Boston University

OpenBU **http://open.bu.edu**

Theses & Dissertations Boston University Theses & Dissertations

2016

Characterizing the Huntington's disease, Parkinson's disease, and pan-neurodegenerative gene expression signature with RNA sequencing

https://hdl.handle.net/2144/17865

Boston University

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

CHARACTERIZING THE HUNTINGTON'S DISEASE, PARKINSON'S DISEASE, AND PAN-NEURODEGENERATIVE GENE EXPRESSION SIGNATURE WITH RNA SEQUENCING

by

ADAM THOMAS LABADORF

B.S., Dickinson College, 2003
M.S., Colorado State University, 2010

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2016

© 2016 by
Adam Thomas Labadorf
All rights reserved

Approved by

First Reader ______________________________
Richard H. Myers, Ph.D.
Professor of Neurology

Second Reader ______________________________
Eric Kolaczyk, Ph.D.
Professor of Mathematics and Statistics

ACKNOWLEDGMENTS

This thesis would not have been possible without the lifelong support, patience, and devotion of my family, for which I am profoundly grateful. The people who stuck with me through the many life altering transitions that led me to this point continue to have my utmost respect, love, and admiration; there are no better friends in the world. To my teachers, professors, and mentors who guided me as I had, and indeed still have, very little idea of how much I don’t know, you are also of fundamental importance to me and who I am, and I could not be more thankful. I also wish to thank my closest mentor, Rick Myers, for his faith, guidance, and confidence in my abilities that helped propel me in times of frustration and difficulty.

CHARACTERIZING THE HUNTINGTON'S DISEASE, PARKINSON'S DISEASE, AND PAN-NEURODEGENERATIVE GENE EXPRESSION SIGNATURE WITH RNA SEQUENCING

ADAM THOMAS LABADORF

Boston University Graduate School of Arts and Sciences

and College of Engineering, 2016

Major Professor: Richard H. Myers, Professor of Neurology

ABSTRACT

Huntington's disease (HD) and Parkinson's disease (PD) are devastating neurodegenerative disorders that are characterized pathologically by degeneration of neurons in the brain and clinically by loss of motor function and cognitive decline in mid to late life. The cause of neuronal degeneration in these diseases is unclear, but both are histologically marked by aggregation of specific proteins in specific brain regions. In HD, fragments of a mutant Huntingtin protein aggregate and cause medium spiny interneurons of the striatum to degenerate. In contrast, PD brains exhibit aggregation of toxic fragments of the alpha synuclein protein throughout the central nervous system and trigger degeneration of dopaminergic neurons in the substantia nigra. Considering the commonalities and differences between these diseases, identifying common biological patterns across HD and PD as well as signatures unique to each may provide significant insight into the molecular mechanisms underlying neurodegeneration as a general process. State-of-the-art high-throughput sequencing technology allows for unbiased, whole genome quantification of RNA molecules within a biological sample that can be used to assess the level of activity, or expression, of thousands of genes simultaneously. In this thesis, I present three studies characterizing the RNA expression profiles of post-mortem HD and PD subjects using high-throughput mRNA sequencing data sets. The first study describes an analysis of differential expression between HD individuals and neurologically normal controls that indicates a widespread increase in immune, neuroinflammatory, and developmental gene expression. The second study expands upon the first study by making methodological improvements and extends the differential expression analysis to include PD subjects, with the goal of comparing and contrasting HD and PD gene expression profiles. This study was designed to identify common mechanisms underlying the neurodegenerative phenotype, transcending those of each unique disease, and has revealed specific biological processes, in particular those related to NFkB inflammation, common to HD and PD. The last study describes a novel methodology for combining mRNA and miRNA expression that seeks to identify associations between mRNA-miRNA modules and continuous clinical variables of interest, including CAG repeat length and clinical age of onset in HD.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iv

ABSTRACT v

TABLE OF CONTENTS vii

LIST OF TABLES x

LIST OF FIGURES xii

Chapter 1. Introduction 1

- Huntington’s Disease 1
- Parkinson’s Disease 3
- HD and PD as Neurodegenerative Diseases 5
- High-throughput mRNA-Seq and miRNA-Seq Data Analysis 7
- Relating mRNA and miRNA data 14
- Sample and Datasets Characteristics 16

Chapter 2. Differential Expression in HD 20

- Introduction 20
- Methods 25
  - Sample Information 25
  - Human Subjects 26
  - mRNA Sample Preparation and Sequencing 26
  - Gene Expression Quantification, Data Cleaning, and DE Analysis 27

DAVID, GO, and MsigDB Enrichment Calculation 28

Association with Clinical Covariates 29

Replication of DE Genes by RT-qPCR in an Independent Sample Set 29

Validation of DE Genes by RT-qPCR...30

Results 31

Widespread Differential Expression Changes Are Observed in HD............31

Immune Response, Development, and Transcriptional Regulation Functions Are Enriched in HD...37

Integrated Geneset Enrichment Analysis Identifies Specific Enriched Functional Categories 39

GO Enrichment Analysis Implicates Development and Immune Response.40

Pathways Involved in Multiple Immune System Processes Are Enriched....42

DE Genes Are Enriched as Targets of Transcription Factors Implicated In HD 43

Integrated Geneset Enrichment Analysis Links Biological Function and Transcriptional Regulation..44

Association of Gene Expression with Clinical Covariates 45

Discussion...47

Acknowledgments...55

Chapter 3. Comparative Analysis of HD and PD Gene Expression 56

Introduction 56

Materials and Methods...58

    - Sample collection, processing, and sequencing 58
    - mRNA-Seq data processing 59
    - Differential expression and assessment of batch effects 60
    - Identification of ND DE genes and enriched gene sets 61
  - Results 62
  - Discussion 72
- Chapter 4: mRNA/miRNA modules associated with clinical features in HD 79
  - Introduction 79
  - Methods 85
  - Results 91
  - Discussion 97
- Chapter 5. Conclusions, Prospective, and Future Work 101
- BIBLIOGRAPHY 107
- CURRICULUM VITAE 138

LIST OF TABLES

Table 1. HD sample statistics. 22

Table 2. Control sample statistics. 24

Table 3. DE genes by significance. 34

Table 4. DE genes by DES. Differential Expression Score (DES) is calculated as (overall mean counts) x abs(log2 FC) x –log10(adjusted p-value) 35

Table 5. DAVID functional clustering. Cluster Function labels were assigned manually by inspecting the terms within the cluster but generally correspond to the name of the most enriched term within the cluster. The (1) Immune response cluster contained 27 distinct terms from across the default genesets used by DAVID. 38

Table 6. Enriched GO Categories. The most enriched GO category GO:sequence-specific DNA binding using the top 25 DE genes ranked by significance. The second most enriched GO category, GO:anterior/posterior pattern specification, was found when considering the top 350 DE genes. .42

Table 7. Protein coding genes associated with clinical covariates. P-values are nominal. 47

Table 8. Sample Statistics 59

Table 9. Top ranked RRA genes. RRA Score can be thought of as a p-value. The remaining columns contain the rank of the corresponding gene in each individual gene list. 67

Table 10. Clinical features and CAG-adjusted derivatives used in this study. 92

Table 11. Module sizes and R2 statistics for all features for the 25 modules. First column is module label. Feature columns report R2 values and evidence scores in parenthesis. 96

Table 12. R2 values and top significantly enriched MsigDB C2 Canonical Pathway gene sets at BH-adjusted p<0.05 for modules. Only the top 5 most enriched gene sets by p-value are reported. 98

LIST OF FIGURES

Figure 1. DE statistics. A) Histogram of log2 fold changes for DE genes showing that 54.8% of the DE genes are overexpressed in HD cases. B) Fraction of up vs. down regulated genes across the gene list ranked by significance. Top and bottom plots are top 500 and remaining genes, respectively. Sliding windows lines plot the fraction up vs. down in the 100 gene window of greater rank than the x coordinate. This plot shows that the most highly differentially expressed genes are predominantly over-expressed in HD relative to control BA9. C) Pie chart shows proportions of biotypes for DE genes according to Ensembl. Protein coding genes are overrepresented among the DE genes. D) Normalized counts for all samples in HD and control for top ten significant genes. Rows are normalized for visualization such that the highest count is equal to 1. These genes are almost exclusively expressed in HD cases. 32

Figure 2. DAVID functional clustering network. Network representation of the DAVID clusters from Table 5. Nodes represent clusters, the size of the node is proportional to the number of unique genes that make up the cluster and numbers within nodes are the number of unique genes mapped to terms in the cluster. Edges between nodes indicate the existence of overlapping genes, where the width is proportional to the percent overlap of genes in the smaller of two connected nodes. The color of nodes and edges is proportional to the average fold change of the genes in the node or edge. .39

Figure 3. Detailed GO enrichment. Top 25 enriched GO categories across all three GO namespaces identified by topGO for different numbers of DE genes. X-axis indicates the number of top genes used for the enrichment in each GO category, e.g. the first column uses the top 25 genes, the second column uses the top 50, and so on. The intensity is proportional to –log10(p-value) from topGO. White dots indicate the gene set with the most significant p-value, concordant with Table 5. This figure shows that the first three GO Categories are defined by genes that are among the top 25 to 150 DE genes in the dataset. GO Categories further down the list are defined by genes whose differential expression is less pronounced between HD and controls. 41

Figure 4. Clustergram of Top Enriched Pathway, TF, and GO terms. Concatenated enrichment profiles for GO, C2, and TF gene set collections, similar to Figure 3, ordered by hierarchical clustering of Euclidean distance between rows. Rows have been normalized by dividing by the row sum for visualization, intensity is proportional to normalized enrichment. Heatmap is partitioned into groups A-E based on hierarchical clustering. Clusters A, B, and C are primarily involved in the immune response and are enriched in gene subsets that include more genes. Clusters D and E are predominantly related to developmental and transcriptional regulation processes. 45

Figure 5. A) Significantly enriched MsigDB C2 Canonical Pathway gene sets for HD, PD, and ND identified by GSEA. Each colored ring segment corresponds to a single enriched gene set. Red (outer), green (middle), and blue (inner) segmented rings indicate whether the HD, PD, or ND DE gene lists, respectively, were significantly enriched for the gene set. Dark and light colored segments indicate up and down regulation (positive, negative GSEA normalized enrichment score), respectively. Black ring around exterior groups gene sets into high level categories as indicated by the two letter code. Gene set name is listed in interior of rings. B) Venn diagram illustrating overlap of significantly enriched gene sets for HD, PD, and ND. All but 24 of the ND significant gene sets were significantly enriched in either HD, PD, or both..66

Figure 6. Box plots of normalized counts for top RRA genes split by condition, RRA rank is in parenthesis. Whiskers extend to 25th and 75th percentile counts, white bars are median counts. With the exception of (12) PITX1, which is driven entirely by HD, all top genes demonstrate substantial differences between both disease conditions and control. 69

Figure 7. A) Venn diagram of HD and PD DE gene list intersection for DE genes adjusted p<0.01. B) Bar chart indicating number of MsigDB C2 Canonical Pathway (CP), C3 miRNA Targets (miR), C3 Transcription Factor Targets (TF), and C5 Gene Ontology (GO) gene sets enriched for the HD unique (HD \ PD), intersection (HD n PD), and PD unique (PD \ HD) genes. For HD \ PD enrichment, 17 redundant or uninformative GO gene sets and 7 TF gene sets for motifs with unknown transcription factors were omitted from the figure results. C) Gene sets enriched for the intersection genes (HD n PD). Adjusted p-values are listed beside gene set name and the originating gene set (CP, miR, TF, or GO) are indicated by color. Gene sets that are groups into boxes share more than 20% of their DE genes and are therefore listed together. 70

Figure 8. Enriched gene sets for the HD unique (HD \ PD) genes from Figure 7 A and reported similarly as in Figure 7 C. Note 17 redundant or uninformative GO gene sets and 7 TF gene sets for motifs with unknown transcription factors were omitted from the figure results. 72

Figure 9. Cartoon of relationships involved with mRNA/miRNA modules. 83

Figure 10. A) Schematic of analysis pipeline. B) Cartoon depiction of PCA dimensionality reduction procedure..87

Figure 11. A) Heatmap of mRNA/miRNA correlations in module 0, mRNA on the y axis and miRNA on the x axis. B) Stacked barplot of the proportions of negative (rho <= -.1), near zero (-.1 < rho < .1), and positive (rho >= .1) correlations for all modules. 93

Figure 12. A) Clustered matrix of member overlap between all modules. The number of overlapping members is listed within each cell, overlaps of zero are omitted. Diagonal entries have been set to zero. B) Boxplots of the R2 distributions from each feature for the true modules, true modules with shuffled features, random modules, and random modules with shuffled features...93

Chapter 1. Introduction

Huntington’s Disease (HD) and Parkinson’s Disease (PD) are devastating neurodegenerative disorders that affect humans in mid to late life, causing progressive loss of motor function, cognitive ability, and ultimately death. Both diseases initially manifest with mild symptoms of motor dysfunction in the form of involuntary movement or bradykinesia (slowness of movement), but patients often develop cognitive deficits and personality changes as the disease progresses. The pathological cause of the diseases is the selective death of neurons in specific brain regions. Despite their similar symptomatology, the mechanisms underlying the pathology for these diseases are distinct in ways that make a comparison of HD and PD an attractive study to gain a better understanding of the molecular processes associated with neurodegeneration as a general phenotype. The studies in this thesis address this question by comparing RNA signatures identified by analyzing high-throughput gene expression measurements using samples extracted from post-mortem brains of individuals who have died of either HD or PD.

Huntington’s Disease

HD is an autosomal dominant neurodegenerative disorder that affects GABAergic medium spiny interneurons in the caudate nucleus and putamen of the brain (Vonsattel et al. 1985). The disease is caused by an expanded tri-nucleotide CAG repeat sequence in the first exon of the Huntingtin (HTT) gene that encodes for a polyglutamine tract in the subsequently translated Htt protein

(MacDonald et al. 1993). Individuals with 40 or more CAG repeats exhibit full penetrance and will eventually develop the disease, while individuals with 36–39 repeats show reduced penetrance (Walker 2007). The normal CAG repeat size range in unaffected individuals is 18–26 (Walker 2007). Individuals with an expanded repeat typically develop and live normally until middle age, on average 40 years of age (R. H. Myers, Marans, and MacDonald 1998), when motor symptoms begin to present. The rapidity of disease progression is inversely correlated with the length of the CAG repeat, where individuals with longer repeat sizes will suffer an earlier age of onset (MacDonald et al. 1993). In extreme cases of 70 or more CAG repeats, individuals exhibit juvenile onset of symptoms within the first decade of life (Richard H. Myers 2004-4). From the age of onset, motor dysfunction steadily progresses, personality and mood changes may emerge, often followed by weight loss, cognitive decline, disability, and eventually death. The average duration of disease from age of onset to death is 15 to 20 years (Richard H. Myers 2004-4). To date, no definitively effective therapies are available that consistently halt or slow the progression of symptoms in HD.

Much is known about the mutant Huntingtin protein (mHtt) that causes HD. The N-terminal fragment containing the expanded polyglutamine tract of mHtt is cleaved by one of several mechanisms and becomes toxic via aberrant interactions with multiple other proteins (Rubinsztein and Carmichael 2003). In particular, mHtt fragments have been shown to interact with transcription factors leading to large scale transcriptional dysregulation (S.-H. Li et al. 2002; A. T. Labadorf and Myers 2015; Cha 2007; Cha 2000), proteins in the NFkB pathway leading to aberrant NFkB activity in multiple central nervous system (CNS) cell types (O'Neill and Kaltschmidt 1997; Meffert et al. 2003; Marcora and Kennedy 2010; Träger et al. 2014; Kaltschmidt et al. 1994–6), calpains and caspases (Rubinsztein and Carmichael 2003; Gafni et al. 2004; Y. J. Kim et al. 2001), and chaperone proteins (Qi and Zhang 2013). In addition to interfering with normal cellular function through aberrant protein-protein interactions, these fragments also form aggregates in the nucleus, dendrites, and synapses of neurons (Rubinsztein and Carmichael 2003) where the number and size of these aggregates correlate with the degree of degeneration (Vonsattel et al. 1985). It is not known conclusively whether mHtt aggregation causes neurodegeneration or whether it is a protective mechanism, but the presence of aggregates are coincident with degenerating neurons cell and animal models (Arrasate and Finkbeiner 2012; Rubinsztein and Carmichael 2003).

Parkinson's Disease

PD is a progressive neurodegenerative disorder that affects dopaminergic neurons in the substantia nigra pars compacta (Shulman, De Jager, and Feany 2011). Symptoms primarily manifest as motor deficits, including involuntary shaking, bradykinesia, rigidity, and difficulty with walking, but individuals may later develop mood and sleep disorders, depression, and dementia at late stages (Dexter and Jenner 2013). As several diseases may manifest with similar symptoms, collectively termed “parkinsonism”, a final diagnosis of PD specifically is made by histological observation of protein aggregates containing the alpha-synuclein protein, termed Lewy bodies, in the brain at autopsy(Shulman, De Jager, and Feany 2011). Both genetics and environment contribute to developing PD, where as much as 60% and 40% of the disease risk is heritable in all families and those excluding known with the most common PD-associated loci, respectively (Hamza and Payami 2010-4), with the remaining risk likely attributed to environmental agents including ageing, drugs, toxins, and pesticides (Allam, Del Castillo, and Navajas 2005; Polito, Greco, and Seripa 2016). As many as 24 genomic loci have been discovered to associate with PD risk (Nalls et al. 2014), most notably in regions encoding the genes MAPT, SNCA (which encodes the alpha-synuclein protein itself), GBA, GAK, and LRRK2.

The mechanisms underlying PD pathology are unclear and appear to be quite varied. Four distinct but related molecular mechanisms have emerged as contributing to neurodegeneration in PD: oxidative stress, mitochondrial dysfunction, altered proteolysis, and inflammatory processes (Dexter and Jenner 2013). Oxidative stress and resulting damage, most likely due to dopamine metabolism and mitochondrial dysfunction, appear to be major consistent contributors to dopaminergic neuronal toxicity (Jenner 2003). Mitochondrial dysfunction, likely caused by both environmental factors and mutations in PD-linked genes, has been associated with cell death in PD via mechanisms involving defects in complex I and IV of the mitochondrial membrane and sensitivity to oxidative stress(Schapira 2008). The presence of incorrectly processed protein aggregates is evidence of proteolytic dysfunction in PD neurons, and alterations in both major proteolytic systems, namely the ubiquitin-proteasome and lysosome systems (McNaught and Jenner 2001). Finally, support for inflammatory process involvement in PD is increasing from gene expression studies in the blood and post-mortem brain tissues of PD patients (Dobbs et al. 1999). See (Dexter and Jenner 2013) for an excellent review of PD molecular pathology.

HD and PD as Neurodegenerative Diseases

HD and PD have a number of similarities and differences that allow interesting comparisons to be drawn concerning the neurodegenerative process. Most obviously, both are progressive neurodegenerative diseases that result in motor and behavioral deficits in mid to late life due to selective degeneration of neurons in the brain. Both diseases are histologically marked by aberrant protein aggregation, suggesting some feature of proteasomal dysfunction is common to both diseases. However, the specific neurons affected, namely medium spiny interneurons of the caudate nucleus and putamen in HD and dopaminergic neurons in the substantia nigra in PD, degenerate in one disease but are spared in the other. This mirror-image property of the degenerative pattern allows the identification of properties specific to each disease as well as mechanisms that may be common to the neurodegenerative phenotype overall. The two diseases are also genetically distinct in the sense that HD is monogenic, where a mutation in a single gene is sufficient to predict disease, where PD is highly heterogeneous with multiple genetic risk factors and environmental influences that modify disease risk. Molecular signatures, e.g. differentially expressed genes, common to both HD and PD may therefore lead to a better understanding of mechanisms that underlie neurodegeneration as a general pathological process. A clearer understanding of the common response to neurodegenerating neurons may have significant impact on therapeutic approaches, and could also have important implications in how we approach therapies for other neurodegenerative diseases of the central nervous system.

A salient question in both HD and PD, and indeed to many diseases, pertains to which genes are involved in the processes underlying pathology. In the past fifteen years, multiple technologies have come available that allow genome-wide quantification of ribonucleic acid molecules (RNAs) in biological samples. One of the most recent of these technologies, so-called high-throughput sequencing, allows unbiased abundance measurements of potentially all RNAs present within a biological sample of interest. Arguably the most important type of RNA molecules is messenger RNA (mRNA), which is transcribed from genes of the genome and are typically translated into protein. A second type of RNA molecule, microRNAs (miRNAs), play an important role in regulating the abundance of mRNA molecules. High-throughput sequencing technologies have been developed to measure the abundance of both mRNAs (mRNA-Seq) and miRNAs (miRNA-Seq). Since mRNAs are the precursors to proteins, their abundance is thought to represent the abundance and, therefore activity, of the proteins that perform most cellular functions. Identifying which mRNA species differ between samples originating from diseased and healthy tissue is therefore an attractive means with which to assess the biological processes that are involved in disease pathogenesis. Analyzing differences in miRNA abundance between disease and control may lead to hypotheses about the causal effects behind the differential abundance of mRNAs, and therefore could lead to experiments that elucidate pathological mechanisms.

To this end, the Myers lab has generated mRNA-Seq and miRNA-Seq datasets from the brains of post-mortem individuals who either died with HD or PD, and neuropathologically normal individuals who died of other causes. The hypothesis underlying these datasets is that the abundance of mRNAs and miRNAs is different between diseased and healthy individuals, and that identifying the specific species implicated by this comparison will yield insight into relevant disease processes. Additionally, as we have datasets from both HD and PD, we are able to compare not only diseased and healthy individuals, but also compare these diseases to one another in pursuit of identifying common neurodegenerative signatures.

High-throughput mRNA-Seq and miRNA-Seq Data Analysis

mRNA-Seq and miRNA-Seq technologies produce data in the form of short digitized biological sequences, termed reads, that represent the nucleotides (nt), i.e. strings of the letters A, C, G, and T symbolizing the nucleotides adenine, cytosine, guanine, and thymine respectively, that make up the molecules in the originating sample. The reads are typically 35–150nt in length, and a single sequencing dataset, or library, may contain as many as 350 million reads. The first task in processing a short read dataset is to identify the biological sequence from which each read originated. For the datasets in these studies, the sequence of each read is compared to the digitized sequence of the human genome, hereafter called the reference genome, in a process called sequence alignment or mapping. The alignment process entails identifying the location(s) where the sequence of a read matches along the linear sequence of the reference. A read aligns to a location in the reference if there is a high degree of similarity between the read sequence and the reference sequence location. In so doing, the originating location of mRNA or miRNA molecules represented by a read may be determined and, in combination with a gene annotation that indicates where in the reference genome each gene is located, reads may be mapped to genes. The overall abundance of the RNA originating from each gene may then be estimated by counting the number of reads that map within the annotated region of the reference for each gene. In principle, this approach results in the relative quantification of RNA and miRNA molecules for all of the genes and miRNAs in the entire genome within a sample.

As described above, the abundance of each mRNA or miRNA is represented by the number of reads, or sequencing counts, that map to each annotated genomic region. By counting the reads for a set of samples against the same annotation, e.g. gene annotation, the sequencing counts for each gene and each sample may be concatenated into a count matrix, where genes are rows and samples are columns. For groups of samples of two types, e.g. HD and healthy control, the counts for a gene within one group may be compared to the counts for the same gene in the other group to assess whether the relative abundance of the mRNAs are different between groups. However, since libraries for different samples may be of different size, e.g. one library has 250M reads while another has 300M, and the counts are in fact proportional to, and not absolute measurements of, the abundance of the originating biological molecules, the counts in this matrix are not directly comparable between samples. To account for this, the counts matrix is subjected to a normalization procedure that attempts to adjust the counts within each sample so that they may be compared to one another.

Several count normalization strategies have been proposed, but most involve identifying a single numeric factor for each sample that will adjust sample counts across samples to make them comparable. The simplest such method is library size level, or total count, normalization, where all counts within a sample are divided in proportion to the overall library size from that sample. This approach, while reasonable, makes the often-violated assumption that highly abundant genes are consistent across samples and therefore performs poorly in practice (Dillies et al. 2013). Another popular approach, related to total count normalization, adjusts the counts of each gene by gene length as well as library size, transforming counts into reads per gene kilobase per million library reads (RPKM) (Mortazavi et al. 2008). This approach has been found to suffer the same biases as total counts and additionally introduces per-gene variance bias leading to more false positives upon differential expression analysis (Dillies et al. 2013). More sophisticated approaches, like that of the edgeR (Robinson, McCarthy, and Smyth 2010) and DESeq2 (Love, Huber, and Anders 2014) packages, assume the majority of genes are not different between samples and show more favorable properties on differential expression sets (Dillies et al. 2013). All of the count matrices in this thesis have therefore been normalized using the DESeq2 method.

After proper count normalization has been performed comes the task of identifying which genes exhibit statistically coherent count behavior. A common pattern in such analyses is to determine which genes demonstrate significantly different abundance between two groups of samples, such as HD vs. healthy control samples. This type of analysis, often termed differential expression (DE) analysis, essentially involves assessing differences in mean normalized counts accounting for variance across samples. Many such methods have been proposed (Soneson and Delorenzi 2013), but the most popular methods at the time of this writing use generalized linear regression, specifically negative binomial (NB) regression, to model gene counts as a function of a binary categorical variable (e.g. HD vs. control). edgeR (Robinson, McCarthy, and Smyth 2010) and DESeq2 (Love, Huber, and Anders 2014) are two such NB regression methods that were designed specifically for the analysis of mRNA-Seq data.

The use of NB regression is motivated by the observation that normalized count data does not follow a Gaussian distribution, in particular because counts are non-negative and the variance and mean of gene counts are empirically not independent, making classical linear regression models inappropriate. NB regression, however, requires the estimation of a dispersion parameter before inference can be performed. Estimation of this parameter increases the computational and statistical complexity of the inference algorithm, and noisy counts can pose difficulty to the estimation procedure, leading to unreliable or spurious results. As an alternative, log-transformed normalized count data are approximately normally distributed, and a number of approaches first perform such a transformation on counts and then apply linear regression-based methods to these transformed counts. A number of such transformations have been proposed, including a moderated log transform (Leek 2014), the VOOM transform in the limma Bioconductor package (Law et al. 2014), and the variance stabilizing transform (VST) in the DESeq2 package (Love, Huber, and Anders 2014). When the number of samples per group is large (>5), these transformation+linear regression based methods perform well (Soneson and Delorenzi 2013). A positive attribute of both NB and transformation+linear regression methods is they allow for statistical adjustment for confounding variables whose effects are not related to the condition of interest, such as age at death in HD vs. control models.

