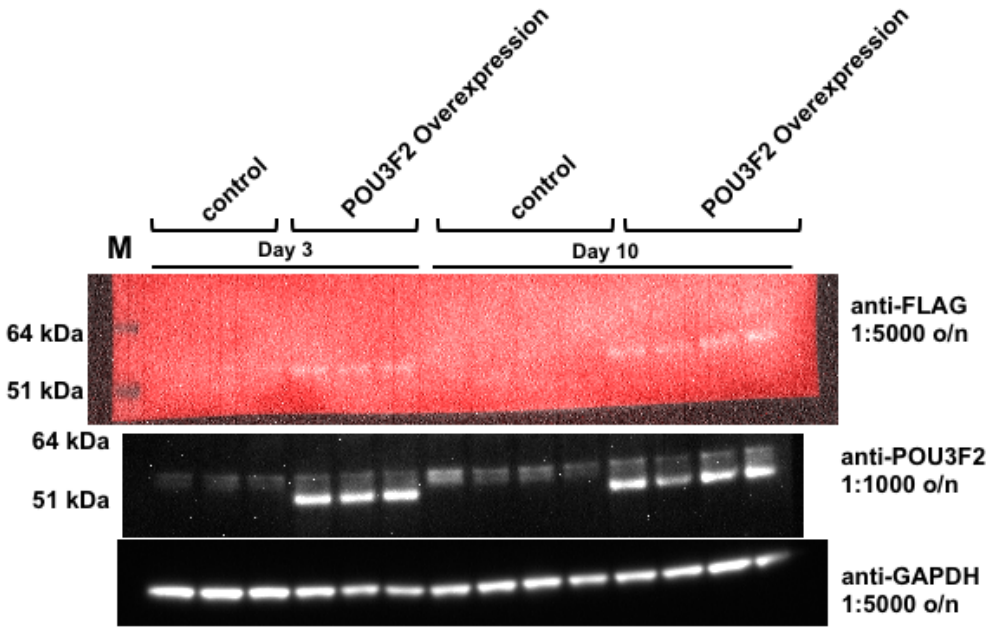


Figure S6. qPCR validation of POU3F2 overexpression in phNSCs

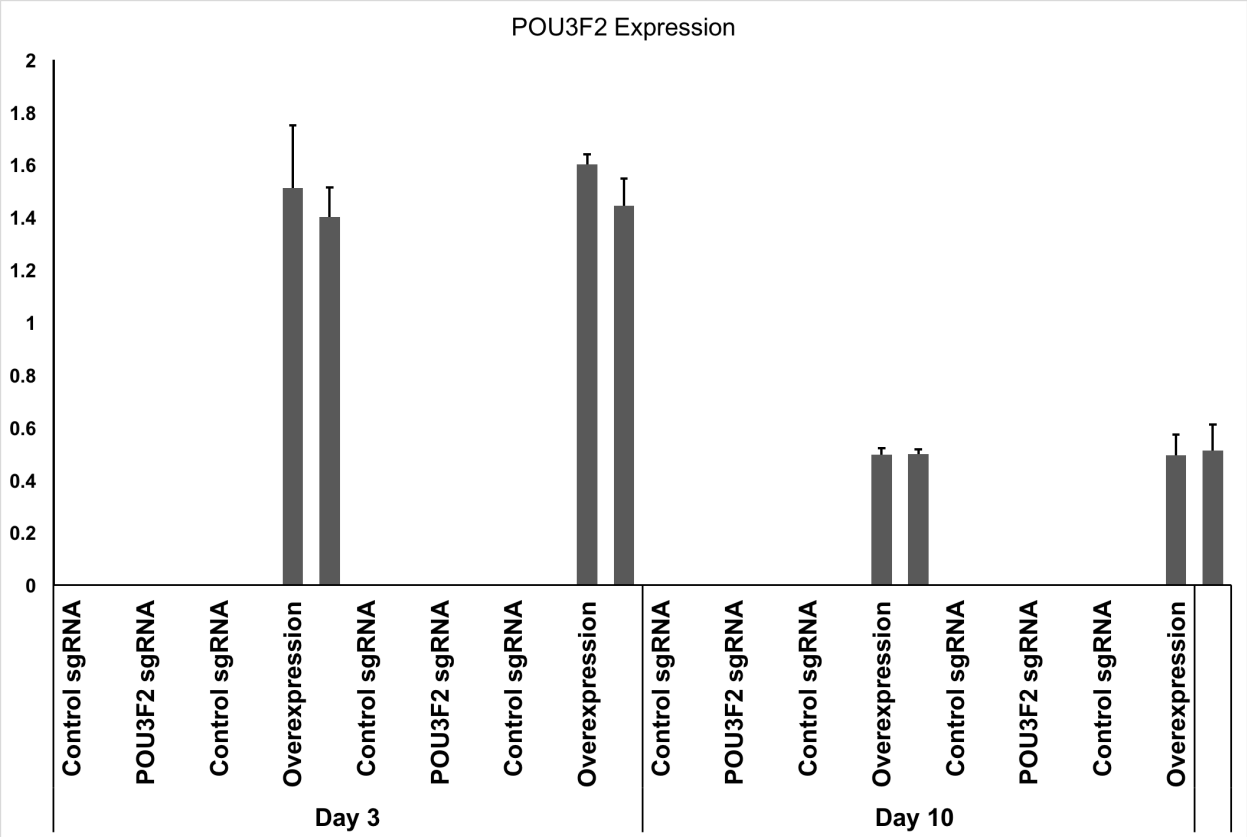


Table S1. Summary statistics from digital genomic footprinting of 15 human brain DNase-seq experiments

Sample ID	Brain Region	mapped reads	DHS regions (min 400 bp)	footprints
ENCSR000EIJ	cerebellum	261386073	166739	51981
ENCSR000EIK	frontal cortex	192478864	53454	11834
ENCSR000EIY	frontal cortex	174008603	43132	8668
ENCSR000ENA	hippocampus	543478591	180260	109701
ENCSR000ENC	cerebellum	98313605	128819	21062
ENCSR000ENE	brain microvascular endothelial cell	97621556	140888	14582
ENCSR000ENF	brain pericyte	56208093	137903	8546
ENCSR000ENL	choroid plexus epithelial cell	376999221	179328	138085
ENCSR224IYD	medulla oblongata	1004733648	227548	98295
ENCSR318PRQ	middle frontal gyrus	593844336	469949	87964
ENCSR475VQD	fetal brain	486381162	218277	76210
ENCSR503HIB	cerebellar cortex	680131240	129976	88658
ENCSR595CSH	fetal brain	719039696	166037	120623
ENCSR706IDL	midbrain	1021179488	300389	249300
ENCSR771DAX	globus pallidus	484128221	94788	36161

Table S2. Enrichments of each TF's target genes for cell type-specific genes. OPC = oligodendrocyte precursor cell; NFO = newly formed oligodendrocyte; MO = myelinating oligodendrocyte. Values are –log10(p-values) for enrichment of each transcription factor's target genes among cell type-specific genes (Fisher's exact test). Cell type-specific genes were derived by applying the pSI algorithm[38] to cell type-specific transcriptome data from Zhang et al. (2014)[37].

Regulator	Astrocytes	Neuron	OPC	NFO	MO	Microglia	Endothelial	Assignment
ARID5A	1.0	0.2	2.4	1.6	0.2	1.6	5.5	Endothelial.Cells
BARHL1	0.9	4.6	3.2	1.5	0.4	0.3	0.6	Neuron
BATF2	6.7	0.0	0.6	1.3	0.2	2.1	7.5	Endothelial.Cells
BCL6B	1.2	0.1	0.8	0.4	0.0	0.6	4.7	Endothelial.Cells
CEBPA	2.0	0.2	1.7	0.1	0.1	9.9	0.4	Microglia
CREB5	0.2	0.0	0.1	3.5	25.0	0.6	2.2	Myelinating.Oligodendrocytes
CUX2	0.7	6.9	1.9	0.8	0.0	0.0	0.7	Neuron
DBX1	0.0	0.1	0.6	0.3	0.3	4.5	0.6	Microglia
DBX2	12.7	0.1	1.3	1.1	0.8	1.8	1.8	Astrocytes
DLX5	0.8	5.7	0.4	0.1	0.4	0.0	0.4	Neuron
DLX6	0.5	6.5	2.8	0.7	0.8	0.0	0.2	Neuron
DMRT2	0.1	0.0	0.2	3.0	9.1	0.3	0.3	Myelinating.Oligodendrocytes
DMRTC2	0.0	0.0	0.1	3.3	6.3	1.0	1.0	Myelinating.Oligodendrocytes
DRGX	2.5	5.7	0.9	0.3	0.0	0.1	0.6	Neuron
E2F1	0.1	1.2	0.4	0.3	6.9	0.2	0.1	Myelinating.Oligodendrocytes
E2F3	0.1	7.2	1.3	0.2	0.0	0.0	0.1	Neuron
EGR3	0.1	9.6	0.6	0.5	0.2	1.5	1.1	Neuron
EGR4	0.2	3.3	0.0	1.6	5.1	3.1	0.2	Myelinating.Oligodendrocytes
EHF	0.1	0.0	0.0	1.4	4.4	0.6	0.0	Myelinating.Oligodendrocytes
ELF1	0.0	0.0	0.1	1.1	9.4	1.5	0.4	Myelinating.Oligodendrocytes
ELF4	0.6	0.0	0.1	0.0	0.4	8.0	0.4	Microglia
EMX1	0.5	4.7	1.5	1.1	0.2	1.4	0.1	Neuron
EPAS1	2.6	0.1	0.7	0.2	0.4	2.3	8.0	Endothelial.Cells
FLI1	0.0	0.0	0.2	0.1	0.5	5.1	4.5	Microglia
FOS	0.9	0.0	0.5	0.2	0.6	10.2	1.8	Microglia
FOXA2	0.1	0.3	0.4	4.9	9.9	0.4	1.9	Myelinating.Oligodendrocytes
FOXC1	0.5	0.0	0.7	0.1	0.4	6.1	37.9	Endothelial.Cells
FOXD4	3.3	0.0	0.8	0.3	0.1	5.0	1.0	Microglia
FOXF1	0.4	0.3	0.2	0.3	0.5	1.2	8.3	Endothelial.Cells
FOXF2	0.6	0.0	2.6	0.8	0.2	2.3	22.2	Endothelial.Cells
FOXH1	0.3	13.5	0.5	0.4	0.0	0.0	0.0	Neuron
FOXJ1	9.8	2.1	2.2	0.1	0.0	2.9	3.2	Astrocytes
FOXL1	0.6	0.3	0.4	0.6	0.1	1.9	5.6	Endothelial.Cells
FOXL2	0.6	0.1	0.3	0.4	0.9	1.6	13.7	Endothelial.Cells
FOXN2	0.3	0.2	0.1	1.2	4.4	0.3	0.1	Myelinating.Oligodendrocytes
FOXO1	4.5	1.9	4.2	1.1	0.9	0.6	3.1	Astrocytes
FOXO4	0.7	0.0	0.4	3.1	16.1	1.1	0.7	Myelinating.Oligodendrocytes
FOXP2	3.6	10.1	2.3	2.5	0.8	0.0	0.5	Neuron
FOXQ1	1.5	0.4	0.3	0.6	0.6	0.2	12.3	Endothelial.Cells
FOXS1	1.0	0.0	1.1	0.0	0.0	4.2	5.6	Endothelial.Cells
GATA2	0.1	0.0	0.1	1.0	3.7	4.0	5.2	Endothelial.Cells
GLI1	4.3	2.9	1.1	0.2	2.3	0.5	0.1	Astrocytes
GLI2	8.1	1.4	0.1	0.1	0.4	0.1	0.1	Astrocytes
GLI3	5.1	0.5	0.5	0.2	0.1	1.3	0.6	Astrocytes
GLIS2	0.3	0.6	0.5	0.1	1.0	5.2	0.8	Microglia
HES1	8.1	0.0	0.6	0.0	0.0	4.7	3.1	Astrocytes
HES5	6.8	6.6	1.7	0.2	0.2	0.5	0.2	Astrocytes
HHEX	1.8	0.0	0.1	0.4	0.8	10.1	11.6	Endothelial.Cells
HIC1	0.3	0.2	0.2	0.0	0.6	0.1	8.9	Endothelial.Cells
HMGA1	0.6	5.0	0.7	0.2	0.0	0.0	0.0	Neuron
HOMEZ	0.3	0.0	0.0	0.4	5.5	1.1	0.3	Myelinating.Oligodendrocytes
HOPX	0.5	8.8	1.2	0.8	0.0	0.1	0.2	Neuron
HOXB2	2.9	5.9	2.2	0.3	0.0	0.3	0.7	Neuron
HOXB4	1.4	0.7	0.3	0.1	0.0	4.8	0.4	Microglia
HOXC6	2.0	6.2	3.2	0.2	0.0	0.0	0.2	Neuron
HOXD1	0.6	0.1	0.3	6.3	17.8	0.4	1.8	Myelinating.Oligodendrocytes
HSFY1	0.1	18.4	4.2	1.0	0.2	0.1	0.1	Neuron

ID1	0.4	0.1	0.7	0.7	0.8	1.8	6.6	Endothelial.Cells
ID3	0.7	0.0	0.4	0.1	1.5	3.7	5.9	Endothelial.Cells
ID4	6.0	1.0	0.6	0.6	0.3	1.1	1.2	Astrocytes
IKZF1	0.1	0.2	0.5	0.2	2.6	12.7	2.6	Microglia
IRF5	0.6	0.0	0.3	0.2	0.0	18.9	2.6	Microglia
IRF6	1.5	0.9	3.1	4.3	0.2	2.6	2.5	Newly.Formed.Oligodendrocyte
IRF7	1.7	5.2	0.8	1.5	0.0	1.6	0.5	Neuron
IRF8	0.6	0.0	0.5	0.3	0.0	13.0	4.4	Microglia
JUNB	1.1	0.2	1.3	0.1	0.0	4.0	4.7	Endothelial.Cells
KLF1	0.2	0.0	0.7	0.5	0.1	4.2	0.0	Microglia
KLF11	1.3	4.9	1.0	0.1	0.0	3.1	1.0	Neuron
KLF15	10.4	0.0	0.2	0.2	5.1	1.5	0.5	Astrocytes
KLF16	0.1	0.0	0.0	0.1	0.1	4.6	0.2	Microglia
KLF2	0.4	0.0	0.2	1.2	0.9	7.9	13.6	Endothelial.Cells
KLF3	0.1	0.0	0.1	1.7	9.0	4.8	0.6	Myelinating.Oligodendrocytes
KLF4	1.3	0.5	0.2	0.2	0.0	4.5	9.0	Endothelial.Cells
KLF5	1.5	6.2	1.0	0.7	0.6	1.0	0.1	Neuron
LHX6	2.4	8.1	0.4	0.4	0.0	0.0	0.4	Neuron
LMO2	0.1	0.0	0.2	0.0	0.1	5.0	7.1	Endothelial.Cells
LYL1	0.0	0.0	2.9	1.3	1.6	18.9	4.5	Microglia
MAFF	2.5	0.1	0.2	0.0	0.0	1.7	4.9	Endothelial.Cells
MAFG	0.0	4.8	0.8	0.9	0.1	0.0	0.0	Neuron
MAFK	0.0	0.0	1.8	0.4	2.3	6.1	1.5	Microglia
MAZ	0.0	8.0	0.0	0.2	0.0	0.1	0.0	Neuron
MEF2A	0.5	8.8	2.4	0.9	0.1	0.0	0.0	Neuron
MEF2C	0.9	11.3	2.6	0.9	0.1	0.0	0.5	Neuron
MESP1	0.0	8.5	0.2	0.1	0.1	0.0	0.1	Neuron
MITF	0.0	0.0	0.6	1.9	8.6	2.5	1.4	Myelinating.Oligodendrocytes
MNT	0.2	6.6	1.1	0.1	0.0	0.6	0.0	Neuron
MSC	0.9	4.5	1.1	0.4	0.3	0.1	0.0	Neuron
MTA3	0.1	4.4	0.8	0.4	0.0	0.0	0.0	Neuron
NEUROD2	1.6	4.6	0.4	0.3	0.0	0.6	1.6	Neuron
NEUROD6	0.9	6.0	2.2	0.7	0.1	1.1	1.0	Neuron
NFKB2	0.4	0.1	0.5	0.3	0.2	6.1	0.7	Microglia
NHLH2	0.5	4.4	1.6	0.2	0.0	0.0	2.5	Neuron
NPAS3	8.5	0.2	1.3	0.8	1.2	0.2	0.0	Astrocytes
OLIG2	0.2	0.2	0.4	1.9	5.4	1.6	0.1	Myelinating.Oligodendrocytes
ONECUT1	0.9	5.1	1.1	0.0	0.0	0.0	0.4	Neuron
OTX1	7.5	4.7	3.9	1.0	0.2	0.6	0.1	Astrocytes
OTX2	4.5	4.1	0.5	0.8	0.3	0.1	0.6	Astrocytes
PAX3	1.9	4.8	1.2	1.0	0.2	0.0	0.3	Neuron
PBX1	1.5	6.1	0.7	0.2	0.0	0.0	0.0	Neuron
POU3F1	6.7	9.6	8.8	0.9	0.1	0.0	1.6	Neuron
POU3F2	11.5	1.3	5.4	2.0	1.2	0.0	4.0	Astrocytes
POU3F4	8.2	8.5	5.4	0.8	0.2	0.0	0.5	Neuron
POU4F2	0.3	4.4	1.4	0.2	0.3	0.1	1.4	Neuron
POU6F2	5.0	6.5	4.7	1.6	0.0	0.0	0.0	Neuron
PPARG	0.7	6.7	0.0	0.3	0.6	0.2	0.5	Neuron
PRRX1	7.7	0.3	4.0	0.8	0.0	1.5	1.4	Astrocytes
REST	1.7	3.5	1.3	2.3	5.5	1.4	1.6	Myelinating.Oligodendrocytes
RFX2	7.3	1.0	0.1	0.1	0.1	1.2	1.1	Astrocytes
RFX4	14.4	0.2	1.4	0.9	0.1	0.6	1.4	Astrocytes
RUNX1	0.8	0.0	0.5	0.1	0.2	16.2	8.8	Microglia
SHOX2	2.5	5.6	1.7	1.0	0.2	0.0	0.5	Neuron
SMAD9	0.1	0.0	0.1	0.6	13.6	3.0	1.2	Myelinating.Oligodendrocytes
SOHLH1	0.0	4.2	0.1	0.0	0.2	0.6	0.0	Neuron
SOX1	10.1	0.8	1.3	0.0	0.6	0.0	0.3	Astrocytes
SOX10	0.1	0.0	1.4	11.2	30.9	1.2	2.0	Myelinating.Oligodendrocytes
SOX11	1.0	5.1	2.8	1.3	0.1	1.5	0.4	Neuron
SOX13	0.8	0.0	0.9	5.4	3.6	2.0	1.6	Newly.Formed.Oligodendrocyte
SOX17	0.4	0.0	0.1	0.3	1.8	3.7	12.3	Endothelial.Cells
SOX18	0.2	0.0	0.7	0.1	2.8	2.2	5.8	Endothelial.Cells
SOX2	21.3	0.0	1.9	0.8	0.9	1.7	2.5	Astrocytes
SOX21	9.5	0.2	4.3	1.3	0.1	0.2	0.2	Astrocytes
SOX3	1.7	0.0	4.5	2.5	2.3	1.1	0.5	Oligodendrocyte.Precursor.Cell
SOX6	9.4	0.2	5.5	0.6	0.0	0.2	1.1	Astrocytes
SOX7	1.1	0.5	0.0	0.4	0.4	0.7	10.2	Endothelial.Cells
SOX8	0.2	0.0	0.3	11.1	21.6	0.4	0.9	Myelinating.Oligodendrocytes

SOX9	37.1	0.0	0.2	0.0	0.0	0.3	2.9	Astrocytes
SPI1	0.1	0.0	1.0	0.0	0.0	21.4	0.5	Microglia
SPIB	0.0	0.5	0.2	0.5	0.2	4.7	0.1	Microglia
SREBF1	1.5	0.0	1.0	0.6	7.5	3.3	0.4	Myelinating.Oligodendrocytes
STAT4	1.1	4.2	2.8	1.7	0.8	0.8	0.7	Neuron
STAT5A	0.2	0.4	0.4	0.0	0.1	6.4	0.7	Microglia
TBX2	0.0	0.2	0.6	0.1	0.2	0.6	8.8	Endothelial.Cells
TBX3	0.0	0.0	0.4	0.1	0.5	2.4	11.5	Endothelial.Cells
TCF21	0.5	4.5	0.4	1.3	1.2	0.0	0.0	Neuron
TCF7L2	2.2	5.2	2.6	0.4	0.9	0.6	0.3	Neuron
TCFL5	0.0	0.0	0.1	2.5	7.5	0.8	0.4	Myelinating.Oligodendrocytes
TFAP2B	0.4	4.3	1.6	1.3	0.7	1.4	0.2	Neuron
TFEB	0.0	0.0	0.0	2.1	15.9	3.1	0.0	Myelinating.Oligodendrocytes
TFEC	0.3	0.0	0.4	0.0	0.7	9.9	0.6	Microglia
TGIF1	0.7	1.8	0.5	0.5	0.1	5.2	0.6	Microglia
THAP10	0.0	0.5	0.1	4.3	2.9	0.3	0.5	Newly.Formed.Oligodendrocyte
THAP8	0.0	0.0	0.1	1.3	6.3	1.2	0.0	Myelinating.Oligodendrocytes
TSHZ3	0.4	4.6	0.2	0.4	0.2	0.3	1.7	Neuron
UNCX	2.6	4.2	1.9	3.1	0.1	0.3	1.3	Neuron
ZEB2	0.0	0.0	0.3	7.2	20.6	0.5	0.6	Myelinating.Oligodendrocytes
ZFHX4	1.4	0.1	0.4	2.2	7.5	1.6	0.4	Myelinating.Oligodendrocytes
ZIC4	0.3	4.8	1.6	0.9	0.2	0.0	0.6	Neuron

Table S3. Post-mortem prefrontal cortex microarray and RNA-seq datasets used for key regulator analysis

Reference	Accession	SCZ	BD	MDD	ASD	AD	Controls
25796564	**GSE53987**	15	17	17	0	0	19
Thomas	**GSE21138**	30	0	0	0	0	29
CMC	**syn5607603**	258	0	0	0	0	279
McMahon	GSE53239	0	10	0	0	0	11
Sibille	GSE54567/GSE54568	0	0	29	0	0	29
Voineagu	GSE28521	0	0	0	16	0	16
Parikshak	Parikshak et al. 2016, Table S3	0	0	0	48	0	49
Schadt-1	GSE33000	0	0	0	0	310	157
Schadt-2	GSE84422	0	0	0	0	17	16

Table S4. Key regulators of prefrontal cortex gene expression changes in bipolar disorder cases vs. controls.