Both negative binomial regression and transformation+linear regression based methods may be sensitive to outlier counts. It is often observed that a single sample may display one or two magnitudes greater counts than other samples in the same gene, posing significant inference challenges for regression-based methods DE methods. A family of nonparametric methods have been proposed (Shi, Chinnaiyan, and Jiang 2015; Bi and Davuluri 2013; Lin, Zhang, and Chen 2014; J. Li and Tibshirani 2013; Tusher, Tibshirani, and Chu 2001) that often use rank-based transformations of counts to identify consistent DE patterns between groups. These methods are robust to outlier counts by definition, and when the number of samples per group is sufficiently large detection of DE has been found to be quite accurate (Soneson and Delorenzi 2013). Non-parametric methods have two important drawbacks. The first is that, since the variance of the counts is not explicitly modeled due to the rank transformation, statistical adjustment for confounding variables is not possible. The second, also related to the rank transformation, is that effect-sizes are no longer based on count abundance but rather on rank, such that genes of both very high and very low abundance are treated equally for detection of DE. DE of very lowly abundant genes may not have significant biological relevance but dilute the genes that are more abundant. Though low relative abundance does not necessarily imply lack of biological significance, this consideration is nonetheless important when choosing a DE method and interpreting the results.

To date, no proposed DE methods are based on logistic regression (LR). The use of LR is widespread in genome wide association studies, where mutations in the genomes of large numbers of subjects are queried for association with a disease or condition of interest, and the LR methodology is equally well understood and established by the community as linear regression. For DE studies that seek to identify genes that have different mean counts between two conditions, LR is an apt model except for one property occasionally encountered in count data. It is sometimes the case that a gene exhibits dramatically different behavior between two conditions (e.g. extremely low counts in healthy control and extremely high in disease), such that the two groups have no counts that overlap. Such a condition is called complete separation, and logistic models fail when posed with parameter inference in these situations. However, two modifications to the classical LR method, Bayes logistic regression (Gelman et al. 2008) and Firth's logistic regression (Firth 1993; Heinze and Schemper 2002), have been proposed that account for complete separation. As recently described (Choi et al 2016, under review), Firth's LR overcomes the problem of complete separation in the analysis of the HD vs. control mRNA-Seq datasets described here and furthermore show favorable statistical properties with respect to type I error rates and statistical power under certain conditions. A novel aspect of the this thesis is the application of Firth's LR to the datasets herein described found in Chapter 4.

In addition to genes that exhibit differential behavior between conditions, identifying genes with expression patterns that correlate with clinically relevant features is also of interest. For example, clinical age of onset is a particularly salient feature in the progression of HD and, despite being partially heritable (Richard H. Myers 2004-4), no common genetic markers have been found to strongly associate with this variable after adjusting for CAG repeat length. Genes whose expression correlates with CAG-adjusted age of onset are of particular interest in HD because these genes may not only inform our understanding of HD neurodegenerative mechanisms but may also give clues to new therapeutics that can modify the onset of symptoms, where presently no such therapies exist. While NB can accomplish this analysis by modeling the counts as a function of the clinical feature, linear regression may be canonically applicable by modeling the feature as a function of the sequencing counts, avoiding the expensive and sensitive NB parameter estimation steps. This approach assumes the clinical feature of interest is normally distributed, which is often the case. Employing a linear model in this way also allows for adjusting the association of counts to the clinical feature by potentially confounding variables.

Relating mRNA and miRNA data

As mentioned previously, miRNA are short RNA molecules, typically 18–22 nucleotides in length, that inhibit mRNA molecules from being translated into protein by targeting specific nucleotide sequences contained within transcribed mRNAs. The mRNA/miRNA relationship is many-to-many, where a single miRNA can target multiple mRNAs and a single mRNA may be targeted by many different miRNAs, resulting in a complex regulatory network that has been implicated in development and disease (W. Zhang et al. 2012; Hiddingh et al. 2014; Cordes and Srivastava 2009; Shenoy and Blelloch 2014). Identifying the mRNA targets of miRNAs is a critical step in understanding the regulatory relationships between these molecules. Several complementary approaches have been proposed for predicting mRNA/miRNA relationships (Lewis, Burge, and Bartel 2005a; Enright et al. 2003; Coronnello et al. 2012; Wong and Wang 2015; Y. Li et al. 2014). Identifying mRNA/miRNA relationships that are different between diseased and healthy tissue may provide important insights into disease processes.

Analysis of mRNA and miRNA abundance measurements using high-throughput transcriptional data revealed that groups of interacting mRNAs and miRNAs, termed mRNA/miRNA modules, often work in concert to regulate specific biological processes and disease (W. Zhang et al. 2012; Hiddingh et al. 2014; Coronnello et al. 2012; Z. Liu et al. 2015; Setty et al. 2012). modules can be detected using transcriptional data from multiple samples by examining the statistical relationship between the abundance of mRNAs and miRNAs. Many mRNA/miRNA module detection approaches have been proposed and generally employ one of four broad strategies. Sequence-based algorithms examine the predicted targets of each miRNA and group mRNAs by overrepresented numbers of shared targets. Expression based approaches use transcriptional data to identify statistical, typically negative, correlations between mRNAs and miRNAs to identify groups of transcriptionally related species. Another approach uses regularized regression based methods, such as LASSO (Tibshirani 1994), to find statistical relationships between groups of mRNAs and miRNAs. And last, the most complex set of algorithms utilizes graphical models, typically implemented using Bayesian statistics, to directly predict the regulatory relationships between mRNAs and miRNAs. See Chapter 4 for more in-depth background regarding modules and module detection methods.

Some of the proposed module detection methods focus on identifying differential module behavior between two conditions of interest. Differences between module definition or activity within a disease, for example, may lead to a more complete understanding of the regulatory underpinnings of the disease. However, to date no algorithms have been proposed that attempt to identify modules that are associated with a continuous feature of interest, such as clinical age of onset in HD. The study in Chapter 4 presents a novel algorithm that takes paired mRNA and miRNA expression data that attempts to identify modules that are associated with continuous variables by integrating established module detection and statistical methods.

Sample and Datasets Characteristics

In total, the studies in this thesis use 29 HD, 29 PD, and 49 neurologically normal control mRNA-Seq samples and 25 HD miRNA-Seq samples generated from post-mortem human brains. Frozen brain tissue from prefrontal cortex Brodmann

Area 9 (BA9) was obtained from the Harvard Brain and Tissue Resource Center McLean Hospital, Belmont MA, the Human Brain and Spinal Fluid Resource Center VA West Los Angeles Healthcare Center (Los Angeles, California) and Banner Sun Health Research Institute (Beach et al. 2008-9) (Sun City, Arizona). For each brain sample, grey matter from the cortical ribbon was dissected by hand with a target mass of 0.08 g and used for RNA extraction. Total RNA was extracted from the samples using established protocols that were submitted for mRNA-Seq and miRNA-Seq sequencing as appropriate on the Illumina HiSeq 2000 platform. There were 80M reads on average per mRNA-Seq dataset, and 10M on average for each miRNA-Seq dataset. More detailed information on sample preparation is found in Chapter 2 methods section titled mRNA Sample Preparation and Sequencing.

A number of clinical features are available for HD and PD. For HD, the four primary features are CAG size (i.e. repeat length), clinical age of onset, and H-V cortical and striatal scores (Hadzi et al. 2012), which are histology-based numerical scores indicating degree of involvement in the cortex and striatum. Since there is a considerable level of correlation between these four covariates, particularly between CAG repeat length and age of onset, we created three new features using age of onset, cortical score, and striatal score to identify modules that are associated with the residual variance of these features after accounting for the contribution of CAG. These clinical features are described in more detail in Chapter 4. For PD, age of clinical onset and dementia status are the only relevant clinical feature available for these samples.

The miRNA data used in this thesis has been previously analyzed and reported in the literature (Hoss et al. 2015; Hoss et al. 2014; Hoss et al. 2016). (Hoss et al. 2015) found a set of miRNAs, in particular miR-10b, were dramatically increased in HD and were associated with clinical features CAG-adjusted age of onset and H-V cortical and striatal scores. The strong association of miR-10b expression and CAG-adjusted age of onset in particular suggest that this miRNA is somehow related to the progression of HD, either as a marker for disease severity or as a directly player in the pathogenic process. miR-10b was also found to be associated with age of onset, though in the opposite direction of effect as in HD, in a follow-up study of the PD miRNA-Seq data (Hoss et al. 2016). These results are strong evidence that miRNA abundance may be a valuable marker for disease progression, and more may yet be learned by jointly analyzing mRNA and miRNA expression data.

This thesis is organized as follows. The study in Chapter 2 attempts to discover differentially expressed genes in HD and elucidate the biological mechanisms underlying transcriptional differences between HD and neuropathologically normal brains. The study in Chapter 3 extends the HD differential expression study to include PD, and makes significant biological and methodological improvements over the HD specific study. Chapter 4 describes an analysis pipeline that seeks to identify and compare and contrast mRNA/miRNA modules in HD and PD as well as use those modules to explain clinical features relevant to each disease. The thesis concludes with conclusions, projections, and future work consequent from the studies described.

Chapter 2. Differential Expression in HD

Introduction

Huntington's Disease (HD) is a devastating neurodegenerative disorder characterized clinically by involuntary choreic movement, personality changes, and premature death (Huntington G. 1872; R. H. Myers, Marans, and MacDonald 1998). The disease is caused by an expanded CAG repeat in the Huntingtin gene (HTT)(MacDonald et al. 1993) that produces selective neuronal loss in the brain (Vonsattel et al. 1985). Individuals commonly present characteristic motor signs in midlife with a mean onset age of 40 years (Richard H. Myers 2004-4). No therapy to date has definitively delayed onset or subsequent progression of these symptoms. Most studies in HD are conducted using model systems, (i.e. cell lines or mouse models) or peripheral human biospecimens such as blood and not in involved brain regions from human HD affected individuals. While collecting and analyzing human post-mortem samples presents challenges, the study of brain regions involved in HD provides relevant insight into the disease pathogenesis.

Although transcriptional dysregulation has been convincingly implicated in HD(Cha 2007; Cha 2000), few genome-wide gene expression studies have targeted affected tissues in post mortem human brain to date. To expand our understanding of alterations in mRNA transcriptomics, we have performed mRNA expression profiling by next-generation sequencing in human post-mortem prefrontal cortex Brodmann area 9 (BA9) in 20 HD and 49 neuropathologically normal individuals using Illumina high-throughput sequencing (See Table 1 and Table 2). Although the primarily affected brain region in HD is the striatum (Vonsattel et al. 1985), neuronal loss of up to 90% by the time of death impedes the interpretation of expression profiles derived from striatal whole tissue homogenate since the cell type distribution is altered from that of corresponding unaffected control tissue. It is well established that the prefrontal cortex is involved in HD pathogenesis (Sotrel et al. 1991; Sotrel et al. 1993) but suffers substantially less neuronal death than striatum (Hoss et al. 2014). The brains used in this study have been comprehensively characterized for pathological involvement through detailed histological examination as previously described (Hadzi et al. 2012), which enables direct interpretation of the results in the physiological context of neurodegeneration. We therefore used whole tissue homogenate from the BA9 region in this study.

Sample ID	PMI	Age of Death	RIN	Age of Onset	Duration	CAG	Grade	H-V Striatal Score	H-V Cortical Score
H_0001	37.25	55	7.1	44	11	45	3	2.661	0.922
H_0002	5.75	69	7.5	63	6	41	3	2.644	1.081
H_0003	20.5	71	7.0	52	19	43	3	2.428	1.707
H_0005	19.15	48	6.9	25	23	48	4	3.820	1.939
H_0006	unk	40	6.2	34	6	51	4	3.522	1.431
H_0007	8	72	8.5	55	17	41	3	2.593	0.849
H_0008	21.3	43	7.4	28	15	49	3	2.701	1.701
H_0009	3.73	68	7.8	45	23	42	3	2.668	1.701
H_0010	6.16	59	8.3	35	24	46	3	2.621	1.200
H_0012	12.75	68	6.0	52	16	42	3	2.661	1.077
H_0013	25.1	57	6.1	40	17	49	3	2.911	1.491
H_0539	14.5	54	6.5	42	12	45	3	2.132	0.401
H_0657	24.3	61	8.1	36	25	45	4	3.290	1.604
H_0658	11	48	7.8	42	6	44	3	2.410	0.978
H_0681	19.06	69	7.0	50	19	42	3	2.484	1.088
H_0695	16.15	55	7.9	36	19	45	4	3.581	2.062
H_0700	15.66	50	8.0	33	17	47	3	2.741	1.202
H_0726	14.75	50	9.2	27	23	48	4	3.598	1.201
H_0740	13.58	75	6.4	60	15	42	3	2.621	2.361
H_0750	16.16	53	6.0	38	15	48	4	3.260	1.010

Table 1. HD sample statistics.

Sample ID	PMI	Age of Death	RIN	mRNA-Seq reads
C_0012	19	66	7.1	118,327,116
C_0013	15	69	7.8	89,478,160
C_0014	21	79	8.0	65,377,604
C_0015	10	61	8.2	123,746,070
C_0016	20	58	8.4	67,758,208
C_0017	21	70	8.2	72,238,818
C_0018	17	66	8.5	64,688,322
C_0020	24	60	7.9	83,696,384
C_0021	26	76	7.3	79,487,172
C_0022	17	61	7.8	73,133,936
C_0023	18	62	6.6	94,493,436
C_0024	26	69	8.7	62,989,822
C_0025	25	61	8.1	55,810,684
C_0026	11	88	7.1	72,581,752
C_0029	13	93	6.4	59,386,108
C_0031	24	53	7.3	73,283,170
C_0032	24	57	8.3	70,994,352
C_0033	15	43	7.5	69,505,712
C_0034	14	71	7.8	65,979,612
C_0035	21	46	7.6	62,300,754
C_0036	17	40	7.5	63,961,372
C_0037	28	44	8.3	60,288,132
C_0038	20	57	7.7	61,019,098
C_0039	15	80	7.3	74,892,650
C_0050	2	74	8.5	85,310,070
C_0053	2	69	8.4	167,044,880
C_0060	2	76	7.5	103,952,680
C_0061	3	78	7.6	95,393,100
C_0062	2	87	8.7	83,773,400
C_0065	2	86	8.7	115,714,502
C_0069	24	54	8.3	128,459,102
C_0070	19	68	6.3	145,087,692
C_0071	21	106	7.6	86,840,836
C_0075	23	52	7.4	99,946,984
C_0076	30	46	8.2	85,890,116
C_0077	21	36	8.5	80,103,722
C_0081	26	55	7.6	82,917,984
C_0082	18	57	7.8	123,118,398
C_0083	32	66	8.4	80,696,360
C_0087	19	64	8.7	77,198,978
C_0002	2	73	7.7	120,108,434
C_0003	2	91	7.9	38,420,004
C_0004	2	82	8.6	75,850,406
C_0005	2	97	9.1	150,661,916
C_0006	5	86	8.6	63,607,838
C_0008	2	91	8.7	66,131,458

C_0009	3	81	6.0	69,284,092
C_0010	2	79	8.4	60,542,776
C_0011	2	63	6.5	93,702,684

Table 2. Control sample statistics.

Statistical analysis of the dataset yielded a large set of 5,480 differentially expressed (DE) genes, which prompted us to develop a novel hypothesis-free geneset enrichment method to categorize the large gene lists into functionally and transcriptionally relevant groups. Our computational analytic approach, using Gene Ontology, biological pathway database, and transcription factor regulatory gene sets, implicates groups of related genes and functions that expose and visually organize the fundamental molecular dysfunctions of the disease. Our computational analytic approach implicates a complex profile of genes related to development, most notably HOX genes, strongly reinforces a fundamental role for neuroinflammation in the HD brain, and expands our understanding of cellular involvement in the disease to implicate all major brain cell type as opposed to one of primarily neuronal degeneration.

Methods

Sample Information

Frozen brain tissue from prefrontal cortex Brodmann Area 9 (BA9) was obtained from the Harvard Brain and Tissue Resource Center McLean Hospital, Belmont MA, the Human Brain and Spinal Fluid Resource Center VA West Los Angeles Healthcare Center (Los Angeles, California) and Banner Sun Health Research Institute (Beach et al. 2008-9) (Sun City, Arizona). Twenty Huntington's disease (HD) samples and forty nine neurologically normal control samples were selected for the study (See Tables 1 and 2). Age at death and RIN were significantly different between cases and controls (p=0.01 and p=0.006, respectively, by Welch two sample t-test). The HD subjects had no evidence of Alzheimer or Parkinson disease comorbidity based on neuropathology reports. All samples were male. Neuropathological information for the HD samples includes the Vonsattel grading (Vonsattel et al. 1985), as well as striatal and cortical scoring recently described by Hadzi et al. (Hadzi et al. 2012). Additionally, CAG repeat size and age at onset were known for the HD samples (Table 1).

Human Subjects

This study has been designated exempt (Protocol # H-28974) by the Boston University School of Medicine Institutional Review Board, as no human subjects were studied and all data are derived from post-mortem human brain specimens.

mRNA Sample Preparation and Sequencing

For each brain sample, grey matter from the cortical ribbon was dissected by hand with a target mass of 0.08 g and used for RNA extraction. 1 ug of RNA was used to construct sequencing libraries using Illumina's TruSeq RNA Sample Prep Kit according to the manufacturer's protocol. All sample dissections and RNA extractions were performed by the same individual. RNA Integrity Number (RIN) was measured by the Agilent Bioanalyzer to assess RNA quality prior to sequencing. In brief, mRNA molecules were polyA selected, chemically fragmented, randomly primed with hexamers, synthesized into cDNA, 3' end-repaired and adenylated, sequencing adapter ligated and PCR amplified. Each adapter-ligated library contained one of twelve TruSeq molecular barcodes. Multiplexed samples were equimolarly pooled into sets of three samples per flowcell lane and sequenced using 2x101bp paired-end runs on Illumina's HiSeq 2000 system at Tufts University sequencing core facility (http://tucf-genomics.tufts.edu/). Demultiplexing and FASTQ file generation (raw sequence read plus quality information in Phred format) were accomplished using Illumina's Consensus Assessment of Sequence and Variation (CASAVA) pipeline. Sequences were aligned against the hg19 reference genome (Lander et al. 2001) using tophat v2.0.6 (D. Kim et al. 2013), with non-default parameters.

Gene Expression Quantification, Data Cleaning, and DE Analysis

Aligned reads were mapped to the Gencode v17 annotation (Harrow et al. 2012) using the htseq-count tool in the HTSeq v0.5.3p9 package (Anders, Pyl, and Huber 2014) with the intersection nonempty strategy. Genes that had less than half of HD and control samples with nonzero counts were filtered from the analysis due to low signal. No samples were identified as outliers, and extreme gene measurements considered outliers were adjusted. Outlier-trimmed raw counts were used in subsequent analyses. DESeq2 (Love, Huber, and Anders 2014) was used to identify DE genes between HD and control, adjusting for age at death binned into intervals 0–45, 46–60, 61–75, and 90+ and a categorical RNA Integrity Number (RIN) variable indicating RIN>7 as covariates. Genes with FDR<0.05 were considered DE.

DAVID, GO, and MsigDB Enrichment Calculation

The DAVID (Huang, Sherman, and Lempicki 2008; Huang, Sherman, and Lempicki 2009-1) functional enrichment clustering tool set to the lowest clustering stringency was used on the top 3000 DE genes to identify groups of enriched functions. DAVID limits the number of genes submitted for analysis to 3000. Clusters were considered significant if the cluster score was greater than –log10(0.05). Separate enrichment analyses were performed using the Gene Ontology (GO) annotation database (Ashburner et al. 2000), the MsigDB (Subramanian et al. 2005) C2 Canonical Pathways gene sets, and the MsigDB C3 Transcription Factor target gene sets. Enrichment was calculated for subsets of top DE genes separately, i.e. enrichment analysis was performed on the top 25 genes, then on the top 50, and so on. GO term enrichment was performed using topGO (Alexa and Rahnenfuhrer 2014) with the “weight01” algorithm and “fisher” statistic, and custom scripts in the R statistical environment (R Development Core Team 2008). Enrichment of MsigDB Canonical Pathways and Transcription Factor genesets was performed with custom R scripts using the “fisher.test” and “p.adjust” routines. Once enrichment profiles for each geneset was computed, the genesets were ranked based on the most significant enrichment found in any gene group. The top 15 most significant geneset enrichment profiles from each database were selected and concatenated into a single enrichment matrix with genesets as rows and gene groups as columns. The rows of this matrix were clustered using agglomerative hierarchical clustering with Ward linkage. Further processing of enrichment results was performed using custom scripts to generate plots in python with matplotlib (Hunter 2007), ipython notebook (Pérez and Granger 2007), and pandas (McKinney 2010).

Association with Clinical Covariates

DESeq2 normalized counts were transformed using the Variance Stabilizing Transform (VST) available in the same package to produce approximately normally-distributed gene expression values. After the normal transformation, the standard linear regression model becomes appropriate for evaluating association with covariates. Linear models predicting VST transformed counts from each clinical covariate after adjusting for RIN were run for each gene in the R statistical environment. P-values were adjusted using the “p.adjust” function in R using the FDR method. To assess which DE genes were associated with H-V cortical score, DESeq2 was used to model read counts as predicted by H-V cortical score adjusting for RIN for each gene, adjusted for multiple hypothesis with the “p.adjust” function in R using the FDR method.

Replication of DE Genes by RT-qPCR in an Independent Sample Set

An independent set of 33 HD and 31 control prefrontal cortex brain samples not used in the sequencing study were subjected to RT-qPCR to replicate the findings of this study. RNA was reverse transcribed using iScript cDNA Synthesis Kit (Bio-Rad). Reverse transcriptase quantitative polymerase chain reaction (RT-qPCR) was carried out for all genes of interest in each sample using TaqMan Gene Expression Assays (Life Technologies) on an ABI 7900HT Real-Time PCR system, according to the manufacturer’s protocol. All probes were human and covered all transcripts: HOXC10 (Assay ID Hs00213579_m1) and NFKBIA (Assay ID Hs00355671_g1) probes were used. Peptidylprolyl isomerase A (PPIA, catalog #4333763F) and beta glucuronidase (GUSB, catalog # 4333767F) were used as endogenous controls. Samples were run in triplicate at 200ng mRNA per reaction. For HOXC10, presence or absence of transcripts was assessed by whether a critical threshold (CT) value was determined or undetermined, respectively, at the threshold chosen by Applied Biosystems SDS software v2.4. For NFKBIA, wells that caused the variance of the corresponding set of replicates to exceed 0.2 were marked as outliers and excluded from the analysis (9 such replicates from unique sample/assay combinations were excluded). To normalize sample input, deltaCT values were calculated for each sample by subtracting the average CT for a target gene by the averaged CT for both control genes. Two sample t-tests assuming equal variance with deltaCT values were used for statistical analysis.

Validation of DE Genes by RT-qPCR

The RNA used in the RT-qPCR was from the same extraction as submitted for sequencing and thus was intended to be a technical validation of the sequencing results. Validation samples were prepared and processed for RT-qPCR in the same manner as the replication samples, described above. All probes were human and covered all transcripts: AHNAK nucleoprotein (AHNAK, Assay ID Hs01102463_m1), paired-like homeodomain (PITX, Assay ID Hs00267528_m1), aquaporin 4 (AQP4, Assay ID Hs00242342_m1), solute carrier family 38, member 2 (SLC38A7C, Assay ID Hs01089954_m1), gap junction protein, alpha 1, 43kDa, (GJA1, Assay ID Hs00748445_s1), and tumor protein p53 inducible nuclear protein 2 (TP53INP2, Assay ID Hs00894008_g1) probes were used. As with the replication study, PPIA and GUSB were used as endogenous controls. Samples were run in triplicate at 30ng per reaction. Wells with critical threshold (CT) values higher than 3 standard deviations were removed from analysis. To normalize sample input, deltaCT values were calculated for each sample by subtracting the average CT for a target gene by the averaged CT for both control genes. Wells that were undeterminable were replaced with the maximum number of cycles (40) in order to calculate deltaCT. Two sample t-tests assuming equal variance with deltaCT values were used for statistical analysis.

Results

Widespread Differential Expression Changes Are Observed in HD

After processing sequencing data to reduce noise, remove outliers, and normalize (see Methods), differential expression (DE) analysis identified 5,480 out of 28,087 confidently expressed genes with significantly altered expression at FDR p-values<0.05 in HD vs. control samples, described in Figure 1. More genes are overexpressed in HD versus control than are underexpressed (3,004 vs. 2,476, Figure 1A), and this effect is consistent across the whole list of DE genes ranked by significance (Figure 1B). 76.7% of the DE genes are protein coding according to the Gencode v17 annotation (Harrow et al. 2012), while the remaining most abundant biotypes include lincRNAs, pseudogenes, and antisense transcripts. A greater portion of DE genes is protein coding when compared to the distribution of biotypes in all 28,087 detectable genes as shown in Figure 1C. Notably, the top DE genes are expressed almost exclusively in HD as illustrated in Figure 1D.

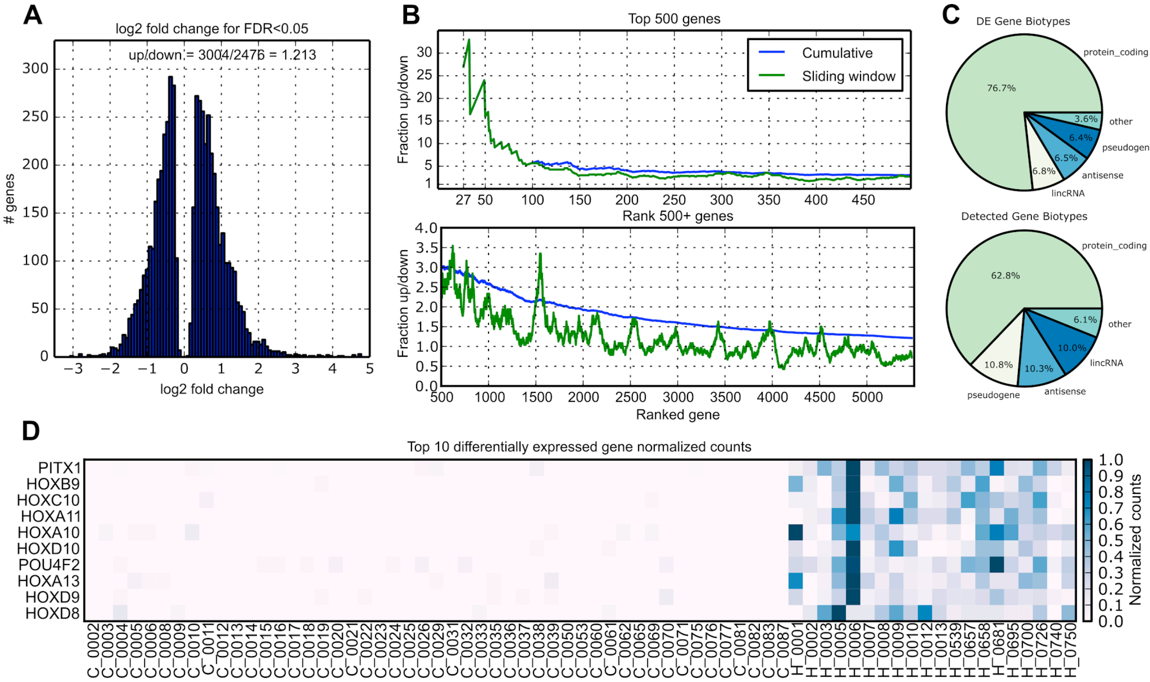


Figure 1. DE statistics. A) Histogram of log2 fold changes for DE genes showing that 54.8% of the DE genes are overexpressed in HD cases. B) Fraction of up vs. down regulated genes across the gene list ranked by significance. Top and bottom plots are top 500 and remaining genes, respectively. Sliding windows lines plot the fraction up vs. down in the 100 gene window of greater rank than the x coordinate. This plot shows that the most highly differentially expressed genes are predominantly over-expressed in HD relative to control BA9. C) Pie chart shows proportions of biotypes for DE genes according to Ensembl. Protein coding genes are overrepresented among the DE genes. D) Normalized counts for all samples in HD and control for top ten significant genes. Rows are normalized for visualization such that the highest count is equal to 1. These genes are almost exclusively expressed in HD cases.

With so many DE genes, it is useful to sort the results in such a manner as to expose meaningful sets of relevant genes. As described in Table 3, the top genes sorted by significance are predominantly located in the Hox clusters and other related developmental genes, a novel result also recently observed for HD in our miRNA study (Hoss et al. 2014). Twenty-four of the 39 HOX genes across all four Hox clusters are DE. The majority of these genes are expressed almost exclusively in HD (see Table 3 and Figure 1D), and consequently attain high significance. However, the relative transcript abundance of these genes is low (e.g. HOXB9 has 8.72 normalized reads on average in the HD samples when the median normalized read count average is 96.6). We sought to identify genes that are both highly expressed and have a large statistically significant difference in expression between HD and control. We created a “differential expression score” (DES) that combines mean expression level, log2 fold change, and statistical significance of differential expression to generate a set of genes that may be relevant to the toxic HD cellular milieu. Table 4 presents the list of the top genes ranked by DES.

Gene Symbol	Overall Mean Counts	HD Mean Counts	Control Mean counts	log2 FC	p-value	padj	DES
PITX1	5.64	18.68	0.32	4.76	9.57E-39	2.69E-34	903.98
HOXB9	2.54	8.72	0.02	4.76	1.63E-25	2.29E-21	249.8732
HOXC10	2.80	9.51	0.06	4.57	2.91E-24	2.72E-20	250.6672
HOXA11	1.96	6.79	0	4.70	3.92E-24	2.75E-20	181.0905
HOXA10	3.49	11.39	0.26	4.27	8.03E-24	4.51E-20	288.5972
HOXD10	2.57	8.77	0.04	4.60	1.35E-23	6.33E-20	227.1957
POU4F2	3.27	10.65	0.26	3.96	3.42E-23	1.37E-19	244.7754
HOXA13	2.45	8.02	0.18	4.16	6.20E-23	2.18E-19	190.9965
HOXD9	2.22	7.18	0.20	3.65	1.22E-18	3.80E-15	117.4429
HOXD8	1.70	5.60	0.12	3.86	2.09E-18	5.88E-15	94.09001
SLC16A12	55.42	167.66	9.60	3.51	4.74E-18	1.11E-14	2717.727
HOXA5	2.19	7.08	0.20	3.87	4.49E-18	1.11E-14	119.0033
HAND1	1.93	6.24	0.18	3.70	1.46E-17	3.16E-14	96.95744
OTP	3.20	9.16	0.76	2.99	3.93E-17	7.88E-14	125.8704
IL17RB	1311.10	2144.33	971.00	1.39	3.80E-16	7.12E-13	22182.16
SLC6A20	173.03	433.28	66.81	2.35	2.49E-15	4.37E-12	4629.918
HOXC6	1.32	4.41	0.06	3.60	4.26E-15	7.04E-12	53.19922
PRKX	604.74	900.29	484.12	1.41	6.22E-15	9.20E-12	9471.658
VNN2	25.74	62.90	10.57	2.49	6.03E-15	9.20E-12	707.7395
HERC2P3	1987.22	3987.18	1170.91	2.06	8.09E-15	1.14E-11	44991.92

Table 3. DE genes by significance.

Gene Symbol	Overall Mean Counts	HD Mean Counts	Control Mean counts	log2 FC	p-value	padj	DES
MBP	180740.9	103940.8	212087.9	-1.14	0.000227	0.003282	513821.5
GFAP	139594.9	147197.9	136491.6	0.74	0.001561	0.013498	194980.9
CLU	98559.44	117016.8	91025.83	0.55	0.000197	0.00296	139030.9
GLUL	61547.89	76676.16	55373.08	0.67	0.000218	0.003176	103210.9
TUBB4A	20856.71	13003.3	24062.19	-0.84	3.44E-08	4.12E-06	94539.22
AQP4	20362.81	27513.91	17443.99	1.09	2.29E-06	0.0001	89094.63
GJA1	13340.95	19835.51	10690.11	1.26	7.06E-08	6.94E-06	86931.93
FAM107A	38970.09	47446.88	35510.18	0.73	0.000164	0.002585	74321.76
SLC38A2	5448.303	9251.666	3895.909	1.31	3.02E-13	2.83E-10	68291.6
SLC1A3	26782.89	35129.11	23376.27	0.85	6.42E-05	0.001294	66171.29
CALM1	83743.27	75824.67	86975.35	-0.34	0.000542	0.006243	64492.7
CALM3	47941.46	38247.79	51898.06	-0.55	0.000424	0.005225	61221.65
AHNAK	9570.149	14157.49	7697.765	1.19	9.48E-08	8.73E-06	57631.07
CTD-2328D6.1	16679.1	5983.11	21044.81	-1.19	0.000217	0.003174	49731.65
NRGN	39663.72	30172.8	43537.57	-0.69	0.002654	0.019734	47221.27
GAS7	15300.17	11322.25	16923.81	-0.69	5.64E-07	3.50E-05	47122.17
TP53INP2	6501.307	3430.574	7754.667	-1.37	8.45E-08	8.02E-06	45652.02
HERC2P3	1987.225	3987.18	1170.917	2.06	8.09E-15	1.14E-11	44991.92
ENO2	25831.65	20005.15	28209.81	-0.57	6.29E-05	0.001273	42930.44
MAP1B	37563.64	29736.15	40758.53	-0.51	0.00057	0.006441	42770.9

Table 4. DE genes by DES. Differential Expression Score (DES) is calculated as (overall mean counts) x abs(log2 FC) x –log10(adjusted p-value)

A number of key proinflammatory genes appear as DE in this dataset. Four of the five NFkB family members NFkB1 (log2 fold change 0.32, q=0.004), NFkB2 (LFC 0.73, q=0.001), RELA (LFC 0.63, q=5.6e-5), and RELB (LFC -0.56, q=0.005) are DE in this dataset. When we examine the 20 interleukin-related genes in the DE gene list, we find that fifteen are cytokine receptors (including IL17RB, IL13RA1, IL4R). However, the cytokines that correspond to these receptors are not DE, nor are TNFalpha or IL6, two primary cytokines of the immune and inflammatory response.

An independent set of 33 HD and 31 control prefrontal cortex brain samples not used in the sequencing study were subjected to Reverse transcriptase quantitative PCR (RT-qPCR) to replicate the findings of two genes found to be DE in this study. HOXC10 and NFKBIA, genes associated with developmental and neuroinflammatory processes, respectively, were chosen for the replication. HOXC10 mRNA species were not detected in any of the control samples, whereas 11 HD samples showed amplified product after 40 PCR cycles (p=0.0002). The presence of HOXC10 mRNA transcripts in HD, and absence in controls, is consistent with the sequencing findings. In the 16 HD and 16 control samples selected for highest mRNA quality, NFKBIA was detected in all samples and, after filtering outlier replicates, was found to be significantly more abundant in HD samples (T=-1.804, p=0.041).

RT-qPCR was used to quantify and orthogonally validate mRNA differential expression from sequencing. Six genes were selected for the study AHNAK, AQP4, SLC38A7C, GJA1, TP53INP2, which had high DES scores, and PITX1, which was the most significantly differentially expressed gene. 21 controls and 15 HD samples from the sequencing study were selected for the assay. Four of the six genes were statistically significant (AHNAK p=0.02; SLC38A7C p=0.01, TP53INP2 p=0.03, PITX1 p=3.4e-10). Two genes did not meet significance (AQP4 p=0.08, GJA1 p=0.08). All differential expression was in the expected direction.

Immune Response, Development, and Transcriptional Regulation Functions Are Enriched in HD

We sought to explore which biological processes are enriched among DE genes in HD. These analyses were performed using the DE list of 5,480 genes ranked by significance. DAVID Functional Enrichment Clustering (Huang, Sherman, and Lempicki 2008; Huang, Sherman, and Lempicki 2009-1) of the top 3000 DE genes (*the DAVID tool restricts the input list size to 3000 genes) identifies numerous biological functions related to immune response, development, cell growth, and transcriptional regulation. Table 5 contains a summary of the enriched clusters identified by DAVID that are significant at a cluster score corresponding to FDR $p<0.05$. DAVID does not enforce mutually exclusive gene membership between GO categories/pathways and thus one finds redundancy in the list of clusters. The themes of immune response, development, and transcriptional regulation are seen as the most consistent functional groups in this analysis. Figure 2 depicts the functional clusters identified by DAVID as a network where nodes are the DE genes underlying the clusters and edges represent common genes between clusters. The cluster with the largest number of genes is immune response with 1,248, followed by skeletal system development with 921.

#	Cluster Function	Cluster Term Keywords	# genes	# terms	score
1	immune response	membrane, plasma, transmembrane, receptor	1248	27	3.764689
2	identical protein binding	protein, activity, identical, function	212	5	3.346027
3	metallothioneins	metal, binding, ion-binding, cluster	33	17	3.338415
4	skeletal system development	morphogenesis, embryonic, regulation, development	577	80	3.186388
5	skeletal system development	regulation, transcription, process, negative	921	76	3.143774
6	gland development	development, gland, mammary, lactation	39	3	2.793014
7	immune system development	myeloid, differentiation, leukocyte, cell	78	11	2.637665
8	pattern specification process	symmetry, determination, pattern, left/right	62	5	2.39939
9	response to oxygen levels	response, oxygen, ovulation, process	54	4	2.374104
10	growth	growth, regeneration, developmental, tissue	52	4	2.325598
11	extracellular matrix	extracellular, matrix, proteinaceous, part	63	4	2.27691
12	cell growth	growth, cell, developmental	36	3	2.222128

Table 5. DAVID functional clustering. Cluster Function labels were assigned manually by inspecting the terms within the cluster but generally correspond to the name of the most enriched term within the cluster. The (1) Immune response cluster contained 27 distinct terms from across the default genesets used by DAVID.

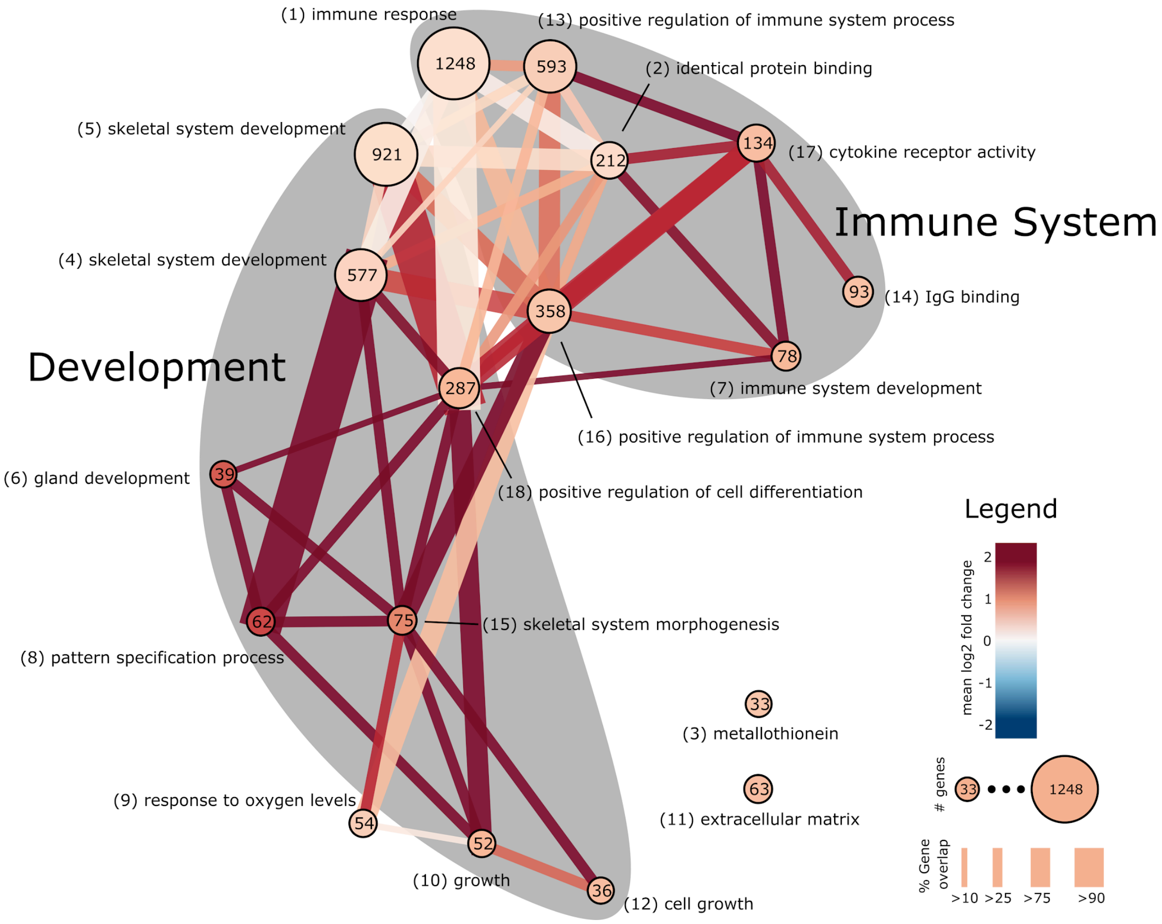

(1) immune response
1248
(13) positive regulation of immune system process
593
(2) identical protein binding
(5) skeletal system development
921
212
134
(17) cytokine receptor activity
Immune System
(4) skeletal system development
577
358
93
(14) IgG binding
Development
78
287
(7) immune system development
(16) positive regulation of immune system process
(6) gland development
39
(18) positive regulation of cell differentiation
Legend
2
1
0
-1
-2
mean log2 fold change
62
75
(15) skeletal system morphogenesis
(8) pattern specification process
33
(3) metallothionein
# genes
33
1248
63
(11) extracellular matrix
(9) response to oxygen levels
54
52
% Gene overlap
>10
>25
>75
>90
(10) growth
36
(12) cell growth

Figure 2. DAVID functional clustering network. Network representation of the DAVID clusters from Table 5. Nodes represent clusters, the size of the node is proportional to the number of unique genes that make up the cluster and numbers within nodes are the number of unique genes mapped to terms in the cluster. Edges between nodes indicate the existence of overlapping genes, where the width is proportional to the percent overlap of genes in the smaller of two connected nodes. The color of nodes and edges is proportional to the average fold change of the genes in the node or edge.

Integrated Geneset Enrichment Analysis Identifies Specific Enriched Functional Categories

The DAVID results, while informative, did not provide sufficiently detailed information to understand how the DE gene list mapped to biological functions. To attain a more fine-grained understanding of the enriched biological functions and characteristics of the DE genes, we next performed a detailed analysis of subsets of the DE gene list using the Gene Ontology (GO) annotation database (Ashburner et al. 2000) and the MsigDB (Subramanian et al. 2005) C2 Canonical Pathways and C3 Transcription Factor target gene sets (see Methods). Briefly, the central idea of the method is to partition the gene list into groups that include increasing numbers of DE genes, where the first group contains the top 25 DE genes, the second group the top 50, and so on for the entire gene list. The last group contains all 5,480 DE genes. Each of these groups is then used to calculate enrichment against each geneset separately using an appropriate statistical method (see below), and then the results from each gene set are concatenated and hierarchically clustered.

GO Enrichment Analysis Implicates Development and Immune Response

GO term enrichment was calculated using topGO (Alexa and Rahnenfuhrer 2014), a tool that uses the GO term hierarchy to identify enrichment of the most biologically specific categories given a gene list. Figure 3 depicts GO term enrichment of ranked subsets of genes ordered by the most significant term across all subsets. Enrichment is only shown for gene subset/term pairs that attain significance at $p<0.05$. In total, 901 biological process (BP) terms, 168 molecular function (MF) terms, and 68 cellular component (CC) terms were found to be significant in at least one of the ranked gene subsets. Performing analysis on subsets of top enriched genes reveals that developmental processes and transcriptional regulation are enriched among the most DE genes, while immune response genes are found throughout the DE gene list. Table 6 contains detailed statistics on the top enriched GO terms. These detailed results are consistent with the cluster results from DAVID and better expose the specific biological functions involved in the DE gene list.

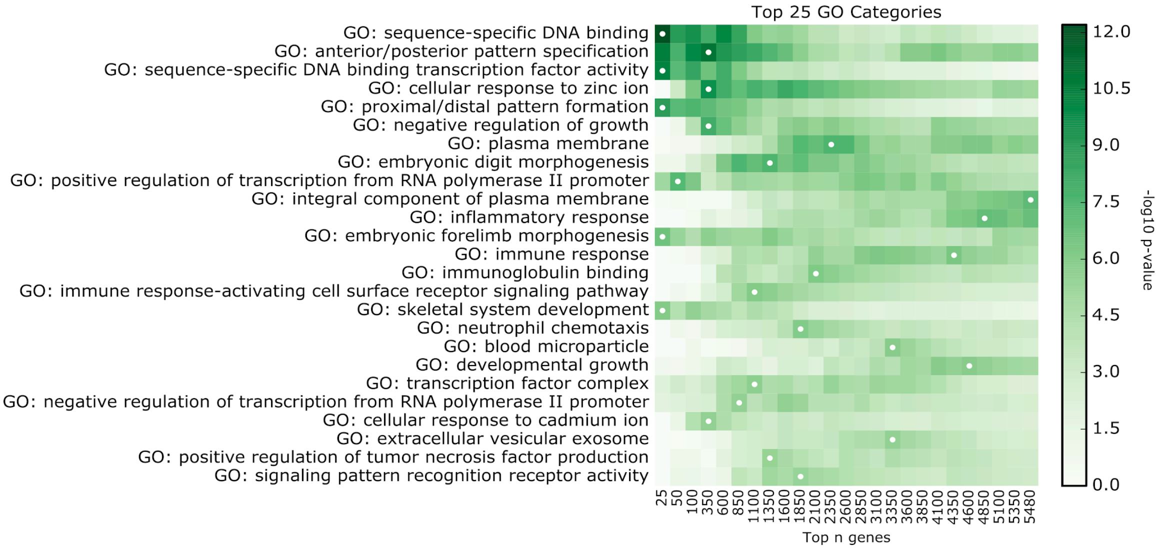


Figure 3. Detailed GO enrichment. Top 25 enriched GO categories across all three GO namespaces identified by topGO for different numbers of DE genes. X-axis indicates the number of top genes used for the enrichment in each GO category, e.g. the first column uses the top 25 genes, the second column uses the top 50, and so on. The intensity is proportional to –log10(p-value) from topGO. White dots indicate the gene set with the most significant p-value, concordant with Table 5. This figure shows that the first three GO Categories are defined by genes that are among the top 25 to 150 DE genes in the dataset. GO Categories further down the list are defined by genes whose differential expression is less pronounced between HD and controls.

GO Category	Top n genes	-log10(p-value)
GO: sequence-specific DNA binding	25	12.211851
GO: anterior/posterior pattern specification	350	10.890978
GO: sequence-specific DNA binding transcription factor activity	25	10.19469
GO: cellular response to zinc ion	350	9.630161
GO: proximal/distal pattern formation	25	8.874839
GO: negative regulation of growth	350	7.983699
GO: plasma membrane	2350	7.603115
GO: embryonic digit morphogenesis	1350	7.542254
GO: positive regulation of transcription from RNA polymerase II promoter	50	7.350002
GO: integral component of plasma membrane	5480	7.167156
GO: inflammatory response	4850	7.057813
GO: embryonic forelimb morphogenesis	25	6.633754
GO: immune response	4350	6.311688
GO: immunoglobulin binding	2100	6.178673
GO: immune response-activating cell surface receptor signaling pathway	1100	6.135233
GO: skeletal system development	25	6.059758
GO: neutrophil chemotaxis	1850	6.038185
GO: blood microparticle	3350	5.968453
GO: developmental growth	4600	5.939322
GO: transcription factor complex	1100	5.636701
GO: negative regulation of transcription from RNA polymerase II promoter	850	5.624786
GO: cellular response to cadmium ion	350	5.593366
GO: extracellular vesicular exosome	3350	5.489169
GO: positive regulation of tumor necrosis factor production	1350	5.481418
GO: signaling pattern recognition receptor activity	1850	5.366574

Table 6. Enriched GO Categories. The most enriched GO category GO:sequence-specific DNA binding using the top 25 DE genes ranked by significance. The second most enriched GO category, GO:anterior/posterior pattern specification, was found when considering the top 350 DE genes.

Pathways Involved in Multiple Immune System Processes Are Enriched

To identify biological pathways as opposed to functional categories, we performed hyper-enrichment of the MsigDB C2 Canonical Pathways using a hypergeometric test on the same ranked subsets of genes as in the GO analysis. These analyses found 538 significantly enriched pathways in at least one gene subset. Enriched Canonical Pathways show a clear immune response and inflammation-related pattern across pathway databases, including Reactome

(Croft et al. 2014-1; Milacic et al. 2012) innate immune system [DOI: 10.3180/REACT_6802.2], KEGG (Kanehisa and Goto 2000) complement and coagulation cascades [hsa04610] and cytokine-cytokine receptor interaction [hsa04060], and PID (Schaefer et al. 2009-1) IL4-mediated signaling events [Pathway id:il4_2pathway] and NFkB canonical pathways [Pathway id:nfkappabcanonicalpathway].

DE Genes Are Enriched as Targets of Transcription Factors Implicated In HD

We next performed transcription factor (TF) target analysis using the MsigDB C3 TF regulation gene set to identify potential regulators responsible for the observed differential expression. 237 TFs were identified as significantly enriched in at least one gene subset. A number of the enriched TFs are known to physically interact with the mutant Htt protein, including SP1 (S.-H. Li et al. 2002) and TBP (van Roon-Mom et al. 2002). The pattern of enrichment for the top TF, MYC-associated zinc finger protein (MAZ), tracks closely with pathways associated with immune response (i.e. both become more enriched as more genes are included) but otherwise has no previous connection with HD. The second most enriched TF is forkhead box O4 (FOXO4). Another notable enriched TF is NFkB, which plays a key role in innate immune response, is critical for glial and neuronal cell function and synaptic signaling (O’Neill and Kaltschmidt 1997) and impairs synaptic transport in the presence of mutant Htt protein (Marcora and Kennedy 2010). Other TFs implicated as potential regulators of the DE genes include NFAT (Hayashida et al. 2010), HSF1 (Neef, Turski, and Thiele 2010), and PU1 (Crotti et al. 2014).

Integrated Geneset Enrichment Analysis Links Biological Function and Transcriptional Regulation

The top fifteen most enriched gene set profiles from each of GO, Canonical Pathways, and Transcription Factors were concatenated and hierarchically clustered to identify which gene sets are enriched in similar DE genes, as shown in Figure 4. The clustering identifies five groups of genesets that correspond primarily to either immune response or developmental functions (A-C, and D-E respectively in Figure 4). Transcription Factor genesets are clustered with pathway and GO genesets, indicating which co-regulated genes are associated with which biological functions. Further remarks on this result are found in the Discussion section.