TF	p.up GSE53239	p.up GSE53987	meta-p up	q up	p.down GSE53239	p.down GSE53987	meta-p down	q down
SOX9	2.1E-24	1.5E-04	2.0E-26	1.5E-23	9.0E-01	1.0E+00	9.9E-01	1.0E+00
SOX2	8.1E-17	3.6E-06	1.5E-20	1.1E-17	3.4E-01	1.0E+00	7.0E-01	1.0E+00
FOXJ1	1.6E-15	2.8E-07	2.3E-20	1.7E-17	9.6E-01	8.9E-01	9.9E-01	1.0E+00
FOXO1	7.9E-07	4.4E-10	1.3E-14	9.3E-12	8.4E-01	9.6E-01	9.8E-01	1.0E+00
PRRX1	5.5E-11	8.7E-06	1.7E-14	1.3E-11	2.4E-01	9.3E-01	5.6E-01	1.0E+00
HMBOX1	2.1E-03	1.0E-08	5.6E-10	4.1E-07	7.5E-01	9.1E-01	9.4E-01	1.0E+00
FOXN2	2.0E-02	7.4E-09	3.6E-09	2.6E-06	9.9E-01	9.7E-01	1.0E+00	1.0E+00
POU3F4	1.5E-06	1.4E-03	4.6E-08	3.3E-05	9.1E-01	8.0E-01	9.6E-01	1.0E+00
POU3F2	6.4E-06	4.4E-03	5.2E-07	3.7E-04	4.7E-01	3.0E-01	4.2E-01	1.0E+00
IRF9	6.2E-04	8.9E-05	9.7E-07	7.0E-04	4.8E-01	9.6E-01	8.2E-01	1.0E+00
FOXN3	2.8E-02	4.9E-06	2.3E-06	1.6E-03	8.5E-01	9.6E-01	9.8E-01	1.0E+00
NPAS3	2.5E-05	8.7E-03	3.6E-06	2.6E-03	7.9E-01	9.2E-01	9.6E-01	1.0E+00
RUNX1	2.7E-03	6.3E-04	2.4E-05	1.7E-02	3.8E-02	1.6E-01	3.8E-02	1.0E+00

Table S5. Key regulators of prefrontal cortex gene expression changes in schizophrenia cases vs. controls.

TF	p.up GSE53987	p.up GSE21138	p.up CMC	meta-p up	q up	p.down GSE53987	p.down GSE21138	p.down CMC	meta-p down	q down
SOX9	6.4E-33	2.9E-24	6.4E-01	9.9E-53	7.4E-50	9.8E-01	8.7E-01	8.6E-01	1.0E+00	1.0E+00
SOX2	3.7E-23	6.7E-14	9.5E-01	8.2E-33	6.1E-30	1.0E+00	1.0E+00	7.8E-01	1.0E+00	1.0E+00
KLF15	2.4E-07	2.2E-14	1.0E+00	5.8E-18	4.3E-15	1.0E+00	1.0E+00	1.8E-02	2.4E-01	1.0E+00
FOXO1	1.1E-12	1.6E-08	4.6E-01	8.9E-18	6.6E-15	1.0E+00	2.1E-01	1.0E+00	8.0E-01	1.0E+00
HES1	1.2E-07	1.9E-11	1.0E+00	2.0E-15	1.5E-12	9.9E-01	1.0E+00	2.8E-01	8.6E-01	1.0E+00
FOXN2	3.5E-08	3.4E-07	6.1E-03	5.2E-14	3.9E-11	8.9E-01	9.9E-01	1.0E+00	1.0E+00	1.0E+00
MAZ	1.0E+00	1.0E+00	9.8E-01	1.0E+00	1.0E+00	1.0E-10	8.7E-01	1.7E-06	1.1E-13	8.0E-11
FOXJ1	7.9E-08	4.9E-09	4.3E-01	1.2E-13	8.7E-11	8.4E-01	2.3E-02	1.0E+00	2.5E-01	1.0E+00
RFX4	1.2E-04	1.1E-11	9.1E-01	7.4E-13	5.5E-10	1.0E+00	7.6E-01	1.5E-02	1.8E-01	1.0E+00
PRRX1	6.5E-09	2.6E-07	1.0E+00	1.0E-12	7.5E-10	8.7E-01	9.0E-01	9.4E-01	1.0E+00	1.0E+00
STAT5B	1.9E-05	9.5E-04	2.2E-07	2.3E-12	1.7E-09	4.4E-01	5.9E-01	9.9E-01	8.4E-01	1.0E+00
ADNP	2.9E-05	2.0E-04	7.4E-07	2.5E-12	1.8E-09	9.9E-01	7.2E-01	1.0E+00	9.9E-01	1.0E+00
ATF2	9.4E-01	1.0E+00	6.1E-02	4.5E-01	1.0E+00	5.4E-07	9.6E-09	1.0E+00	3.0E-12	2.2E-09
IKZF2	5.5E-06	2.4E-03	1.8E-06	1.3E-11	9.4E-09	1.0E+00	9.2E-01	1.0E+00	1.0E+00	1.0E+00
TEAD1	4.1E-09	2.4E-04	1.7E-01	7.7E-11	5.7E-08	2.8E-01	4.8E-01	8.5E-01	6.3E-01	1.0E+00
ZEB1	3.6E-06	8.9E-07	3.6E-01	4.7E-10	3.4E-07	7.4E-01	3.2E-01	8.2E-01	7.7E-01	1.0E+00
NFE2L2	7.2E-07	3.2E-06	5.9E-01	5.5E-10	4.0E-07	9.4E-01	8.8E-01	8.3E-01	9.9E-01	1.0E+00
DBX2	7.6E-07	8.7E-06	2.2E-01	5.8E-10	4.3E-07	9.8E-01	5.1E-01	1.6E-01	5.3E-01	1.0E+00
HMBOX1	2.0E-07	1.4E-04	2.7E-01	2.5E-09	1.9E-06	5.8E-01	5.6E-01	1.0E+00	9.0E-01	1.0E+00
NKX2-1	2.0E-05	1.5E-05	5.1E-02	5.3E-09	3.8E-06	7.9E-01	7.8E-01	9.5E-01	9.8E-01	1.0E+00
ARID4B	3.8E-07	6.6E-02	6.5E-04	5.6E-09	4.1E-06	4.7E-01	8.8E-01	1.0E+00	9.4E-01	1.0E+00
RFX2	3.4E-03	5.1E-09	1.0E+00	5.8E-09	4.2E-06	9.6E-01	9.9E-01	4.8E-02	4.1E-01	1.0E+00
ATRX	1.7E-04	8.3E-02	1.6E-06	7.7E-09	5.5E-06	8.0E-01	2.7E-02	1.0E+00	2.6E-01	1.0E+00
HES6	9.9E-01	1.0E+00	1.0E+00	1.0E+00	1.0E+00	2.3E-03	9.8E-01	1.1E-08	8.1E-09	6.0E-06
ZBTB7A	1.0E+00	9.7E-01	9.9E-01	1.0E+00	1.0E+00	1.5E-03	9.9E-01	1.8E-08	8.7E-09	6.4E-06
SREBF1	6.5E-06	4.2E-06	1.0E+00	8.8E-09	6.3E-06	9.2E-01	1.0E+00	3.0E-02	3.0E-01	1.0E+00
MEIS2	3.8E-05	1.3E-06	8.0E-01	1.2E-08	8.9E-06	7.0E-01	2.3E-02	7.4E-01	1.8E-01	1.0E+00
SOX21	8.7E-07	2.7E-04	2.3E-01	1.6E-08	1.2E-05	9.8E-01	7.5E-02	9.0E-01	4.9E-01	1.0E+00
RUNX1	1.3E-03	5.2E-08	1.0E+00	2.1E-08	1.5E-05	9.0E-01	9.6E-01	9.3E-01	1.0E+00	1.0E+00
NFIC	1.0E+00	3.9E-01	1.0E+00	9.3E-01	1.0E+00	7.5E-03	1.3E-01	1.2E-07	3.2E-08	2.4E-05
LIN54	4.6E-04	4.6E-01	1.6E-06	8.6E-08	6.2E-05	3.6E-01	1.2E-01	1.0E+00	4.0E-01	1.0E+00
TAF1	1.3E-04	9.8E-05	6.3E-02	1.9E-07	1.4E-04	5.6E-01	7.3E-01	9.9E-01	9.4E-01	1.0E+00
TFE3	9.9E-01	6.7E-01	1.0E+00	9.9E-01	1.0E+00	1.2E-03	8.2E-01	1.0E-06	2.2E-07	1.7E-04
NR2F1	6.4E-01	9.3E-01	9.9E-01	9.8E-01	1.0E+00	5.2E-05	1.4E-02	1.4E-03	2.3E-07	1.7E-04
TFAP2C	1.0E+00	8.2E-01	9.8E-01	1.0E+00	1.0E+00	5.1E-03	4.9E-01	4.0E-07	2.4E-07	1.8E-04
NR1H2	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	3.7E-02	5.8E-01	6.5E-08	3.2E-07	2.4E-04
TRPS1	1.6E-05	1.5E-04	9.1E-01	4.9E-07	3.5E-04	9.9E-01	2.4E-01	5.9E-01	6.8E-01	1.0E+00
ID4	1.2E-03	2.1E-06	9.4E-01	5.1E-07	3.6E-04	9.4E-01	8.0E-01	1.8E-02	2.0E-01	1.0E+00
NPAS3	1.1E-06	4.9E-03	8.9E-01	1.0E-06	7.2E-04	9.7E-01	9.1E-01	9.2E-03	1.4E-01	1.0E+00
JUN	3.0E-03	6.0E-05	2.9E-02	1.1E-06	7.6E-04	8.3E-02	5.1E-01	9.6E-01	3.8E-01	1.0E+00
ZHX1	1.4E-03	2.1E-03	2.0E-03	1.2E-06	8.6E-04	8.7E-01	2.4E-01	9.9E-01	7.9E-01	1.0E+00
LYL1	6.9E-04	1.1E-05	1.0E+00	1.4E-06	1.0E-03	9.6E-01	9.9E-01	1.0E-01	6.0E-01	1.0E+00
PATZ1	3.5E-03	2.4E-06	9.1E-01	1.5E-06	1.0E-03	9.7E-01	1.0E+00	2.0E-01	7.8E-01	1.0E+00
FOXO4	1.8E-04	3.0E-04	1.4E-01	1.5E-06	1.1E-03	7.5E-01	7.4E-01	8.3E-01	9.6E-01	1.0E+00
ONECUT1	7.3E-03	3.7E-01	3.5E-06	1.8E-06	1.3E-03	9.7E-01	2.8E-01	9.9E-01	8.5E-01	1.0E+00
IRF9	8.5E-05	1.4E-04	8.3E-01	1.8E-06	1.3E-03	9.8E-01	6.0E-01	8.7E-01	9.7E-01	1.0E+00
MLX	1.0E+00	1.0E+00	9.9E-01	1.0E+00	1.0E+00	8.7E-06	8.7E-01	1.4E-03	2.0E-06	1.5E-03
SMAD9	1.0E-05	1.3E-03	1.0E+00	2.3E-06	1.7E-03	1.0E+00	9.7E-01	6.3E-02	4.7E-01	1.0E+00
NR1H3	9.5E-01	7.9E-01	1.0E+00	1.0E+00	1.0E+00	2.8E-02	8.3E-01	6.4E-07	2.7E-06	2.0E-03
MEF2A	7.4E-01	1.5E-01	1.8E-06	2.7E-05	1.9E-02	3.3E-06	6.3E-03	1.0E+00	3.6E-06	2.7E-03
ARID1A	3.5E-05	3.3E-02	1.8E-02	3.7E-06	2.6E-03	9.2E-01	7.2E-01	9.0E-01	9.8E-01	1.0E+00
TSHZ3	1.0E+00	9.9E-01	3.3E-01	9.0E-01	1.0E+00	1.7E-04	3.9E-02	3.5E-03	4.0E-06	2.9E-03
KLF9	9.9E-01	1.3E-01	8.7E-01	6.3E-01	1.0E+00	3.7E-03	7.9E-01	9.4E-06	4.6E-06	3.4E-03
POU3F4	3.0E-05	5.9E-03	1.6E-01	4.8E-06	3.3E-03	8.5E-01	4.9E-01	8.6E-01	9.1E-01	1.0E+00
KLF13	1.0E+00	3.6E-01	9.2E-01	9.0E-01	1.0E+00	1.2E-02	9.5E-01	2.7E-06	5.2E-06	3.8E-03
NCOA1	1.0E+00	6.8E-01	1.0E+00	9.9E-01	1.0E+00	5.1E-05	5.7E-01	1.5E-03	6.9E-06	5.0E-03
ARID4A	1.1E-02	4.2E-01	9.8E-06	7.3E-06	5.1E-03	7.3E-01	9.8E-01	9.9E-01	9.9E-01	1.0E+00
PPARA	7.7E-05	6.1E-04	9.9E-01	7.5E-06	5.3E-03	9.9E-01	9.1E-01	1.0E+00	1.0E+00	1.0E+00
FOXN3	1.7E-02	3.1E-02	1.1E-04	9.3E-06	6.5E-03	2.2E-01	4.1E-01	1.0E+00	5.7E-01	1.0E+00
POU3F2	2.4E-03	7.7E-04	3.4E-02	9.7E-06	6.8E-03	7.4E-01	4.5E-01	9.7E-01	8.9E-01	1.0E+00
NR2F6	1.0E+00	9.8E-01	1.0E+00	1.0E+00	1.0E+00	5.7E-03	9.6E-01	1.5E-05	1.2E-05	8.8E-03
SP3	1.5E-02	8.3E-03	8.0E-04	1.5E-05	1.0E-02	5.5E-01	7.5E-01	1.8E-02	1.3E-01	1.0E+00

FOXC1	3.5E-02	3.5E-06	8.8E-01	1.6E-05	1.1E-02	1.6E-01	9.8E-01	6.0E-01	5.8E-01	1.0E+00
FOXO3	1.0E-04	2.0E-01	6.9E-03	2.0E-05	1.4E-02	1.4E-01	2.1E-01	1.0E+00	3.1E-01	1.0E+00
JDP2	9.6E-01	9.1E-01	1.2E-01	6.0E-01	1.0E+00	3.5E-04	1.2E-02	4.7E-02	2.7E-05	1.9E-02
SMAD1	2.6E-05	2.5E-02	3.6E-01	3.1E-05	2.1E-02	9.5E-01	8.6E-01	3.6E-01	8.7E-01	1.0E+00
FOXH1	9.1E-01	9.8E-01	1.2E-01	6.2E-01	1.0E+00	6.9E-04	4.7E-04	8.6E-01	3.6E-05	2.6E-02
ESRRG	9.8E-01	1.0E+00	7.3E-01	1.0E+00	1.0E+00	5.9E-02	1.9E-03	3.3E-03	4.6E-05	3.3E-02
BPTF	3.6E-04	1.4E-02	7.8E-02	4.7E-05	3.2E-02	9.5E-01	9.5E-01	9.9E-01	1.0E+00	1.0E+00

Table S6. Key regulators of prefrontal cortex gene expression changes in Alzheimer's disease cases vs. controls.

TF	p.up GSE84422	p.up GSE33000	meta-p up	q up	p.down GSE84422	p.down GSE33000	meta-p dn	q down
ATF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00	8.9E-04	9.2E-49	9.8E-50	6.8E-47
TBPL1	1.0E+00	1.0E+00	1.0E+00	1.0E+00	4.2E-04	4.0E-47	2.0E-48	1.4E-45
PATZ1	1.3E-08	2.0E-31	2.4E-37	1.6E-34	1.0E+00	1.0E+00	1.0E+00	1.0E+00
TFEB	1.1E-15	9.7E-23	8.9E-36	6.2E-33	1.0E+00	1.0E+00	1.0E+00	1.0E+00
KLF15	2.3E-05	1.1E-30	2.0E-33	1.4E-30	9.9E-01	1.0E+00	1.0E+00	1.0E+00
SP2	8.5E-12	9.0E-24	6.1E-33	4.2E-30	1.0E+00	1.0E+00	1.0E+00	1.0E+00
FOXH1	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.4E-03	6.8E-28	6.7E-29	4.6E-26
KLF2	5.5E-06	2.3E-20	7.5E-24	5.1E-21	1.0E+00	1.0E+00	1.0E+00	1.0E+00
MAFG	9.7E-01	1.0E+00	1.0E+00	1.0E+00	4.7E-02	1.3E-22	3.3E-22	2.3E-19
EWSR1	3.3E-05	3.0E-19	5.3E-22	3.6E-19	9.9E-01	1.0E+00	1.0E+00	1.0E+00
MEF2A	1.0E+00	1.0E+00	1.0E+00	1.0E+00	4.0E-05	1.7E-18	3.5E-21	2.4E-18
HMGA1	1.0E+00	1.0E+00	1.0E+00	1.0E+00	5.0E-07	1.4E-16	3.8E-21	2.6E-18
HOPX	1.0E+00	1.0E+00	1.0E+00	1.0E+00	2.2E-02	1.6E-20	1.8E-20	1.3E-17
HES1	1.5E-03	6.5E-19	4.7E-20	3.2E-17	9.9E-01	1.0E+00	1.0E+00	1.0E+00
MESP1	1.0E+00	1.0E+00	1.0E+00	1.0E+00	3.9E-02	1.1E-19	2.0E-19	1.4E-16
KLF3	4.5E-06	2.1E-15	4.5E-19	3.1E-16	9.9E-01	1.0E+00	1.0E+00	1.0E+00
LYL1	1.6E-03	9.5E-17	6.9E-18	4.7E-15	9.8E-01	1.0E+00	1.0E+00	1.0E+00
CREB5	6.8E-05	1.5E-14	4.2E-17	2.9E-14	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HSF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00	7.7E-03	3.7E-16	1.2E-16	8.0E-14
ZHX3	7.0E-05	7.2E-14	2.1E-16	1.4E-13	9.7E-01	1.0E+00	1.0E+00	1.0E+00
SREBF1	7.4E-08	7.8E-11	2.4E-16	1.6E-13	9.9E-01	1.0E+00	1.0E+00	1.0E+00
FOXD4L3	1.0E+00	1.0E+00	1.0E+00	1.0E+00	2.8E-02	3.9E-16	4.4E-16	3.0E-13
STAT3	1.4E-02	6.4E-15	3.3E-15	2.2E-12	9.9E-01	1.0E+00	1.0E+00	1.0E+00
ELF1	1.8E-05	1.5E-10	9.3E-14	6.2E-11	9.9E-01	1.0E+00	1.0E+00	1.0E+00
MEF2C	9.5E-01	1.0E+00	1.0E+00	1.0E+00	6.0E-03	5.3E-13	1.1E-13	7.3E-11
FOXO4	3.7E-04	1.2E-11	1.5E-13	1.0E-10	5.6E-01	1.0E+00	8.8E-01	1.0E+00
MTA3	9.9E-01	1.0E+00	1.0E+00	1.0E+00	9.8E-04	2.0E-11	6.3E-13	4.3E-10
PBX1	1.0E+00	1.0E+00	1.0E+00	1.0E+00	3.4E-03	3.1E-11	3.2E-12	2.2E-09
GLIS3	2.5E-02	7.9E-12	6.0E-12	4.0E-09	1.0E+00	1.0E+00	1.0E+00	1.0E+00
GATAD1	5.4E-04	7.5E-10	1.2E-11	8.0E-09	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SMAD9	3.8E-03	6.1E-10	6.4E-11	4.3E-08	7.5E-01	1.0E+00	9.7E-01	1.0E+00
FOXP2	4.0E-01	9.9E-01	7.6E-01	1.0E+00	7.9E-06	5.1E-07	1.1E-10	7.5E-08
ELF4	1.8E-02	7.9E-10	3.7E-10	2.4E-07	9.9E-01	1.0E+00	1.0E+00	1.0E+00
SOX5	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.4E-03	1.2E-08	4.5E-10	3.0E-07
NPAS3	3.7E-03	6.1E-09	5.8E-10	3.8E-07	9.7E-01	1.0E+00	1.0E+00	1.0E+00
HDX	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.1E-05	3.5E-06	9.9E-10	6.7E-07
ZHX2	2.4E-04	1.7E-07	1.0E-09	6.6E-07	9.9E-01	1.0E+00	1.0E+00	1.0E+00
TCF3	7.5E-04	6.0E-08	1.1E-09	7.4E-07	8.5E-01	1.0E+00	9.9E-01	1.0E+00
LHX6	9.4E-01	1.0E+00	1.0E+00	1.0E+00	4.9E-04	1.1E-07	1.4E-09	9.1E-07
ID3	1.7E-02	3.7E-09	1.6E-09	1.0E-06	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ZBTB33	9.7E-01	9.9E-01	1.0E+00	1.0E+00	1.2E-06	9.1E-05	2.7E-09	1.8E-06
POU3F1	9.5E-01	1.0E+00	1.0E+00	1.0E+00	2.1E-04	9.5E-07	4.7E-09	3.2E-06
LHX2	9.3E-01	9.9E-01	1.0E+00	1.0E+00	2.1E-04	1.0E-06	4.8E-09	3.2E-06
MXI1	1.4E-05	1.8E-05	5.8E-09	3.8E-06	9.5E-01	1.0E+00	1.0E+00	1.0E+00
STAT4	9.5E-01	1.0E+00	1.0E+00	1.0E+00	1.3E-02	3.1E-08	8.8E-09	5.9E-06
KLF1	6.1E-03	7.1E-08	9.7E-09	6.4E-06	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ZFX	5.7E-05	1.1E-05	1.4E-08	9.0E-06	7.0E-01	1.0E+00	9.5E-01	1.0E+00
KLF13	1.2E-07	5.2E-03	1.4E-08	9.2E-06	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HLF	9.3E-01	1.0E+00	1.0E+00	1.0E+00	2.1E-02	4.0E-08	1.9E-08	1.3E-05
EP300	5.0E-04	2.2E-06	2.4E-08	1.6E-05	9.3E-01	1.0E+00	1.0E+00	1.0E+00
MZF1	4.9E-04	2.6E-06	2.8E-08	1.8E-05	1.0E+00	1.0E+00	1.0E+00	1.0E+00
RARG	7.8E-03	1.7E-07	2.9E-08	1.9E-05	8.4E-01	1.0E+00	9.9E-01	1.0E+00
KLF11	3.9E-04	5.5E-06	4.5E-08	2.9E-05	9.9E-01	1.0E+00	1.0E+00	1.0E+00
PAX9	4.0E-03	8.5E-07	7.0E-08	4.6E-05	9.9E-01	1.0E+00	1.0E+00	1.0E+00
NFIX	9.9E-07	6.1E-03	1.2E-07	7.9E-05	9.7E-01	1.0E+00	1.0E+00	1.0E+00
TCFL5	3.1E-04	3.2E-05	2.0E-07	1.3E-04	1.0E+00	1.0E+00	1.0E+00	1.0E+00
NFIA	3.1E-03	3.7E-06	2.2E-07	1.4E-04	8.0E-01	9.9E-01	9.8E-01	1.0E+00
FOXQ1	4.1E-02	2.9E-07	2.3E-07	1.5E-04	7.5E-01	1.0E+00	9.7E-01	1.0E+00
FOXJ3	1.0E+00	9.1E-01	1.0E+00	1.0E+00	1.6E-04	1.8E-04	5.1E-07	3.4E-04
E2F6	4.6E-03	6.9E-06	5.8E-07	3.7E-04	9.9E-01	1.0E+00	1.0E+00	1.0E+00
NR3C2	9.8E-01	1.0E+00	1.0E+00	1.0E+00	1.0E-02	7.7E-06	1.4E-06	9.2E-04
NFKB1	1.2E-03	1.1E-04	2.2E-06	1.4E-03	1.0E+00	1.0E+00	1.0E+00	1.0E+00

ARNTL2	9.5E-01	1.0E+00	1.0E+00	1.0E+00	3.7E-02	4.2E-06	2.6E-06	1.7E-03
THAP6	3.9E-03	4.6E-05	2.9E-06	1.9E-03	2.0E-01	1.0E+00	5.2E-01	1.0E+00
ONECUT2	1.0E+00	9.9E-01	1.0E+00	1.0E+00	8.0E-03	2.4E-05	3.1E-06	2.0E-03
DBX1	1.1E-05	1.8E-02	3.3E-06	2.0E-03	9.8E-01	1.0E+00	1.0E+00	1.0E+00
SOX18	1.7E-02	1.6E-05	4.3E-06	2.7E-03	9.8E-01	1.0E+00	1.0E+00	1.0E+00
HOMEZ	2.7E-03	1.1E-04	4.7E-06	2.9E-03	9.7E-01	1.0E+00	1.0E+00	1.0E+00
ZBTB16	9.1E-01	1.0E+00	1.0E+00	1.0E+00	2.1E-03	1.4E-04	4.9E-06	3.2E-03
SOX10	7.2E-03	6.0E-05	6.7E-06	4.2E-03	9.4E-01	1.0E+00	1.0E+00	1.0E+00
CREB3L4	4.3E-03	2.7E-04	1.7E-05	1.1E-02	9.6E-01	1.0E+00	1.0E+00	1.0E+00
LMO2	2.0E-02	6.8E-05	2.0E-05	1.2E-02	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HSF1	1.1E-02	2.8E-04	4.3E-05	2.6E-02	8.9E-01	1.0E+00	9.9E-01	1.0E+00
FOXK2	9.5E-01	9.7E-01	1.0E+00	1.0E+00	2.1E-03	1.8E-03	4.9E-05	3.2E-02
LIN54	9.8E-01	1.0E+00	1.0E+00	1.0E+00	2.0E-02	2.0E-04	5.3E-05	3.4E-02
PRKRIR	9.8E-01	9.7E-01	1.0E+00	1.0E+00	1.7E-02	2.9E-04	6.4E-05	4.2E-02
ZBTB7B	2.3E-02	2.1E-04	6.6E-05	4.0E-02	9.6E-01	9.9E-01	1.0E+00	1.0E+00
SMAD7	1.2E-02	4.7E-04	7.5E-05	4.5E-02	5.1E-01	1.0E+00	8.5E-01	1.0E+00

Table S7. Key regulators of prefrontal cortex gene expression changes in autism spectrum disorder cases vs. controls.