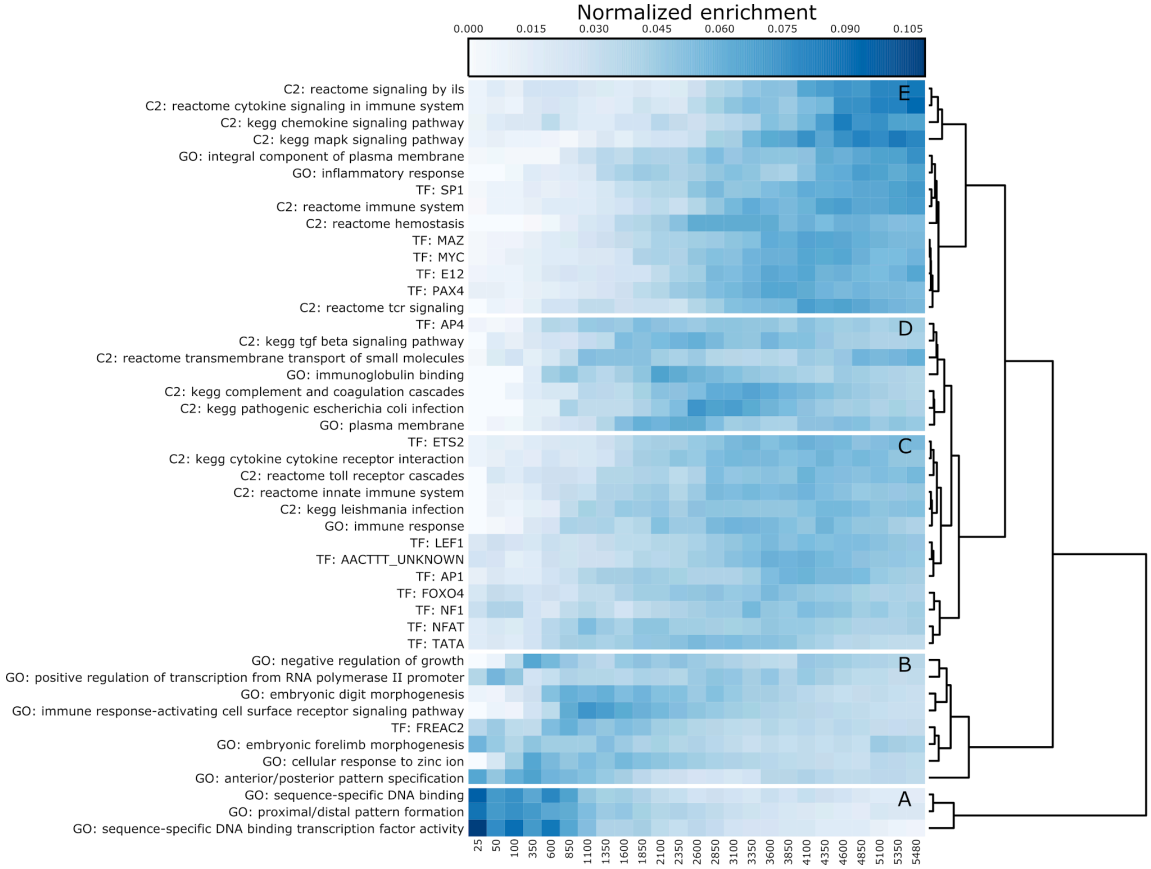


Figure 4. Clustergram of Top Enriched Pathway, TF, and GO terms. Concatenated enrichment profiles for GO, C2, and TF gene set collections, similar to Figure 3, ordered by hierarchical clustering of Euclidean distance between rows. Rows have been normalized by dividing by the row sum for visualization, intensity is proportional to normalized enrichment. Heatmap is partitioned into groups A-E based on hierarchical clustering. Clusters A, B, and C are primarily involved in the immune response and are enriched in gene subsets that include more genes. Clusters D and E are predominantly related to developmental and transcriptional regulation processes.

Association of Gene Expression with Clinical Covariates

Genes whose expression is associated with CAG-adjusted age at onset are potential genetic factors that modify the presentation of disease independent of CAG repeat length, though in the presence of the mutation, and thus may be useful as a biomarker in identifying patients at risk of early onset. Therefore, to identify genetic factors that may modify clinical covariates, each of the 28,087 confidently expressed genes was analyzed for association with CAG repeat length, CAG-adjusted residual age at onset, and scores representing cortical and striatal involvement using the Hadzi-Vonsattel (H-V) method (Hadzi et al. 2012). Due to the significant association between age at onset and CAG repeat length, a CAG-adjusted residual age at onset variable was constructed with the model from Djousse et al (Djousse et al. 2004-6) and used to test for association (see Methods).

Association was assessed using a linear regression model predicting normalized, normally-transformed counts (see Methods) from each covariate separately, adjusting for RNA integrity number RIN. No gene associations reached genome-wide significance after multiple hypothesis adjustment, though many reached nominal significance as described in Table 7. We did not find any significant association between gene expression in HD brains and either the striatal or cortical H-V involvement scores. While this may be a consequence of the relatively small sample size of twenty HD brains studied here, it is also worth noting that these brains exhibited a wide range of cortical (from 0.401 to 2.361) and striatal (from 2.132 to 3.820) involvement on the H-V scale. To identify potential confounding in the DE gene list by cortical involvement, we analyzed the DE gene counts to identify any with significant association with H-V cortical score (see Methods). None of the DE genes attained significance after multiple hypothesis adjustment, indicating the DE gene results are not confounded by cortical involvement.

CAG Repeat Length			CAG Adjusted Onset			Cortical involvement score		
Gene	beta	p-value	Gene	beta	p-value	Gene	beta	p-value
C2CD3	-0.07	0.0001	CAPN8	0.58	0.0001	STRADB	0.27	7.8E-05
NPBWR1	0.22	0.0002	ARSF	-0.58	0.0004	ABCF3	-0.36	0.00038
GPR142	-0.13	0.0002	BICD2	-0.22	0.0004	BARD1	0.88	0.00042
CEP95	-0.09	0.0004	MYB	-0.68	0.0007	TMEM190	0.58	0.00051
C18orf42	0.20	0.0005	GDF5	0.68	0.0012	GLUD1	0.62	0.00052
NNAT	0.17	0.0006	KLHL40	0.53	0.0014	F2R	1.01	0.00054
OFD1	-0.10	0.0006	PODNL1	0.55	0.0015	FAM64A	-1.01	0.00054
SOX1	0.11	0.0006	CRELD2	0.30	0.0017	SDC4	1.06	0.00055
PCDH8	0.23	0.0007	PLEK2	-0.63	0.0018	RIN2	0.82	0.00067
NAA20	0.06	0.0007	ZNF398	-0.25	0.0018	ANGPTL4	1.49	0.00075
SH3TC2	-0.24	0.0008	EPS8L2	0.38	0.0025	STOX1	0.70	0.00078
RWDD2B	0.10	0.0008	PAX5	-0.64	0.0025	DLK2	-0.77	0.00089
IGF1	0.19	0.0008	GATSL1	-0.42	0.0028	WWOX	0.44	0.00099
PAPL	-0.21	0.0008	ICMT	-0.24	0.0031	RFC5	-0.32	0.00102
DST	-0.13	0.0008	NPY2R	-0.78	0.0032	DPH2	-0.30	0.00112
C1orf131	-0.06	0.0008	POLA2	0.33	0.0034	ETNPPL	0.98	0.00118
GDNF	-0.15	0.0009	PRPSAP1	0.24	0.0035	PON2	0.71	0.00135
PDCD2	0.034	0.0009	TTC16	0.45	0.0036	ELP4	0.60	0.00136
NCKAP5	-0.14	0.0010	C3orf52	-0.56	0.0036	MYADM	-0.40	0.00143
FAM194A	0.16	0.0010	FAM127C	0.19	0.0040	NR5A1	-0.65	0.0014

Table 7. Protein coding genes associated with clinical covariates. P-values are nominal.

Discussion

We conducted mRNA transcriptional analyses in HD and control brains to identify altered gene expression profiles in this disease. To our knowledge, these are the first reported results from a gene expression analysis of high-throughput mRNA sequencing from post-mortem human HD and control brains. Widespread DE genes strongly implicate immune response, transcriptional dysregulation, and extensive developmental processes across all primary brain cell types (i.e. astrocytes, oligodendrocytes, microglia, and neurons). The genes from the DES-ranked list in Table 4 reveal a variety of disease related processes, implicating genetic signatures for different brain cell types as well as genes heavily associated with brain injury and neurodegeneration. The top two DES-ranked genes, MBP (myelin basic protein) and GFAP (glial fibrillary acidic protein), are typical markers used to identify oligodendrocytes and reactive astrocytes, respectively (Baumann and Pham-Dinh 2001). These proteins have also been implicated in immune processes, blood-brain barrier permeability, and response to injury in the central nervous system (Baumann and Pham-Dinh 2001; D'Aversa et al. 2013-4; Lumpkins et al. 2008). The next highest DES-ranked gene, CLU (clusterin), is associated with clearance of cellular debris, lipid recycling, apoptosis, and, as a stress-induced secreted chaperone protein, has been genetically associated with late-onset Alzheimer's disease (Jones and Jomary 2002). GLUL (glutamate-ammonia ligase) is a glutamine synthetase found primarily in astrocytes in the brain and is involved in neuron protection from excitotoxicity through the conversion of ammonia and glutamate to glutamine (Suárez, Bodega, and Fernández 2002). Alteration in TUBB4A (tubulin beta-4A chain), a major component of microtubules, has been associated with neurodegenerative diseases caused by hypomyelination with atrophy of the basal ganglia and cerebellum (Blumkin et al. 2014). AQP4 (aquaporin) is a specific marker for astrocytic endfeet and has been linked to Ca2+ induced edema (Thrane et al. 2011). ENO2 (ennolase), a neuron-lineage-specific gene ranked 19th by DES, has been identified as a marker for ischemic brain injury (Cronberg et al. 2011). Although it is not included in the top list, the analysis also identified CD40, a protein uniquely expressed in activated microglia for antigen presentation in the brain (Ponomarev, Shriver, and Dittel 2006). Together, these genes suggest a systemic response in all brain cell types to stress and brain injury.

While some of the differences in gene expression that are observed in our studies are almost certainly a consequence of alterations in the cellular distribution in HD due to the loss of neuronal cells and the reactive response to degeneration in the HD brain, it is important to note that we did not find that the levels of gene expression in HD brains were related to the extent of cortical involvement. Specifically, while the HD samples in this study range from very low (H-V cortical score 0.401) to very high (H-V cortical score 2.361) levels of cortical involvement, levels of differentially expressed genes were not found to be significantly associated with H-V cortical score. Because the H-V cortical score comprehensively characterizes the level of involvement and cellular architecture of the HD brains studied, these findings suggest that the differentially expressed genes are not simply a reflection of altered distribution of cell types in the samples studied.

DAVID functional clustering analysis identified a number of functionally related clusters with overlapping genes. The network in Figure 2 illustrates that the immune system and developmental clusters are highly interrelated in their underlying genes, suggesting a link between these cellular processes. The detailed analysis of different gene subsets for enrichment of GO, Canonical Pathways, and Transcription Factors affords some insight into this relationship as illustrated in Figure 4. The top fifteen most enriched gene set profiles from each collection were concatenated and hierarchically clustered to identify which gene sets are enriched in similar DE genes. The clustering identifies five distinct clusters that are functionally organized into coherent groups (labeled A-E in Figure 4). Clusters A, B, and C are primarily involved in the immune response and are enriched in gene subsets that include more genes. Transcription factors SP1, MAZ, MYC, E12, and PAX4 are enriched in similar sets of DE genes that are also involved in inflammatory and immune response, suggesting these functions are transcriptionally related. Clusters D and E are predominantly related to developmental and transcriptional regulation processes, and are clustered with transcription factor FREAC2 (Forkhead Box F2, also known as FOXF2) which, as a member of the forkhead family of transcription factors, is potentially implicated in development, organogenesis, regulation of metabolism, and immune system processes (Jackson et al. 2010).

The strong implication of immune response and neuroinflammation in this study is consistent with prior reports as a critical aspect of the human response to HD (Ellrichmann et al. 2013; Silvestroni et al. 2009; Björkqvist et al. 2008). The set of DE genes is highly enriched for multiple immune system processes, including both innate and adaptive immune response, implicating a tissue-wide immune response at multiple cellular levels. The presence of the proinflammatory genes NFkB and interleukins (IL8, IL9, IL15, IL18) is strong indication of an innate immune response and is previously reported in the HD literature (Ellrichmann et al. 2013; Silvestroni et al. 2009; Björkqvist et al. 2008).

Except for our recent miRNA finding (Hoss et al. 2014), the Hox locus has not previously been implicated in HD in model or human systems. The extent of altered developmental genes is quite striking and affords no immediate interpretation since the enriched developmental processes seem to be specific to cell types that have no obvious role in the central nervous system (i.e. skeletal, limb morphogenesis, etc.). This apparently non-specific developmental enrichment might therefore be a consequence of profound transcriptional changes related to the extreme inflammatory stress experienced by the affected brain regions as well as transcriptional dysregulation due aberrant interactions between TFs and mutant HTT protein fragments. It is still unclear whether a subset or if all brain cell types are responsible for this signal, and elucidation of the source of the developmental gene transcription may provide further insight into the cell type specificity of transcriptional dysregulation.

This dataset suggests the calpain family of proteolytic proteins plays a role in HD. Calpains have a direct role in the cleavage of mutant Htt into toxic fragments (Gafni and Ellerby 2002) and the inhibition of these proteins leads to decreased neuronal toxicity in in vitro settings (Gafni et al. 2004). Three calpains, CAPN2, CAPN7, and CAPN11, are significantly DE in this dataset, where 2 and 7 are highly abundant and up-regulated in HD while 11 shows low expression and is down regulated. Calpains are typically activated by elevated intracellular Ca+2 levels (Goll et al. 2003) and there is significant evidence in this dataset that genes responsive to calcium and other ionic metals are activated. Four of the eight calmodulin related genes (CALM1, CALM2, CALML3, CALML4) are DE in the dataset, and are all significantly down regulated with the exception of CALML4 (LFC -0.55, -0.35, -0.97, 0.42, respectively). Calcium plays a key role in apoptotic phagocytosis and the inflammatory response (Gronski et al. 2009; Razzell et al. 2013), processes that are strongly implicated in this dataset, and disrupted calcium concentration has been implicated in HD and neurodegeneration in general (Giacomello, Hudec, and Lopreiato 2011; Wojda, Salinska, and Kuznicki 2008). Among the enriched GO categories are calcium-dependent protein binding, calcium-dependent phospholipid binding, cellular response to cadmium ion, and cellular response to zinc ion. Metallothioneins appear as one of the most enriched DAVID functional clustering results, with nearly every metallothionein 1 subtype DE in the dataset (all except MT1B). Altogether, this dataset strongly implicates the presence of metal ion disequilibrium in the HD context. Though the presence of ion disequilibrium is strongly implicated by this study, it is unclear whether this effect is a cause or a consequence of the toxic effects of mutant Htt.

A popular hypothesis asserts that mitochondrial dysfunction contributes to neurodegeneration in HD (Damiano et al. 2010; Costa and Scorrano 2012; Schapira et al. 08/2014). Dysregulation of mitochondrial function in HD is thought to be induced by disrupted cytoplasmic Ca2+ concentrations (Damiano et al. 2010) which lead to alterations in bioenergetic processes and mitochondrial morphology (Costa and Scorrano 2012). Several of the signals observed in this study suggest an imbalance in calcium ion homeostasis in the human HD brain as described above, which supports the hypothesis that mitochondrial dysfunction is implicated in human HD. However, none of the mitochondrial genes are DE in this dataset.

In contrast to this study, Hodges et al (Hodges et al. 2006) found no detectable gene expression changes for HD in post mortem BA9 tissue. Nonetheless, there are consistencies between our findings. First, although overall gene expression was observed to be down regulated in the striatum for Hodges et al, the distribution of fold changes for BA9 in both studies indicate overall up regulation. Second, and more significantly, there is suggestive overlap of enriched biological processes between the two datasets across brain regions. Specifically, they observed that central nervous system and neuronal developmental genes, ion transport, microtubule, and vesicle-related processes were enriched, signals also observed in this study.

The discovery of thousands of statistically significant differences in gene expression presented a major challenge to the interpretation of this dataset. The DAVID analysis, which is specifically designed to interpret large gene lists, was not sufficiently detailed to readily provide insight about which genes were involved in which functions, nor did the tool organize its output in a way that presents how different enriched genesets are related. The method developed here addresses both of these issues, and allows the use of different statistical enrichment methods, as appropriate, for different gene sets. It also combines and visualizes the enrichment information in such a way as to facilitate generating specific hypotheses concerning which genes are related through their enrichment profiles. The link between genes that are regulated by TFs known to interact with mHtt fragments and their immunological functions (Figure 4 cluster A) proposes a mechanism by which mHtt may play a toxic role to cells, namely via transcriptionally altering genes involved in the immune response. FOXF2 was also identified as a TF that is potentially responsible for aspects of both the inflammatory and developmental gene expression changes (Figure 4 cluster D). These insights were not obvious from the DAVID results, demonstrating the utility of our novel analytical methodology.

These data represent the most comprehensive characterization of genome-wide gene expression in human HD subjects to date. The broad scope of changes across biological functions and cell types establishes HD as a systemic disease of the brain, implicating not only neurons but also the primary glial cell types. This new molecular evidence supports previous imaging-based observations of cortical and whole-brain structural changes in HD (Selemon, Rajkowska, and Goldman-Rakic 2004; Squitieri et al. 2009). The immune response is intrinsically intercellular in its activation and function, cued by the complex interaction of stressed neurons and the reactive glial cells of the central nervous system immune response. This brings into focus the importance of considering the HD brain as a whole organ, and important advances in understanding and mitigating HD pathogenesis may be gained by developing and studying models of these complex multi-cellular interactions. In particular, in vitro studies of human-derived neuronal HD cell line models and HD mouse models cannot capture the complexity of the human brain microenvironment, an especially important point for mouse models due to the compelling differences between the human and murine inflammatory response (Seok et al. 2013). It remains to be shown precisely which cell types are responsible for which aspects of the biological response observed in this study. Similarly, it is not known how the immune and developmental DE genes are related, and whether some complex combination of these genes can be shown to modulate clinical features of disease, in particular age of onset. It is conceivable that subjects with a different or more extreme immune response may experience neurodegeneration differently than others, and we hypothesize that this avenue of research will yield important advances in our understanding of HD pathogenesis.

Acknowledgments

We would like to acknowledge Ms. Jayalakshmi Mysore for her help with the HD postmortem brains and the HTT CAG genotyping. This work was supported by the Jerry McDonald HD Research Fund (RHM) and Public Health Service, National Institutes of Health grants, National Institute of Neurological disorders and stroke, R01-NS073947 (RHM) and R01-NS32765 (MEM) and PHY-1444389 NSF-EArly-concept Grants for Exploratory Research (EAGER). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 3. Comparative Analysis of HD and PD Gene Expression

Introduction

Huntington's Disease (HD) and Parkinson's Disease (PD) are progressive neurodegenerative disorders that are marked by common histological patterns and, to a lesser degree, clinical symptoms. Broadly, both diseases are characterized clinically by loss of motor function and cognitive decline, and histologically by aberrant protein aggregation in neurons and the selective degeneration of neurons in contrasting patterns in the brain (Dexter and Jenner 2013; Vonsattel et al. 1985). However, HD patients begin presenting symptoms on average by the age of 40 (R. H. Myers, Marans, and MacDonald 1998), where most PD patients onset is age 60 on average (Dexter and Jenner 2013), and both the proteins that aggregate and the types of neurons affected in the diseases are distinct. Specifically in HD, a polyglutamine expansion in the Htt protein encoded by a mutant HTT gene causes Htt protein fragments to aggregate, a process that the medium spiny neurons of the caudate nucleus and putamen regions in the brain are particularly vulnerable (Vonsattel et al. 1985). In PD, toxic fragments of the aSyn protein encoded by the SNCA gene aggregate, for which dopaminergic neurons of the substantia nigra are vulnerable (Dexter and Jenner 2013). Thus, protein aggregates are coincident with degenerating neurons in both diseases, but the contribution of the protein aggregates to substantively non-overlapping neuronal cell type and brain region vulnerability is unclear (Arrasate and Finkbeiner 2012; Rubinsztein and Carmichael 2003).

Consequently, proteolytic dysfunction is hypothesized to be a common mechanism to both diseases, but the extent to which other shared biological processes may underlie their neurodegenerative pathology is remains to be shown.

Transcriptional dysregulation has been observed in both HD and PD (Cha 2007; Elstner et al. 2011). Transcription, neuroinflammation, and developmental processes have been shown to be dysregulated in the brains of HD individuals (A. Labadorf et al. 2015), while inflammation and mitochondrial dysfunction were observed to be altered in the brains of PD individuals (Dumitriu et al. 2012). However, a systematic comparison of the transcriptional signatures of HD and PD has not been performed to date, and those genes and biological processes common to both diseases, if any, remain to be determined. To address this question, we sought to identify genes that are consistently differentially expressed (DE) in the post-mortem brains of HD and PD human subjects compared to neuropathologically normal control brains using mRNA-Seq. We hypothesize that common altered genes and pathways in HD and PD will elucidate the mechanistic underpinnings of the neurodegenerative process. This study presents the results of a comparison of DE genes for each of HD and PD versus controls analyzed separately. In addition, in order to identify consistent effects with lower effect size across diseases, an analysis was performed where the HD and PD datasets are concatenated as a single category (neurodegenerative disease, ND) and compared with controls. DE genes are determined using a specialized form of logistic regression as described in (Choi et al, under review), which better controls type I errors when compared to negative binomial based DE detection methods.

Materials and Methods

Sample collection, processing, and sequencing

The HD, PD, and control samples used in this study are those previously described in our past work (A. Labadorf et al. 2015; Dumitriu et al. 2012). An additional nine HD brain samples were included in this study beyond those in (A. Labadorf et al. 2015), including two HD gene positive asymptomatic individuals, obtained from the Harvard Brain Tissue Resource Center. All samples underwent the same tissue dissection and RNA extraction sample preparation protocol performed by the same individuals. Briefly, RNA was extracted from the prefrontal cortex of postmortem brains of HD and PD subjects, as well as neuropathologically normal controls. RNA was poly-A selected and subjected to mRNA sequencing on the Illumina HiSeq 2000 platform. Sample statistics are contained in Table 8. See (A. Labadorf et al. 2015) and (Dumitriu et al. 2012) for more detailed information about sample preparation.

Sample type	N	Mean (SD) Age at death	Mean (SD) PMI	Mean (SD) RIN
HD	29	60.5 (11.4)	16.4 (7.8)	7.1 (1.2)
PD	29	77.5 (8.9)	11.1 (9.7)	7.0 (0.7)
Control	49	68.6 (15.8)	14.6 (9.5)	7.8 (0.7)

Table 8. Sample Statistics

mRNA-Seq data processing

Each FASTQ file containing mRNA sequences was first trimmed for sequence quality using the sickle software package (Joshi NA, Fass NJ 2011) with default arguments. The resulting short reads were aligned against the hg38 build of the human reference genome using the STAR aligner v2.4.0h1 (Dobin et al. 2013) with permissive multimapping parameters (200 maximum alignments – outFilterMultimapNmax 200) and otherwise parameter values suggested in the STAR manual. Multimapped reads were assigned unique alignment locations using the ORMAN software package (Dao et al. 2014). Aligned reads were counted against GENCODE v21 annotation (Harrow et al. 2012)using the HTSeq package v0.6.1p1 (Anders, Pyl, and Huber 2014). Read counts for all samples were normalized using the method described in the DESeq2 package v1.10.1 (Love, Huber, and Anders 2014) and outlier counts were trimmed using the strategy described in (A. Labadorf et al. 2015). Since the original were poly-A selected, only genes with biotypes known to be polyadenylated (i.e. ‘protein_coding’, ‘lincRNA’, ‘processed_transcript’, ’sense_intronic’, ’sense_overlaping’, ’IG_V_gene’, ’IG_D_gene’, ’IG_J_gene’, ’IG_C_gene’, ’TR_V_gene’, ’TR_D_gene’, ’TR_J_gene’, and ’TR_C_gene’) as annotated by Ensembl BioMart (Kinsella et al. 2011) downloaded on May 27th, 2015. To avoid spurious results due to low abundance, genes were further filtered if more than half of the ND or control samples had zero counts.

Differential expression and assessment of batch effects

DE genes were identified using Firth’s logistic (FL) regression (Firth 1993; Heinze and Schemper 2002) applied to mRNA-Seq data as described in (Choi et al, under review). Briefly, in contrast to negative binomial regression models like edgeR (Robinson, McCarthy, and Smyth 2010) and DESeq2 (Love, Huber, and Anders 2014), this method models a binomial status variable (e.g. case vs. control) as a function of gene counts and any other potentially confounding variables (RIN value, PMI, etc.). Classical logistic regression has historically not been used to determine DE genes because of the so-called “complete separation” problem, where model parameter estimation fails when there is perfect or nearly perfect separation of a predictor with respect to a binomial variable (e.g. one condition has extremely high read counts and the other has very low read counts). FL regression addresses this issue by using a modified likelihood function to enable reliable parameter estimation, and has other statistical advantages with respect to type I error rates and power. Note the DE statistic from FL regression is log odds ratio (LOR) of case versus controls, that is, positive LOR indicates greater mRNA abundance in case and negative LOR indicates greater abundance in control. All reported p-values are Benjamini-Hochberg (BH) (Benjamini and Hochberg 1995)adjusted unless noted otherwise. See (Choi et al, under review) for further information on this method applied to mRNA-Seq data.

The sequencing datasets in this study were sequenced in five separate batches. To evaluate whether there was evidence of batch effects confounding the identification of DE genes, we ran three statistical models and compared beta estimates. Within each of HD and PD, we ran a full FL model of case versus control against all case counts without a batch variable (FWoB), a full FL model with a categorical variable corresponding to batch (FWB), and separate models of case versus control within each batch and then meta-analyzed the beta estimates using a random-effect meta-analysis method (META). Consistent differences in gene-wise count beta estimates of FWoB and FWB models would indicate evidence of confounding, while the META model is free of confounding by batch by design. There was no evidence of a systemic effect of beta estimate differences between the three models. We therefore concluded that batch was not a significant confounder of DE between case and control and did not include a batch variable in the DE models.

Identification of ND DE genes and enriched gene sets

DE genes were identified as those with BH adjusted p-values < 0.01 from the Firth's logistic regression models of HD vs. control, PD vs. control, and ND vs. control models, yielding three independent DE gene lists. Read counts for each gene were scaled to have a mean of zero and standard deviation of one to obtain standardized regression coefficients, which makes coefficients comparable across genes. All controls were used in each model. Gene set enrichment analysis was performed on each gene list ranked by read counts beta coefficient using the GSEA (Subramanian et al. 2005) implementation in the DOSE software package (Yu et al. 2015) against the MsigDB Canonical Pathway (C2) geneset database (Subramanian et al. 2005). GSEA enrichment was computed using the complete list of genes irrespective of significance ranked by standardized beta coefficient of the count variable. The robust rank aggregation (RRA) algorithm (Kolde et al. 2012) was used to identify individual genes that were consistently altered across these gene lists. Briefly, RRA is a probabilistic, non-parametric, rank-based method for detecting genes ranked consistently better than expected under the null hypothesis of uncorrelated inputs in the presence of noise and outliers. The genes identified as most significant by RRA are the most likely to be implicated in the general ND phenotype.

In addition to producing independent HD and PD DE gene lists, we sought to functionally characterize the genes that are uniquely significant to each disease as well as those in common. To accomplish this, the DE genes from HD and PD were intersected, partitioning the genes into HD-specific, PD-specific, and DE genes common to the two gene lists. Each of these partitioned gene lists were then subjected to gene set enrichment on the MsigDB Canonical Pathway (C2), Transcription Factor Targets (C3), and Gene Ontology (C5) gene set databases (Subramanian et al. 2005) using a hypergeometric test.

Results

Firth's logistic (FL) regression identified 2427, 1949, and 4843 significantly DE genes for HD, PD, and ND, respectively, at q-value < 0.01. Gene set enrichment analysis of MsigDB C2 gene sets identified 226, 199, and 250 gene sets significantly enriched at q-value < 0.05 for HD, PD, and ND, respectively. Due to the large number of DE genes in each dataset, we focus exclusively on the GSEA enrichment results here.

There was a high degree of overlap between the significantly enriched gene sets of HD and PD. 145 gene sets were significantly enriched in both DE gene lists, while 81 and 54 gene sets were uniquely significant in HD and PD, respectively. When a pathway was enriched in more than one list, the pathway was always, without exception, enriched in the same direction, either positively (genes are more abundant in disease) or negatively (genes are less abundant in disease). There were 24 gene sets uniquely significant in ND. Figure 5 depicts the enriched gene sets grouped by high-level biological category for HD, PD, and ND.

We make several observations of Figure 5A. First, the plurality of enriched gene sets across all three data sets are related to immune processes (IM) and are with few exceptions positively enriched. Pathways related to neuronal processes (NE) are largely negatively enriched and there is a subset of these gene sets that are exclusively enriched in HD. With the exception of DNA damage, all remaining biological categories are represented for both HD and PD. DNA damage related pathways (DN) are unique to the PD dataset and are negatively enriched. Multiple apoptosis (AP), developmental (DE), transcription/translation (TR), and transcription factor target (TF) gene sets are also enriched in all three gene lists.

There were 83 gene sets that did not fit cleanly into a single category (OT), which notably include pathways related to endocytosis (KEGG_ENDOCYTOSIS), signaling (BIOCARTA_MAPK_PATHWAY, PID_RAS_PATHWAY, REACTOME_SIGNALING_BY_EGFR_IN_CANCER, REACTOME_PHOSPHOLIPASE_C_MEDIATED_CASCADE), cellular adhesion and extracellular matrix (KEGG_FOCAL_ADHESION, KEGG_CELLULAR_ADHESION_MOLECULES_CAM, REACTOME_COLLAGEN_FORMATION, REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION), glycans (KEGG_O_GLYCAN_BIOSYNTHESIS, PID_GLYPICAN_1PATHWAY, KEGG_GLYCOSAMINOGLYCAN_BIOSYNTEHSIS_HEPARAN_SULFATE), and metabolism (KEGG_GALACTOSE_METABOLISM, REACTOME_INTEGRATION OF_ENERGY_METABOLISM, REACTOME_INSULIN_SYNTHESIS_AND_PROCESSING).

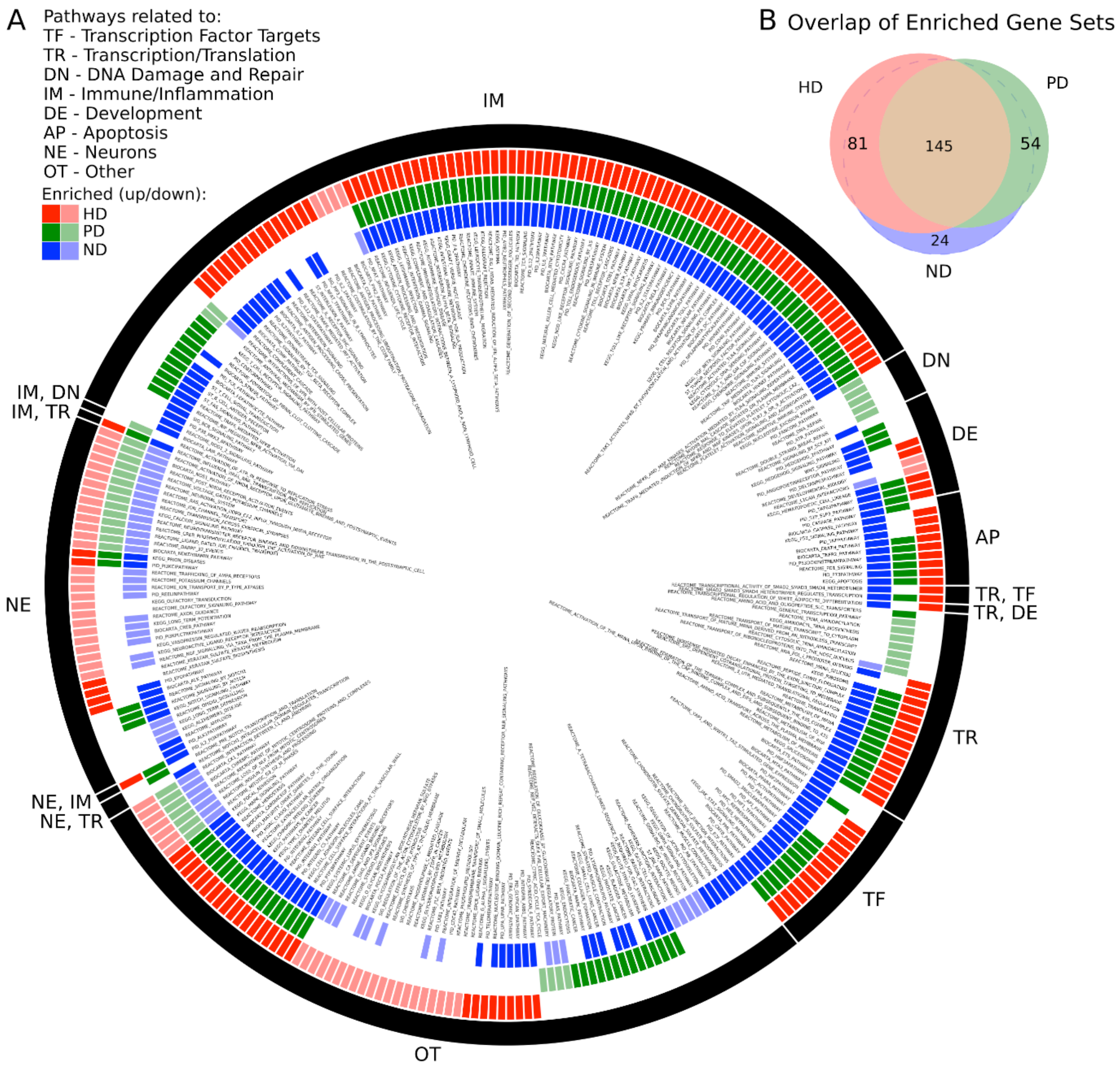


Figure 5. A) Significantly enriched MsigDB C2 Canonical Pathway gene sets for HD, PD, and ND identified by GSEA. Each colored ring segment corresponds to a single enriched gene set. Red (outer), green (middle), and blue (inner) segmented rings indicate whether the HD, PD, or ND DE gene lists, respectively, were significantly enriched for the gene set. Dark and light colored segments indicate up and down regulation (positive, negative GSEA normalized enrichment score), respectively. Black ring around exterior groups gene sets into high-level categories as indicated by the two letter code. Gene set name is listed in interior of rings. B) Venn diagram illustrating overlap of significantly enriched gene sets for HD, PD, and ND. All but 24 of the ND significant gene sets were significantly enriched in either HD, PD, or both.

RRA identified 1353 genes with a score < 0.01. The top ten genes identified by RRA as most highly ranked across all three gene lists are reported in Table 9. The rank of each gene in the individual gene lists are also reported in the table, showing that most genes are relatively highly ranked across all three studies as expected.

Symbol	Description	RRA Rank	RRA Score	HD Rank	PD Rank	ND Rank
ENSG00000272403	no description	1	4.93e-9	25	15	1
SPR	sepiapterin reductase	2	1.13e-7	70	38	5
DDIT4	DNA-damage-inducible transcript 4	3	1.27e-7	74	60	22
TRIP10	thyroid hormone receptor interactor 10	4	1.87e-7	84	59	7
TNFRSF10D	tumor necrosis factor receptor superfamily, member 10d	5	2.30e-7	90	55	20
PRMT6	Protein arginine methyltransferase 6	6	2.54e-7	29	93	10
GPSM3	G-protein signaling modulator 3	7	2.62e-7	81	98	11
GPCPD1	Glycerophosphocholine phosphodiesterase 1	8	2.79e-7	13	96	2
GPR4	G protein-coupled receptor 4	9	2.97e-7	98	75	24
NFKBIA	Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, Alpha	10	3.35e-7	11	103	3

Table 9. Top ranked RRA genes. RRA Score can be thought of as a p-value. The remaining columns contain the rank of the corresponding gene in each individual gene list.

The most consistently ranked gene is RP1-93H18.7 (ENSG00000272403.1), a lncRNA, which was removed from Ensembl starting at version GRCh38.p2, but shows consistent transcription in these data. This gene is directly downstream of the gene DSE (dermatan sulfate epimerase), which is also DE in both HD and PD, is involved in embryonic development (Stachtea et al. 2015; Habicher et al. 2015), and has been related to the immune response in cancer patients (Mizukoshi et al. 2012). Deficiencies in the second ranked gene, SPR (sepiapterin reductase), have been linked to DOPA-responsive dystonia (Wijemanne and Jankovic 2015). The third gene, DDIT4 (DNA-Damage-Inducible Transcript 4), is a multifunctional gene which, via its inhibition of the mammalian target of rapamycin complex 1 (mTORC1), regulates in cell growth, proliferation, and survival (Dennis et al. 2013), controls p53/TP53-mediated apoptosis in response to DNA damage(Cam et al. 2014; Vadysirisack et al. 2011), and plays a role in neurodegeneration, neuronal death and differentiation, and neuron migration during embryonic brain development (Romaní-Aumedes et al. 2014; Canal et al. 2014; Ota et al. 2014; Malagelada et al. 2011). TRIP10 (thyroid hormone receptor interactor 10), another multi-functional gene, is involved in insulin signaling (Chang, Chiang, and Saltiel 2013), endocytosis (Feng et al. 2010), and structures specific to monocyte-derived cells (Linder et al. 2000). TNFRSF10D (Tumor Necrosis Factor Receptor Superfamily, Member 10d, Decoy With Truncated Death Domain) inhibits certain types of apoptosis and may play a role in NfkB pathway (Degli-Esposti et al. 1997).

Figure 6 illustrates the differences in normalized counts for the top genes identified by RRA. With the exception of (12) PITX1, which is driven entirely by HD, all top genes demonstrate substantial differences between both disease conditions and control.

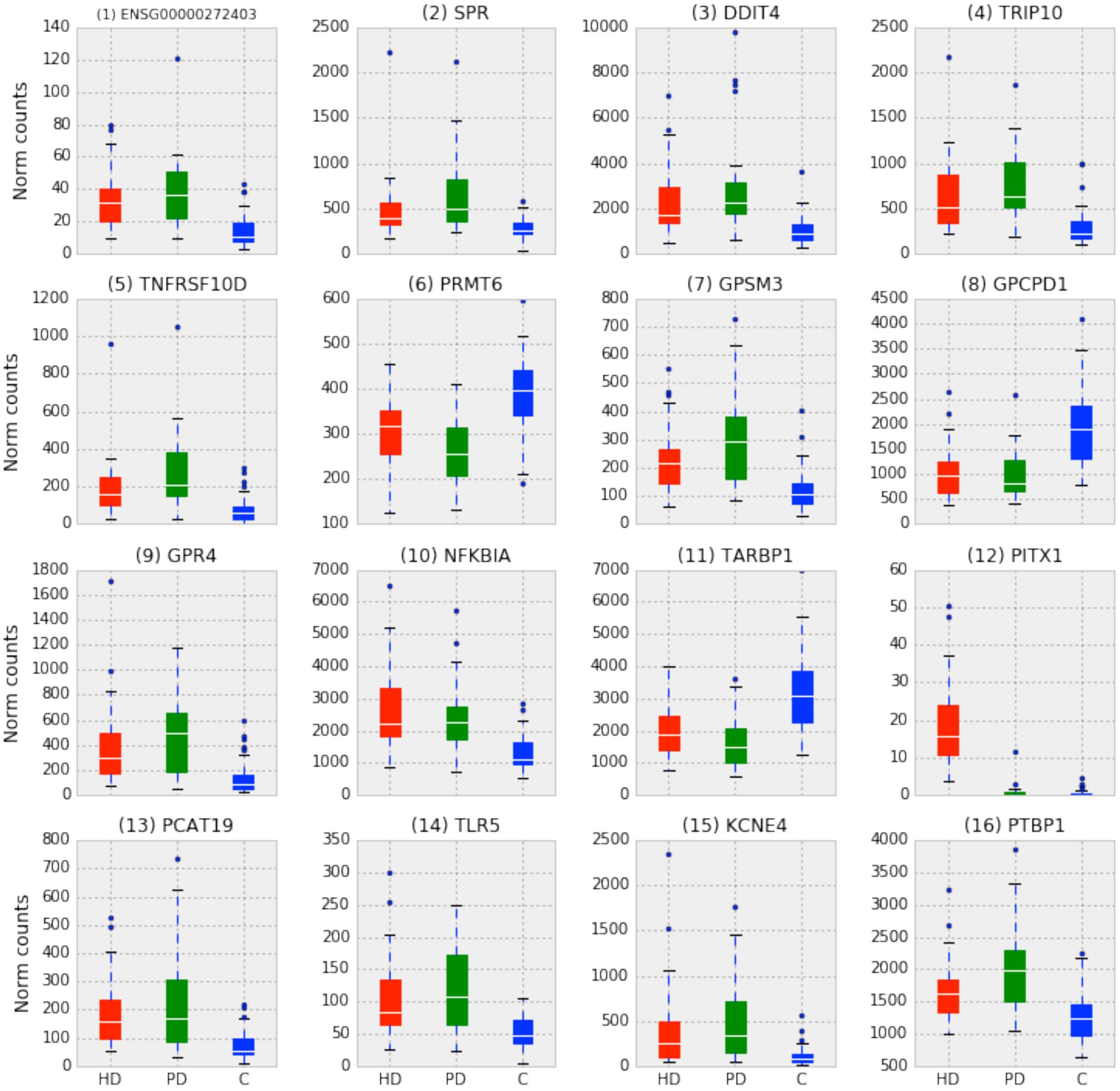
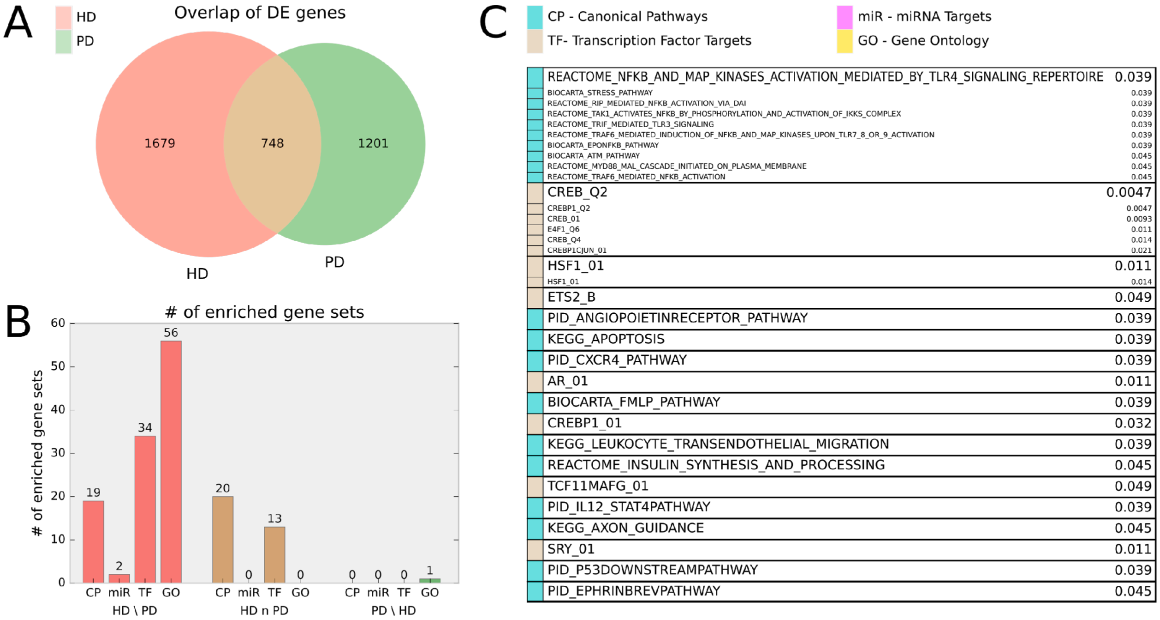


Figure 6. Box plots of normalized counts for top RRA genes split by condition, RRA rank is in parenthesis. Whiskers extend to 25th and 75th percentile counts, white bars are median counts. With the exception of (12) PITX1, which is driven entirely by HD, all top genes demonstrate substantial differences between both disease conditions and control.

Finally, we examined the significant DE genes from HD and PD for intersection. Figure 7 illustrates the overlap of DE genes between diseases and describes gene set enrichment results for the intersection. Figure 8 contains the enrichment results for the HD unique genes.

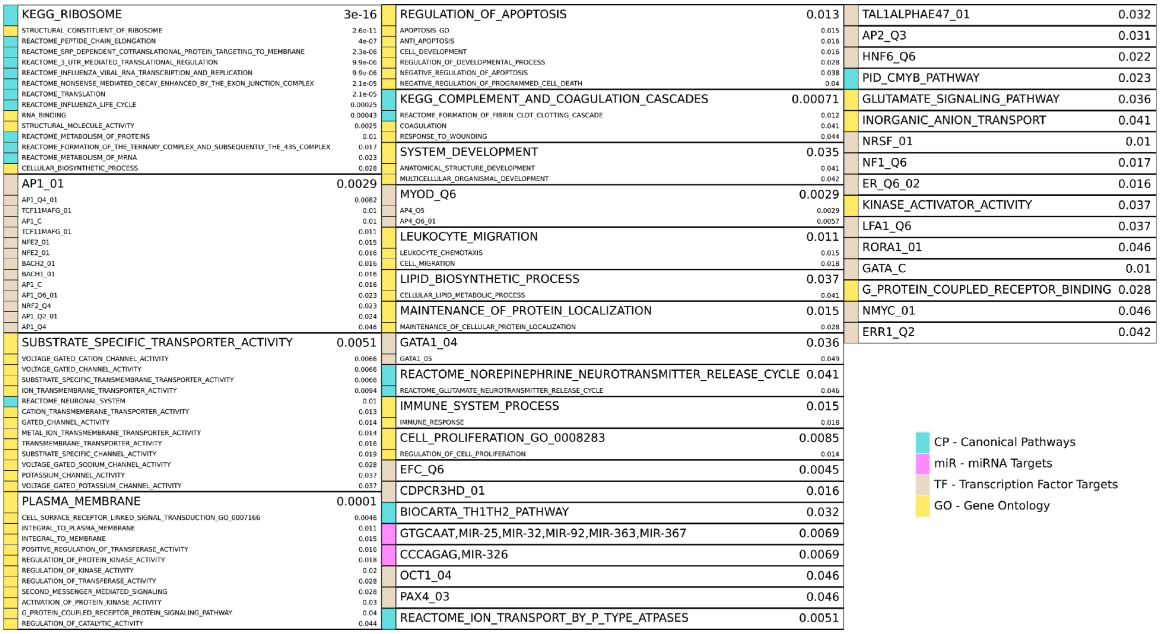


🟦	REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR4_SIGNALING_REPERTOIRE	0.039
🟦	BIOCARTA_STRESS_PATHWAY	0.039
🟦	REACTOME_RIP_MEDIATED_NFKB_ACTIVATION_VIA_DAI	0.039
🟦	REACTOME_TAK1_ACTIVATES_NFKB_BY_PHOSPHORYLATION_AND_ACTIVATION_OF_IKKS_COMPLEX	0.039
🟦	REACTOME_TRIF_MEDIATED_TLR3_SIGNALING	0.039
🟦	REACTOME_TRAF6_MEDIATED_INDUCTION_OF_NFKB_AND_MAP_KINASES_UPON_TLR7_8_OR_9_ACTIVATION	0.039
🟦	BIOCARTA_EPONFKB_PATHWAY	0.039
🟦	BIOCARTA_ATM_PATHWAY	0.045
🟦	REACTOME_MYD88_MAL_CASCADE_INITIATED_ON_PLASMA_MEMBRANE	0.045
🟦	REACTOME_TRAF6_MEDIATED_NFKB_ACTIVATION	0.045
🟫	CREB_Q2	0.0047
🟫	CREBP1_Q2	0.0047
🟫	CREB_01	0.0093
🟫	E4F1_Q6	0.011
🟫	CREB_Q4	0.014
🟫	CREBP1CJUN_01	0.021
🟫	HSF1_01	0.011
🟫	HSF1_01	0.014
🟫	ETS2_B	0.049
🟦	PID_ANGIOPOIETINRECEPTOR_PATHWAY	0.039
🟦	KEGG_APOPTOSIS	0.039
🟦	PID_CXCR4_PATHWAY	0.039
🟫	AR_01	0.011
🟦	BIOCARTA_FMLP_PATHWAY	0.039
🟫	CREBP1_01	0.032
🟦	KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	0.039
🟦	REACTOME_INSULIN_SYNTHESIS_AND_PROCESSING	0.045
🟫	TCF11MAFG_01	0.049
🟦	PID_IL12_STAT4PATHWAY	0.039
🟦	KEGG_AXON_GUIDANCE	0.045
🟫	SRY_01	0.011
🟦	PID_P53DOWNSTREAMPATHWAY	0.039
🟦	PID_EPHRINBREVPATHWAY	0.045

Figure 7. A) Venn diagram of HD and PD DE gene list intersection for DE genes adjusted p<0.01. B) Bar chart indicating number of MsigDB C2 Canonical Pathway (CP), C3 miRNA Targets (miR), C3 Transcription Factor Targets (TF), and C5 Gene Ontology (GO) gene sets enriched for the HD unique (HD \ PD), intersection (HD n PD), and PD unique (PD \ HD) genes. For HD \ PD enrichment, 17 redundant or uninformative GO gene sets and 7 TF gene sets for motifs with unknown transcription factors were omitted from the figure results. C) Gene sets enriched for the intersection genes (HD n PD). Adjusted p-values are listed beside gene set name and the originating gene set (CP, miR, TF, or GO) are indicated by color. Gene sets that are groups into boxes share more than 20% of their DE genes and are therefore listed together.

As shown in Figure 7A, there were 748 DE genes in common between HD and PD, while 1679 and 1201 DE genes were unique to HD and PD, respectively. When the genes from each partition were analyzed for enrichment against MsigDB C2 Canonical Pathway (CP), C3 miRNA Targets (miR), C3 Transcription Factor Targets (TF), and C5 Gene Ontology (GO) gene sets using a hypergeometric test, the HD unique genes showed much more enrichment for all four gene set categories than the other two gene partitions (Figure 7B), with 111 gene sets significantly enriched in total. By comparison, the intersection genes were enriched for 33 gene sets, while PD was enriched for only one (GO:0031570 DNA_INTEGRITY_CHECKPOINT, p=0.049) despite having a comparable number of DE genes to the HD unique set (1201 vs. 1679).

Figure 7C lists the gene sets enriched in the intersecting DE genes, where gene sets that share > 20% of their DE genes are grouped together. Multiple gene sets related to nuclear factor kappa-light-chain-enhancer of activated B cells, NFkB, and transcription factor cAMP response element-binding protein, CREB, targets are enriched in the intersection genes. Other transcription factor targets including heat shock transcription factor (HSF1), Protein C-ets-2 (ETS2), androgen receptor 1 (AR1), the Nuclear Factor, Erythroid 2-Like 1/V-Maf Avian Musculoaponeurotic Fibrosarcoma Oncogene Homolog G complex (TCF11/MAFG), and Sex Determining Region Y (SRY) are also enriched. Apoptosis related gene sets (KEGG_APOPTOSIS, PID_P53DOWNSTREAMPATHWAY), inflammatory gene sets (PID_CXCR4_PATHWAY, KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION, PID_IL12_STAT4PATHWAY), pathways related to angiogenesis/axon guidance (PID_ANGIOPOIETINRECEPTOR_PATHWAY, PID_EPHRINBREVPATHWAY, KEGG_AXON_GUIDANCE), and insulin synthesis were observed.