TF	p.up GSE28521	p.up Parikshak	meta-p up	q up	p.down GSE28521	p.down Parikshak	meta-p down	q down
SOX9	2.9E-05	9.7E-38	2.7E-40	1.7E-37	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ATF2	1.0E+00	1.0E+00	1.0E+00	1.0E+00	6.3E-09	4.1E-18	1.6E-24	1.0E-21
HES1	1.2E-06	3.5E-15	2.1E-19	1.3E-16	1.0E+00	1.0E+00	1.0E+00	1.0E+00
FOXJ1	7.8E-03	1.0E-18	3.8E-19	2.4E-16	4.2E-01	1.9E-01	2.8E-01	1.0E+00
ESRRG	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.2E-06	7.5E-15	4.3E-19	2.7E-16
RFX4	3.9E-05	3.2E-13	5.0E-16	3.2E-13	8.7E-01	9.0E-01	9.7E-01	1.0E+00
PATZ1	2.8E-10	2.8E-06	2.8E-14	1.8E-11	9.9E-01	1.0E+00	1.0E+00	1.0E+00
FOXH1	9.9E-01	1.0E+00	1.0E+00	1.0E+00	2.1E-08	9.0E-08	6.7E-14	4.3E-11
FOXD4L3	9.6E-01	1.0E+00	1.0E+00	1.0E+00	1.3E-05	1.1E-09	4.6E-13	2.9E-10
RFX2	4.8E-04	5.6E-11	8.6E-13	5.5E-10	7.9E-01	9.3E-01	9.6E-01	1.0E+00
ESRRA	7.2E-01	9.5E-01	9.4E-01	1.0E+00	9.4E-03	1.2E-11	3.4E-12	2.1E-09
RUNX1	3.1E-02	6.8E-12	6.3E-12	4.0E-09	7.8E-01	1.0E+00	9.7E-01	1.0E+00
FOXC1	1.9E-04	2.2E-09	1.2E-11	7.9E-09	1.0E+00	1.0E+00	1.0E+00	1.0E+00
LYL1	3.0E-04	1.6E-09	1.4E-11	8.8E-09	1.0E+00	1.0E+00	1.0E+00	1.0E+00
HLF	9.9E-01	5.6E-01	8.8E-01	1.0E+00	3.1E-03	2.6E-10	2.3E-11	1.5E-08
TBPL1	1.0E+00	1.0E+00	1.0E+00	1.0E+00	6.3E-04	2.3E-09	4.1E-11	2.6E-08
DBX2	1.8E-02	4.3E-10	2.0E-10	1.3E-07	1.0E+00	1.0E+00	1.0E+00	1.0E+00
JUN	9.3E-04	9.0E-09	2.2E-10	1.4E-07	8.9E-01	8.2E-01	9.6E-01	1.0E+00
MKX	4.3E-01	6.5E-01	6.4E-01	1.0E+00	4.5E-06	4.7E-06	5.4E-10	3.4E-07
ELF4	2.5E-04	1.1E-07	7.3E-10	4.6E-07	9.8E-01	1.0E+00	1.0E+00	1.0E+00
SREBF1	2.7E-07	1.2E-04	8.0E-10	5.0E-07	9.9E-01	1.0E+00	1.0E+00	1.0E+00
KLF15	5.4E-06	5.4E-05	6.7E-09	4.2E-06	1.0E+00	1.0E+00	1.0E+00	1.0E+00
GLIS3	4.0E-02	1.0E-08	9.2E-09	5.8E-06	1.0E+00	1.0E+00	1.0E+00	1.0E+00
KLF12	1.0E+00	9.9E-01	1.0E+00	1.0E+00	2.9E-06	2.4E-04	1.6E-08	9.8E-06
MEF2A	9.1E-01	9.6E-01	9.9E-01	1.0E+00	4.4E-02	2.0E-08	1.9E-08	1.2E-05
SPDEF	9.6E-01	1.0E+00	1.0E+00	1.0E+00	4.0E-03	6.2E-07	5.2E-08	3.3E-05
SOX3	1.8E-02	1.9E-07	6.9E-08	4.3E-05	9.6E-01	9.2E-01	9.9E-01	1.0E+00
TLX2	1.0E+00	9.2E-01	1.0E+00	1.0E+00	3.3E-06	1.0E-03	6.9E-08	4.3E-05
SCRT1	7.7E-01	9.7E-01	9.6E-01	1.0E+00	4.9E-02	7.2E-08	7.3E-08	4.5E-05
PPARA	2.1E-02	3.6E-07	1.5E-07	9.4E-05	8.6E-01	9.9E-01	9.9E-01	1.0E+00
CLOCK	1.0E+00	1.0E+00	1.0E+00	1.0E+00	5.5E-03	2.7E-06	2.9E-07	1.8E-04
DDIT3	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.2E-02	1.3E-06	3.0E-07	1.9E-04
ZNF423	1.4E-02	2.7E-06	6.7E-07	4.1E-04	4.4E-01	9.8E-01	7.9E-01	1.0E+00
JDP2	4.7E-01	8.6E-01	7.7E-01	1.0E+00	8.0E-03	5.3E-06	7.6E-07	4.7E-04
SOX1	2.8E-02	2.3E-06	1.1E-06	7.0E-04	9.9E-01	9.7E-01	1.0E+00	1.0E+00
BATF2	2.2E-02	3.1E-06	1.2E-06	7.2E-04	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SRY	2.2E-01	8.6E-02	9.5E-02	1.0E+00	2.3E-02	3.3E-06	1.3E-06	8.2E-04
STAT3	8.1E-04	9.7E-05	1.4E-06	8.4E-04	9.9E-01	1.0E+00	1.0E+00	1.0E+00
GSX1	6.8E-03	1.7E-05	2.0E-06	1.2E-03	9.6E-01	9.9E-01	1.0E+00	1.0E+00
KLF16	1.1E-05	1.2E-02	2.2E-06	1.4E-03	1.0E+00	1.0E+00	1.0E+00	1.0E+00
ETV4	4.4E-02	4.5E-06	3.2E-06	2.0E-03	4.9E-01	8.2E-01	7.7E-01	1.0E+00
KLF2	2.7E-04	8.2E-04	3.6E-06	2.2E-03	9.2E-01	1.0E+00	1.0E+00	1.0E+00
STAT4	5.9E-01	7.5E-01	8.0E-01	1.0E+00	9.9E-05	2.5E-03	3.9E-06	2.4E-03
SOHLH1	7.0E-01	8.7E-01	9.1E-01	1.0E+00	3.9E-03	7.0E-05	4.4E-06	2.7E-03
TCF7L1	2.4E-01	5.7E-02	7.3E-02	1.0E+00	2.8E-04	1.1E-03	4.9E-06	3.1E-03
HDX	9.1E-01	9.8E-01	9.9E-01	1.0E+00	9.7E-03	3.7E-05	5.6E-06	3.5E-03
PRDM1	2.9E-02	1.4E-05	6.3E-06	3.8E-03	2.3E-01	3.6E-01	2.9E-01	1.0E+00
TSHZ3	9.7E-01	6.3E-01	9.1E-01	1.0E+00	1.1E-03	7.9E-04	1.3E-05	8.0E-03
TEAD1	1.6E-02	5.4E-05	1.3E-05	8.0E-03	4.2E-01	2.2E-01	3.1E-01	1.0E+00
HES5	9.0E-03	1.5E-04	2.0E-05	1.2E-02	9.0E-01	8.8E-01	9.8E-01	1.0E+00
HSFY1	7.7E-01	8.1E-01	9.2E-01	1.0E+00	2.3E-03	6.3E-04	2.1E-05	1.3E-02
ZEB1	2.4E-02	1.2E-04	4.1E-05	2.4E-02	9.7E-01	9.9E-01	1.0E+00	1.0E+00
HSF1	6.0E-03	6.5E-04	5.3E-05	3.2E-02	1.0E+00	9.9E-01	1.0E+00	1.0E+00
HEYL	4.9E-02	8.3E-05	5.4E-05	3.2E-02	5.3E-02	9.3E-01	2.0E-01	1.0E+00
NFAT5	8.7E-01	1.0E+00	9.9E-01	1.0E+00	1.5E-02	2.8E-04	5.4E-05	3.4E-02
RFXANK	1.7E-04	2.6E-02	6.1E-05	3.6E-02	1.0E+00	1.0E+00	1.0E+00	1.0E+00
SATB1	9.2E-01	9.8E-01	9.9E-01	1.0E+00	5.8E-03	8.3E-04	6.4E-05	4.0E-02
SIX4	7.4E-01	3.7E-01	6.3E-01	1.0E+00	7.7E-03	8.0E-04	8.0E-05	4.9E-02

Table S8. Key regulator TFs at GWAS risk loci

PubMed ID	First Author	Disease Trait	SNP	P-value	TRN Key Regulator Analysis	TF
23562540	Cruchaga C	Alzheimer's disease biomarkers	rs514716	3.00E-09	AD	GLIS3
24618891	Muhleisen	Bipolar disorder	rs12202969	1.00E-08	SCZ,BD	POU3F2
25056061	Ripke S	Schizophrenia	rs8082590	2.00E-08	SCZ,AD,ASD	SREBF1
24162737	European Alzheimer's Disease Initiative (EADI)	Alzheimer's disease (late onset)	rs190982	3.00E-08	AD	MEF2C
27329760	Hou L	Bipolar disorder	rs1487441	3.00E-08	SCZ,BD	POU3F2
26198764	Goes FS	Schizophrenia	rs9398171	4.00E-08	SCZ	FOXO3
26198764	Goes FS	Schizophrenia	rs12883788	3.00E-07	SCZ,BD,AD	NPAS3
26830138	Herold C	Alzheimer disease and age of onset	rs3931397	6.00E-07	AD	NR3C2
20889312	Wang KS	Bipolar disorder and schizophrenia	rs11880706	2.00E-06	SCZ,ASD	RFX2
26821981	Koga AT	antipsychotic drug dosage in schizophrenia or schizoaffective disorder	rs75905933	4.00E-06	SCZ,AD	KLF13
20713499	Huang J	Schizophrenia, bipolar disorder and depression (combined)	rs4982029	4.00E-06	SCZ,BD,AD	NPAS3
18711365	Ferreira MA	Bipolar disorder	rs8015959	5.00E-06	SCZ,BD,AD	NPAS3
27770636	Mez J	Late-onset Alzheimer's disease	rs148003968	5.00E-06	AD	TFEB
23453885	Smoller JW	Autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia (combined)	rs7565792	7.00E-06	BD,SCZ	FOXN2
19023125	Potkin SG	Brain imaging in schizophrenia (interaction)	rs9491640	9.00E-06	SCZ,BD	POU3F2

Table S9. TFBS-disrupting, SCZ-associated SNPs.

SCZ Risk Locus (Ripke et al. 2014)			Predicted Functional SNP						Predicted Target Gene		Predicted Regulator		Support from eQTLs					
Location	best SNP	p min.	snp	chr	pos	a1	a2	scz2.or	scz2.pvalue	Gene	TSS Distance	TF	cor(TF, target gene)	proxy_snp	r2	D'	eQTL.study	eQTL.tissue	eQTL.pvalue
chr2:57943593-58502192	rs11682175	1.47E-11	rs13384219	chr2	58134458	A	G	1.12	2.1E-05	*VRK2*	-336	***POU3F2***	0.67	rs13384219	1	1	Westra2013	Whole_Blood	2.0E-03
												POU3F3	0.39						
chr2:198148577-198835577	rs6434928	2.06E-11	rs12618612	chr2	198171058	T	G	1.06	3.7E-08	*ANKRD44*	-4839	*PRRX1*	0.46	rs4850787	0.97	0.99	GTEx2015_wTestis		7.2E-18
												HESX1	0.40						
	rs13668836		chr2	198172148	T	G	0.94	1.7E-08	*ANKRD44*	-3754	*HOXD1*	0.56	rs34773166	0.96	0.99	GTEx2015_wTestis		5.0E-27	
chr2:225334096-225467796	rs11685299	1.12E-08	rs3768886	chr2	225450161	C	G	0.94	5.8E-08	*CUL3*	30	*ETV1*	0.67	rs16866061	0.92	0.99	Westra2013	Whole_Blood	5.7E-08
												ETV5	0.44						
												SPDEF	0.35						
												ELF4	0.00						
chr3:135807405-136615405	rs7432375	7.26E-11	rs9845788	chr3	135914715	A	G	0.95	6.0E-07	*MSL2*	-1371	*NR2C1*	0.63	NA					
												E2F6	0.59						
												SP4	0.57						
												KLF11	0.53						
												SP3	0.49						
												SP1	0.45						
												KLF9	0.45						
												SP2	0.42						
												KLF12	0.33						
												KLF5	-0.15						
chr4:170357552-170646052	rs10520163	1.47E-09	rs28535010	chr4	170532858	T	C	1.05	6.0E-05	*CLCN3*	-926	*HOXD1*	0.63	rs6553446	0.98	-1	CMC	DLPFC	1.4E-03
												GBX2	0.23						
										NEK1	-922	*CDX2*	0.24	rs34587213	0.91	0.96	CMC	DLPFC	1.1E-07
chr5:137598121-137948092	rs3849046	4.67E-09	rs3756766	chr5	137673167	A	C	1.07	2.2E-08	*FAM53C*	5536	*TFAP2C*	0.56	rs10900851	0.99	-1	Westra2013	Whole_Blood	2.5E-08
	rs154069		chr5	137876920	A	G	0.95	2.3E-07	*ETF1*	-2077	*NR1H3*	0.64	rs11738995	0.94	-0.97	Westra2013	Whole_Blood	1.7E-10	
												NCOA1	0.71						
												SREBF2	0.68						
												NR2F6	0.62						
												ARNT2	0.54						
												AHR	0.49						
												MLX	0.49						
												TFE3	0.44						
												ARNT	0.12						
	rs11746692		chr5	137946355	A	C	1.06	4.6E-07	*CTNNA1*	-313	*NR1H2*	0.33	rs891953	0.99	0.99	CMC	DLPFC	1.1E-04	
chr7:104598064-105063064	rs6466055	1.13E-09	rs624384724	chr7	105049431	A	T	1.06	2.1E-05	*SRPK2*	9668	*ARID5B*	0.67	NA					
												POU3F2	0.64						
												HOXB13	0.55						
												HOXB7	0.55						
												ISX	0.39						
												HOXC12	0.18						
												ARID3B	-0.28						
												PAX6	-0.35						
chr10:104423800-105165583	rs11191419	6.20E-19	rs79780963	chr10	104952499	T	C	0.86	7.5E-16	*NT5C2*	-564	*KLF7*	0.61	rs11191580	0.95	0.98	Westra2013	Whole_Blood	8.9E-57
												SP1	0.60						
												KLF3	0.57						
												SP3	0.46						
												SP4	0.38						
												SP2	0.27						
												EGR4	-0.08						
chr12:123448113-123909113	rs2851447	1.86E-14	rs1716162	chr12	123642538	A	T	0.93	2.1E-08	*PITPNM2*	7974	*RREB1*	-0.06	rs1716162	1	1	GTEx2015_wThyroid		2.5E-12
	rs941306		chr12	123715113	T	C	0.93	6.6E-08	*C12orf65*	-2364	*TBPL1*	0.41	rs10772998	0.99	0.99	GTEx2015_wArtery_Aorta		1.5E-07	
chr17:17722402-18030202	rs8082590	1.77E-08	rs7359509	chr17	17741875	A	G	1.06	1.4E-06	*SREBF1*	1546	***KLF15***	0.67	rs3183702	0.95	0.98	Westra2013	Whole_Blood	9.4E-91
												SP2	0.41						
												KLF9	0.07						
												EGR4	-0.17						
	rs4925119		chr17	17742904	A	G	1.05	5.9E-06	*SREBF1*	2575	***KLF15***	0.67	rs7359509	0.94	0.98	Westra2013	Whole_Blood	9.4E-91	
												SP2	0.41						
												KLF9	0.07						
												EGR4	-0.17						
	rs6502618		chr17	17746741	A	G	1.05	2.1E-06	*SREBF1*	6411	*RFXANK*	0.59	rs3183702	0.93	0.97	Westra2013	Whole_Blood	9.4E-91	
												RFX4	0.50						
chr22:42315744-42689414	rs6002655	1.71E-09	rs133376	chr22	42466905	T	C	1.05	3.0E-06	*FAM109B*	-3354	*KLF14*	0.67	rs1063392	0.94	0.99	CMC	DLPFC	4.0E-08
												KLF16	0.48						
												KLF2	0.32						
												EGR2	0.16						

Table S10. TOPO cloning analysis of POU3F2 sgRNA editing at target site.

Clone ID	Sequence Result
1	*reference*
2	*edited*
3	*reference*
4	*edited*
5	*edited*
6	*reference*
7	*edited*
8	*edited*
9	*edited*
10	*reference*
11	*reference*
12	*reference*
13	*edited*
14	*edited*
15	*reference*
16	*edited*
17	*reference*
18	*reference*
19	*edited*
20	*reference*
21	*reference*
22	*edited*
23	*reference*
24	*edited*
25	*edited*
26	*reference*
27	*reference*
28	*reference*
29	*reference*
30	*reference*
31	*edited*
32	*reference*
33	*edited*

% REFERENCE	% EDITED
0.55	**0.45**

5.9 Notes

This work was performed in collaboration with Dani E. Bergey, Bijoya Basu, Cory C. Funk, Paul Shannon, Leroy Hood, Nathan D. Price, Carlo Colantuoni, and Seth A. Ament

AUTHOR CONTRIBUTIONS

SAA and JRP designed the experiments. JRP, BB and DB performed the experiments. SAA, JRP, DB, CF, CC, NP, and LH analyzed the data. SAA, JP, and CC wrote the paper.

ACKNOWLEDGMENTS

Patrick Paddison, Yu Ding and Chad Toledo (Fred Hutchinson Cancer Research Center) provided human neural stem cells. Elizabeth Gray, Daniel Stetson, and Kathleen Pestal (University of Washington) provided LentiCRISPR plasmids and protocols. John Kelsoe and Tatyana Shekhtman (University of California, San Diego) provided genomic DNA for promoter cloning of *VRK2*. Nathaniel Peters (W.M. Keck Microscopy Center, University of Washington) provided immunofluorescence imaging and advice. Gene Robinson provided helpful discussion and comments on an early version of this manuscript. This work was supported by a National Science Foundation Graduate Research Fellowship (JRP); **a** NARSAD Young Investigator Award from the Brain and Behavior Research Foundation (SAA), the NIGMS Center for Systems Biology at the Institute for Systems Biology (P50 GM076547; LH, NDP), and by the Big Data for Discovery Science Center of the NIH Big Data to Knowledge program (NDP, LH).

6 Metabolic network analysis of Huntington's disease

6.1 Abstract

Huntington's disease (HD) is a fatal neurodegenerative disease caused by a CAG trinucleotide repeat expansion in the *HTT* gene, with age-at-onset inversely proportional to the number of CAG repeats. Differences in metabolism and brain gene expression are among the earliest detectable changes in HD and in HD mouse models. We investigate changes in the expression of genes encoding metabolic enzymes using a large, previously-published dataset with RNA-seq from an allelic series of HD knock-in mice, sampled at three time points: two, six and ten months of age. In addition, transcriptomic data were available from 13 distinct tissues in 6-month-old $Htt^{Q175/+}$ mice. Metabolic network analysis using PathWave suggested CAG-repeat length dependent gene expression changes in 44 of 79 metabolic pathways from KEGG in mouse striatum. The urea cycle pathway and cholesterol biosynthesis pathway were among the top differentially expressed metabolic pathways in our analysis. We found differentially expressed pathways (DEPs) in nine of the thirteen tissues surveyed. We used CAG-repeat length dependent genes and oPPOSUM to discover 20 transcription factors predicted to regulate metabolic genes in the mouse striatum. Our study thus identifies *Htt* CAG-repeat length dependent changes in metabolic pathways, and makes predictions about the regulators of differentially expressed metabolic genes.

6.2 Introduction

Huntington's Disease (HD) is a progressive, neurodegenerative disease caused by an expanded CAG tract in the *Huntingtin* (*HTT*) gene [1,2]. HD inheritance follows an autosomal dominant pattern, and individuals carrying a copy of mutant *HTT* with a CAG-repeat length of 40 or more will develop the disease in their lifetime. Reduced penetrance is observed in individuals with 36 to 39 CAG repeats. Both the severity and age of onset of HD are inversely proportional to *HTT* CAG-repeat length [1]. Clinically, HD results in a variety of symptoms that include motor, affective, and cognitive dysfunction [1]. While mutant *HTT* is widely accepted as the cause of HD [3-5], molecular pathogenesis of HD remains poorly understood.

One unresolved question involves the mechanism of tissue selectivity for HD pathology. Mutant *HTT* is ubiquitously expressed across all body tissues. However, pathological changes occur most dramatically in the striatum, a subcortical region of the forebrain with well-characterized roles in motor control, reward, and mood [6]. Specifically, cell death and other forms of pathology occur primarily in striatal medium spiny neurons, and significantly less so to other neuronal types in various other brain regions [7].

Several studies have reported effects of the HD mutation on cell metabolism and survival pathways [8], among a wide range of other disease-perturbed cellular processes. Metabolic defects induced by mutant *HTT* are thought to contribute significantly to increased cell death especially in the striatum [9]. A number of studies have pointed to the existence of a brain energy deficit during the early stages of HD [10]. Altered brain energy homeostasis, with decreased ATP levels in the striatum have been demonstrated in R6/2 mice [11]. Critically, this study demonstrated an inverse correlation between striatal ATP levels and the HD CAG repeat length, the critical driver of pathogenesis in human HD [11].