🟦	**KEGG_RIBOSOME**	**3e-16**
🟨	STRUCTURAL_CONSTITUENT_OF_RIBOSOME	2.6e-11
🟦	REACTOME_PEPTIDE_CHAIN_ELONGATION	4e-07
🟦	REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	2.3e-06
🟦	REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION	9.9e-06
🟦	REACTOME_INFLUENZA_VIRAL_RNA_TRANSCRIPTION_AND_REPLICATION	9.9e-06
🟦	REACTOME_NONSENSE_MEDIATED_DECAY_ENHANCED_BY_THE_EXON_JUNCTION_COMPLEX	2.1e-05
🟦	REACTOME_TRANSLATION	2.1e-05
🟦	REACTOME_INFLUENZA_LIFE_CYCLE	0.00025
🟨	RNA_BINDING	0.00043
🟨	STRUCTURAL_MOLECULE_ACTIVITY	0.0025
🟦	REACTOME_METABOLISM_OF_PROTEINS	0.01
🟦	REACTOME_FORMATION_OF_THE_TERNARY_COMPLEX_AND_SUBSEQUENTLY_THE_43S_COMPLEX	0.017
🟦	REACTOME_METABOLISM_OF_MRNA	0.023
🟨	CELLULAR_BIOSYNTHETIC_PROCESS	0.028
🟫	**AP1_01**	**0.0029**
🟫	AP1_Q4_01	0.0082
🟫	TCF11MAFG_01	0.01
🟫	AP1_C	0.01
🟫	TCF11MAFG_01	0.011
🟫	NFE2_01	0.015
🟫	NFE2_01	0.016
🟫	BACH2_01	0.016
🟫	BACH1_01	0.016
🟫	AP1_C	0.016
🟫	AP1_Q6_01	0.023
🟫	NRF2_Q4	0.023
🟫	AP1_Q2_01	0.024
🟫	AP1_Q4	0.046
🟨	**SUBSTRATE_SPECIFIC_TRANSPORTER_ACTIVITY**	**0.0051**
🟨	VOLTAGE_GATED_CATION_CHANNEL_ACTIVITY	0.0066
🟨	VOLTAGE_GATED_CHANNEL_ACTIVITY	0.0066
🟨	SUBSTRATE_SPECIFIC_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.0066
🟨	ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.0094
🟦	REACTOME_NEURONAL_SYSTEM	0.01
🟨	CATION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.013
🟨	GATED_CHANNEL_ACTIVITY	0.014
🟨	METAL_ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.014
🟨	TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.016
🟨	SUBSTRATE_SPECIFIC_CHANNEL_ACTIVITY	0.019
🟨	VOLTAGE_GATED_SODIUM_CHANNEL_ACTIVITY	0.028
🟨	POTASSIUM_CHANNEL_ACTIVITY	0.037
🟨	VOLTAGE_GATED_POTASSIUM_CHANNEL_ACTIVITY	0.037
🟨	**PLASMA_MEMBRANE**	**0.0001**
🟨	CELL_SURFACE_RECEPTOR_LINKED_SIGNAL_TRANSDUCTION_GO_0007166	0.0046
🟨	INTEGRAL_TO_PLASMA_MEMBRANE	0.011
🟨	INTEGRAL_TO_MEMBRANE	0.015
🟨	POSITIVE_REGULATION_OF_TRANSFERASE_ACTIVITY	0.016
🟨	REGULATION_OF_PROTEIN_KINASE_ACTIVITY	0.018
🟨	REGULATION_OF_KINASE_ACTIVITY	0.02
🟨	REGULATION_OF_TRANSFERASE_ACTIVITY	0.028
🟨	SECOND_MESSENGER_MEDIATED_SIGNALING	0.028
🟨	ACTIVATION_OF_PROTEIN_KINASE_ACTIVITY	0.03
🟨	G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	0.04
🟨	REGULATION_OF_CATALYTIC_ACTIVITY	0.044
🟨	**REGULATION_OF_APOPTOSIS**	**0.013**
🟨	APOPTOSIS_GO	0.015
🟨	ANTI_APOPTOSIS	0.016
🟨	CELL_DEVELOPMENT	0.016
🟨	REGULATION_OF_DEVELOPMENTAL_PROCESS	0.028
🟨	NEGATIVE_REGULATION_OF_APOPTOSIS	0.038
🟨	NEGATIVE_REGULATION_OF_PROGRAMMED_CELL_DEATH	0.04
🟦	**KEGG_COMPLEMENT_AND_COAGULATION_CASCADES**	**0.00071**
🟦	REACTOME_FORMATION_OF_FIBRIN_CLOT_CLOTTING_CASCADE	0.012
🟨	COAGULATION	0.041
🟨	RESPONSE_TO_WOUNDING	0.044
🟨	**SYSTEM_DEVELOPMENT**	**0.035**
🟨	ANATOMICAL_STRUCTURE_DEVELOPMENT	0.041
🟨	MULTICELLULAR_ORGANISMAL_DEVELOPMENT	0.042
🟫	**MYOD_Q6**	**0.0029**
🟫	AP4_Q5	0.0029
🟫	AP4_Q6_01	0.0057
🟨	**LEUKOCYTE_MIGRATION**	**0.011**
🟨	LEUKOCYTE_CHEMOTAXIS	0.015
🟨	CELL_MIGRATION	0.018
🟨	**LIPID_BIOSYNTHETIC_PROCESS**	**0.037**
🟨	CELLULAR_LIPID_METABOLIC_PROCESS	0.041
🟨	**MAINTENANCE_OF_PROTEIN_LOCALIZATION**	**0.015**
🟨	MAINTENANCE_OF_CELLULAR_PROTEIN_LOCALIZATION	0.028
🟫	**GATA1_04**	**0.036**
🟫	GATA1_05	0.049
🟦	**REACTOME_NOREPINEPHRINE_NEUROTRANSMITTER_RELEASE_CYCLE**	**0.041**
🟦	REACTOME_GLUTAMATE_NEUROTRANSMITTER_RELEASE_CYCLE	0.046
🟨	**IMMUNE_SYSTEM_PROCESS**	**0.015**
🟨	IMMUNE_RESPONSE	0.018
🟨	**CELL_PROLIFERATION_GO_0008283**	**0.0085**
🟨	REGULATION_OF_CELL_PROLIFERATION	0.014
🟫	**EFC_Q6**	**0.0045**
🟫	**CDPCR3HD_01**	**0.016**
🟦	**BIOCARTA_TH1TH2_PATHWAY**	**0.032**
🟪	**GTGCAAT,MIR-25,MIR-32,MIR-92,MIR-363,MIR-367**	**0.0069**
🟪	**CCCAGAG,MIR-326**	**0.0069**
🟫	**OCT1_04**	**0.046**
🟫	**PAX4_03**	**0.046**
🟦	**REACTOME_ION_TRANSPORT_BY_P_TYPE_ATPASES**	**0.0051**
🟫	**TAL1ALPHAE47_01**	**0.032**
🟫	**AP2_Q3**	**0.031**
🟫	**HNF6_Q6**	**0.022**
🟦	**PID_CMYB_PATHWAY**	**0.023**
🟨	**GLUTAMATE_SIGNALING_PATHWAY**	**0.036**
🟨	**INORGANIC_ANION_TRANSPORT**	**0.041**
🟫	**NRSF_01**	**0.01**
🟫	**NF1_Q6**	**0.017**
🟫	**ER_Q6_02**	**0.016**
🟨	**KINASE_ACTIVATOR_ACTIVITY**	**0.037**
🟫	**LFA1_Q6**	**0.037**
🟫	**RORA1_01**	**0.046**
🟫	**GATA_C**	**0.01**
🟨	**G_PROTEIN_COUPLED_RECEPTOR_BINDING**	**0.028**
🟫	**NMYC_01**	**0.046**
🟫	**ERR1_Q2**	**0.042**

🟦 CP - Canonical Pathways
🟪 miR - miRNA Targets
🟫 TF - Transcription Factor Targets
🟨 GO - Gene Ontology

Figure 8. Enriched gene sets for the HD unique (HD \ PD) genes from Figure 7 A and reported similarly as in Figure 7 C. Note 17 redundant or uninformative GO gene sets and 7 TF gene sets for motifs with unknown transcription factors were omitted from the figure results.

HD specific enrichments are shown in Figure 8, where a broad spectrum of biological processes is implicated in the HD-unique DE genes. The most striking enriched gene set is KEGG_RIBOSOME, with many other related gene sets involved in translation and molecular metabolism similarly enriched. Multiple gene sets that share > 20% of their DE genes are associated with Jun Proto-Oncogene (AP1), BTB And CNC Homology 1, Basic Leucine Zipper Transcription Factor 1 and 2 (BACH1, BACH2), and NRF2/TCF11 are also implicated. Other strongly implicated biological processes are ion channel activity, plasma membrane and signaling, apoptosis, immune system and inflammatory processes, developmental genes, neuron-related signaling pathways, many transcription factors, and two families of miRNAs. Only one gene set (TFC11MAFG_01) was enriched in more than one of the three gene partitions. The remainder of the enriched gene sets, and indeed most of the biological processes, are distinct between the three gene partitions.

Discussion

To the authors' knowledge, this study presents the first comprehensive comparative analysis of DE gene expression from HD, PD, and ND in post-mortem human brains assessed with mRNA-Seq. The comparison of HD and PD in particular is motivated by the fact that these diseases can be viewed as mirror-images of each other. GABAergic medium spiny interneurons, which compose most of the neurons in the striatum and selectively die in HD but are spared in PD, project directly into the substantia nigra and coordinate motor activity throughout the brain via dopamine-induced signaling (Yager et al. 2015). Dopaminergic neurons in the substantia nigra, on the other hand, which also are important in coordinating motor activity as well as arousal, reinforcement, and reward (Schultz 2007), selectively degenerate in PD but are spared in HD. It was observed in a study of 523 HD subjects that the incidence of PD in this cohort was lower than that of the general population, though both HD and PD individuals develop Alzheimer's disease at the same rate (Hadzi et al. 2012), suggesting the selective death of medium spiny neurons might be neuroprotective of dopaminergic neuron death. Given the intimate neurological link between the affected neurons in HD and PD, and the mutual exclusivity of their degeneration, this comparison poses a very interesting contrast to identify common responses to neurodegeneration that are not confounded by neuron type. Unfortunately, a direct comparison of neurons in these regions of postmortem human brains is not possible, precisely due to this mirror-image pathology. The choice of the BA9 brain region is motivated by the fact that, due to degeneration, the affected neurons are largely missing from the striatum and substantia nigra in HD and PD, respectively, whereas BA9 is generally free of involvement in both diseases (Hadzi et al. 2012; Braak et al. 2003; Halliday, Del Tredici, and Braak 2006). Because the primarily affected neurons in HD and PD do not exist in BA9, the biological processes implicated by this analysis are likely in part the response to, rather than direct cause of, the respective diseases. Nonetheless, the remarkable consistency between HD and PD observed in this analysis points to important mechanisms that further our understanding of neurodegenerative disease as a general process.

The biological processes implicated by DE gene lists identified from each condition separately are compellingly similar. From Figure 5, we see that the majority of enriched biological pathways are common and that they are invariably perturbed in the same direction in both diseases. Furthermore, combining HD and PD data into an ND condition does not yield significantly more novel biological insights. This remarkable consistency between the pathway enrichment results suggest that the underlying molecular responses to neurodegeneration in HD and PD may be more similar than they are different, despite their different pathological underpinnings. Of particular significance is the strong positive enrichment of immune and inflammatory pathways, which have been convincingly implicated in both diseases separately (A. Labadorf et al. 2015; Kwan et al. 2012; Crotti et al. 2014; Ellrichmann et al. 2013; Dexter and Jenner 2013; Dobbs et al. 1999; Jenner 2003; Allen Reish and Standaert 2015), but the compelling similarity of these signatures between HD and PD revealed by this analysis has not been illustrated to date.

The negative regulation of neuron-related pathways is also noteworthy, since the BA9 brain region, from which these samples are derived, is not known to be heavily involved in either of these diseases. Despite the lack of clear and consistent neurodegeneration in this brain region, the widespread biological pathways shown to be affected in this analysis strongly suggest neurons in BA9 do indeed experience a common set of effects in the neuropathology for HD and PD.

Many of the individual genes identified by RRA as most consistently different in HD, PD, and ND have previously been the focus of studies in neurodegeneration. The second highest ranked gene SPR has been the focus of significant study in PD and is related to the PARK3 gene locus (Sharma et al. 2011; Tobin et al. 2007), but has not been previously implicated in HD. Inhibition of DNA-damage inducible transcript 4 (DDIT4/RTP801/REDD1) has been associated with neuroprotection in PD models and patients (Malagelada et al. 2010) and is involved with mutant Huntingtin-induced cell death (Martín-Flores et al. 2015). Thyroid hormone receptor interactor 10 (TRIP10) has been shown to interact directly with mutant huntingtin (Holbert et al. 2003), and while it not known to play a role in PD pathology, its elevated mRNA abundance in these PD samples suggest it may indeed be implicated. Other top genes have also been implicated in neurodegeneration: tumor necrosis factor receptor superfamily 10D (TNFRSF10D) (López-Gómez et al. 2011; Frenkel 2015), protein arginine methyltransferase 6 (PRMT6) (Scaramuzzino et al. 2015), and toll-like receptor 5 (TLR5) (Arroyo et al. 2011). Further investigation of this list of genes is likely to yield novel insights into the mechanisms of neurodegeneration.

The intersection of DE genes between HD and PD also affords important insight into genes relevant to fundamental neurodegenerative processes. Most notably, pathways related to NFkB and transcriptional targets of CREB are prominent in the enrichment results. The NFkB pathway is prominent in both HD (Marcora and Kennedy 2010; Träger et al. 2014) and PD pathology (Flood et al. 2011; Ghosh et al. 2007) through its central role in inflammatory signaling. CREB is directly targeted by brain derived-neurotrophic factor (BDNF) (Pizzorusso et al. 2000), an important trophic factor in the brain. Both BDNF (Zuccato et al. 2008), and CREB (Choi et al. 2009; Obrietan and Hoyt 2004) have been directly implicated in HD pathology, while CREB is also believed to play a critical role in dopamine receptor-mediated nuclear signaling (Andersson, Konradi, and Cenci 2001), and disruption of the protein's function leads to neurodegeneration (Devi and Ohno 2014; Mantamadiotis et al. 2002). The specific inflammation-related gene sets (HSF1 transcription factor targets, CXCR4, IL12) suggests there is some specificity in the aspects of the pan-neurodegenerative neuroimmune response. Recent studies in both HD and PD have focused on the role of insulin sensitivity and metabolism in patients (Block et al. 2010; Aviles-Olmos et al. 2013; Russo et al. 2013), supporting the role of insulin synthesis as an enriched biological pathway in the common gene list. While the enrichment of apoptosis-related pathways was not surprising, pathways related to angiopoietin, ephrin, and axon guidance suggest that biological processes related to neuronal plasticity are active in both of these diseases and may even indicate that neuroprotective or neuroregenerative processes are a component of the neurodegenerative response.

These data also point to compelling differences between HD and PD. Interestingly, two groups of genes, DNA damage and repair and tRNA related processes, seem to be uniquely perturbed and negatively enriched in PD. The DNA damage and repair gene set enrichment may be a reflection of mitochondrial DNA damage. In PD, dopaminergic neurons of the substantia nigra (though not cortical neurons) were found to be particularly vulnerable to mitochondrial DNA damage (Sanders et al. 2014), and Lewy body pathology, the histological hallmark of PD, is associated with mitochondrial DNA damage (Müller et al. 2013). More generally, mitochondrial DNA damage and oxidative stress are associated with several neurodegenerative diseases including PD, Alzheimer's disease (Moreira et al. 2010), and ALS (Coppedè 2011). There is evidence supporting the involvement of aminoacyl tRNA synthetases in neurological disease, including ALS, leukoencephalopathy, and PD (Park, Schimmel, and Kim 2008).

In HD, there is a number of uniquely perturbed gene sets related to glycan biosynthesis and metabolism are negatively regulated, and these pathways have not been previously implicated in HD. The 1687 HD-unique DE genes are enriched for many gene sets across a broad spectrum of biological processes, including mRNA and protein metabolism, ion channel activity, signaling and kinase activity, apoptosis, immune response, and development. Other, more specific gene sets related to a large number of transcription factors further support the observation of transcriptional dysregulation in HD (Cha 2000). The specificity of these enriched TF gene sets is quite striking, as the targeted DE genes appear to be largely disjoint between them, suggesting potential, specific causes of the dysregulated transcriptional effects in HD. The enrichment of two miRNA families are also particularly relevant in light of recent reports of miRNA dysregulation in HD (Hoss et al. 2015; Hoss et al. 2016).

It is interesting to note the disparity in enrichment between the HD and PD unique DE genes. Though the numbers of unique DE genes are comparable, the large number of enriched gene sets in HD stands in sharp contrast to the almost total absence of enrichment in PD. This result implies that the DE genes in HD are more consistently related to one another than in PD. One possible, and potentially important explanation for this is that HD is a much more homogeneous disease than PD. It is well established that PD has a significant sporadic component (Lesage and Brice 2012), caused by a combination of genetic and environmental factors. The relative heterogeneity of PD may make finding consistently effective treatments difficult, and the absence of biological enrichment in specific pathways, other than those common to both diseases, from this analysis may be a reflection of an underlying molecular basis for this effect. It may be that, given sufficient sample size, coherent subgroups of patients may be identified by examining patterns in their gene expression using datasets such as those analyzed here. On the other hand, despite extensive molecular characterization of HD, effective, widely available therapies for HD have remained elusive despite the relative homogeneity of the disease process among HD patients.

These findings have important implications on our understanding of the neurodegenerative disease process. The significant involvement of the inflammatory pathways in both diseases in an area not thought to be directly involved in disease pathogenesis suggests the response to neurodegeneration is widespread throughout the brain. NFkB in particular appears to be a major player, which is well supported in the HD and PD literature. It is unclear whether the neuroinflammatory response is protective, deleterious, or both from these data, but investigation into the role these processes play, and the potential therapeutic value of modulating them, should be made a high priority.

Chapter 4: mRNA/miRNA modules associated with clinical features in HD

Introduction

The abundance of mRNA transcripts in a cell is regulated by many mechanisms. microRNAs (miRNAs) are short RNA molecules, typically 18–22 nucleotides in length, that inhibit mRNA molecules from being translated into protein by targeting specific nucleotide sequences contained within transcribed mRNAs. Specifically, the first 7 nucleotides of each miRNA, termed the seed sequence, base pair with complementary nucleotide sequences found in mRNA transcripts. This short RNA duplex typically occurs in the 3' untranslated region (3'UTR) of mRNA molecules and regulates mRNA expression by inhibiting the ribosomal complex during translation or marking the mRNA for degradation (Ling, Fabbri, and Calin 2013). The mRNA/miRNA relationship is many-to-many, where a single miRNA can target multiple mRNAs and a single mRNA may be targeted by many different miRNAs, resulting in a complex regulatory network that has been implicated in important biological processes, including development and disease (W. Zhang et al. 2012; Hiddingh et al. 2014; Cordes and Srivastava 2009; Shenoy and Blelloch 2014).

Identifying the mRNA targets of miRNAs is a critical step in understanding the regulatory relationships between these molecules. Several complementary approaches have been proposed for predicting mRNA/miRNA relationships, including sequence-based predictions obtained by scanning 3'UTR sequences for conserved miRNA seed sequences (Lewis, Burge, and Bartel 2005b), modeling the thermodynamic properties of the mRNA/miRNA duplex (Enright et al. 2003), modeling the 2 dimensional hairpin structure of precursor miRNAs (Rehmsmeier et al. 2004), and by identifying statistical relationships (e.g. Pearson correlation) between the abundance of mRNAs and miRNAs across multiple samples (Gennarino et al. 2009). Bioinformatic mRNA/miRNA target predictions are then used to design wet lab experiments to validate putative relationships, which further improves the confidence of computational target predictions and algorithms. Approaches that combine prediction information across methods have also been proposed that produce mRNA/miRNA target predictions that are more consistently validated in experimental settings than any one method alone (Kozomara and Griffiths-Jones 2011; Le et al. 2015). As a result, mRNA/miRNA prediction databases for multiple organisms are available (Kozomara and Griffiths-Jones 2011; Kozomara and Griffiths-Jones 2014) that aid in the analysis and interpretation of transcriptional mRNA and miRNA data.

Analysis of mRNA and miRNA abundance measurements using high-throughput transcriptional data revealed that groups of interacting mRNAs and miRNAs, termed mRNA/miRNA modules, often work in concert to regulate specific biological processes (W. Zhang et al. 2012; Hiddingh et al. 2014; Coronnello et al. 2012; Z. Liu et al. 2015; Setty et al. 2012). mRNA/miRNA modules can be detected using transcriptional data from multiple samples by examining the statistical relationship between the abundance of mRNAs and miRNAs. Specifically, since the regulatory effect of miRNAs on mRNAs is typically inhibitory, we expect the abundance of miRNAs and their target mRNAs to be inversely correlated, since a greater abundance of a miRNA should result in a greater degradation of its targets. Therefore, the most consistent negatively correlated mRNA/miRNA pairs are the most likely candidates for direct miRNA regulatory relationships. However, indirect regulatory relationships may result in positive correlation between miRNAs and their targets. For example, if a miRNA targets an mRNA that encodes for a transcriptional inhibitor protein, the relationship between the miRNA and the targets of the transcriptional inhibitor will be positive, since decreasing the abundance of an inhibitor increases the abundance of its targets. In this instance, the miRNA in question does indeed have a regulatory relationship with the increased mRNAs, though not in the expected (i.e. negative) direction, and should therefore be considered a member of a module with those mRNAs. Figure 9 contains an illustration of the mRNA/miRNA module concept.

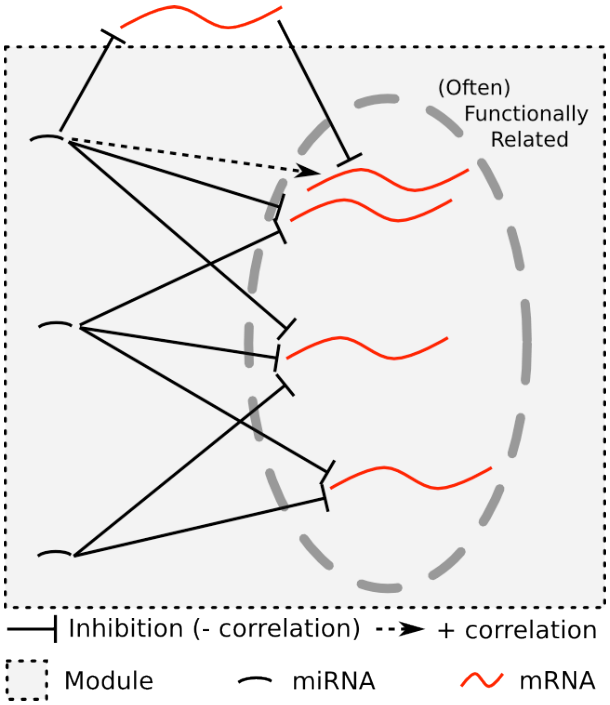


Figure 9. Cartoon of relationships involved with mRNA/miRNA modules.

Many mRNA/miRNA module detection approaches have been proposed. The earliest method proposed to detect miRNA-mRNA modules was by Yoon & Micheli (Yoon and Micheli 2005), where sequence complementarity, free energy estimation, and evolutionary conservation were used in combination to define putative miRNA-mRNA interactions. (Joung et al. 2007) used coevolutionary learning and estimation-of-distribution algorithms to combine miRNA expression data and binding information with the goal of finding correlated sets of miRNAs and mRNAs. (Tran, Satou, and Ho 2008) used rule induction to identify miRNA-mRNA modules. All pairwise mRNA expression value correlations were computed to split the dataset into “similar” and “dissimilar” classes with respect to each mRNA. (Peng et al. 2009) calculated all pairwise miRNA-mRNA correlations from expression data, filtered those correlations by an FDR threshold identified by randomization and combined the thresholded correlation matrix with predicted target information to create an adjacency matrix of miRNA-mRNA pairs. (Joung and Fei 2009) proposed a hierarchical probabilistic graphical model that uses predicted miRNA targeting information and mRNA expression values to infer regulatory modules. (B. Liu, Li, and Tsykin 2009) defined the miRNA-mRNA regulatory network as a bipartite graph composed of bicliques (completely connected subgraphs of at least m miRNAs and n mRNAs) using predicted target information and miRNA and mRNA expression profiles. (B. Liu et al. 2010) created an algorithm inspired by Correspondence Latent Dirichlet Allocation, which is useful for modeling the joint probability distribution of a continuous and a discrete random variable. (S. Zhang et al. 2011) used miRNA target predictions, miRNA and mRNA expression data, and protein protein interaction data to infer functional miRNA-mRNA modules. (Lu et al. 2011) used Lasso regression to estimate the effect of TargetScan and PicTar predicted targets on mRNA by including targeting miRNAs as variables in the Lasso model to predict mRNA expression. (J. Zhang et al. 2012) proposed a semi-supervised model that uses differentially expressed miRNAs and mRNAs to infer functional miRNA-mRNA regulatory modules between two conditions with a probabilistic topic model, similar to that of (B. Liu et al. 2010). (Bryan et al. 2014) combined miRNA and mRNA expression values into expression correlation matrix, where negatively correlated relationships are considered direct and positive relationships are considered putative indirect regulatory relationships. Most recently, (Y. Li et al. 2014) combined miRNA expression, mRNA expression, and miRNA-mRNA target info to identify miRNA regulatory modules by calculating synergistic miRNA-mRNA interactions.

Several of the proposed module detection methods [Peng 2009; Joung & Fei 2009b; Liu 2009; J. Zhang 2012] focus on using expression data to identify modules that show different behavior between two conditions, for example healthy versus diseased samples. The rationale behind identifying condition-specific mRNA/miRNA modules is that dysregulated mRNA/miRNA relationships may be a mechanistic underpinning of a disease, or might be useful as a biomarker for disease treatment efficacy. In these studies, modules are evaluated for association with a categorical variable of interest (e.g. disease vs. healthy) but to date no algorithm has been proposed to identify association of modules to a continuous variable, for example fasting glucose levels in the blood, or clinical age of onset of neurodegenerative diseases. This chapter presents a novel approach to identify modules from paired miRNA and mRNA expression profiles combined with predicted target information that exhibit a statistical relationship with a continuous quantity of interest. The algorithm is applied to paired, high-throughput mRNA and miRNA sequencing datasets from individuals who have died of Huntington's disease to evaluate how well the expression of module mRNAs and miRNAs can predict CAG repeat length, age of onset, and degree of neurodegeneration.

Methods

The proposed method incorporates a priori miRNA-mRNA target prediction information, paired miRNA and mRNA expression data, and continuous variable information, as depicted in **Figure 10**A. The algorithm requires as input an mRNA expression matrix of S samples x N mRNAs, a miRNA expression matrix of S samples x M miRNAs, one S-length vector of continuous variables, hereafter called the feature, and a set of predicted miRNA/mRNA target pairs. The miRNA matrix is filtered to include only miRNAs found in the predicted target set. The mRNA and miRNA matrices and the predicted targets are used as input to the miRMAP program (Bryan et al. 2014) to detect miRNA/mRNA modules. Briefly, the miRMAP algorithm computes either Pearson or Spearman correlation of all mRNA/miRNA pairs and weights the correlation by the predicted target information. Since it is not necessarily appropriate to expect strictly linear relationships between miRNAs and their targets, Spearman correlation is used for this study. Correlations may be positive or negative, allowing for both direct and indirect regulatory relationships within modules. After constructing the weighted mRNA/miRNA correlations, the resulting matrix is subjected to a biclustering algorithm that “seeds and extends” modules based on highly correlated mRNA/miRNA pairs. The algorithm allows the same mRNAs/miRNAs to participate in more than one module and the number of modules returned is specified by the user, which is set to 25 for this study. The modules identified by miRMAP are collections of correlated mRNAs and miRNAs, termed module members, that are used in downstream analysis.