A variety of putative mechanisms that might underlie the low energy state in HD brain have also been described. These include abnormal glycolysis [12], impaired oxidative phosphorylation [13], and defective mitochondrial metabolism [14]. Other studies in HD models have implicated altered cholesterol and fatty acid synthesis [15,16], increased oxidative damage [17], and defective thermoregulation and mitochondrial oxidation [18]. Investigating the timing and impact of changes in these metabolic pathways in a genetically accurate mouse model of HD is valuable in helping understand proximal mechanisms.

It has also been shown in multiple studies of human post-mortem tissue as well as mouse models of HD that transcriptional changes occur early and throughout pathogenesis [3,19,20]. One approach to parsing large changes in gene expression is to make predictions about transcription factor (TF) drivers of these changes. Validated TF drivers of metabolic gene expression change could lead to novel therapeutics that target TFs and thus mitigate these HD-induced metabolic changes.

Here, we use published [3] transcriptomic data to carry out pathway-level metabolic analysis of gene expression changes as a result of CAG expansion in the murine *Htt* gene and in HD patients. The mouse dataset includes an allelic series of knock-in mice harboring six different heterozygous mutant *Htt* CAG-repeat lengths ($Htt^{Q20/+}$, $Htt^{Q80/+}$, $Htt^{Q92/+}$, $Htt^{Q111/+}$, $Htt^{Q140/+}$, and $Htt^{Q175/+}$) from three different mouse ages– two, six, and ten months. Striatum, cortex, and liver were sequenced at each time point. In addition to these three tissues, the dataset also includes a tissue survey at six months of 11 additional tissues. Our analyses of these data not only allowed us to discover CAG-length dependent metabolic pathways in this model, but it also enabled us to predict some of the transcriptional regulators that potentially drive these pathway-level changes.

6.3 Methods

All statistical analyses and plotting were carried out using Python 3.4.3 and R 3.1.3. All code written for this study is publicly available in a github repository at:

https://github.com/rmhariharan/HD-metabolism

Mouse gene expression data

We obtained mouse transcriptome sequencing (RNA-seq) data from GEO dataset GSE65776 [3]. Briefly, we used FPKM values from analysis of these RNA-seq data for striatum, cerebral cortex, and liver of two, six and ten month-old mice. The mice used in that study harbor one endogenous wild type *Htt* allele and a second mutant allele. The mutant allele is a knock-in of human *mHtt* whose exon 1 carries one of six different CAG-repeat or Q lengths (which encode for polyglutamine): $Htt^{+/+}$ (Q7/Q7), $Htt^{Q20/+}$, $Htt^{Q80/+}$, $Htt^{Q92/+}$, $Htt^{Q111/+}$, $Htt^{Q140/+}$, and $Htt^{Q175/+}$. We also used data from their mouse tissue survey: $Htt^{+/+}$ and $Htt^{Q175/+}$ at six months of age. Eight individual mice were used for each CAG length per mouse age. The tissue survey at six months of age includes 11 additional tissues including 5 brain regions (brainstem, cerebellum, hippocampus, hypothalamus plus thalamus, and corpus callosum) and 6 peripheral tissues (white gonadal adipose, white intestinal adipose, brown adipose, skin, heart and gastrocnemius muscle). For tissue survey data analyses, $Htt^{+/+}$ were compared against $Htt^{Q175/+}$ mice for each tissue. We then extracted expression values for all metabolic genes defined in the iMM1415 model [21]. The iMM1415 model is a functional metabolic model consisting of 1415 genes encoding 2,212 gene-associated reactions, and 99 sub-systems or pathways. The model remains the most complete mouse metabolic model to date. FPKM values were then log 10 transformed.

Human gene expression data

We utilized a publicly available microarray gene expression data from GSE3790 to compare metabolic pathway changes between mouse models and human disease. Briefly, GSE3790 includes human postmortem data from HD patients and controls assayed using Affymetrix U133A and B microarrays. It consists of 36 samples from caudate nucleus of HD patients and 30 age- and sex matched controls [20].

Identification of differentially expressed metabolic pathways using PathWave

We ran PathWave 2.1 [22] on log transformed FPKM data to discover differentially expressed metabolic pathways (DEPs). PathWave integrates transcriptomic data with network topology to identify both global (pathway level) and local (regulatory shifts within pathways) dysregulations of gene expression. Local regulatory changes include both up and downregulation of genes within an individual pathway, and takes into account the order of metabolic reactions. PathWave therefore not only provides multiple testing corrected *P*-values for each metabolic pathway, but also the details of up- and down-regulated reactions within these pathways.

For each of the three time points (two, six, and ten-months of age) four different groups of long CAG-length mice ($Htt^{Q92/+}$, $Htt^{Q111/+}$, $Htt^{Q140/+}$, and $Htt^{Q175/+}$) were compared against baseline controls ($Htt^{+/+}$ pooled together with $Htt^{Q20/+}$). Because pathway-level changes are reliant on significant differential gene expression change, we chose to exclude the Q80 mouse comparison due to the low number of differentially expressed genes reported by Langfelder et al. To analyze the tissue survey data, for each tissue, long CAG-length mice ($Htt^{Q175/+}$) were compared against wild-type mice ($Htt^{+/+}$).

Here, we ran PathWave using the following parameters: preprocessed pathways was set to either “mmu” or “hsa” depending on whether mouse or human data was being analyzed, number of permutations was set to 10,000, a filter size of three was used, and a multiple

testing corrected *P*-value cut-off of 1 was set. We also restricted PathWave to only test metabolic pathways for mouse defined in KEGG. Of the 1415 genes present in the iMM1415 genome-scale mouse model, only 10 of these genes were not present in KEGG. These 10 genes are pseudogenes. These parameters allowed us to calculate a *P*-value for each KEGG metabolic pathways. Log10 transformed *P*-values were then plotted as a heatmap using *seaborn* (version 0.7.0), a python visualization package.

Identification of transcriptional regulators of CAG-repeat length dependent metabolic genes

We extracted all metabolic genes that showed CAG-repeat length dependent increase or decrease in fold change at each time-point. Specifically, a gene was identified as CAG-repeat length dependent if their fold change values correlated with the four Q-length comparisons. We then used these genes as input for oPPOSUM 3.0, an online tool to detect over-represented conserved transcription factor binding sites [23]. We then mined oPPOSUM output to select sets of target genes of all transcription factor motif matches with a stringent Z-score cut-off of greater than 4.00.

Robustness testing of PathWave results

We also carried out leave-one-out robustness testing to determine whether PathWave results were driven by single (outlier) samples. For every Q-length comparison, we ran PathWave with n-1 samples (n being the total number of samples at that Q length and time point available for analysis) to calculate the number of DEPs. This was repeated iteratively in a sliding window manner to calculate the number of DEPs from all combinations of n-1 samples.

6.4 Results

Metabolic pathway changes in mouse striatum as a result of *Htt* CAG expansion

To understand metabolic network topology changes that result from *Htt* CAG repeat expansion, we applied PathWave to expression data from mouse striatum sampled at 2, 6 and 10 months of age. Our choice of using PathWave for this analysis was motivated by (a) its emphasis on biological mechanisms via metabolic network integration, and (b) its increased sensitivity over GSEA [22]. This allowed us to discover several striking trends in the data.

Using PathWave we tested 79 KEGG metabolic pathways for differential expression, and found that 44 of these pathways demonstrate CAG-repeat length dependent changes (Figure 1). In line with the expectation that increasing CAG length is positively correlated with gene expression changes, we found the least number of differentially expressed pathways (DEPs) at $P \leq 0.05$ in Q92 mice and the most DEPs in Q175 mice. We also noticed a similar trend in the ten month old mice (Figure 1a). Leave-one-out robustness testing, in which we iteratively removed one sample from each Q length and time point and re-calculated the significance, suggested that these DEP numbers are not critically dependent on specific mouse samples (Figure S1).

Gene changes within metabolic pathways

Each KEGG metabolic pathway that is differentially expressed can be viewed in higher resolution at the gene level to understand particular contributors to overall network signal. We provide more detailed KEGG diagrams with gene expression changes for each gene component for 15 DEPs from the mouse striatum comparison (Figures S2 through S17). We also provide gene expression based heatmaps for all 44 metabolic pathways exhibiting CAG-repeat length dependence, which allow us to identify key gene-level expression changes (Figures S18 through S56).

Metabolic changes in mouse striatum over time

To understand early, proximal changes to the transcriptome that result from the HD mutation, we defined three distinct groups of pathways based on the mouse age at which a given pathway shows CAG-length dependent increase and reaches statistical significance ($P \leq 0.05$, Figure 1b). Thus, we have pathways showing CAG-length dependence in both six- and ten-month old mice (group I), only at the six-month time point (group II), or only at the ten-month point (group III). Specifically, we report 14 group I, 13 group II, and 17 group III metabolic pathways.

Interestingly, two metabolic pathways, (i) alanine, aspartate, and glutamate metabolism, and (ii) sphingolipid metabolism, have marginally sub-threshold *P*-values in Q175 mice at two months. There are 16 and 15 total reactions annotated in KEGG to these pathways out of which 9 and 7 reactions were identified as significantly differentially regulated at $P \leq 0.05$ by PathWave. This could implicate these two metabolic pathways as some of the first to become dysregulated and thus more proximal to mutant *Htt*-driven changes.

Metabolic changes in cortex and liver tissue

We next analyzed metabolic pathway changes in mouse cortex and liver, which were also sampled at 2, 6 and 10 months of age. In cortex, a total of 69 DEPs were identified at any time point / allele comparison. 22 of these DEPs were discovered in 10-month-old Q175 mice, while relatively little signal was observed in two-month-old mice. However, in contrast to 44 CAG- and age-dependent DEPs in striatum, just two of these 69 DEPs -- i. starch and sucrose metabolism and ii. cysteine and methionine metabolism -- showed consistent dysregulation both with increasing CAG length and across the six and ten-month time points.

In liver, a total of 15 DEPs reached statistical significance in any time point and allele comparison (Figure 2). However, none of these pathways increased in significance with

increasing CAG length or age. Thus, we find these results to be less meaningful than DEPs identified in striatum or cortex because HD pathogenesis is known to be CAG-length and time dependent.

Comparison of striatal metabolic changes between human and mouse

To understand metabolic changes from mouse data that were reflective of human disease, we performed a parallel analysis on transcriptomic data from human post-mortem tissue of HD patient caudate nucleus and controls (Hodges et al 2006). We found 56 DEPs that reached significance. Among these, we report the top ten metabolic DEPs in the human striatum identified by PathWave together with their differential expression status across various mouse Q lengths (Table 1). Among the top ten pathways, we have maximal overlap between the mouse Q175 striatum comparison at six months, followed by Q140 at six months. Interestingly, citrate cycle, a central carbon pathway which is a DEP in human striatum, also seems to be significantly perturbed in the majority of mouse Q lengths at both six and ten months.

Metabolic consequences of mutant *Htt* in 13 mouse tissues

PathWave analysis of RNA-seq data from wild type and Q175 mice at six months of age also identified several trends (Figure 3). We were able to identify DEPs in nine out of 13 tissues tested. In addition to striatum, cerebral cortex, and liver, we found DEPs in cerebellum, brain stem, gonadal adipose, hippocampus, hypothalamus, and adipose brown tissue of Q175 mice. Also, we found the maximum number of DEPs in striatum, followed by cerebral cortex. In adipose white tissue and heart muscle, we did not find any metabolic pathways with significant differences.

Computational prediction of transcription factor drivers of metabolic genes

To predict potential transcription factor drivers of metabolic gene expression change, we used CAG-repeat length dependent metabolic genes and applied oPPOSUM, a tool that detects over-representation of sequence motifs recognized by transcription factors [23]. This analysis found an overlapping set of TFs (10) that were conserved in regulating metabolic genes at 2, 6 and 10 month time points (Figure 4). A total of 25 TFs had enriched targets at 2 months, and 38 TFs had enriched targets at the later two time points.

Many of the TFs themselves experience changes in expression with increasing CAG length (Figure 4). In order to understand which TFs fall into this category, we next restricted our motif enrichment analysis to CAG-repeat length dependent TFs and metabolic genes and identified a smaller subset of transcriptional regulators of metabolism at two, six, and ten months, respectively (Figure 5, Supplemental Table 1).

We wanted to characterize a particular TF *Egr1* in more detail in order to understand how it's expression and the expression of predicted gene targets are changing. We plotted the Log 2-fold change of *Egr1* for various CAG-repeat length comparisons (Figure 5a) and *Egr1* target genes identified by oPPOSUM that show CAG-repeat length dependent changes (Figure 5d). *Egr1* experiences CAG-repeat length dependent repression and is predicted to regulate more than 100 metabolic gene targets in the mouse brain striatum (Figure 5).

6.5 Discussion

We carried out comprehensive, transcriptome-based analyses of metabolic changes using data from an allelic and tissue series of HD mouse models [3]. We investigated the metabolic changes associated with various *Htt* CAG-repeat lengths in three different mouse tissues across three different time points (two, six, and ten-month old mice). Consistent with known tissue vulnerabilities in HD, we found the greatest number of metabolic changes in

striatum, the next greatest in cortex, and very few changes in liver and other tissues. Several metabolic pathways that change in a CAG-length dependent manner in the striatum are also significantly dysregulated in a human post-mortem study of HD cases vs controls. Furthermore, we predicted transcriptional regulators that are likely to drive metabolic gene expression change in mutant *Htt* expressing mouse tissues.

We discovered that more than half of all tested metabolic pathways in mouse striatum exhibit CAG-repeat length dependence in at least one of the two later time points. This underscores the significant effect of mutant *Htt* in altering brain striatal metabolism, which is the tissue primarily affected in HD. From the tissue pathology perspective, at least some of these metabolic changes are likely to be etiological. But we are currently unable to distinguish metabolic changes etiologic to, and reactive to, tissue pathology. Some of the more specific metabolic changes that we discovered are discussed next.

Alanine, aspartate, and glutamate metabolism is one of the metabolic pathways that showed CAG-repeat length dependence in mouse striatum from our study (Figures 1 and S2, S3). Brain glutamate and aspartate are powerful excitatory neurotransmitters, and their importance to HD pathology has been investigated [24]. HD has been viewed as a disease characterized by abnormal excitotoxicity [25]. Also, abnormal levels of all three metabolites, alanine, aspartate, and glutamate have been previously observed in several human and mouse metabolomic studies [26-32]. Our results demonstrate early differential expression of the alanine, aspartate, and glutamate pathway.

Two of the most significant differentially expressed genes in the longer CAG-repeat length mouse striata are *Cps1* and *Asl* which are part of the alanine, aspartate, and glutamate pathways (Figure S2). Interestingly, these two genes are also part of the urea cycle (module M00029 in KEGG). A recent study had reported severely elevated urea levels in the brains of

HD patients [33]. In another study of metabolic profiles of HD sheep models reported that significant changes in urea along with other metabolites could be used to predict presymptomatic transgenic animals from controls [34]. Lastly, urea cycle deficiency has been reported in the R6/2 HD mouse model [32]. Our observation of urea cycle changes in knock-in mouse models of HD offer additional evidence that it is worth further exploring the role of these metabolic pathways in HD pathology.

Another metabolic pathway with strong CAG-repeat length dependence in the mouse striatum is steroid biosynthesis pathway (Figures 1, S4). Cholesterol biosynthetic pathway dysfunction in HD has been described earlier [15]. Here, we observed CAG-repeat length dependence for the expression of *Cyp51* in six month old mice, a gene that was reported as transcriptionally altered in HD striatum in the Valenza *et al* study. In addition, we discovered two other genes, *Msmo1* and *Cyp27b1*, both involved in cholesterol biosynthesis, that also exhibit CAG-repeat length dependence in six month old mouse striatum. Our study thus extends the earlier one, and implicates additional genes as contributors to altered cholesterol biosynthesis.

We were able to predict several early transcription factor drivers that may underlie some of the metabolic changes we observe in the mouse striatum. The five TF regulators of metabolism at two months may represent some of the earliest drivers of pathogenesis in this mouse model of HD. Downstream cascades triggered by these events can be prevented by interventions targeted at more proximal mechanisms effecting metabolic gene expression, early in the process. Thus, these five TFs warrant further exploration.

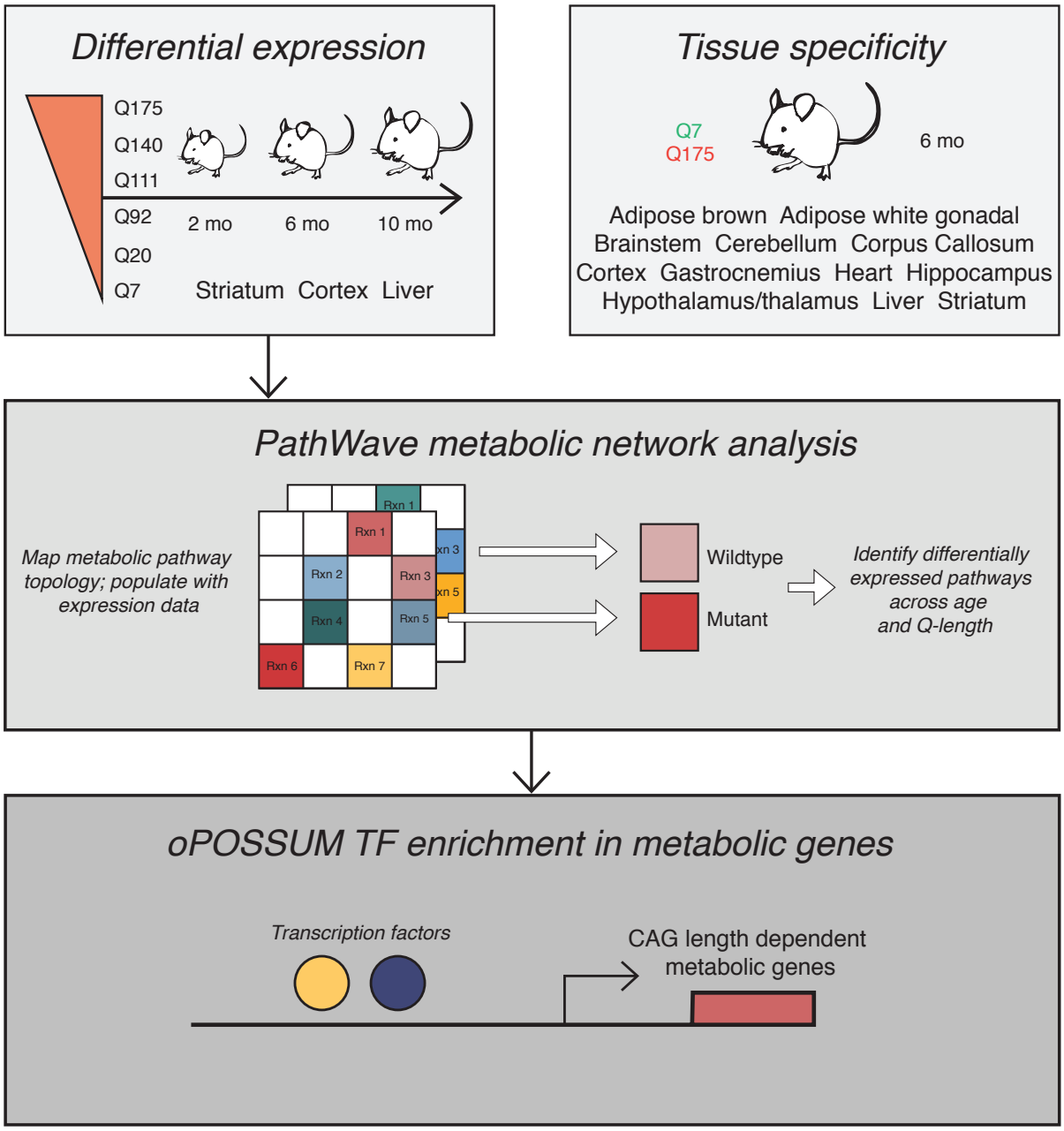
A key question in the HD field is addressing the spectrum of differential toxicities brought about by the mutant HTT protein [35]. This is especially relevant to understanding the acute brain striatal toxicity that mutant HTT seems to possess. Based on our study, we put forth

multiple hypotheses to explain how metabolic dysregulation effects neuronal death: (1) Increased striatal burden of metabolic dysregulation. This is supported by our finding an abundance of DEPs in the striatum as compared to other tissues. The increased metabolic stress may trigger cell death cascades in this tissue. The exact nature and form of cell death in neurodegenerative disease remains obscure [36]. However, metabolic stress might trigger apoptosis in neurons, and a component of apoptotic cell death has been implicated in neuronal cell death in degenerative disease [37]. This is to be differentiated from oxidative stress which has been suggested to be causally linked to neuronal cell death in the striatum [38]. (2) Specific metabolic pathways become differentially expressed earlier and more severely in striatum as compared to other tissues. There seems to be some evidence for this from our study since we are unable to detect any DEPs in two- and six-month cortex and liver. (3) A third hypothesis could be that while metabolic dysregulation occurs generally across multiple tissues as a result of the HTT mutation, other tissues are much more effective than the striatum in re-establishing homeostasis.

In summary, our analyses support previously observed metabolic abnormalities in the mouse striatum, and also identified several novel perturbed pathways. Importantly, we put forward a hypothesis for transcriptional drivers of metabolic gene expression change based on differentially expressed TFs that may underlie some of these metabolic abnormalities.

6.6 Figure Legends

Graphical Abstract.



Sl #	Pathway	Q92 (6m)	Q111 (6m)	Q140 (6m)	Q175 (6m)	Q92 (10m)	Q111 (10m)	Q140 (10m)	Q175 (10m)	Human
1	Glycosaminoglycan biosynthesis - chondroitin sulfate	1.6	0.0	2.5	4.8	0.0	0.0	0.5	0.0	8.0
2	beta-Alanine metabolism	0.0	2.7	0.8	2.8	0.0	0.8	0.0	1.2	7.6
3	Histidine metabolism	0.0	0.0	2.3	7.4	0.0	1.4	0.4	1.6	6.9
4	Valine, leucine and isoleucine degradation	0.0	0.0	1.4	2.7	0.0	2.1	0.9	3.3	6.9
5	Glycerophospholipid metabolism	0.0	3.8	1.9	3.1	0.0	0.0	1.1	0.0	6.6
6	Citrate cycle (TCA cycle)	1.0	4.2	5.9	5.6	0.0	2.7	3.4	4.5	6.4
7	Glutathione metabolism	0.4	2.4	4.9	6.7	0.0	0.6	2.2	1.5	6.2
8	Amino sugar and nucleotide sugar metabolism	0.0	2.3	2.7	3.5	0.5	1.9	1.8	3.2	5.9
9	Thiamine metabolism	0.0	1.9	6.3	3.4	0.0	1.0	2.2	2.5	5.9
10	Purine metabolism	0.0	0.0	0.0	0.0	0.0	1.0	1.5	0.9	5.7

Table 1. Top ten metabolic DEPs in the Hodges post-mortem data identified by PathWave. -log 10 P values are shown for these as well as for various allelic series Q length comparisons. A Q length comparison column was included in the table only if at least one of the pathways had a PathWave calculated multiple testing corrected P values ≤ 0.05 (~ -log 10 P value of 1.30, marked in red).

Figure 1. Heatmap (b) of –log10 multiple testing corrected *P*-values for each metabolic pathway from PathWave analysis of mouse brain striatal transcriptomic data. For each of the three time points (two, six, and ten-months of age) four different CAG-repeat length mice (Q92, Q111, Q140, and Q175) were compared against baseline controls (Q7s pooled together with Q20s). n = 8 for each CAG-repeat length per mouse age. Number of pathways meeting statistical significance ($P \leq 0.05$) in each comparison is shown at the top (a). Three distinct groups of pathways (left) were defined based on which mouse age a given pathway shows CAG-repeat length dependent increase in statistical significance: pathways showing CAG-repeat length dependence in (I) both six- and ten-month old mice or, (II) only at the six-month time point or, (III) only at the ten-month point. Results from PathWave analyses of post-mortem human striatal data (Hodges *et al*, 2006) are also shown in the last column.