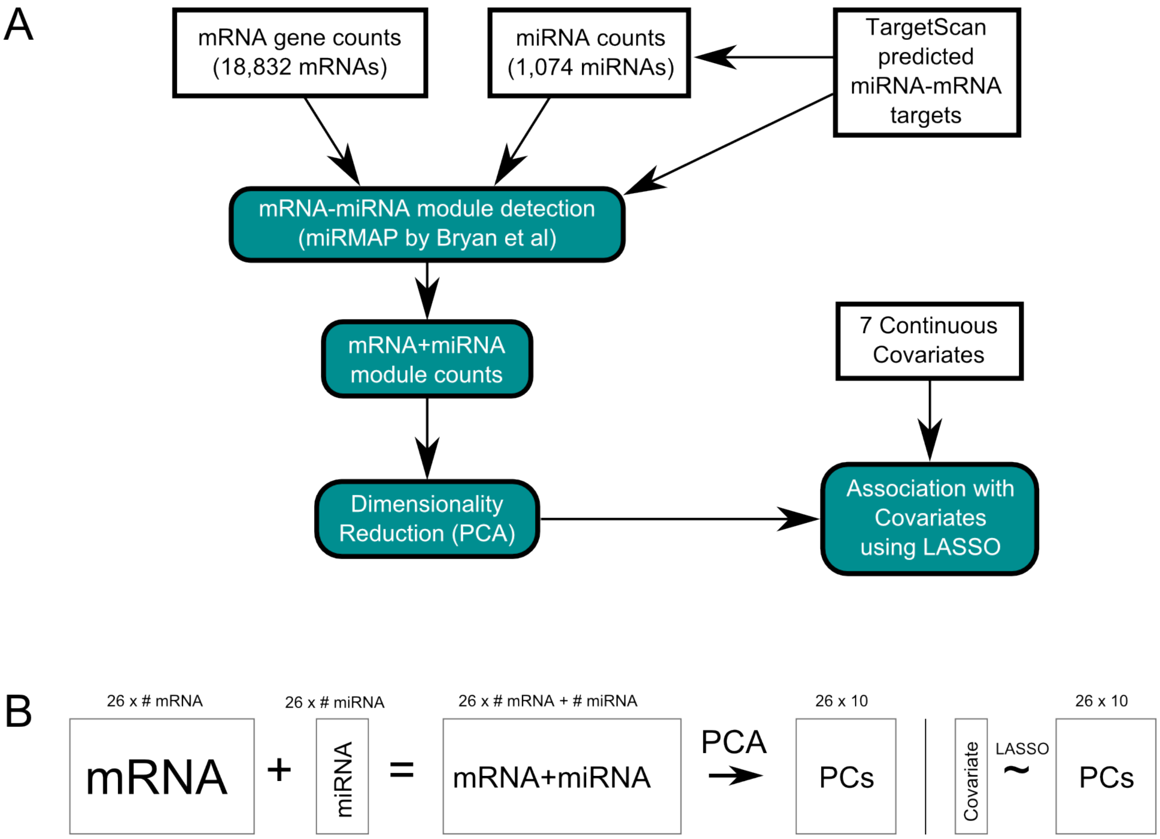

A
mRNA gene counts (18,832 mRNAs)
miRNA counts (1,074 miRNAs)
TargetScan predicted miRNA-mRNA targets
mRNA-miRNA module detection (miRMAP by Bryan et al)
mRNA+miRNA module counts
7 Continuous Covariates
Dimensionality Reduction (PCA)
Association with Covariates using LASSO
B
26 x # mRNA
26 x # miRNA
26 x # mRNA + # miRNA
26 x 10
26 x 10
mRNA
+
miRNA
=
mRNA+miRNA
PCA
PCs
Covariate
LASSO
~
PCs

Figure 10. A) Schematic of analysis pipeline. B) Cartoon depiction of PCA dimensionality reduction procedure.

The hypothesis underlying this methodology is that the statistical variance of a feature may be better explained by the expression of module members than otherwise unrelated mRNAs and miRNAs. To test this hypothesis, we use LASSO regression (Tibshirani 1994) to model a feature as a function of expression values from module members and compare the explained variance (model R2) to LASSO regressions using the same number of randomly selected mRNAs and miRNAs. The LASSO regressions are conducted as follows. Expression values for the members of each module identified by miRMAP are extracted from the original expression matrices and concatenated into a per-module matrix of size S samples x P, where P is the total number of members in the module. When sample size is small, as is often the case in gene expression studies, the number of variables (i.e. module members) in the model may exceed the number of samples, which motivates the choice of LASSO as opposed to classical linear regression. Also, by definition, the columns of the module expression matrix are highly collinear, since they were selected on the basis of high correlation. Such a matrix poses difficulty for meaningful parameter estimation in LASSO regression, so a dimensionality reduction procedure using is employed. Principal Component Analysis (PCA) is conducted on the module expression matrix from above, and the first ten components are retained. The projection of each sample from the original matrix is then computed for these ten components, creating a matrix of size S x 10, where each column contains the projection of a single component for all samples. The columns of this matrix are thus orthogonal by definition and eliminate the collinearity of the module expression values while maintaining a large proportion of the variance across columns. The feature is normalized to have zero mean and unit variance before applying LASSO regression. Normalizing the feature ensures that the scale of the variable does not impact the R2 value of the fitted model, as we are interested only in the explanatory relationship between the feature and PCA-projected expression values. **Figure 10**B illustrates the PCA procedure described above.

To assess whether the module member expression explains more of the feature variance than unrelated mRNAs and miRNAs, a randomization procedure is performed. For each module, random mRNAs and miRNAs in the same numbers as the module members are selected from the original expression matrices. The expression values are concatenated together as for the true modules and the PCA-LASSO regression procedure described above is performed on this random module expression matrix using the feature, retaining the R2 value. This randomization procedure is performed 1000 times for each original module, forming a distribution of “null” R2 statistics. The R2 statistic for the true module is then compared to this distribution of R2 statistics, counting the number of times a module with random members exceeded the true R2. By dividing this count by the number of random trials, we arrive at an empirical evidence score that represents how well the true module better explains the feature than random modules. Though the evidence score lacks the statistical properties to be a p-value, the score is analogous to a significance value and therefore modules with an evidence score of less than 0.01 or 0.05 may be considered pseudo-significant. For brevity, such modules shall be described as significant in this text. The mRNAs within significant modules are subjected to gene set enrichment analysis using the MsigDB C2 Canonical Pathway (Subramanian et al. 2005) gene set database using a hypergeometric test, where a gene set is considered significantly enriched if it achieves Benjamini-Hochberg adjusted p-value < 0.05.

In addition to the evidence score randomization procedure above, additional permutations of the data are applied to the algorithm to better interpret the results from the true modules. It is possible that there is latent structure within the expression data that causes association with the feature. For example, if genome-wide transcriptional dysregulation is associated with the severity of a disease, it may be that any combination of mRNAs and miRNAs can explain the feature simply on account of this relationship, potentially confounding the interpretation of the association of a true module. To account for this, true and randomized module analyses as described above are run after the feature is randomly shuffled, breaking any potential relationship between the sample and the feature. These analyses represent the different “null” distribution describing the expected R2 statistics when there is no relationship between sample expression and the feature overall.

The above methodology was applied to a set of 26 HD individuals with paired mRNA-Seq and miRNA-Seq dataset. The mRNA-Seq datasets were aligned against the hg38 human reference genome using STAR (Dobin et al. 2013) and counted against Gencode v21 (Harrow et al. 2012) gene annotation using htseq-count in the HTSeq package (Anders, Pyl, and Huber 2014). Genes that had zero counts in any sample were filtered, resulting in 18,832 distinct genes left for module detection. The miRNA-Seq data was aligned against the hg38 human genome using bowtie (Langmead et al. 2009) and counted against the miRBase v21 miRNA annotation (Kozomara and Griffiths-Jones 2014) using htseq-count in the HTSeq package (Anders, Pyl, and Huber 2014). The TargetScan v6.2 mRNA/miRNA conserved target database (Lewis, Burge, and Bartel 2005b; Grimson et al. 2007) was used to filter the miRNAs, resulting in 1,074 distinct miRNAs left for module detection.

Seven continuous clinical features were evaluated for association with modules for these HD samples, as described in Table 10. The four primary features are CAG size (i.e. repeat length), clinical age of onset, and H-V cortical and striatal scores (Hadzi et al. 2012), which are histology-based numerical scores indicating degree of involvement in the cortex and striatum, respectively. There is a considerable level of correlation between these four covariates, particularly between CAG repeat length and age of onset, since individuals with longer CAG repeat lengths experience more severe pathology and typically have a younger age of onset. However, CAG does not explain any of these features perfectly, and substantial variance remains after accounting for the statistical contribution of CAG. Therefore, we created three new features using age of onset, cortical score, and striatal score to identify modules that are associated with the residual variance of these features after accounting for the contribution of CAG. CAG-adjusted age of onset was created using results reported in (Djoussé et al. 2003). CAG-adjusted cortical and striatal scores were created by taking the residuals of a linear regression modeling each of these features as a function of CAG repeat length.

Covariate	Min	Mean (std)	Max
Onset	25	44.6 (11.5)	70
CAG Size	40	44.5 (2.6)	51
Striatal Score	1.4	2.6 (0.6)	3.8
Cortical Score	0.4	1.2 (0.5)	2.8
CAG-adj Onset	-1.4	0.4 (0.8)	1.6
CAG-adj Striatal	-1.3	-0.1 (0.7)	1.4
CAG-adj Cortical	-1.4	-0.08 (0.9)	1.9

Table 10. Clinical features and CAG-adjusted derivatives used in this study.

Results

Table 11 contains the number of module members and R2 values for the 25 true modules identified by the PCA-LASSO analysis. Except for module 6, all modules had the same number of mRNAs and miRNAs (50, and 5, respectively). These numbers correspond to two user-specified parameters to miRMAP, namely the maximum number of mRNAs and minimum number of miRNAs allowed in any module. That nearly all modules were limited to this size suggests that there is a high level of correlation among mRNAs and a relatively lower level of correlation among miRNAs in these datasets. When we examined the correlations of mRNAs to miRNAs within each module, we found that a much larger proportion of mRNA/miRNA pairs were positively correlated than negatively correlated. Figure 11A is a heatmap depicting the correlation values for module 0, where red and blue cells represent positive and negative correlations, respectively. This pattern is consistent across all modules as depicted in Figure 11B.

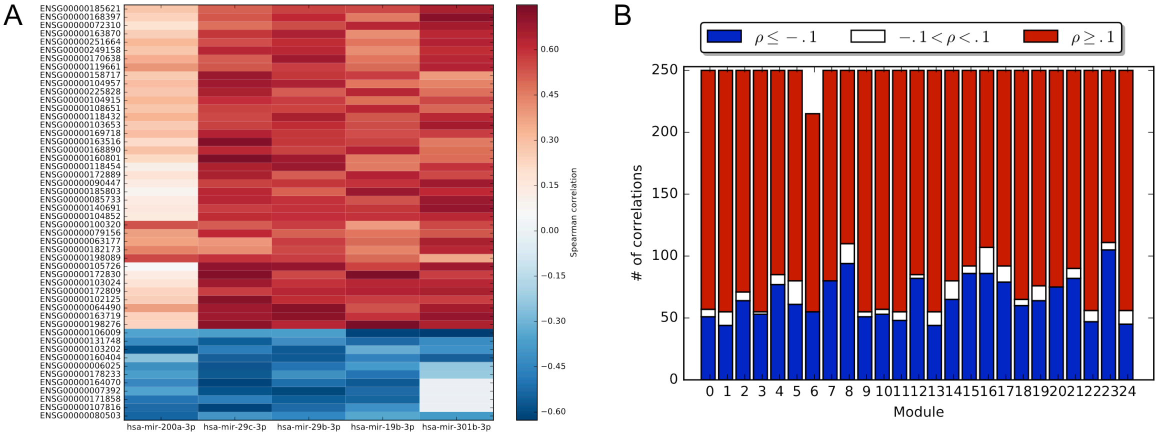


Figure 11. A) Heatmap of mRNA/miRNA correlations in module 0, mRNA on the y axis and miRNA on the x axis. B) Stacked barplot of the proportions of negative (rho <= -.1), near zero (-.1 < rho < .1), and positive (rho >= .1) correlations for all modules.

Since the miRMAP algorithm does not require members to be disjoint between modules, some modules contained many of the same members, as depicted in Figure 12A. The two most similar modules (6 and 8) share 27 of their members (24 mRNAs and 3 miRNAs), and all but two modules (13 and 22), share at least 15 of their members with another module.

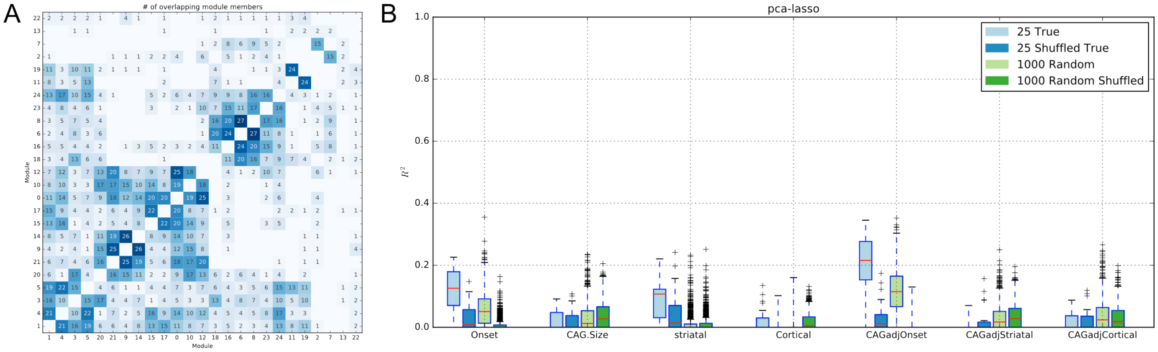


Figure 12. A) Clustered matrix of member overlap between all modules. The number of overlapping members is listed within each cell, overlaps of zero are omitted. Diagonal entries have been set to zero. B) Boxplots of the R2 distributions from each feature for the true modules, true modules with shuffled features, random modules, and random modules with shuffled features.

The R2 values were varied across the modules and features, with onset and CAG-adjusted onset consistently having the highest R2 values whereas cortical and especially CAG-adjusted striatal score R2 values were very close to 0. Figure 12B is a boxplot of the R2 values from Table 11 for the true modules, true modules with shuffled features, random modules, and random modules with shuffled features. The R2 values for true modules explaining CAG-adjusted onset are the most consistently higher than either of the random modules, followed by onset, striatal score, and cortical score. CAG size, CAG-adjusted striatal score, and CAG-adjusted cortical score were explained no better by true modules than random. As expected, when the relationship between expression and feature was broken by shuffling the feature, most of the explanatory power of the true modules reduced to nearly zero for all features.

	# mRNA	# miRNA	Onset	CAG.Size	striatal	Cortical	CAOnset	CAStriatal	CACortical
0	50	5	0.12 (0.14)	0.00 (0.60)	0.06 (0.08)	0.00 (1.00)	0.21 (0.11)	0.00 (1.00)	0.00 (1.00)
1	50	5	0.17 (0.04)	0.00 (1.00)	0.12 (0.02)	0.03 (0.04)	0.35 (0.00)	0.00 (1.00)	0.00 (1.00)
2	50	5	0.03 (0.61)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.06 (0.77)	0.00 (1.00)	0.00 (1.00)
3	50	5	0.13 (0.12)	0.00 (1.00)	0.12 (0.02)	0.03 (0.03)	0.28 (0.01)	0.00 (1.00)	0.00 (1.00)
4	50	5	0.18 (0.03)	0.00 (0.58)	0.11 (0.02)	0.00 (1.00)	0.30 (0.01)	0.00 (1.00)	0.00 (1.00)
5	50	5	0.14 (0.09)	0.00 (1.00)	0.16 (0.01)	0.05 (0.01)	0.30 (0.01)	0.00 (1.00)	0.00 (1.00)
6	43	5	0.07 (0.38)	0.00 (1.00)	0.07 (0.06)	0.00 (1.00)	0.09 (0.62)	0.00 (1.00)	0.00 (1.00)
7	50	5	0.01 (0.79)	0.05 (0.25)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.01 (0.53)	0.08 (0.21)
8	50	5	0.06 (0.46)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.12 (0.45)	0.00 (1.00)	0.00 (1.00)
9	50	5	0.19 (0.02)	0.09 (0.11)	0.12 (0.02)	0.00 (1.00)	0.20 (0.14)	0.02 (0.50)	0.09 (0.16)
10	50	5	0.21 (0.01)	0.07 (0.19)	0.15 (0.01)	0.02 (0.06)	0.25 (0.04)	0.00 (1.00)	0.06 (0.26)
11	50	5	0.06 (0.45)	0.00 (1.00)	0.12 (0.02)	0.11 (0.01)	0.22 (0.10)	0.02 (0.49)	0.00 (1.00)
12	50	5	0.20 (0.02)	0.06 (0.21)	0.13 (0.02)	0.00 (1.00)	0.24 (0.05)	0.00 (1.00)	0.06 (0.27)
13	50	5	0.00 (1.00)	0.00 (1.00)	0.03 (0.15)	0.13 (0.00)	0.02 (0.92)	0.07 (0.16)	0.00 (1.00)
14	50	5	0.19 (0.02)	0.06 (0.23)	0.11 (0.02)	0.00 (1.00)	0.24 (0.05)	0.00 (1.00)	0.05 (0.31)
15	50	5	0.11 (0.17)	0.00 (1.00)	0.01 (0.24)	0.00 (1.00)	0.18 (0.18)	0.00 (1.00)	0.00 (1.00)
16	50	5	0.07 (0.37)	0.00 (1.00)	0.01 (0.28)	0.00 (1.00)	0.15 (0.31)	0.00 (1.00)	0.00 (1.00)
17	50	5	0.08 (0.35)	0.00 (1.00)	0.01 (0.28)	0.00 (1.00)	0.22 (0.10)	0.00 (1.00)	0.00 (1.00)
18	50	5	0.03 (0.64)	0.00 (1.00)	0.09 (0.04)	0.00 (0.11)	0.12 (0.46)	0.00 (1.00)	0.00 (1.00)
19	50	5	0.08 (0.30)	0.00 (1.00)	0.13 (0.02)	0.08 (0.01)	0.25 (0.04)	0.00 (0.60)	0.00 (1.00)
20	50	5	0.22 (0.01)	0.05 (0.28)	0.19 (0.00)	0.04 (0.02)	0.30 (0.01)	0.00 (1.00)	0.04 (0.41)
21	50	5	0.18 (0.03)	0.06 (0.21)	0.12 (0.02)	0.00 (1.00)	0.21 (0.11)	0.00 (1.00)	0.06 (0.28)
22	50	5	0.23 (0.01)	0.04 (0.33)	0.22 (0.00)	0.08 (0.01)	0.34 (0.00)	0.00 (1.00)	0.02 (0.50)
23	50	5	0.14 (0.10)	0.01 (0.51)	0.06 (0.07)	0.00 (1.00)	0.21 (0.12)	0.00 (1.00)	0.01 (0.60)
24	50	5	0.16 (0.05)	0.00 (1.00)	0.10 (0.03)	0.00 (1.00)	0.31 (0.01)	0.00 (1.00)	0.00 (1.00)

Table 11. Module sizes and R2 statistics for all features for the 25 modules. First column is module label. Feature columns report R2 values and evidence scores in parenthesis.

Evidence scores for each module for each feature are found in parenthesis of Table 11. Modules with evidence scores < 0.05 are considered interesting since they explain greater feature variance than 95% of the random modules evaluated. No modules for CAG size, CAG-adjusted striatal, or CAG-adjusted cortical passed this threshold, while 10, 15, 8, and 10 modules had passing evidence scores for onset, striatal score, cortical score, and CAG-adjusted onset features, respectively. Table 12 contains R2 values and top enriched gene sets for each module that was found to be significant in at least one feature.

	Onset	striatal	Cortical	CAGadjOnset	# enriched	Enrich
1	0.17	0.11	0.02	0.34	154	KEGG_RIBOSOME, REACTOME_NONSENSE_MEDIATED_DECAY_ENHANCED_BY_THE_EXON_JUNCTION_COMPLEX, REACTOME_INFLUENZA_VIRAL_RNA_TRANSCRIPTION_AND_REPLICATION, REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE, REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION
3	NA	0.12	0.03	0.27	2	PID_ALK1PATHWAY, REACTOME_ACTIVATED_NOTCH1_TRANSMITS_SIGNAL_TO_THE_NUCLEUS
4	0.17	0.10	NA	0.30	17	KEGG_RIBOSOME, REACTOME_NONSENSE_MEDIATED_DECAY_ENHANCED_BY_THE_EXON_JUNCTION_COMPLEX, REACTOME_PEPTIDE_CHAIN_ELONGATION, REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE, REACTOME_SYNTHESIS_OF_PIPS_AT_THE_PLASMA_MEMBRANE
5	NA	0.16	0.05	0.30	0	no enrichment
9	0.19	0.12	NA	NA	29	KEGG_GLIOMA, REACTOME_CIRCADIAN_REPRESSION_OF_EXPRESSION_BY_REV_ERBA, REACTOME_RORA_ACTIVATES_CIRCADIAN_EXPRESSION, PID_HDAC_CLASSIII_PATHWAY, KEGG_PYRUVATE_METABOLISM
10	0.21	0.14	NA	0.25	0	no enrichment
11	NA	0.12	0.10	NA	14	REACTOME_SIGNALING_BY_NOTCH4, REACTOME_SIGNALING_BY_NOTCH2, REACTOME_SIGNALING_BY_NOTCH3, REACTOME_ACTIVATED_NOTCH1_TRANSMITS_SIGNAL_TO_THE_NUCLEUS, REACTOME_SIGNALING_BY_NOTCH
12	0.19	0.12	NA	NA	1	PID_RXR_VDR_PATHWAY
13	NA	NA	0.13	NA	2	PID_RANBP2PATHWAY, PID_BETACATENIN_NUC_PATHWAY
14	0.19	0.10	NA	0.24	2	REACTOME_SYNTHESIS_OF_PIPS_AT_THE_PLASMA_MEMBRANE, PID_MYC_PATHWAY
18	NA	0.08	NA	NA	0	no enrichment
19	NA	0.12	0.08	0.25	162	REACTOME_SIGNALING_BY_NOTCH2, REACTOME_SIGNALING_BY_NOTCH3, REACTOME_SIGNALING_BY_NOTCH4, PID_ILK_PATHWAY, REACTOME_SIGNALING_BY_CONSTITUTIVELY_ACTIVE_EGFR
20	0.22	0.19	0.04	0.29	3	PID_RXR_VDR_PATHWAY,

						REACTOME_RORA_ACTIVATES_CIRCADIAN_EXPRESSION, REACTOME_ELONGATION_ARREST_AND_RECOVERY
21	0.17	0.12	NA	NA	0	no enrichment
22	0.22	0.22	0.07	0.33	17	KEGG_P53_SIGNALING_PATHWAY, REACTOME_MYOGENESIS, REACTOME_BMAL1_CLOCK_NPAS2_ACTIVATES_CIRCADIAN_EXPRESSION, REACTOME_CELL_CELL_JUNCTION_ORGANIZATION, KEGG_CIRCADIAN_RHYTHM_MAMMAL
24	0.15	0.10	NA	0.30	72	KEGG_RIBOSOME, BIOCARTA_GATA3_PATHWAY, PID_RXR_VDR_PATHWAY, REACTOME_PEPTIDE_CHAIN_ELONGATION, REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE

Table 12. R2 values and top significantly enriched MsigDB C2 Canonical Pathway gene sets at BH-adjusted p<0.05 for modules. Only the top 5 most enriched gene sets by p-value are reported.

Module 1, 20, and 22 are significantly associated with all four of the features in the table. Striatal score is significantly associated with the most modules (15), followed by onset and CAG-adjusted onset (10), and cortical score (8). Taken together, the modules are most associated with biological processes related to the ribosome, transcription and translation, vitamin D metabolism, circadian rhythm expression, apoptosis, extracellular matrix organization, and NOTCH signaling. Module 13 is exclusively significant for cortical score and is enriched for RANBP2 and beta catenin pathways. Four modules (5, 10, 18 and 21) were not enriched for any gene sets but explained a comparable amount of feature variance as other modules that did have significant enrichment.

Discussion

We have presented a novel methodology for identifying mRNA/miRNA modules that are associated with a continuous variable and applied the technique to paired mRNA-Seq and miRNA-Seq datasets from post-mortem human HD brains. The method combines existing tools (miRMAP, PCA, LASSO regression) into a pipeline that addresses a number of major challenges posed in elucidating the relationships between mRNA and miRNA abundance data and phenotypic features. First, using paired mRNA/miRNA data from the same individuals allows more accurate estimation of relationships between mRNA and miRNA abundance than by comparing unrelated datasets or computational approaches alone. Second, the miRMAP algorithm was chosen in part because it considers indirect (i.e. positively correlated) relationships as well as the expected negative correlations. This is important, since most of the mRNA/miRNA relationships identified are indeed positively correlated, and still the mRNAs within many of the modules are significantly enriched in biological processes, suggesting they are functionally related. Third, the use of PCA alleviates the problem of high collinearity among the regressors, which module members exhibit by definition, but still retains the majority of variance among the expression variables. And finally, using LASSO regression addresses the problem of having relatively few samples compared to the number of variables in the model, a common situation in gene expression studies.

The analysis identified many modules that significantly explained a number of the clinical features. Of particular interest are the modules associated with CAG-adjusted age of onset, most of which explained much more variance in the feature than randomly chosen modules. This quantity represents the variance in age of onset that cannot be explained by CAG repeat size, which is the primary determinant of disease severity. With few exceptions, GWAS studies of HD patients have yet to find genetic markers that are associated with residual age of onset after accounting for CAG size, making the finding in this study particularly compelling. What is potentially more interesting is the lack of modules found to be significantly associated with CAG size itself. We note that the original onset feature, unadjusted for CAG, is also significantly explained by a number of modules, but that the overall R2 values are less than those for CAG-adjusted onset. This suggests that the contribution of CAG to onset may involve genes that are less functionally related, since the modules that explain CAG-adjusted onset are significantly enriched in a number of biological processes. This may in turn suggest that the molecular responses to CAG and those driving age of onset are distinct, and that those genes which are driving the variance in age of onset beyond CAG may yet be identified. It is also interesting to note that the CAG-adjusted striatal and cortical scores are not better explained by any module than random. This suggests that the relationship between CAG and age of onset is somehow different than that with the histological markers of neurodegeneration in HD.