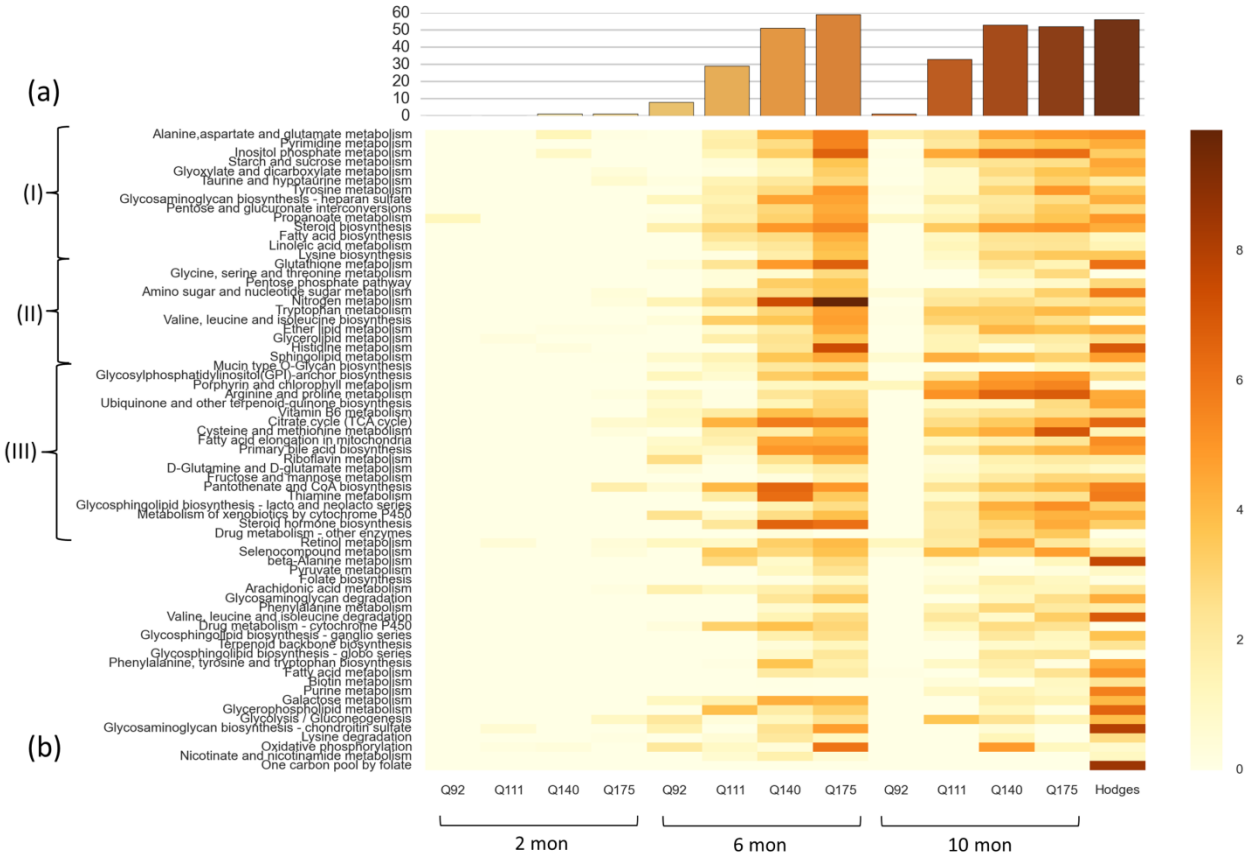


Figure 2. Heatmap (b) of –log10 multiple testing corrected *P*-values for each metabolic pathway from PathWave analysis of mouse brain cortex and liver transcriptomic data. For each of the three time points (two, six, and ten-months of age) four different CAG-length mice (Q92, Q111, Q140, and Q175) were compared against baseline controls (Q7s pooled together with Q20s). n = 8 for each CAG-repeat length per mouse age. Number of pathways meeting statistical significance ($P \leq 0.05$) in each comparison is shown at the top (a). Only the top two pathways show CAG-repeat length dependence and only in brain cortex.

Figure 3. Heatmap (b) of –log10 multiple testing corrected *P*-values for each metabolic pathway from PathWave analysis of transcriptomic data from 13 mouse tissues. For each tissue, long CAG- repeat length mice (Q175s) were compared against wild-type mice (Q7s). n = 8 for each CAG-repeat length per mouse age. Number of pathways meeting statistical significance ($P \leq 0.05$) in each comparison is shown at the top (a).

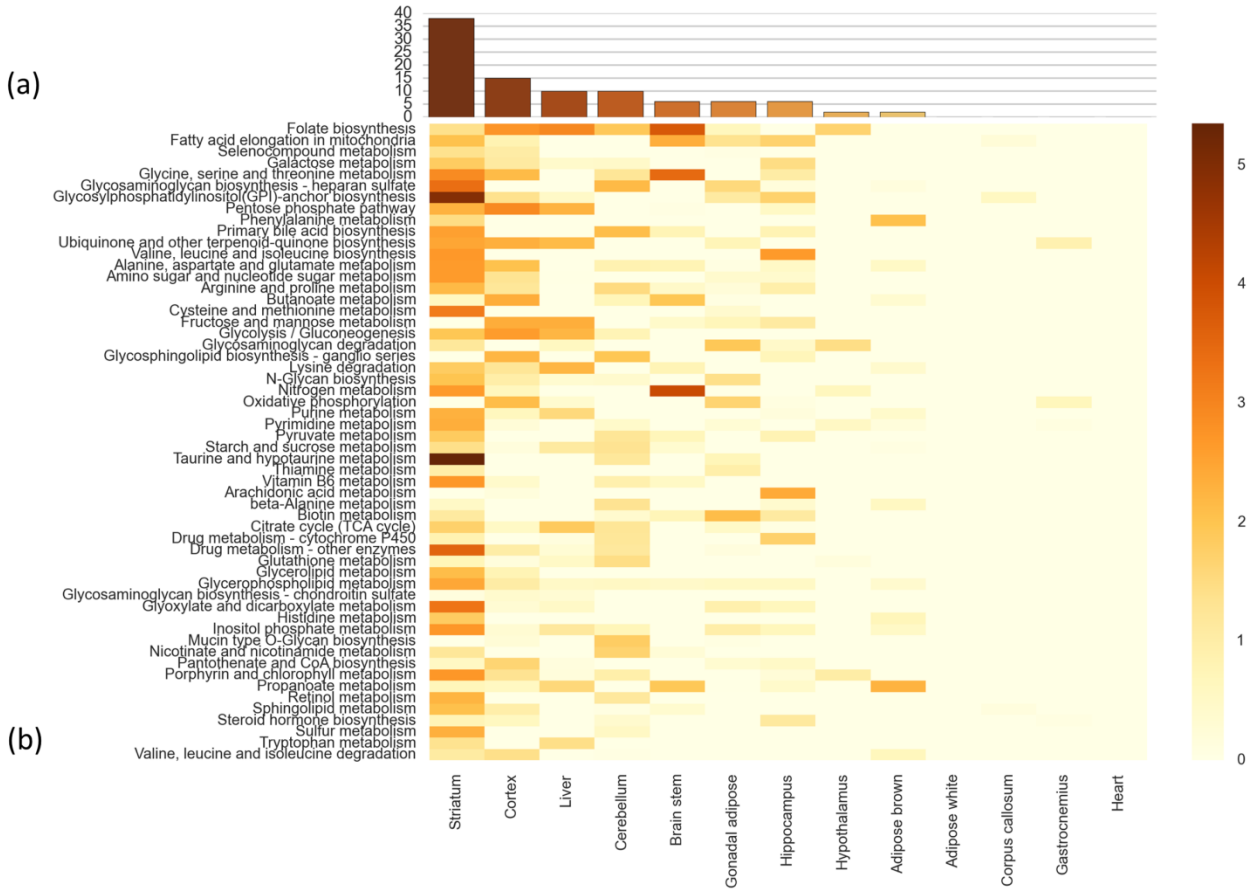


Figure 4. Venn diagram representing overlap of significant TFs (Z-score>4.0) with enrichment for binding sites in CAG-length dependent metabolic genes from 2, 6, and 10 month time points in mouse striatum.

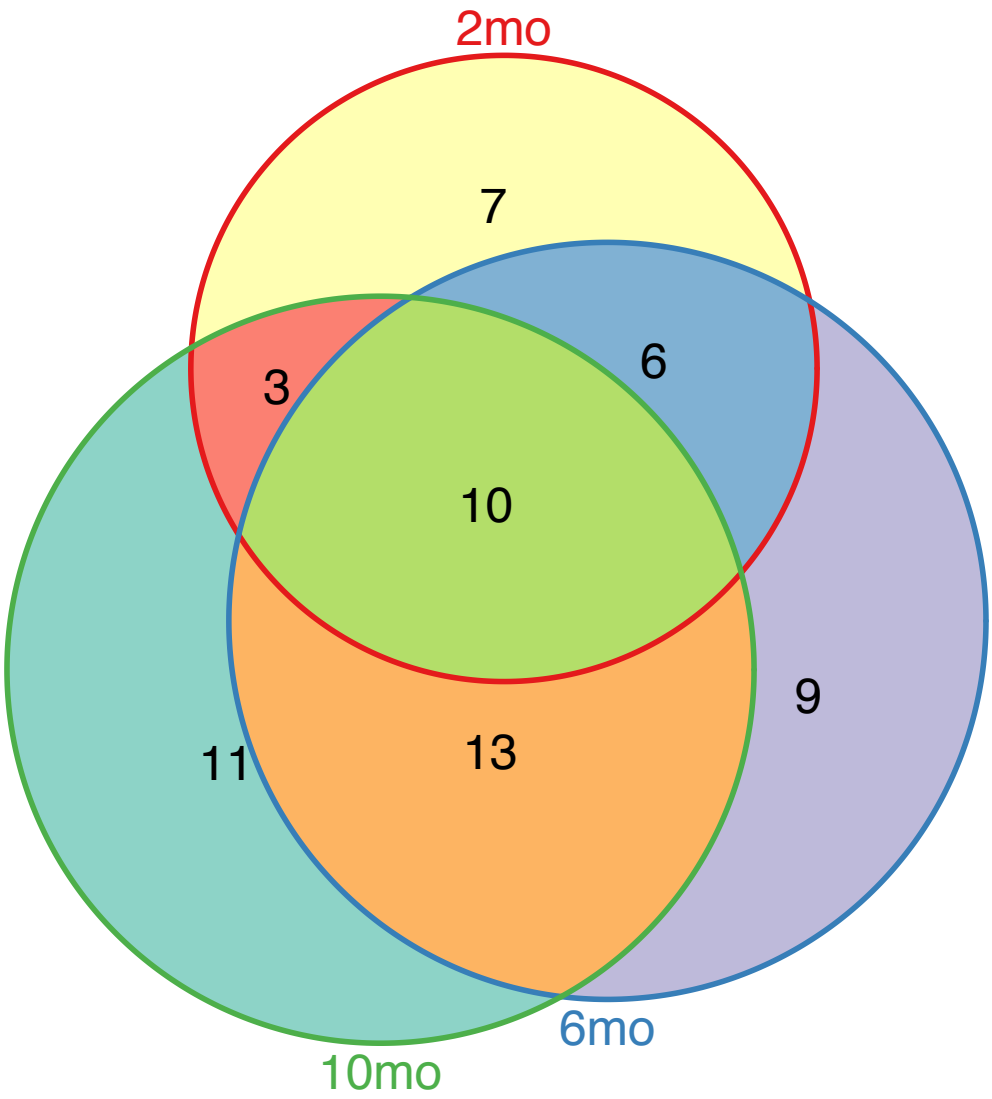


Figure 5. Heatmap of log 2 fold changes of transcription factors (TFs) predicted to at least partly regulate transcription of metabolic genes in the allelic series mouse striatum. Log 2 fold changes for these gene transcripts were obtained from the Langfelder *et al* study [3]. Y-axis: CAG-repeat length dependent TFs from the three different time points (2m, 6m, and 10m), X-axis: Mouse striatal comparisons of each CAG-repeat length against baseline from which log 2 fold changes were calculated in that study.

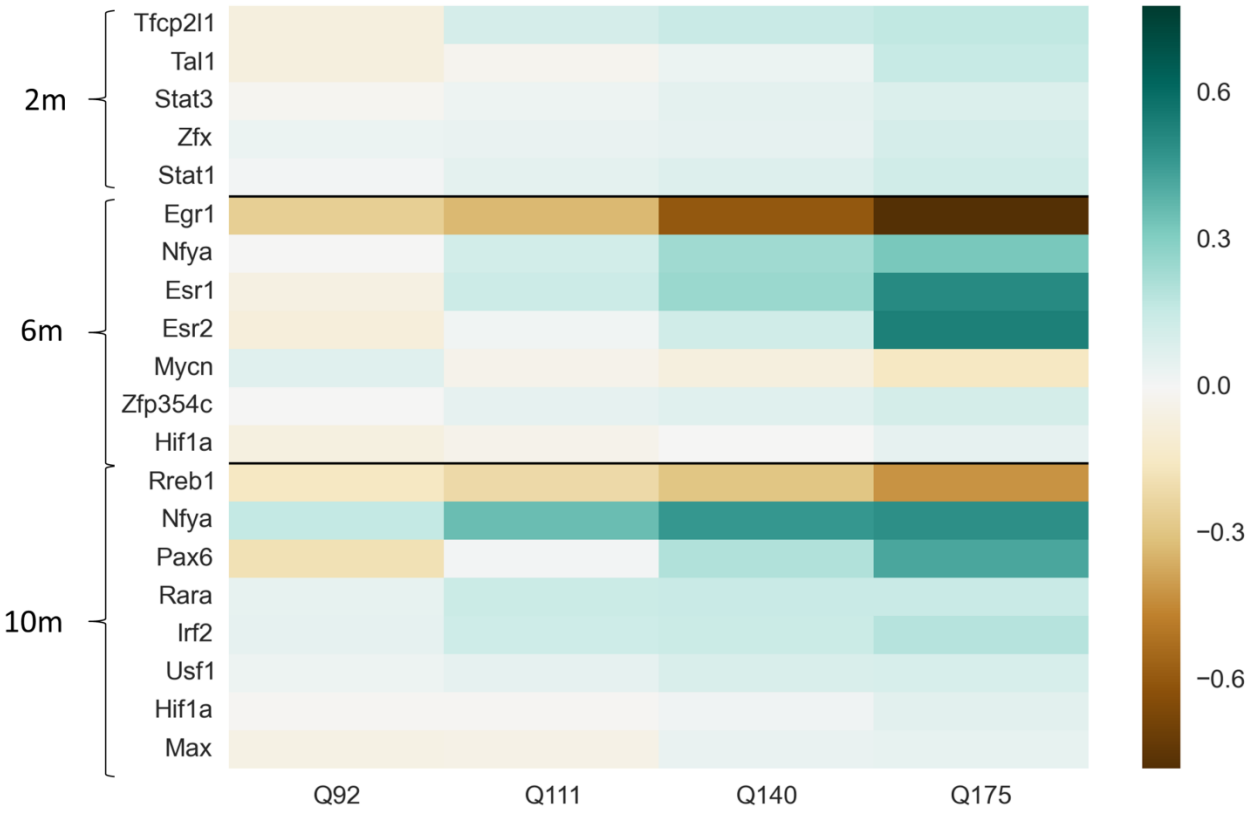


Figure 6. Log 2 fold change of (a) Egr1 TF for various CAG-repeat length comparisons with Q20 as the baseline, (b) Egr1 target genes identified by oPPOSUM that show CAG-repeat length dependence. Egr1 binding motif Logos (c), and distance from peak center (d).

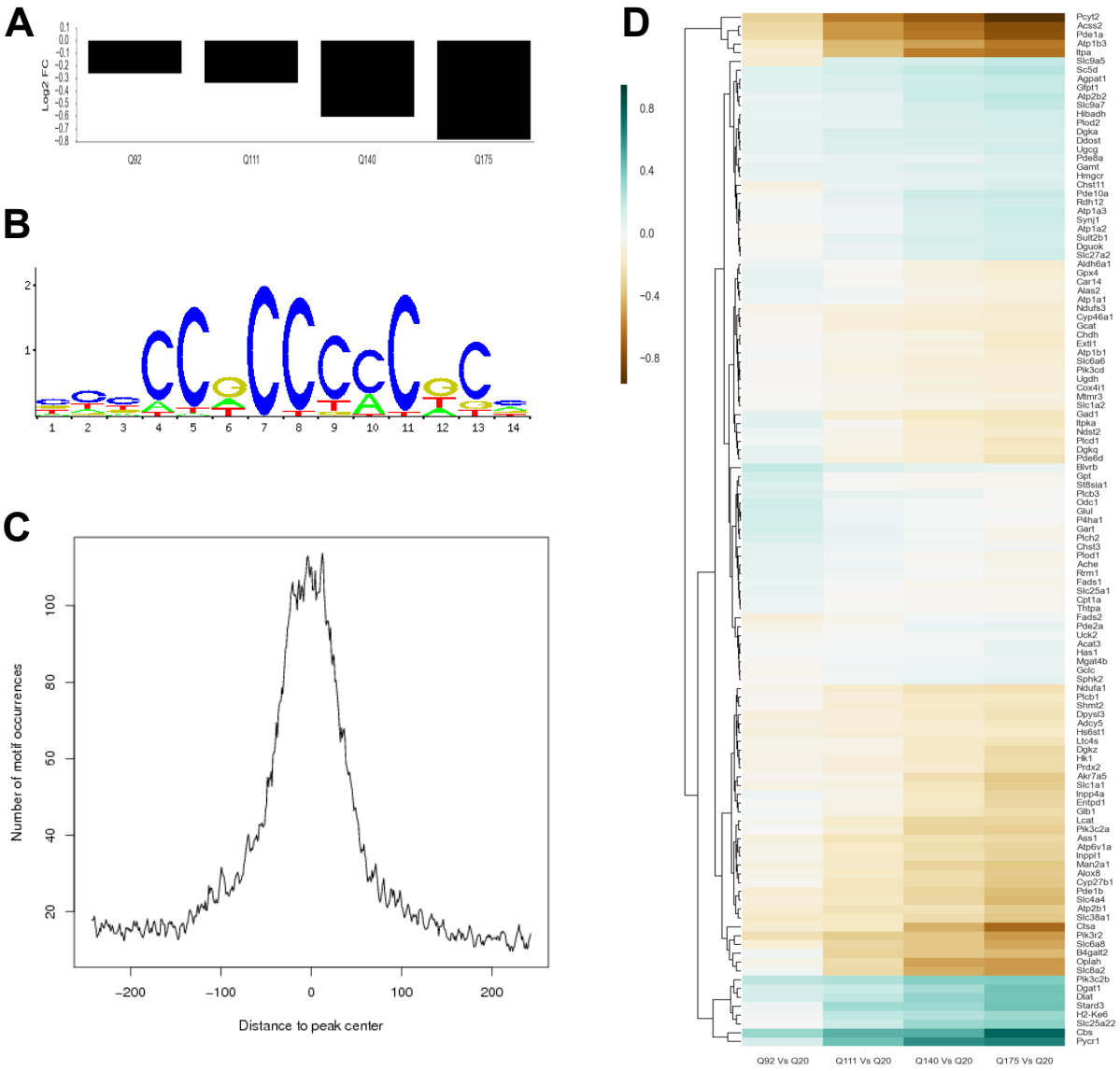

A
Log2 FC
Q92
Q111
Q140
Q175
B
C
Number of motif occurrences
Distance to peak center
D
Q92 Vs Q20
Q111 Vs Q20
Q140 Vs Q20
Q175 Vs Q20

Figure S1. Number of differentially expressed pathways that meet statistical significance ($P \leq 0.05$) during leave-one-out robustness testing. PathWave was used on transcriptomic data from three different mouse tissues to identify differentially expressed metabolic pathways. For each of the three time points (two, six, and ten-months of age) four different CAG-repeat length mice (Q92, Q111, Q140, and Q175) were compared against baseline controls (Q7s pooled together with Q20s). Eight samples were available for each CAG length per mouse age. For each comparison, we systematically held back each of the 16 samples, and calculated the number of differentially expressed metabolic pathways meeting statistical significance ($P \leq 0.05$). Results from PathWave analysis of all 16 combinations per comparison are plotted for each tissue and mouse age.

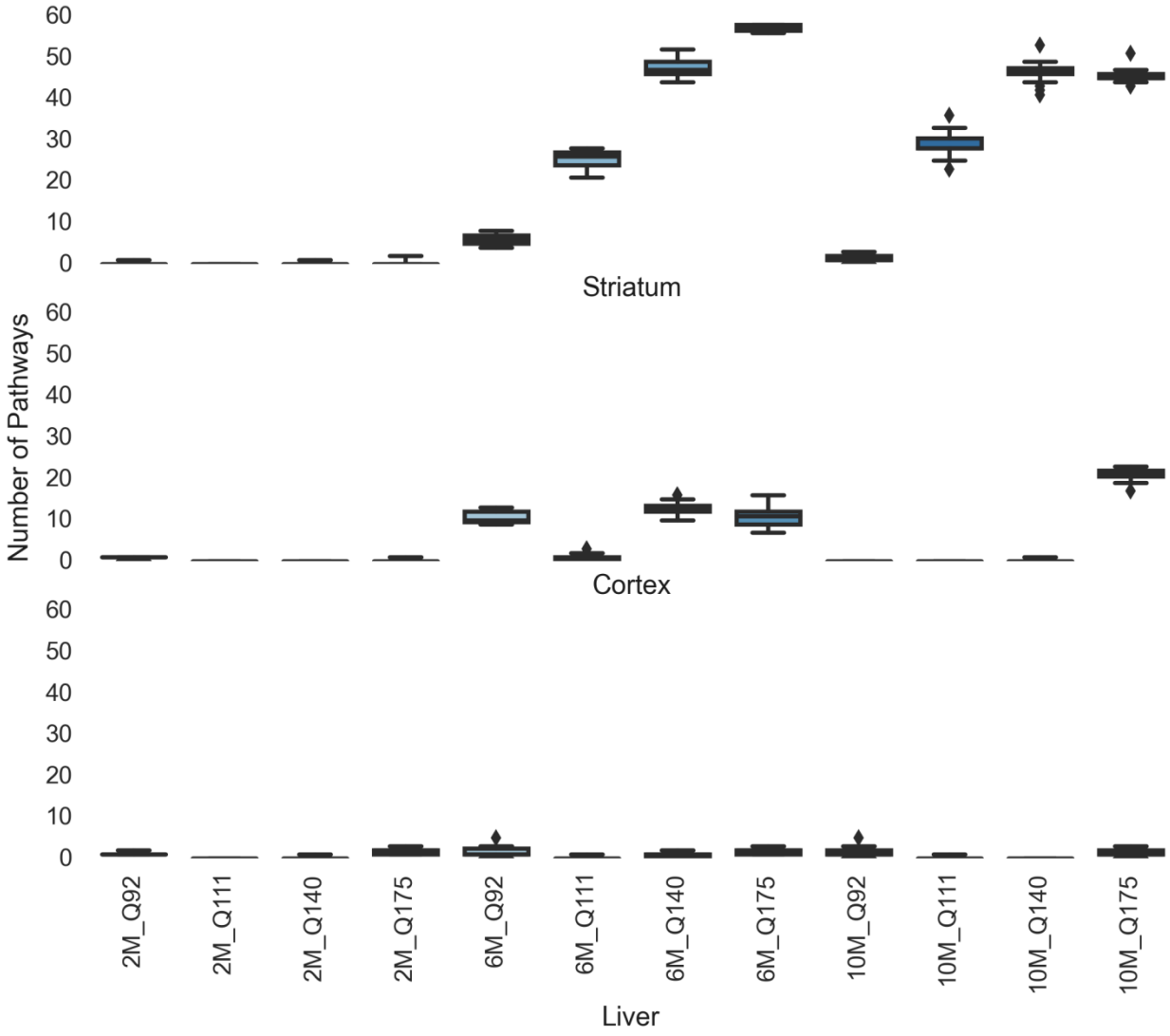


Table S1.

SL #	TF gene name	Z score
1	Tcfcp2l1	8.9
2	Tal1	6.2
3	Stat3	5.5
4	Zfx	5.1
5	Stat1	4.8
6	Egr1	6.3
7	Nfya	7.0
8	Esr1	8.4
9	Esr2	6.1
10	Mycn	4.2
11	Zfp354c	9.1
12	Hif1a	5.7
13	Rreb1	4.3
14	Nfya	8.0
15	Pax6	4.2
16	Rara	4.6
17	Irf2	4.4
18	Usf1	4.7
19	Hif1a	6.3
20	Max	7.0

Table 1. TF gene names and their Z-scores as identified by oPPOSUM regulating metabolic genes in a CAG-repeat length dependent manner. Gene expression data from two month (TFs 1 through 5), six month (TFs 6 through 12) and ten month (TFs 13 through 20) allelic series mouse striatum were used to identify these TFs.

Figures S2 through S17. KEGG metabolic pathway diagrams from PathWave analysis of baseline control mouse striatum (Q7s + Q20s) against longer CAG-repeat length ones (Q140s) at six months of age. Color codes for enzyme expression are as follows: upregulated in green, downregulated in red, no significant change in grey, and no evidence in white.
Figures S18 through S56. Heatmaps of differentially expressed genes in each metabolic pathway in mouse striatum. For each of the three time points (two, six, and ten-months of age) four different CAG-repeat length mice (Q92, Q111, Q140, and Q175) were compared against baseline controls (Q7s pooled together with Q20s). n = 8 for each CAG-repeat length per time point. Scaled fold changes are shown here, red indicates upregulation while blue is downregulation.

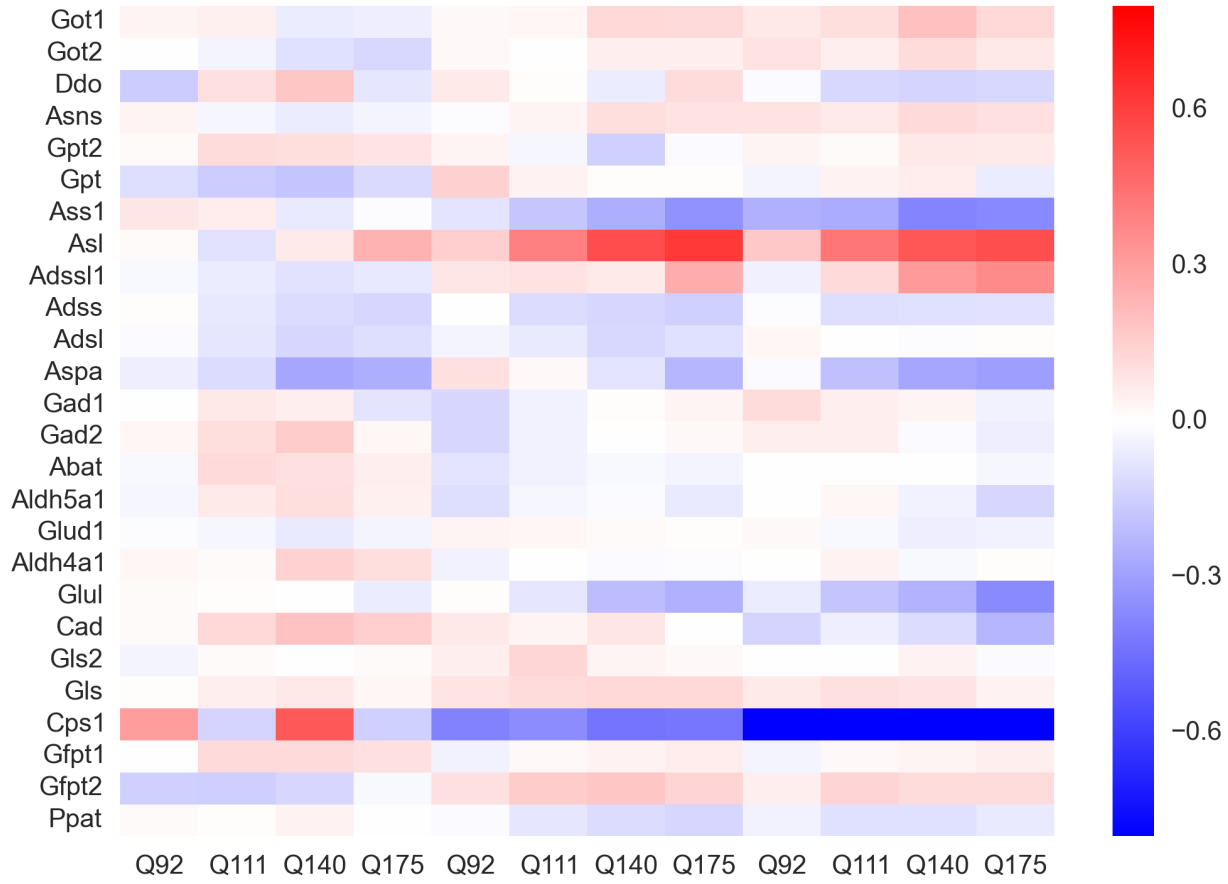

Got1
Got2
Ddo
Asns
Gpt2
Gpt
Ass1
Asl
Adssl1
Adss
Adsl
Aspa
Gad1
Gad2
Abat
Aldh5a1
Glud1
Aldh4a1
Glul
Cad
Gls2
Gls
Cps1
Gfpt1
Gfpt2
Ppat
Q92 Q111 Q140 Q175 Q92 Q111 Q140 Q175 Q92 Q111 Q140 Q175
0.6
0.3
0.0
−0.3
−0.6

Figure S3.

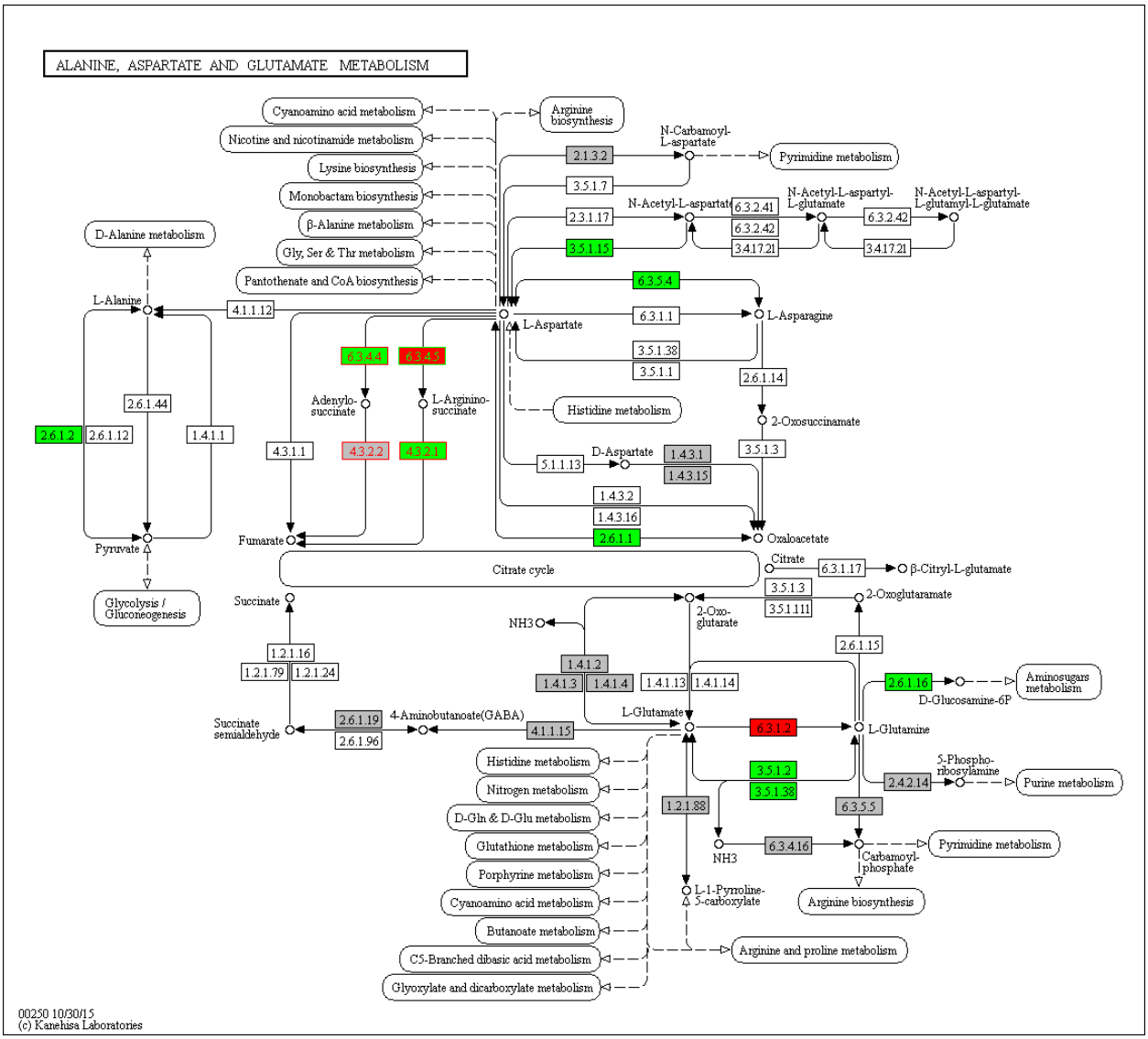
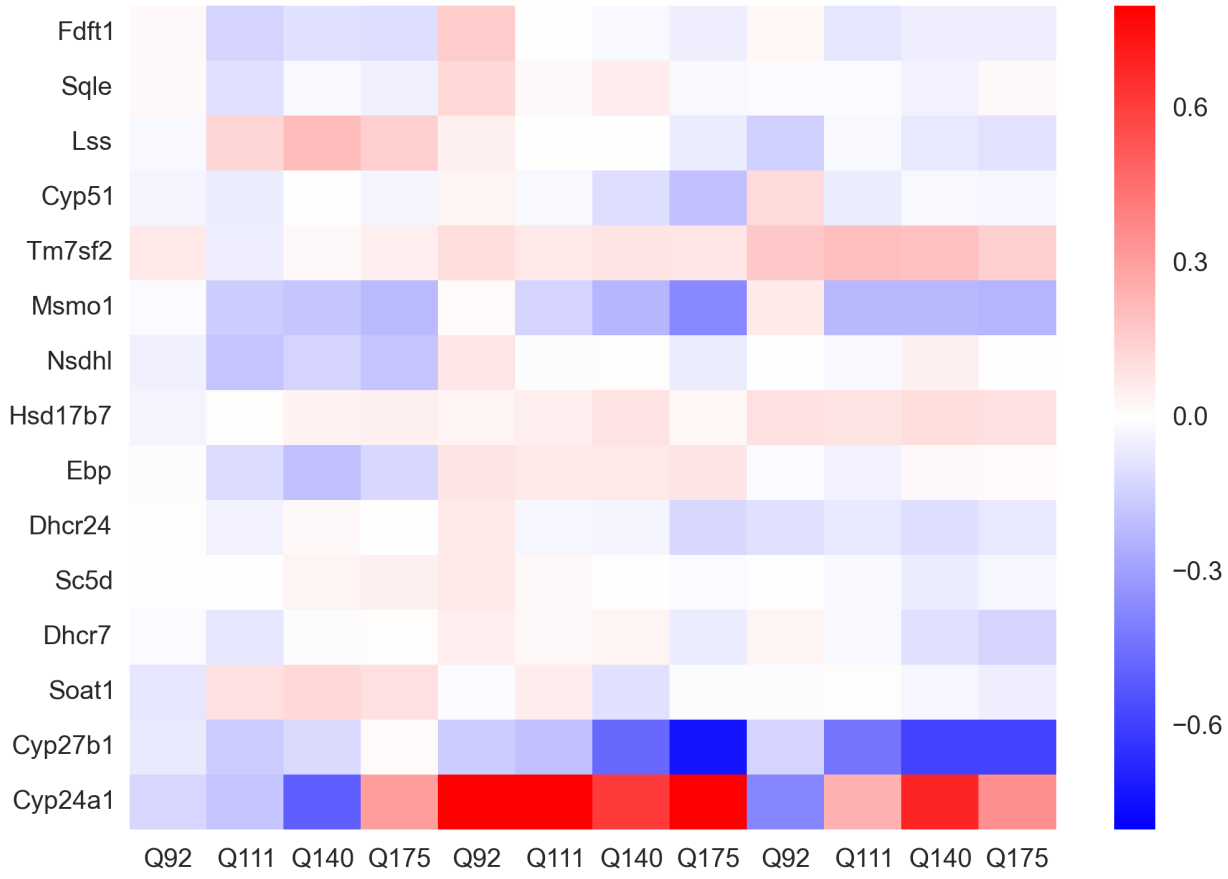

ALANINE, ASPARTATE AND GLUTAMATE METABOLISM
Cyanoamino acid metabolism
Nicotine and nicotinamide metabolism
Lysine biosynthesis
Monobactam biosynthesis
β-Alanine metabolism
Gly, Ser & Thr metabolism
Pantothenate and CoA biosynthesis
Arginine biosynthesis
N-Carbamoyl-L-aspartate
Pyrimidine metabolism
N-Acetyl-L-aspartate
N-Acetyl-L-aspartyl-L-glutamate
N-Acetyl-L-aspartyl-L-glutamyl-L-glutamate
D-Alanine metabolism
L-Alanine
L-Aspartate
L-Asparagine
Adenylo-succinate
L-Arginino-succinate
Histidine metabolism
2-Oxosuccinamate
D-Aspartate
Fumarate
Oxaloacetate
Citrate cycle
Citrate
β-Citryl-L-glutamate
Pyruvate
Glycolysis / Gluconeogenesis
Succinate
2-Oxo-glutarate
2-Oxoglutaramate
NH3
Succinate semialdehyde
4-Aminobutanoate(GABA)
L-Glutamate
L-Glutamine
Aminosugars metabolism
D-Glucosamine-6P
5-Phospho-ribosylamine
Purine metabolism
Histidine metabolism
Nitrogen metabolism
D-Gln & D-Glu metabolism
Glutathione metabolism
Porphyrine metabolism
Cyanoamino acid metabolism
Butanoate metabolism
C5-Branched dibasic acid metabolism
Glyoxylate and dicarboxylate metabolism
L-1-Pyrroline-5-carboxylate
Arginine and proline metabolism
Carbamoyl-phosphate
Pyrimidine metabolism
Arginine biosynthesis
00250 10/30/15
(c) Kanehisa Laboratories

Figure S4.


Fdft1
Sqle
Lss
Cyp51
Tm7sf2
Msmo1
Nsdhl
Hsd17b7
Ebp
Dhcr24
Sc5d
Dhcr7
Soat1
Cyp27b1
Cyp24a1
Q92
Q111
Q140
Q175
Q92
Q111
Q140
Q175
Q92
Q111
Q140
Q175
0.6
0.3
0.0
−0.3
−0.6

6.7 Notes

This work was performed in collaboration with Ramkumar Hariharan, Dani E Bergey, Jeffrey B Carroll, Seth A Ament, Leroy Hood and Nathan D Price.

7 Conclusion

7.1 Summary

Understanding how networks of genes, transcription factors, and variation in regulatory loci contribute to human disease is a central tenant of modern genetics. The generation of large genomic datasets including global transcript levels, regions of open chromatin, and binding sites of sequence specific TFs (among many other types of sequencing data), has allowed us to integrate information across these inputs and outputs of gene expression differences and ask several questions addressed in this thesis. Among these questions, in the context of HD, what gene expression changes happen first and what are the dynamics of those changes over time/age? How do these longitudinal changes differ across different genetic backgrounds? In the context of HD and psychiatric disorders, can we predict TF drivers of complex gene expression change? Using genetic and transcriptomic information, can we rank and validate these predictions? Importantly, can we distill central hypotheses about functional disease mechanisms from large, global genomic datasets?

My thesis addresses the first question posed; how does a monogenic (HD mutation) affect gene expression longitudinally on a weekly time scale. Importantly, I observed robust differences in gene expression change across four different genetic backgrounds. This emphasized the need to incorporate genetic variation in studies of human disease in mouse and animal models; and connects the potentially protective effects specific genetic variants in the mouse to those identified in a human GWAS study (GEM consortium, 2015).

My work goes on to examine how specific TFs might be mediating early gene expression changes in HD using the integration of DNaseI footprinting and RNA-seq datasets. In this

study, we were able to rank predicted gene regulatory drivers based on the enrichment of differentially expressed genes in sets of genes predicted to be regulated by a given TF. Using these *in silico* hypotheses, I went on to map the global genomic binding sites of a particular driver TF, SMAD3, from *in vivo* mouse striatal tissue. The findings from this ChIP-seq data validated our predictions about gene targets of SMAD3, and found a global decrease in the binding of SMAD3 in HD mutant samples.

My work goes on to apply a similar framework to understanding gene regulatory drivers in psychiatric disorders. In this study, we integrate GWAS information in addition to enrichment of differentially expressed genes in TF modules. Furthermore, we predict which SNPs might be functional by intersecting them with TF binding sites. These *in silico* predictions generate a set of hypotheses about contributors to brain gene expression change, and I go on to validate several of these both *in vitro* using primary human neural stem cells and luciferase reporter assays.

This work establishes new hypotheses for the involvement of gene networks, TF drivers, and regulatory variation in neurologic disease, and sets the stage for many directions of follow-up studies.

7.2 Discussion and Future Directions

Improving TRNs

The methodology applied in this thesis incorporates footprinting information and co-expression of TFs and their target genes to predict directed edges in a network of >700 TFs and >10,000 target genes. Our current model is constrained to TF binding sites within +/- 10 kilobases from any given transcription start site. Therefore, this model is a survey of proximal promoter interactions, and could be strengthened by the addition of distal enhancer inputs.

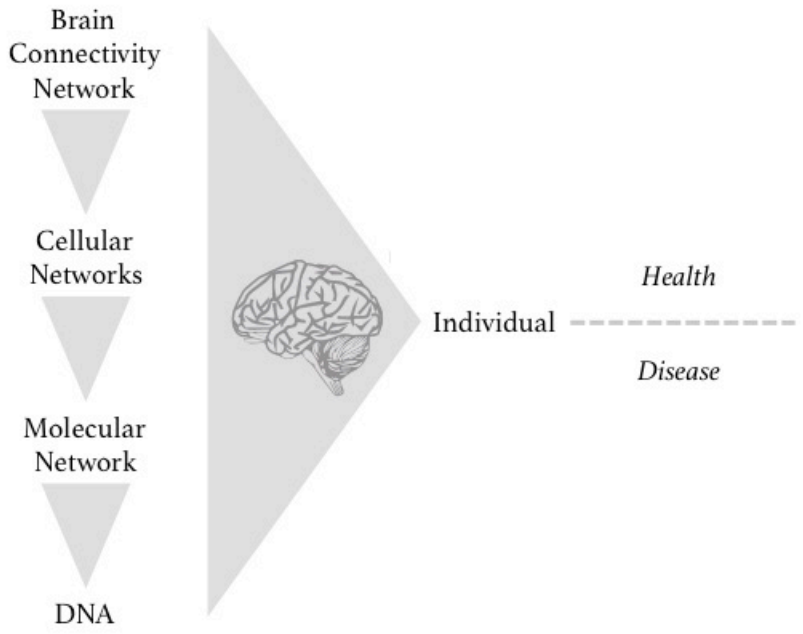
Regulation of genes is controlled through many biological mechanisms in addition to TF interaction with promoters and enhancers. The incorporation of miRNAs into our model could further improve our understanding of the system.

Connecting networks to disease pathology and phenotype

My work attempts to connect TF networks and hypotheses to disease pathology in a few ways. It should be noted that the connection established is primarily one of validating existing data and differentially expressed genes. Disease pathology is of course, a multi-scalar phenotype consisting of complex changes that occur on many different levels: gene networks, cellular networks, and tissue-level interactions which are then expressed in the form of an individual's symptoms such as behavior or movement disorders. Making the connection from gene network to neurologic symptom within the current framework of my research and the model systems applied is immensely challenging. However, moving beyond the validation-style studies I have accomplished is critical in connecting gene regulation to disease. I believe moving into human-derived organoids is one potential approach to getting closer to cellular

network phenotypes. New animal models of disease could also help with our understanding of behavior phenotypes assayed as a result of genetic changes.

Figure. Multiple scales to understanding brain disease pathology



Functional exploration of disease-associated regulatory networks and interactions

For my future work, I would like to demonstrate hierarchies of contributing factors to complex gene expression changes in brain diseases. Thus far, computational models are able to make predictions about regulatory loci or transcription factor drivers that might be contributing to gene expression changes. We are able to rank these predictions based on significance, but we do not have an experimental system in which to validate these lists appropriately. I would like to develop experimental pipelines for screening these predictions that not only tests the contributions of individual loci or TFs but also combinations of these that might be synergistic. The ultimate goal of delineating a hierarchy of drivers in disease will be to help direct therapeutic interventions. Importantly, I hope that the 'read-out' from

this approach will be to point us to which sites or TFs might have the highest impact as druggable targets, or most proximal to mechanistic changes.

Treating networks

A recurring criticism of the generation of gene network hypotheses about disease is the challenge in treating or targeting these networks. One important direction for the research presented here is an investigation of the ability to correct misregulation of specific networks using therapeutics; for example, using small-molecule drug screens in a disease-relevant *in vitro* or *in vivo* system.

8 References

8.1 Introduction References

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012 Sep 6;489(7414):101–8.
2. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27–30.
3. Eddy JA, Hood L, Price ND, Geman D. Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC). PLoS Comput Biol. 2010 May 27;6(5):e1000792.
4. Haynes BC, Maier EJ, Kramer MH, Wang PI, Brown H, Brent MR. Mapping functional transcription factor networks from gene expression data. Genome Res. 2013 Aug;23(8):1319–28.
5. Brent MR. Past Roadblocks and New Opportunities in Transcription Factor Network Mapping. Trends Genet. 2016 Nov;32(11):736–50.
6. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods. 2009 Apr;6(4):283–9.
7. Boyle AP, Song L, Lee B-K, London D, Keefe D, Birney E, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011 Mar;21(3):456–64.
8. Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, et al. Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. Cell Rep. 2014 Sep 25;8(6):2015–30.
9. Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science. 2014 Nov 21;346(6212):1007–12.
10. Wilken MS, Brzezinski JA, La Torre A, Siebenthall K, Thurman R, Sabo P, et al. DNase I hypersensitivity analysis of the mouse brain and retina identifies region-specific regulatory elements. Epigenetics Chromatin. 2015 Feb 28;8:8.
11. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012 Sep 6;489(7414):83–90.
12. Vierstra J, Reik A, Chang K-H, Stehling-Sun S, Zhou Y, Hinkley SJ, et al. Functional footprinting of regulatory DNA. Nat Methods. 2015 Oct;12(10):927–30.
13. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012 Sep 7;337(6099):1190–5.
14. Gusmao EG, Allhoff M, Zenke M, Costa IG. Analysis of computational footprinting methods for DNase sequencing experiments. Nat Methods. 2016 Apr;13(4):303–9.
15. Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res. 2013 Nov;41(21):e201.

16. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol. 2014 Jan 19;32(2):171–8.
17. Anderson WF, Ohlendorf DH, Takeda Y, Matthews BW. Structure of the cro repressor from bacteriophage lambda and its interaction with DNA. Nature. 1981 Apr 30;290(5809):754–8.
18. Dynan WS, Tjian R. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. Cell. 1983 Nov;35(1):79–87.
19. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009 Apr;10(4):252–63.
20. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D108–10.
21. Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res. 2009 Jan;37(Database issue):D77–82.
22. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2010 Jan;38(Database issue):D105–10.
23. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat Rev Genet. 2012 Jul 10;13(8):537–51.
24. Krystal JH, State MW. Psychiatric disorders: diagnosis to therapy. Cell. 2014 Mar 27;157(1):201–14.

8.2 Chapter 3 References

1. Bates J., Tabrizi S. and Jones L. eds. (2014) Huntington's Disease Oxford University Press, USA.

2. MacDonald M., Ambrose C. and Duyao M. (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.

3. Duyao M., Ambrose C., Myers R., Novelletto A., Persichetti F., Frontali M., Folstein S., Ross C., Franz M., Abbott M., *et al.* (1993) Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat. Genet.*, **4**, 387–392.

4. Andrew S.E.S., Goldberg Y., Kremer B., Paul Goldberg Y., Kremer B., Telenius H., Theilmann J., Adam S., Starr E., Squitieri F., *et al.* (1993) The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nat. Genet.*, **4**, 398–403.

5. Langbehn D.R., Brinkman R.R., Falush D., Paulsen J.S. and Hayden M.R. (2004) A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin. Genet.*, **65**, 267–77.

6. Lee J.-M., Ramos E.M., Lee J.-H., Gillis T., Mysore J.S., Hayden M.R., Warby S.C., Morrison P., Nance M., Ross C.A., *et al.* (2012) CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology*, **78**, 690–5.

7. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium (2015) Identification of

Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell*, **162**, 516–26.

8. Vonsattel J.P., Myers R.H., Stevens T.J., Ferrante R.J., Bird E.D. and Richardson E.P. (1985) Neuropathological classification of Huntington's disease. *J. Neuropathol. Exp. Neurol.*, **44**, 559–77.

9. Smith R., Brundin P. and Li J.-Y. (2005) Synaptic dysfunction in Huntington's disease: a new perspective. *Cell. Mol. Life Sci.*, **62**, 1901–1912.

10. Möller T. (2010) Neuroinflammation in Huntington's disease. *J. Neural Transm.*, **117**, 1001–1008.

11. Valenza M., Rigamonti D., Goffredo D., Zuccato C., Fenu S., Jamot L., Strand A., Tarditi A., Woodman B., Racchi M., *et al.* (2005) Dysfunction of the cholesterol biosynthetic pathway in Huntington's disease. *J. Neurosci.*, **25**, 9932–9.

12. Koroshetz W.W.J., Jenkins B.B.G., Rosen B.R. and Beal M.F. (1997) Energy metabolism defects in Huntington's disease and effects of coenzyme Q10. *Ann. Neurol.*, **41**, 160–165.

13. Zuccato C., Valenza M. and Cattaneo E. (2010) Molecular mechanisms and potential therapeutical targets in Huntington's disease. *Physiol. Rev.*, **90**, 905–81.

14. Hodges A., Strand A.D., Aragaki A.K., Kuhn A., Sengstag T., Hughes G., Elliston L.A., Hartog C., Goldstein D.R., Thu D., *et al.* (2006) Regional and cellular gene expression changes in human Huntington's disease brain. *Hum. Mol. Genet.*, **15**, 965–77.

15. Seredenina T. and Luthi-Carter R. (2012) What have we learned from gene expression profiles in Huntington's disease? *Neurobiol. Dis.*, **45**, 83–98.

16. Langfelder P., Cantle J.P., Chatzopoulou D., Wang N., Gao F., Al-Ramahi I., Lu X.-H., Ramos E.M., El-Zein K., Zhao Y., *et al.* (2016) Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. *Nat. Neurosci.*, **19**, 623–33.

17. Giralt A., Puigdellívol M., Carretón O., Paoletti P., Valero J., Parra-Damas A., Saura C.A., Alberch J. and Ginés S. (2012) Long-term memory deficits in Huntington's disease are associated with reduced CBP histone acetylase activity. *Hum. Mol. Genet.*, **21**, 1203–16.

18. Hölter S.M., Stromberg M., Kovalenko M., Garrett L., Glasl L., Lopez E., Guide J., Götz A., Hans W., Becker L., *et al.* (2013) A broad phenotypic screen identifies novel phenotypes driven by a single mutant allele in Huntington's disease CAG knock-in mice. *PLoS One*, **8**, e80923.

19. Yhnell E., Dunnett S.B. and Brooks S.P. (2016) A Longitudinal Motor Characterisation of the HdhQ111 Mouse Model of Huntington's Disease. *J. Huntingtons. Dis.*, **5**, 149–61.

20. Kuhn A., Goldstein D.R., Hodges A., Strand A.D., Sengstag T., Kooperberg C., Becanovic K., Pouladi M.A., Sathasivam K., Cha J.-H.J., *et al.* (2007) Mutant huntingtin's effects on striatal gene expression in mice recapitulate changes observed in human Huntington's disease brain and do not differ with mutant huntingtin length or wild-type huntingtin dosage. *Hum. Mol. Genet.*, **16**, 1845–61.

21. Hwang D., Lee I.Y., Yoo H., Gehlenborg N., Cho J.-H., Petritis B., Baxter D., Pitstick R., Young R., Spicer D., *et al.* (2009) A systems approach to prion disease. *Mol. Syst. Biol.*, **5**, 252.

22. Lee J.-M., Pinto R.M., Gillis T., St Claire J.C. and Wheeler V.C. (2011) Quantification of age-dependent somatic CAG repeat instability in Hdh CAG knock-in mice reveals different

expansion dynamics in striatum and liver. *PLoS One*, **6**, e23647.

23. Lloret A., Dragileva E., Teed A., Espinola J., Fossale E., Gillis T., Lopez E., Myers R.H., MacDonald M.E. and Wheeler V.C. (2006) Genetic background modifies nuclear mutant huntingtin accumulation and HD CAG repeat instability in Huntington's disease knock-in mice. *Hum. Mol. Genet.*, **15**, 2015–24.

24. Wheeler V.C., Auerbach W., White J.K., Srinidhi J., Auerbach A., Ryan A., Duyao M.P., Vrbanac V., Weaver M., Gusella J.F., *et al.* (1999) Length-dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse. *Hum. Mol. Genet.*, **8**, 115–22.

25. White J.K., Auerbach W., Duyao M.P., Vonsattel J.P., Gusella J.F., Joyner A.L. and MacDonald M.E. (1997) Huntingtin is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion. *Nat. Genet.*, **17**, 404–10.

26. Wheeler V.C., White J.K., Gutekunst C.A., Vrbanac V., Weaver M., Li X.J., Li S.H., Yi H., Vonsattel J.P., Gusella J.F., *et al.* (2000) Long glutamine tracts cause nuclear localization of a novel form of huntingtin in medium spiny striatal neurons in HdhQ92 and HdhQ111 knock-in mice. *Hum. Mol. Genet.*, **9**, 503–13.

27. Gutekunst C.A., Li S.H., Yi H., Mulroy J.S., Kuemmerle S., Jones R., Rye D., Ferrante R.J., Hersch S.M. and Li X.J. (1999) Nuclear and neuropil aggregates in Huntington's disease: relationship to neuropathology. *J. Neurosci.*, **19**, 2522–34.

28. Wheeler V.C., Gutekunst C.-A., Vrbanac V., Lebel L.-A., Schilling G., Hersch S., Friedlander R.M., Gusella J.F., Vonsattel J.-P., Borchelt D.R., *et al.* (2002) Early phenotypes that presage late-onset neurodegenerative disease allow testing of modifiers in Hdh CAG knock-in mice. *Hum. Mol. Genet.*, **11**, 633–40.

29. Pinto R.M., Dragileva E., Kirby A., Lloret A., Lopez E., St Claire J., Panigrahi G.B., Hou C., Holloway K., Gillis T., *et al.* (2013) Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. *PLoS Genet.*, **9**, e1003930.

30. Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–9.

31. Dougherty J.D., Schmidt E.F., Nakajima M. and Heintz N. (2010) Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.*, **38**, 4218–30.

32. Doyle J.P., Dougherty J.D., Heiman M., Schmidt E.F., Stevens T.R., Ma G., Bupp S., Shrestha P., Shah R.D., Doughty M.L., *et al.* (2008) Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell*, **135**, 749–62.

33. Collaborative Cross Consortium C.C., Aylor D.L., Valdar W., Foulds-Mathes W., Buus R.J., Verdugo R.A., Ayroles J.F., Carbone M.A., Stone E.A., Jordan K.W., *et al.* (2012) The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics*, **190**, 389–401.

34. Svenson K., Gatti D., Valdar W. and Welsh C. (2012) High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics*, **190**, 437–47.

35. Cha J.H. (2000) Transcriptional dysregulation in Huntington's disease. *Trends Neurosci.*, **23**, 387–92.

36. Sugars K.L., Brown R., Cook L.J., Swartz J. and Rubinsztein D.C. (2004) Decreased cAMP response element-mediated transcription: an early event in exon 1 and full-length cell models of Huntington's disease that contributes to polyglutamine pathogenesis. *J. Biol. Chem.*, **279**, 4988–99.

37. Thomas E.A. (2006) Striatal specificity of gene expression dysregulation in Huntington's disease. *J. Neurosci. Res.*, **84**, 1151–64.

38. Benn C.L., Sun T., Sadri-Vakili G., McFarland K.N., DiRocco D.P., Yohrling G.J., Clark T.W., Bouzou B. and Cha J.-H.J. (2008) Huntingtin Modulates Transcription, Occupies Gene Promoters In Vivo, and Binds Directly to DNA in a Polyglutamine-Dependent Manner. *J. Neurosci.*, **28**, 10720–33.

39. Freiman R.N. and Tjian R. (2002) Neurodegeneration. A glutamine-rich trail leads to transcription factors. *Science*, **296**, 2149–50.

40. Zhai W., Jeong H., Cui L., Krainc D. and Tjian R. (2005) In vitro analysis of huntingtin-mediated transcriptional repression reveals multiple transcription factor targets. *Cell*, **123**, 1241–53.

41. Seong I.S., Woda J.M., Song J.-J., Lloret A., Abeyrathne P.D., Woo C.J., Gregory G., Lee J.-M., Wheeler V.C., Walz T., *et al.* (2010) Huntingtin facilitates polycomb repressive complex 2. *Hum. Mol. Genet.*, **19**, 573–83.

42. Kovalenko M., Dragileva E., St. Claire J., Gillis T., Guide J.R., New J., Dong H., Kucherlapati R., Kucherlapati M.H., Ehrlich M.E., *et al.* (2012) Msh2 Acts in Medium-Spiny Striatal Neurons as an Enhancer of CAG Instability and Mutant Huntingtin Phenotypes in Huntington's Disease Knock-In Mice. *PLoS One*, **7**, e44273.

43. Carpenter A.E., Jones T.R., Lamprecht M.R., Clarke C., Kang I., Friman O., Guertin D.A., Chang J., Lindquist R.A., Moffat J., *et al.* (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, **7**, R100.

44. Lee J.-M., Zhang J., Su A.I., Walker J.R., Wiltshire T., Kang K., Dragileva E., Gillis T., Lopez E.T., Boily M.-J., *et al.* (2010) A novel approach to investigate tissue-specific trinucleotide repeat instability. *BMC Syst. Biol.*, **4**, 29.

45. Johnson W.E., Li C. and Rabinovic A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–27.

46. Smyth G.K.G. (2005) Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer-Verlag, New York, pp. 397–420.

47. Benjamini Y., Drai D. and Elmer G. (2001) Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.*, **125**, 279–284.

8.3 Chapter 4 References

1. Alexandrov V, Brunner D, Menalled LB, Kudwa A, Watson-Johnson J, Mazzella M, Russell I, Ruiz MC, Torello J, Sabath E, Sanchez A, Gomez M, Filipov I, Cox K, Kwan M, Ghavami A, Ramboz S, Lager B, Wheeler VC, Aaronson J, et al (2016) Large-scale phenome analysis defines a behavioral signature for Huntington's disease genotype in

mice. *Nat. Biotechnol.* **34:** 838–44
2. Arlotta P, Molyneaux BJ, Jabaudon D, Yoshida Y & Macklis JD (2008) Ctip2 controls the differentiation of medium spiny neurons and the establishment of the cellular architecture of the striatum. *J. Neurosci.* **28:** 622–32
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM & Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–9
4. Bailey TL & Machanick P (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.* **40:** e128
5. Battaglia G, Cannella M, Riozzi B, Orobello S, Maat-Schieman ML, Aronica E, Busceti CL, Ciarmiello A, Alberti S, Amico E, Sassone J, Sipione S, Bruno V, Frati L, Nicoletti F & Squitieri F (2011) Early defect of transforming growth factor β1 formation in Huntington's disease. *J. Cell. Mol. Med.* **15:** 555–71
6. Becanovic K, Pouladi MA, Lim RS, Kuhn A, Pavlidis P, Luthi-Carter R, Hayden MR & Leavitt BR (2010) Transcriptional changes in Huntington disease identified using genome-wide expression profiling and cross-platform analysis. *Hum. Mol. Genet.* **19:** 1438–52
7. Benn CL, Sun T & Sadri-Vakili G Huntingtin Modulates Transcription, Occupies Gene Promoters In Vivo, and Binds Directly to DNA in a Polyglutamine-Dependent Manner. **28:**
8. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS & Thorsson V (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* **7:** R36
9. Carty N, Berson N, Tillack K, Thiede C, Scholz D, Kottig K, Sedaghat Y, Gabrysiak C, Yohrling G, von der Kammer H, Ebneth A, Mack V, Munoz-Sanjuan I & Kwak S (2015) Characterization of HTT inclusion size, location, and timing in the zQ175 mouse model of Huntington's disease: an in vivo high-content imaging study. *PLoS One* **10:** e0123527
10. Deng YP, Wong T, Bricker-Anthony C, Deng B & Reiner A (2013) Loss of corticostriatal and thalamostriatal synaptic terminals precedes striatal projection neuron pathology in heterozygous Q140 Huntington's disease mice. *Neurobiol. Dis.* **60:** 89–107
11. Dickey AS, Pineda V V, Tsunemi T, Liu PP, Miranda HC, Gilmore-Hall SK, Lomas N, Sampat KR, Buttgereit A, Torres M-JM, Flores AL, Arreola M, Arbez N, Akimov SS, Gaasterland T, Lazarowski ER, Ross CA, Yeo GW, Sopher BL, Magnuson GK, et al (2015) PPAR-δ is repressed in Huntington's disease, is required for normal neuronal function and can be targeted therapeutically. *Nat. Med.* **22:** 37–45
12. Dougherty JD, Schmidt EF, Nakajima M & Heintz N (2010) Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.* **38:** 4218–30
13. Doyle JP, Dougherty JD, Heiman M, Schmidt EF, Stevens TR, Ma G, Bupp S, Shrestha P, Shah RD, Doughty ML, Gong S, Greengard P & Heintz N (2008) Application of a

translational profiling approach for the comparative analysis of CNS cell types. *Cell* **135:** 749–62

14. Durrenberger PF, Fernando FS, Kashefi SN, Bonnert TP, Seilhean D, Nait-Oumesmar B, Schmitt A, Gebicke-Haerter PJ, Falkai P, Grünblatt E, Palkovits M, Arzberger T, Kretzschmar H, Dexter DT & Reynolds R (2015) Common mechanisms in neurodegeneration and neuroinflammation: a BrainNet Europe gene expression microarray study. *J. Neural Transm.* **122:** 1055–68
15. Fossale E, Seong IS, Coser KR, Shioda T, Kohane IS, Wheeler VC, Gusella JF, MacDonald ME & Lee J-M (2011) Differential effects of the Huntington's disease CAG mutation in striatum and cerebellum are quantitative not qualitative. *Hum. Mol. Genet.* **20:** 4258–67
16. Friedman J, Hastie T & Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*
17. Friedman N, Linial M, Nachman I & Pe'er D (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7:** 601–20
18. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, et al (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* **489:** 91–100
19. Giles P, Elliston L, Higgs G V, Brooks SP, Dunnett SB & Jones L (2012) Longitudinal analysis of gene expression and behaviour in the HdhQ150 mouse model of Huntington's disease. *Brain Res. Bull.* **88:** 199–209
20. Grant CE, Bailey TL & Noble WS (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–8
21. Haury A-C, Mordelet F, Vera-Licona P & Vert J-P (2012) TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst. Biol.* **6:** 145
22. Hodges A, Strand AD, Aragaki AK, Kuhn A, Sengstag T, Hughes G, Elliston LA, Hartog C, Goldstein DR, Thu D, Hollingsworth ZR, Collin F, Synek B, Holmans PA, Young AB, Wexler NS, Delorenzi M, Kooperberg C, Augood SJ, Faull RLM, et al (2006) Regional and cellular gene expression changes in human Huntington's disease brain. *Hum. Mol. Genet.* **15:** 965–77
23. Hsiao H-Y, Chen Y-C, Chen H-M, Tu P-H & Chern Y (2013a) A critical role of astrocyte-mediated nuclear factor-κB-dependent inflammation in Huntington's disease. *Hum. Mol. Genet.* **22:** 1826–42
24. Hsiao H-Y, Chen Y-C, Chen H-M, Tu P-H & Chern Y (2013b) A critical role of astrocyte-mediated nuclear factor-κB-dependent inflammation in Huntington's disease. *Hum. Mol. Genet.* **22:** 1826–42
25. Hume MA, Barrera LA, Gisselbrecht SS & Bulyk ML (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **43:** D117-22
26. Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes (2003) *Nat. ...*

27. Huynh-Thu VA, Irrthum A, Wehenkel L & Geurts P (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5:**
28. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T & Taipale J (2013) DNA-binding specificities of human transcription factors. *Cell* **152:** 327–39
29. Kandasamy M, Reilmann R, Winkler J, Bogdahn U & Aigner L (2011) Transforming Growth Factor-Beta Signaling in the Neural Stem Cell Niche: A Therapeutic Target for Huntington's Disease. *Neurol. Res. Int.* **2011:** 124256
30. Kuhn A, Goldstein DR, Hodges A, Strand AD, Sengstag T, Kooperberg C, Becanovic K, Pouladi MA, Sathasivam K, Cha J-HJ, Hannan AJ, Hayden MR, Leavitt BR, Dunnett SB, Ferrante RJ, Albin R, Shelbourne P, Delorenzi M, Augood SJ, Faull RLM, et al (2007) Mutant huntingtin's effects on striatal gene expression in mice recapitulate changes observed in human Huntington's disease brain and do not differ with mutant huntingtin length or wild-type huntingtin dosage. *Hum. Mol. Genet.* **16:** 1845–61
31. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR & Ma'ayan A (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26:** 2438–44
32. Langfelder P, Cantle JP, Chatzopoulou D, Wang N, Gao F, Al-Ramahi I, Lu X-H, Ramos EM, El-Zein K, Zhao Y, Deverasetty S, Tebbe A, Schaab C, Lavery DJ, Howland D, Kwak S, Botas J, Aaronson JS, Rosinski J, Coppola G, et al (2016) Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. *Nat. Neurosci.* **19:** 623–33
33. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9:** 357–9
34. Li L, Liu H, Dong P, Li D, Legant WR, Grimm JB, Lavis LD, Betzig E, Tjian R & Liu Z (2016) Real-time imaging of Huntingtin aggregates diverting target search and gene transcription. *Elife* **5:** 1–29
35. Luthi-Carter R, Strand A, Peters NL, Solano SM, Hollingsworth ZR, Menon AS, Frey AS, Spektor BS, Penney EB, Schilling G, Ross CA, Borchelt DR, Tapscott SJ, Young AB, Cha JH & Olson JM (2000) Decreased expression of striatal signaling genes in a mouse model of Huntington's disease. *Hum. Mol. Genet.* **9:** 1259–71
36. MacDonald M, Ambrose C & Duyao M (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*
37. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ & Stolovitzky G (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods* **9:** 796–804
38. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R & Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1:** S7
39. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen C, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A &

Wasserman WW (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42:** D142-7

40. Matys V, Kel-Margoulis O V, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE & Wingender E (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34:** D108-10
41. Morton AJ, Wood NI, Hastings MH, Hurelbrink C, Barker RA & Maywood ES (2005) Disintegration of the sleep-wake cycle and circadian timing in Huntington's disease. *J. Neurosci.* **25:** 157–63
42. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, et al (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489:** 83–90
43. Niewiadomska-Cimicka A, Krzyżosiak A, Ye T, Podleśny-Drabiniok A, Dembélé D, Dollé P & Krężel W (2016) Genome-wide Analysis of RARβ Transcriptional Targets in Mouse Striatum Links Retinoic Acid Signaling with Huntington's Disease and Other Neurodegenerative Disorders. *Mol. Neurobiol.*
44. Orsi GA, Kasinathan S, Zentner GE, Henikoff S & Ahmad K (2015) Mapping regulatory factors by immunoprecipitation from native chromatin. *Curr. Protoc. Mol. Biol.* **110:** 21.31.1-25
45. Parker JA, Vazquez-Manrique RP, Tourette C, Farina F, Offner N, Mukhopadhyay A, Orfila A-M, Darbois A, Menet S, Tissenbaum HA & Neri C (2012) Integration of β-catenin, sirtuin, and FOXO signaling protects from mutant huntingtin toxicity. *J. Neurosci.* **32:** 12630–40
46. Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C & Ott S (2013) Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41:** e201
47. Plaisier CL, O'Brien S, Bernard B, Reynolds S, Simon Z, Toledo CM, Ding Y, Reiss DJ, Paddison PJ & Baliga NS (2016) Causal Mechanistic Regulatory Network for Glioblastoma Deciphered Using Systems Genetics Network Analysis. *Cell Syst.*
48. Reiss DJ, Plaisier CL, Wu W-J & Baliga NS (2015) cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism. *Nucleic Acids Res.* **43:** e87
49. Ring KL, An MC, Zhang N, O'Brien RN, Ramos EM, Gao F, Atwood R, Bailus BJ, Melov S, Mooney SD, Coppola G & Ellerby LM (2015) Genomic Analysis Reveals Disruption of Striatal Neuronal Development and Therapeutic Targets in Human Huntington's Disease Neural Stem Cells. *Stem cell reports* **5:** 1023–38
50. Robinson MD, McCarthy DJ & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–40
51. Rothe T, Deliano M, Wójtowicz AM, Dvorzhak A, Harnack D, Paul S, Vagner T, Melnick I, Stark H & Grantyn R (2015) Pathological gamma oscillations, impaired dopamine release, synapse loss and reduced dynamic range of unitary glutamatergic

synaptic transmission in the striatum of hypokinetic Q175 Huntington mice. *Neuroscience* **311:** 519–38

52. Seong IS, Woda JM, Song J-J, Lloret A, Abeyrathne PD, Woo CJ, Gregory G, Lee J-M, Wheeler VC, Walz T, Kingston RE, Gusella JF, Conlon RA & MacDonald ME (2010) Huntingtin facilitates polycomb repressive complex 2. *Hum. Mol. Genet.* **19:** 573–83
53. Seredenina T & Luthi-Carter R (2012) What have we learned from gene expression profiles in Huntington's disease? *Neurobiol. Dis.* **45:** 83–98
54. Shirasaki DI, Greiner ER, Al-Ramahi I, Gray M, Boontheung P, Geschwind DH, Botas J, Coppola G, Horvath S, Loo JA & Yang XW (2012) Network organization of the huntingtin proteomic interactome in mammalian brain. *Neuron* **75:** 41–57
55. Singhrao SK, Neal JW, Morgan BP & Gasque P (1999) Increased complement biosynthesis by microglia and complement activation on neurons in Huntington's disease. *Exp. Neurol.* **159:** 362–76
56. Skene PJ, Illingworth RS, Webb S, Kerr ARW, James KD, Turner DJ, Andrews R & Bird AP (2010) Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Mol. Cell* **37:** 457–68
57. Tabrizi SJ, Scahill RI, Owen G, Durr A, Leavitt BR, Roos RA, Borowsky B, Landwehrmeyer B, Frost C, Johnson H, Craufurd D, Reilmann R, Stout JC, Langbehn DR & TRACK-HD Investigators (2013) Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet. Neurol.* **12:** 637–49
58. Tang B, Becanovic K, Desplats PA, Spencer B, Hill AM, Connolly C, Masliah E, Leavitt BR & Thomas EA (2012) Forkhead box protein p1 is a transcriptional repressor of immune signaling in the CNS: implications for transcriptional dysregulation in Huntington disease. *Hum. Mol. Genet.* **21:** 3097–111
59. Thomas EA, Coppola G & Desplats PA The HDAC inhibitor 4b ameliorates the disease phenotype and transcriptional abnormalities in Huntington's disease transgenic mice. **105:**
60. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (* ...
61. Valenza M, Rigamonti D, Goffredo D, Zuccato C, Fenu S, Jamot L, Strand A, Tarditi A, Woodman B, Racchi M, Mariotti C, Di Donato S, Corsini A, Bates G, Pruss R, Olson JM, Sipione S, Tartari M & Cattaneo E (2005) Dysfunction of the cholesterol biosynthetic pathway in Huntington's disease. *J. Neurosci.* **25:** 9932–9
62. Vonsattel JP, Myers RH, Stevens TJ, Ferrante RJ, Bird ED & Richardson EP (1985) Neuropathological classification of Huntington's disease. *J. Neuropathol. Exp. Neurol.* **44:** 559–77
63. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, Scott LJ & Sartor MA (2014) ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucleic Acids Res.* **42:** e105
64. Wheeler VC, Auerbach W, White JK, Srinidhi J, Auerbach A, Ryan A, Duyao MP, Vrbanac V, Weaver M, Gusella JF, Joyner AL & MacDonald ME (1999) Length-

dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse. *Hum. Mol. Genet.* **8:** 115–22
65. Wheeler VC, White JK, Gutekunst CA, Vrbanac V, Weaver M, Li XJ, Li SH, Yi H, Vonsattel JP, Gusella JF, Hersch S, Auerbach W, Joyner AL & MacDonald ME (2000) Long glutamine tracts cause nuclear localization of a novel form of huntingtin in medium spiny striatal neurons in HdhQ92 and HdhQ111 knock-in mice. *Hum. Mol. Genet.* **9:** 503–13
66. Wingender E, Schoeps T & Dönitz J (2013) TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* **41:** D165-70
67. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, et al (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515:** 355–64
68. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, Phatnani HP, Guarnieri P, Caneda C, Ruderisch N, Deng S, Liddelow SA, Zhang C, Daneman R, Maniatis T, Barres BA & Wu JQ (2014) An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *J. Neurosci.* **34:** 11929–47
69. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W & Liu XS (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9:** R137
70. Zuccato C, Belyaev N, Conforti P, Ooi L, Tartari M, Papadimou E, MacDonald M, Fossale E, Zeitlin S, Buckley N & Cattaneo E (2007) Widespread disruption of repressor element-1 silencing transcription factor/neuron-restrictive silencer factor occupancy at its target genes in Huntington's disease. *J. Neurosci.* **27:** 6972–83
71. Zuccato C, Tartari M, Crotti A, Goffredo D, Valenza M, Conti L, Cataudella T, Leavitt BR, Hayden MR, Timmusk T, Rigamonti D & Cattaneo E (2003) Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nat. Genet.* **35:** 76–83

8.4 Chapter 5 References

1. Zhang, B. *et al.* Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease. *Cell* **153,** 707–720 (2013).
2. Hodges, A. *et al.* Regional and cellular gene expression changes in human Huntington's disease brain. *Hum. Mol. Genet.* **15,** 965–77 (2006).
3. Akula, N. *et al.* RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder. *Mol. Psychiatry* **19,** 1179–1185 (2014).

4. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19,** 1442–1453 (2016).
5. Seifuddin, F. *et al.* Systematic review of genome-wide gene expression studies of bipolar disorder. *BMC Psychiatry* **13,** 213 (2013).
6. Torkamani, A., Dean, B., Schork, N. J. & Thomas, E. A. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. *Genome Res.* **20,** 403–412 (2010).
7. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474,** 380–4 (2011).
8. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506,** 185–90 (2014).
9. Hauberg, M. E. *et al.* Large-Scale Identification of Common Trait and Disease Variants Affecting Gene Expression. *Am. J. Hum. Genet.* **100,** 885–894 (2017).
10. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–5 (2012).
11. Gusev, A. *et al.* Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *Am. J. Hum. Genet.* **95,** 535–552 (2014).
12. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat. Neurosci.* **18,** 199–209 (2015).
13. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515,** 209–215 (2014).
14. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–7 (2014).
15. Hou, L. *et al.* Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Hum. Mol. Genet.* **25,** 3383–3394 (2016).
16. Sklar, P. *et al.* Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43,** 977–83 (2011).
17. Mühleisen, T. W. *et al.* Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat. Commun.* **5,** 3339 (2014).
18. Kamnasaran, D., Muir, W. J., Ferguson-Smith, M. A. & Cox, D. W. Disruption of the neuronal PAS3 gene in a family affected with schizophrenia. *J. Med. Genet.* **40,** 325–32 (2003).
19. Torkamani, A., Dean, B., Schork, N. J. & Thomas, E. A. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. *Genome Res.* **20,** 403–12 (2010).
20. Langfelder, P. *et al.* Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. *Nat. Neurosci.* **19,** 623–33 (2016).
21. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538,** 523–527 (2016).
22. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6,** 283–9 (2009).
23. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90 (2012).
24. Zaharia, M. *et al.* Faster and More Accurate Sequence Alignment with SNAP. (2011).
25. Boyle, A. P., Guinney, J., Crawford, G. E. & Furey, T. S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24,** 2537–8 (2008).
26. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41,** e201 (2013).

27. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42,** D142-7 (2014).
28. Kulakovskiy, I. V. *et al.* HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.* **44,** D116–D125 (2016).
29. Pachkov, M., Balwierz, P. J., Arnold, P., Ozonov, E. & van Nimwegen, E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.* **41,** D214–D220 (2013).
30. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489,** 391–399 (2012).
31. Ding, S.-L. *et al.* Comprehensive cellular-resolution atlas of the adult human brain. *J. Comp. Neurol.* **524,** 3127–3481 (2016).
32. Greenfield, A., Hafemeister, C. & Bonneau, R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics* **29,** 1060–7 (2013).
33. Haury, A.-C., Mordelet, F., Vera-Licona, P. & Vert, J.-P. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst. Biol.* **6,** 145 (2012).
34. Bonneau, R. *et al.* The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* **7,** R36 (2006).
35. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9,** 796–804 (2012).
36. Rubenstein, J. L. (John L. & Rakic, P. *Patterning and cell type specification in the developing CNS and PNS : comprehensive developmental neuroscience*. (Academic Press, 2013).
37. Zhang, Y. *et al.* An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *J. Neurosci.* **34,** 11929–47 (2014).
38. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.* **38,** 4218–30 (2010).
39. Zhou, Q., Choi, G. & Anderson, D. J. The bHLH transcription factor Olig2 promotes oligodendrocyte differentiation in collaboration with Nkx2.2. *Neuron* **31,** 791–807 (2001).
40. Olson, J. M. *et al.* NeuroD2 is necessary for development and survival of central nervous system neurons. *Dev. Biol.* **234,** 174–87 (2001).
41. Wu, S.-X. *et al.* Pyramidal neurons of upper cortical layers generated by NEX-positive progenitor cells in the subventricular zone. *Proc. Natl. Acad. Sci.* **102,** 17172–17177 (2005).
42. Ray, R. D. & Zald, D. H. Anatomical insights into the interaction of emotion and cognition in the prefrontal cortex. *Neurosci. Biobehav. Rev.* **36,** 479–501 (2012).
43. Bicks, L. K., Koike, H., Akbarian, S. & Morishita, H. Prefrontal Cortex and Social Cognition in Mouse and Man. *Front. Psychol.* **6,** 1805 (2015).
44. Hercher, C., Chopra, V. & Beasley, C. L. Evidence for morphological alterations in prefrontal white matter glia in schizophrenia and bipolar disorder. *J. Psychiatry Neurosci.* **39,** 376–85 (2014).
45. Narayanan, M. *et al.* Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. *Mol. Syst. Biol.* **10,** 743 (2014).
46. Salat, D. H., Kaye, J. A. & Janowsky, J. S. Selective preservation and degeneration within the prefrontal cortex in aging and Alzheimer disease. *Arch. Neurol.* **58,** 1403–8 (2001).
47. Glausier, J. R., Fish, K. N. & Lewis, D. A. Altered parvalbumin basket cell inputs in the dorsolateral prefrontal cortex of schizophrenia subjects. *Mol. Psychiatry* **19,** 30–36 (2014).
48. Weinberger, D. & Berman, K. Physiologic dysfunction of dorsolateral prefrontal cortex in schizophrenia: I. Regional cerebral blood flow evidence. *Arch. Gen.* (1986).

49. Reinhart, V. *et al.* Evaluation of TrkB and BDNF transcripts in prefrontal cortex, hippocampus, and striatum from subjects with schizophrenia, bipolar disorder, and major depressive disorder. *Neurobiol. Dis.* **77,** 220–227 (2015).
50. Akula, N. *et al.* RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder. *Mol. Psychiatry* **19,** 1179–85 (2014).
51. Chang, L.-C. *et al.* A Conserved BDNF, Glutamate- and GABA-Enriched Gene Module Related to Human Depression Identified by Coexpression Meta-Analysis and DNA Variant Genome-Wide Association Studies. *PLoS One* **9,** e90980 (2014).
52. Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540,** 423–427 (2016).
53. Wang, M. *et al.* Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer?s disease. *Genome Med.* **8,** 104 (2016).
54. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).
55. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).
56. Ward, L. D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44,** D877–D881 (2016).
57. Glusman, G., Caballero, J., Mauldin, D. E., Hood, L. & Roach, J. C. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* **27,** 3216–7 (2011).
58. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (80-.).* **337,** 1190–1195 (2012).
59. Steen, V. M. *et al.* Genetic evidence for a role of the SREBP transcription system and lipid biosynthesis in schizophrenia and antipsychotic treatment. *Eur. Neuropsychopharmacol.* (2016). doi:10.1016/j.euroneuro.2016.07.011
60. Kochunov, P. & Hong, L. E. Neurodevelopmental and Neurodegenerative Models of Schizophrenia: White Matter at the Center Stage. *Schizophr. Bull.* **40,** 721–728 (2014).
61. Dominguez, M. H., Ayoub, A. E. & Rakic, P. POU-III Transcription Factors (Brn1, Brn2, and Oct6) Influence Neurogenesis, Molecular Identity, and Migratory Destination of Upper-Layer Cells of the Cerebral Cortex. *Cereb. Cortex* **23,** 2632–2643 (2013).
62. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155,** 1008–21 (2013).
63. Bakken, T. E. *et al.* A comprehensive transcriptional map of primate brain development. *Nature* **535,** 367–375 (2016).
64. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci.* **112,** 7285–7290 (2015).
65. van de Leemput, J. *et al.* CORTECON: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. *Neuron* **83,** 51–68 (2014).
66. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
67. Shalizi, A. *et al.* A Calcium-Regulated MEF2 Sumoylation Switch Controls Postsynaptic Differentiation. *Science (80-.).* **311,** 1012–1017 (2006).
68. Pickard, B. S. *et al.* Interacting haplotypes at the NPAS3 locus alter risk of schizophrenia and bipolar disorder. *Mol. Psychiatry* **14,** 874–84 (2009).
69. Glaser, B. *et al.* Identification of a potential Bipolar risk haplotype in the gene encoding

the winged-helix transcription factor RFX4. *Mol. Psychiatry* **10,** 920–927 (2005).
70. Marenco, S. & Weinberger, D. R. The neurodevelopmental hypothesis of schizophrenia: following a trail of evidence from cradle to grave. *Dev. Psychopathol.* **12,** 501–27 (2000).
71. Owen, M. J., O'Donovan, M. C., Thapar, A. & Craddock, N. Neurodevelopmental hypothesis of schizophrenia. *Br. J. Psychiatry* **198,** (2011).
72. Suvà, M. L. *et al.* Reconstructing and Reprogramming the Tumor-Propagating Potential of Glioblastoma Stem-like Cells. *Cell* **157,** 580–594 (2014).
73. Niu, W. *et al.* SOX2 Reprograms Resident Astrocytes into Neural Progenitors in the Adult Brain. *Stem Cell Reports* **4,** 780–794 (2015).
74. Marchetto, M. C. *et al.* Altered proliferation and networks in neural cells derived from idiopathic autistic individuals. *Mol. Psychiatry* **22,** 820–835 (2017).
75. Belinson, H. *et al.* Prenatal β-catenin/Brn2/Tbr2 transcriptional cascade regulates adult social and stereotypic behaviors. *Mol. Psychiatry* **21,** 1417–1433 (2016).
76. Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science (80-.).* **347,** 1465–1470 (2015).
77. Koohy, H., Down, T. A., Spivakov, M. & Hubbard, T. A Comparison of Peak Callers Used for DNase-Seq Data. *PLoS One* **9,** e96303 (2014).
78. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–8 (2011).
79. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res.* **43,** W39–W49 (2015).
80. Wingender, E., Schoeps, T. & Dönitz, J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* **41,** D165-70 (2013).
81. glmnet: Lasso and elastic-net regularized generalized linear models. *R Packag. ...* (2009).
82. Sun, Y. *et al.* Long-term tripotent differentiation capacity of human neural stem (NS) cells in adherent culture. *Mol. Cell. Neurosci.* **38,** 245–58 (2008).
83. Elkabetz, Y. & Studer, L. Human ESC-derived Neural Rosettes and Neural Stem Cell Progression. *Cold Spring Harb. Symp. Quant. Biol.* **73,** 377–387 (2008).
84. Ward, L. D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44,** D877–D881 (2016).

8.5 Chapter 6 References

1 Ross, C. A. & Tabrizi, S. J. Huntington's disease: from molecular pathogenesis to clinical treatment. *Lancet Neurol* **10**, 83-98, doi:10.1016/S1474-4422(10)70245-3 (2011).
2 A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* **72**, 971-983 (1993).
3 Langfelder, P. *et al.* Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. *Nat Neurosci* **19**, 623-633, doi:10.1038/nn.4256 (2016).
4 Myers, R. H. *et al.* De novo expansion of a (CAG)n repeat in sporadic Huntington's disease. *Nat Genet* **5**, 168-173, doi:10.1038/ng1093-168 (1993).
5 Persichetti, F. *et al.* Huntington's disease CAG trinucleotide repeats in pathologically confirmed post-mortem brains. *Neurobiol Dis* **1**, 159-166 (1994).

6 Imarisio, S. *et al.* Huntington's disease: from pathology and genetics to potential therapies. *Biochem J* **412**, 191-209, doi:10.1042/BJ20071619 (2008).
7 Ehrlich, M. E. Huntington's disease and the striatal medium spiny neuron: cell-autonomous and non-cell-autonomous mechanisms of disease. *Neurotherapeutics* **9**, 270-284, doi:10.1007/s13311-012-0112-2 (2012).
8 Labbadia, J. & Morimoto, R. I. Huntington's disease: underlying molecular mechanisms and emerging concepts. *Trends Biochem Sci* **38**, 378-385, doi:10.1016/j.tibs.2013.05.003 (2013).
9 Sawa, A. Mechanisms for neuronal cell death and dysfunction in Huntington's disease: pathological cross-talk between the nucleus and the mitochondria? *J Mol Med (Berl)* **79**, 375-381 (2001).
10 Mochel, F. & Haller, R. G. Energy deficit in Huntington disease: why it matters. *J Clin Invest* **121**, 493-499, doi:10.1172/JCI45691 (2011).
11 Mochel, F. *et al.* Early alterations of brain cellular energy homeostasis in Huntington disease models. *J Biol Chem* **287**, 1361-1370, doi:10.1074/jbc.M111.309849 (2012).
12 Powers, W. J. *et al.* Selective defect of in vivo glycolysis in early Huntington's disease striatum. *Proc Natl Acad Sci U S A* **104**, 2945-2949, doi:10.1073/pnas.0609833104 (2007).
13 Milakovic, T. & Johnson, G. V. Mitochondrial respiration and ATP production are significantly impaired in striatal cells expressing mutant huntingtin. *J Biol Chem* **280**, 30773-30782, doi:10.1074/jbc.M504749200 (2005).
14 Cui, L. *et al.* Transcriptional repression of PGC-1alpha by mutant huntingtin leads to mitochondrial dysfunction and neurodegeneration. *Cell* **127**, 59-69, doi:10.1016/j.cell.2006.09.015 (2006).
15 Valenza, M. *et al.* Dysfunction of the cholesterol biosynthetic pathway in Huntington's disease. *J Neurosci* **25**, 9932-9939, doi:10.1523/JNEUROSCI.3355-05.2005 (2005).
16 Block, R. C., Dorsey, E. R., Beck, C. A., Brenna, J. T. & Shoulson, I. Altered cholesterol and fatty acid metabolism in Huntington disease. *J Clin Lipidol* **4**, 17-23, doi:10.1016/j.jacl.2009.11.003 (2010).
17 Acuna, A. I. *et al.* A failure in energy metabolism and antioxidant uptake precede symptoms of Huntington's disease in mice. *Nat Commun* **4**, 2917, doi:10.1038/ncomms3917 (2013).
18 Weydt, P. *et al.* Thermoregulatory and metabolic defects in Huntington's disease transgenic mice implicate PGC-1alpha in Huntington's disease neurodegeneration. *Cell Metab* **4**, 349-362, doi:10.1016/j.cmet.2006.10.004 (2006).
19 Ament, S. A. *et al.* High resolution time-course mapping of early transcriptomic, molecular and cellular phenotypes in Huntington's disease CAG knock-in mice across multiple genetic backgrounds. *Human Molecular Genetics* **26**, 913-922, doi:10.1093/hmg/ddx006 (2017).
20 Hodges, A. *et al.* Regional and cellular gene expression changes in human Huntington's disease brain. *Human molecular genetics* **15**, 965-977, doi:10.1093/hmg/ddl013 (2006).
21 Sigurdsson, M. I., Jamshidi, N., Steingrimsson, E., Thiele, I. & Palsson, B. O. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC Syst Biol* **4**, 140, doi:10.1186/1752-0509-4-140 (2010).
22 Piro, R. M. *et al.* Network topology-based detection of differential gene regulation and regulatory switches in cell metabolism and signaling. *BMC Syst Biol* **8**, 56, doi:10.1186/1752-0509-8-56 (2014).

23 Kwon, A. T., Arenillas, D. J., Worsley Hunt, R. & Wasserman, W. W. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 (Bethesda)* **2**, 987-1002, doi:10.1534/g3.112.003202 (2012).
24 Behrens, P. F., Franz, P., Woodman, B., Lindenberg, K. S. & Landwehrmeyer, G. B. Impaired glutamate transport and glutamate-glutamine cycling: downstream effects of the Huntington mutation. *Brain* **125**, 1908-1922 (2002).
25 Shin, J. Y. *et al.* Expression of mutant huntingtin in glial cells contributes to neuronal excitotoxicity. *J Cell Biol* **171**, 1001-1012, doi:10.1083/jcb.200508072 (2005).
26 Joyner, P. M., Matheke, R. M., Smith, L. M. & Cichewicz, R. H. Probing the metabolic aberrations underlying mutant huntingtin toxicity in yeast and assessing their degree of preservation in humans and mice. *J Proteome Res* **9**, 404-412, doi:10.1021/pr900734g (2010).
27 Reynolds, N. C., Jr., Prost, R. W. & Mark, L. P. Heterogeneity in 1H-MRS profiles of presymptomatic and early manifest Huntington's disease. *Brain Res* **1031**, 82-89, doi:10.1016/j.brainres.2004.10.030 (2005).
28 Taylor-Robinson, S. D. *et al.* Proton magnetic resonance spectroscopy in Huntington's disease: evidence in favour of the glutamate excitotoxic theory. *Mov Disord* **11**, 167-173, doi:10.1002/mds.870110209 (1996).
29 Tkac, I., Dubinsky, J. M., Keene, C. D., Gruetter, R. & Low, W. C. Neurochemical changes in Huntington R6/2 mouse striatum detected by in vivo 1H NMR spectroscopy. *J Neurochem* **100**, 1397-1406, doi:10.1111/j.1471-4159.2006.04323.x (2007).
30 Tsang, T. M. *et al.* Metabolic characterization of the R6/2 transgenic mouse model of Huntington's disease by high-resolution MAS 1H NMR spectroscopy. *J Proteome Res* **5**, 483-492, doi:10.1021/pr0502440 (2006).
31 Underwood, B. R. *et al.* Huntington disease patients and transgenic mice have similar pro-catabolic serum metabolite profiles. *Brain* **129**, 877-886, doi:10.1093/brain/awl027 (2006).
32 Chiang, M.-C. *et al.* Dysregulation of C/EBPα by mutant Huntingtin causes the urea cycle deficiency in Huntington's disease. *Human Molecular Genetics* **16**, 483-498, doi:10.1093/hmg/ddl481 (2007).
33 Patassini, S. *et al.* Identification of elevated urea as a severe, ubiquitous metabolic defect in the brain of patients with Huntington's disease. *Biochem Biophys Res Commun* **468**, 161-166, doi:10.1016/j.bbrc.2015.10.140 (2015).
34 Skene, D. J. *et al.* Metabolic profiling of presymptomatic Huntington's disease sheep reveals novel biomarkers. *Scientific Reports* **7**, 43030, doi:10.1038/srep43030
35 Fossale, E. *et al.* Differential effects of the Huntington's disease CAG mutation in striatum and cerebellum are quantitative not qualitative. *Hum Mol Genet* **20**, 4258-4267, doi:10.1093/hmg/ddr355 (2011).
36 Jellinger, K. A. Cell death mechanisms in neurodegeneration. *J Cell Mol Med* **5**, 1-17 (2001).
37 Rajda, C., Pukoli, D., Bende, Z., Majlath, Z. & Vecsei, L. Excitotoxins, Mitochondrial and Redox Disturbances in Multiple Sclerosis. *Int J Mol Sci* **18**, doi:10.3390/ijms18020353 (2017).
38 Kumar, A. & Ratan, R. R. Oxidative Stress and Huntington's Disease: The Good, The Bad, and The Ugly. *J Huntingtons Dis* **5**, 217-237, doi:10.3233/JHD-160205 (2016).