The analysis identified multiple modules that were significantly enriched in gene sets for specific biological processes that may be related to HD pathology. In particular, processes related to the ribosome and protein translation are enriched in multiple modules that are associated with age of onset, CAG-adjusted age of onset, and striatal and cortical score, suggesting that the mRNAs in these modules may influence, or be affected by, the neurodegenerative process. Another consistently implicated biological process across modules is NOTCH signaling, a process not previously implicated in HD. A third consistently enriched biological process relates to circadian rhythm, and there is growing evidence that circadian rhythm and sleep cycle disorder is a symptom of, and possible contributor to, HD pathology (Morton 2013). Together, these results present strong evidence that mRNA/miRNA modules may indeed reflect aspects of HD pathology. Further investigation into which specific genes are driving the association of these modules to the clinical features is likely to yield insight into the molecular mechanisms underlying HD.

One shortcoming of this analysis is that it may be challenging to ascertain which genes within modules are the largest contributors to the overall association with features. This is due to the PCA data reduction step, where instead of individual genes in the model, associations are made against principal component projections of the module member expression values. Model variables therefore represent weightings of all module members, which makes the interpretation of the results somewhat challenging. Other methods that use PCA in this way have been proposed, e.g. WGCNA (Langfelder and Horvath 2008), which may be helpful in this regard.

Chapter 5. Conclusions, Prospective, and Future Work

The studies presented in this thesis constitute the most comprehensive characterization of transcriptional signatures of HD and PD in post mortem human brains to date. The use of mRNA-Seq libraries allows for unbiased, genome-wide assessment of differences in mRNA species abundance, and therefore provides both potent hypothesis-generation opportunities as well as a detailed dataset that may be queried for evidence supporting specific hypotheses. The wealth of biological information produced by these high-throughput datasets is valuable in this regard but also presents significant challenges in how to best interpret large amounts of biological information. In both differential expression, studies, for example, there were on the order of thousands of differentially expressed genes, and even after gene set enrichment analysis, which attempts to condense gene lists into more manageable and interpretable numbers, we found hundreds of biological processes involved in neurodegeneration. The task of interpreting the biology implicated by these studies, and translating that knowledge into potentially actionable next steps, comprised much of the effort toward effectively present these results. Nonetheless, this analysis suggests coherent biological processes, some specific and others broad, that are implicated in distinguishing between neurodegenerated and healthy tissues as also related to clinical features of the diseases.

The biological processes most prominently implicated in HD and PD are neuroinflammatory and immune responses, which have considerable supporting evidence from prior studies [CITE], that seem to be reflected in both diseases in consistent ways. In particular, pathways related to NFkB activation are consistently perturbed in the pan-neurodegenerative phenotype observed in HD and PD. The involvement of NFkB in both diseases is interesting in several respects. First, NFkB is a critical pathway in cellular responses to stress, including inflammatory and immune response and is expressed in nearly all human tissue types [CITE]. Second, it functions as a transcriptional activator and perturbations to the NFkB pathway have been associated with multiple diseases [CITE]. While there is no direct evidence that there are alterations to the NFkB system itself other than increased activity in these datasets, such activity may nonetheless contribute to the environment of cellular stress observed in neurodegeneration. And third, NFkB pathways have been implicated specifically in the central nervous system as modulators of synaptic plasticity, learning, and other critical brain-related functions [CITE]. More generally, the datasets implicate gene signatures that span the broader spectrum of the inflammatory and immune response. In both diseases, pathways related to both innate and adaptive immune response are enriched, highlighted by multiple toll-like, interleukin, and other cytokine receptor signaling pathways. Thus, these processes likely play a significant role in neurodegeneration, but whether and how this role is protective, deleterious, or both remains to be conclusively shown.

Other biological processes implicated by the datasets suggest a more direct mechanistic role in neurodegeneration. Pathways related to neurons are consistently negatively enriched in both diseases. While it may be tempting to interpret this as reflective of a loss of neurons due to degeneration, it is important to note that the brain region studied is not known to be significantly involved in either HD or PD, and therefore not likely to be due to local neurodegeneration. It may be that these neurons are stressed in the same way, but to a lesser extent, than those primarily affected, or alternatively that they are unaffected by the primary effects of the disease but suffer in response to the consequences of neurodegenerating neurons elsewhere in the brain. Another explanation might be that, in response to the toxic microenvironment of the neurodegeneration, the activity of immune cells in the brain is increased in proportion to the number of neurons, resulting in an apparent down regulation of neuronal genes. However, this hypothesis is challenged by the observation that not all of the implicated neuronal gene sets are negatively enriched, and neuron-specific pathways related to PI3K and prion response are increased, suggesting that neurons are not substantially decreased relative to glia in neurodegenerative disease compared with healthy tissue. Fortunately, other enriched pathways may provide insight into the mechanisms underlying these differences and inspire new focused experiments to shed light on these aspects of neurodegenerative pathology.

A greater understanding of how and why the differentially expressed genes associated with neurodegeneration are controlled will lead to insights that may result in better therapies for neurodegeneration. Multiple transcription factors are implicated by the differentially expressed genes in both HD and PD, suggesting direct causal transcriptional mechanisms underlying the observed gene expression patterns. Further analysis of these specific transcription factors and their targets is an obvious first step in this pursuit, but studying other modulators of RNA transcriptional abundance may also improve our understanding of the mechanistic underpinnings of the disease. As described in this thesis, by combining miRNA abundance data with mRNA abundance data to identify regulatory patterns that explain clinical features, suggestive relationships of mRNAs to disease progression in HD were revealed. Specifically, groups of mRNAs and miRNAs whose collective abundances could explain as much as 30% or more of the age of onset after adjusting for the contribution of CAG-repeat size. Genetic or genomic factors explaining CAG-adjusted age of onset in particular have been elusive to date, and the presence of explanatory factors within gene and miRNA expression data is encouraging. It is remarkable that, by simply combining the mRNA and miRNA data from the same individuals, statistical (and potentially regulatory) patterns between mRNAs and miRNAs are so associated with clinical features, considering the analysis did not incorporate any information about genes characteristic of HD. It remains to be shown whether similar relationships between these molecular species exist in healthy tissues, but comparing and contrasting these regulatory patterns in this way may lead to a better understanding of why some HD subjects experience more severe degeneration than others with the same mutational burden.

The results of these studies highlight the considerable challenges neurodegenerative diseases pose to identifying successful therapies. The differential expression studies in HD (Chapter 2) and HD vs. PD (Chapter 3) show that the transcriptional differences between degenerated normal brains are vast and span a broad spectrum of biological processes. Considering the severe consequences of neuronal degeneration and the long duration of these diseases, it is somewhat unsurprising that the extent and breadth of these differences are so dramatic. As post-mitotic cells, neurons may be sensitive to small but chronic deleterious effects that can compound over time, making detection of strong causal molecular markers of disease elusive. The incremental effect of stressors leading to degeneration also pose difficulties for identifying and administering neuroprotective therapies, since the long-term studies required to assess efficacy are costly and logistically challenging. Furthermore, while studying human samples (vis a vis in vitro or animal models) provides a direct view into the neurodegenerative phenotype, it is difficult if not impossible to separate causal versus consequential patterns in data generated from post-mortem tissues.

Though significant challenges remain, these studies suggest a number of clear next steps to further our understanding of neurodegeneration. The most important may be to thoroughly investigate the role of inflammation and immune response in the neurodegenerated brain. One particular challenge in this pursuit is to find an appropriate biological model of neuroinflammation, since common animal models, most notably mice, exhibit significantly different immune response capabilities than humans [CITE] and in vitro systems that mimic the multicellular immune response are complex and can be difficult to culture. To follow up on the exciting finding that certain mRNA species explain clinical features, a high priority should be placed on assessing these relationships in peripheral material of living HD subjects, such as blood, cerebrospinal fluid, or other available tissues. Should similar relationships exist between clinical features and mRNA abundance measurements extracted from these tissues, the potential benefit of a reliable biomarker for progression in HD would be of immense significance for making clinical decisions and assessing the efficacy of HD therapies. These are but two immediate follow-up studies suggested by these data, but further mining of the expansive results presented in this thesis are certain to many potentially fruitful avenues of discovery.

BIBLIOGRAPHY

Alexa, Adrian, and Jorg Rahnenfuhrer. 2014. *topGO: Enrichment Analysis for Gene Ontology* (version 2.14.0).

Allam, Mohamed Farouk, Amparo Serrano Del Castillo, and Rafael Fernández-Crehuet Navajas. 2005. “Parkinson’s Disease Risk Factors: Genetic, Environmental, or Both?” *Neurological Research* 27 (2): 206–8.

Allen Reish, Heather E., and David G. Standaert. 2015. “Role of α-Synuclein in Inducing Innate and Adaptive Immunity in Parkinson Disease.” *Journal of Parkinson’s Disease* 5 (1): 1–19.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2014. “HTSeq - A Python Framework to Work with High-Throughput Sequencing Data.” *bioRxiv*, August, 002824.

Andersson, M., C. Konradi, and M. A. Cenci. 2001. “cAMP Response Element-Binding Protein Is Required for Dopamine-Dependent Gene Expression in the Intact but Not the Dopamine-Denervated Striatum.” *Journal of Neuroscience* 21 (24): 9930–43.

Arrasate, Montserrat, and Steven Finkbeiner. 2012. “Protein Aggregates in Huntington’s Disease.” *Experimental Neurology* 238 (1): 1–11.

Arroyo, Daniela S., Javier A. Soria, Emilia A. Gaviglio, Maria C. Rodriguez-Galan, and Pablo Iribarren. 2011. “Toll-like Receptors Are Key Players in Neurodegeneration.” *International Immunopharmacology* 11 (10): 1415–21.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology.” *Nature Genetics* 25 (1): 25–29.

Aviles-Olmos, Iciar, Patricia Limousin, Andrew Lees, and Thomas Foltynie. 2013. “Parkinson’s Disease, Insulin Resistance and Novel Agents of Neuroprotection.” *Brain* 136 (Pt 2): 374–84.

Baumann, Nicole, and Danielle Pham-Dinh. 2001. “Biology of Oligodendrocyte and Myelin in the Mammalian Central Nervous System.” *Physiological Reviews* 81 (2): 871–927.

Beach, Thomas G., Lucia I. Sue, Douglas G. Walker, Alex E. Roher, Lihfen Lue, Linda Vedders, Donald J. Connor, Marwan N. Sabbagh, and Joseph Rogers. 2008-9. “The Sun Health Research Institute Brain Donation Program: Description and Experience, 1987–2007.” *Cell and Tissue Banking* 9 (3): 229–45.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Bi, Yingtao, and Ramana V. Davuluri. 2013. “NPEBseq: Nonparametric Empirical Bayesian-Based Procedure for Differential Expression Analysis of RNA-Seq Data.” *BMC Bioinformatics* 14 (August): 262.

Björkqvist, Maria, Edward J. Wild, Jenny Thiele, Aurelio Silvestroni, Ralph Andre, Nayana Lahiri, Elsa Raibon, et al. 2008. “A Novel Pathogenic Pathway of

Immune Activation Detectable before Clinical Onset in Huntington's Disease." *The Journal of Experimental Medicine* 205 (8): 1869–77.

Block, Robert C., E. Ray Dorsey, Christopher A. Beck, J. Thomas Brenna, and Ira Shoulson. 2010. "Altered Cholesterol and Fatty Acid Metabolism in Huntington Disease." *Journal of Clinical Lipidology* 4 (1): 17–23.

Blumkin, Lubov, Ayelet Halevy, Dominique Ben-Ami-Raichman, Dvir Dahari, Ami Haviv, Cohen Sarit, Dorit Lev, Marjo S. van der Knaap, Tally Lerman-Sagie, and Esther Leshinsky-Silver. 2014. "Expansion of the Spectrum of TUBB4A-Related Disorders: A New Phenotype Associated with a Novel Mutation in the TUBB4A Gene." *Neurogenetics* 15 (2): 107–13.

Braak, Heiko, Kelly Del Tredici, Udo Rüb, Rob A. I. de Vos, Ernst N. H. Jansen Steur, and Eva Braak. 2003. "Staging of Brain Pathology Related to Sporadic Parkinson's Disease." *Neurobiology of Aging* 24 (2): 197–211.

Bryan, Kenneth, Marta Terrile, Isabella M. Bray, Raquel Domingo-Fernandéz, Karen M. Watters, Jan Koster, Rogier Versteeg, and Raymond L. Stallings. 2014. "Discovery and Visualization of miRNA–mRNA Functional Modules within Integrated Data Using Bicluster Analysis." *Nucleic Acids Research* 42 (3): e17–e17.

Cam, Maren, Hemant K. Bid, Linlin Xiao, Gerard P. Zambetti, Peter J. Houghton, and Hakan Cam. 2014. "p53/TAp63 and AKT Regulate Mammalian Target of Rapamycin Complex 1 (mTORC1) Signaling through Two Independent Parallel Pathways in the Presence of DNA Damage." *The Journal of*

Biological Chemistry 289 (7): 4083–94.

Canal, Mercè, Joan Romaní-Aumedes, Núria Martín-Flores, Víctor Pérez-Fernández, and Cristina Malagelada. 2014. “RTP801/REDD1: A Stress Coping Regulator That Turns into a Troublemaker in Neurodegenerative Disorders.” *Frontiers in Cellular Neuroscience* 8 (October): 313.

Cha, Jang-Ho J. 2000. “Transcriptional Dysregulation in Huntington’s Disease.” *Trends in Neurosciences* 23 (9): 387–92.

———. 2007. “Transcriptional Signatures in Huntington’s Disease.” *Progress in Neurobiology* 83 (4): 228–48.

Chang, Louise, Shian-Huey Chiang, and Alan R. Saltiel. 2013. “TC10α Is Required for Insulin-Stimulated Glucose Uptake in Adipocytes.” *Endocrinology*, July. Endocrine Society. doi:10.1210/en.2006-1167.

Choi, Yun-Sik, Boyoung Lee, Hee-Yeon Cho, Iza B. Reyes, Xin-An Pu, Takaomi C. Saido, Kari R. Hoyt, and Karl Obrietan. 2009. “CREB Is a Key Regulator of Striatal Vulnerability in Chemical and Genetic Models of Huntington’s Disease.” *Neurobiology of Disease* 36 (2): 259–68.

Coppedè, Fabio. 2011. “An Overview of DNA Repair in Amyotrophic Lateral Sclerosis.” *Scientific World Journal* 11 (October): 1679–91.

Cordes, Kimberly R., and Deepak Srivastava. 2009. “MicroRNA Regulation of Cardiovascular Development.” *Circulation Research* 104 (6): 724–32.

Coronnello, Claudia, Ryan Hartmaier, Arshi Arora, Luai Huleihel, Kusum V. Pandit, Abha S. Bais, Michael Butterworth, et al. 2012. “Novel Modeling of

Combinatorial miRNA Targeting Identifies SNP with Potential Role in Bone Density.” *PLoS Computational Biology* 8 (12): e1002830.

Costa, Veronica, and Luca Scorrano. 2012. “Shaping the Role of Mitochondria in the Pathogenesis of Huntington’s Disease.” *The EMBO Journal* 31 (8): 1853–64.

Croft, David, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, et al. 2014-1. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 42 (D1): D472–77.

Cronberg, T., M. Rundgren, E. Westhall, E. Englund, R. Siemund, I. Rosén, H. Widner, and H. Friberg. 2011. “Neuron-Specific Enolase Correlates with Other Prognostic Markers after Cardiac Arrest.” *Neurology* 77 (7): 623–30.

Crotti, Andrea, Christopher Benner, Bilal E. Kerman, David Gosselin, Clotilde Lagier-Tourenne, Chiara Zuccato, Elena Cattaneo, Fred H. Gage, Don W. Cleveland, and Christopher K. Glass. 2014. “Mutant Huntingtin Promotes Autonomous Microglia Activation via Myeloid Lineage-Determining Factors.” *Nature Neuroscience* 17 (4): 513–21.

Damiano, Maria, Laurie Galvan, Nicole Déglon, and Emmanuel Brouillet. 2010. “Mitochondria in Huntington’s Disease.” *Biochimica et Biophysica Acta* 1802 (1): 52–61.

Dao, Phuong, Ibrahim Numanagić, Yen-Yi Lin, Faraz Hach, Emre Karakoc, Nilgun Donmez, Colin Collins, Evan E. Eichler, and S. Cenk Sahinalp. 2014. “ORMAN: Optimal Resolution of Ambiguous RNA-Seq Multimappings in the

Presence of Novel Isoforms." *Bioinformatics* 30 (5): 644–51.

D'Aversa, Teresa G., Eliseo A. Eugenin, Lillie Lopez, and Joan W. Berman. 2013-4. "Myelin Basic Protein Induces Inflammatory Mediators from Primary Human Endothelial Cells and Blood-Brain Barrier Disruption: Implications for the Pathogenesis of Multiple Sclerosis." *Neuropathology and Applied Neurobiology* 39 (3): 270–83.

Degli-Esposti, M. A., W. C. Dougall, P. J. Smolak, J. Y. Waugh, C. A. Smith, and R. G. Goodwin. 1997. "The Novel Receptor TRAIL-R4 Induces NF-kappaB and Protects against TRAIL-Mediated Apoptosis, yet Retains an Incomplete Death Domain." *Immunity* 7 (6): 813–20.

Dennis, Michael D., Nora K. McGhee, Leonard S. Jefferson, and Scot R. Kimball. 2013. "Regulated in DNA Damage and Development 1 (REDD1) Promotes Cell Survival during Serum Deprivation by Sustaining Repression of Signaling through the Mechanistic Target of Rapamycin in Complex 1 (mTORC1)." *Cellular Signalling* 25 (12): 2709–16.

Devi, Latha, and Masuo Ohno. 2014. "PERK Mediates eIF2α Phosphorylation Responsible for BACE1 Elevation, CREB Dysfunction and Neurodegeneration in a Mouse Model of Alzheimer's Disease." *Neurobiology of Aging* 35 (10): 2272–81.

Dexter, David T., and Peter Jenner. 2013. "Parkinson Disease: From Pathology to Molecular Disease Mechanisms." *Free Radical Biology and Medicine* 62 (September): 132–44.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2013. “A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis.” *Briefings in Bioinformatics* 14 (6): 671–83.

Djousse, Luc, Beth Knowlton, Michael R. Hayden, Elisabeth W. Almqvist, Ryan R. Brinkman, Christopher A. Ross, Russel L. Margolis, et al. 2004-6. “Evidence for a Modifier of Onset Age in Huntington Disease Linked to the HD Gene in 4p16.” *Neurogenetics* 5 (2): 109–14.

Dobbs, R. J., A. Charlett, A. G. Purkiss, S. M. Dobbs, C. Weller, and D. W. Peterson. 1999. “Association of Circulating TNF-α and IL-6 with Ageing and Parkinsonism.” *Acta Neurologica Scandinavica* 100 (1): 34–41.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Dumitriu, Alexandra, Jeanne C. Latourelle, Tiffany C. Hadzi, Nathan Pankratz, Dan Garza, John P. Miller, Jeffery M. Vance, Tatiana Foroud, Thomas G. Beach, and Richard H. Myers. 2012. “Gene Expression Profiles in Parkinson Disease Prefrontal Cortex Implicate FOXO1 and Genes under Its Transcriptional Regulation.” *PLoS Genetics* 8 (6): e1002794.

Ellrichmann, Gisa, Christiane Reick, Carsten Saft, and Ralf A. Linker. 2013. “The Role of the Immune System in Huntington’s Disease.” *Journal of Immunology Research* 2013 (July): e541259.

Elstner, Matthias, Christopher M. Morris, Katharina Heim, Andreas Bender, Divya Mehta, Evelyn Jaros, Thomas Klopstock, Thomas Meitinger, Douglass M. Turnbull, and Holger Prokisch. 2011. “Expression Analysis of Dopaminergic Neurons in Parkinson’s Disease and Aging Links Transcriptional Dysregulation of Energy Metabolism to Cell Death.” *Acta Neuropathologica* 122 (1): 75–86.

Enright, Anton J., Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S. Marks. 2003. “MicroRNA Targets in Drosophila.” *Genome Biology* 5 (1): R1.

Feng, Yanming, Sean M. Hartig, John E. Bechill, Elisabeth G. Blanchard, Eva Caudell, and Seth J. Corey. 2010. “The Cdc42-Interacting Protein-4 (CIP4) Gene Knock-out Mouse Reveals Delayed and Decreased Endocytosis.” *Journal of Biological Chemistry* 285 (7): 4348–54.

Firth, David. 1993. “Bias Reduction of Maximum Likelihood Estimates.” *Biometrika* 80 (1): 27–38.

Flood, Patrick M., Li Qian, Lynda J. Peterson, Feng Zhang, Jing-Shan Shi, Hui-Ming Gao, and Jau-Shyong Hong. 2011. “Transcriptional Factor NF-κB as a Target for Therapy in Parkinson’s Disease.” *Parkinson’s Disease* 2011 (March): 216298.

Frenkel, Dan. 2015. “A New TRAIL in Alzheimer’s Disease Therapy.” *Brain* 138 (Pt 1): 8–10.

Gafni, Juliette, and Lisa M. Ellerby. 2002. “Calpain Activation in Huntington’s Disease.” *Journal of Neuroscience* 22 (12): 4842–49.

Gafni, Juliette, Evan Hermel, Jessica E. Young, Cheryl L. Wellington, Michael R. Hayden, and Lisa M. Ellerby. 2004. “Inhibition of Calpain Cleavage of Huntingtin Reduces Toxicity Accumulation of Calpain/Caspase Fragments in the Nucleus.” *Journal of Biological Chemistry* 279 (19): 20211–20.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. “A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models.” *Annals of Applied Statistics* 2 (4): 1360–83.

Gennarino, Vincenzo Alessandro, Marco Sardiello, Raffaella Avellino, Nicola Meola, Vincenza Maselli, Santosh Anand, Luisa Cutillo, Andrea Ballabio, and Sandro Banfi. 2009. “MicroRNA Target Prediction by Expression Analysis of Host Genes.” *Genome Research* 19 (3): 481–90.

Ghosh, Anamitra, Avik Roy, Xiaojuan Liu, Jeffrey H. Kordower, Elliott J. Mufson, Dean M. Hartley, Sankar Ghosh, R. Lee Mosley, Howard E. Gendelman, and Kalipada Pahan. 2007. “Selective Inhibition of NF-kappaB Activation Prevents Dopaminergic Neuronal Loss in a Mouse Model of Parkinson’s Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (47): 18754–59.

Giacomello, Marta, Roman Hudec, and Raffaele Lopreiato. 2011. “Huntington’s Disease, Calcium, and Mitochondria.” *BioFactors* 37 (3): 206–18.

Goll, Darrel E., Valery F. Thompson, Hongqi Li, Wei Wei, and Jinyang Cong. 2003. “The Calpain System.” *Physiological Reviews* 83 (3): 731–801.

Grimson, Andrew, Kyle Kai-How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. 2007. “MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing.” *Molecular Cell* 27 (1): 91–105.

Gronski, Matthew A., Jason M. Kinchen, Ignacio J. Juncadella, Nathalie C. Franc, and Kodi S. Ravichandran. 2009. “An Essential Role for Calcium Flux in Phagocytes for Apoptotic Cell Engulfment and the Anti-Inflammatory Response.” *Cell Death and Differentiation* 16 (10): 1323–31.

Habicher, Judith, Tatjana Haitina, Inger Eriksson, Katarina Holmborn, Tabea Dierker, Per E. Ahlberg, and Johan Ledin. 2015. “Chondroitin / Dermatan Sulfate Modification Enzymes in Zebrafish Development.” *PLoS One* 10 (3): e0121957.

Hadzi, Tiffany C., Audrey E. Hendricks, Jeanne C. Latourelle, Kathryn L. Lunetta, L. Adrienne Cupples, Tammy Gillis, Jayalakshmi Srinidhi Mysore, et al. 2012. “Assessment of Cortical and Striatal Involvement in 523 Huntington Disease Brains.” *Neurology* 79 (16): 1708–15.

Halliday, G. M., K. Del Tredici, and H. Braak. 2006. “Critical Appraisal of Brain Pathology Staging Related to Presymptomatic and Symptomatic Cases of

Sporadic Parkinson’s Disease.” *Journal of Neural Transmission. Supplementum*, no. 70: 99–103.

Hamza, Taye H., and Haydeh Payami. 2010-4. “The Heritability of Risk and Age at Onset of Parkinson’s Disease after Accounting for Known Genetic Risk Factors.” *Journal of Human Genetics* 55 (4): 241–43.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. “GENCODE: The Reference Human Genome Annotation for The ENCODE Project.” *Genome Research* 22 (9): 1760–74.

Hayashida, Naoki, Mitsuaki Fujimoto, Ke Tan, Ramachandran Prakasam, Toyohide Shinkawa, Liangping Li, Hitoshi Ichikawa, Ryosuke Takii, and Akira Nakai. 2010. “Heat Shock Factor 1 Ameliorates Proteotoxicity in Cooperation with the Transcription Factor NFAT.” *The EMBO Journal* 29 (20): 3459–69.

Heinze, Georg, and Michael Schemper. 2002. “A Solution to the Problem of Separation in Logistic Regression.” *Statistics in Medicine* 21 (16): 2409–19.

Hiddingh, Lotte, Rajiv S. Raktoe, Judith Jeuken, Esther Hulleman, David P. Noske, Gertjan J. L. Kaspers, W. Peter Vandertop, Pieter Wesseling, and Thomas Wurdinger. 2014. “Identification of Temozolomide Resistance Factors in Glioblastoma via Integrative miRNA/mRNA Regulatory Network Analysis.” *Scientific Reports* 4 (June): 5260.

Hodges, Angela, Andrew D. Strand, Aaron K. Aragaki, Alexandre Kuhn, Thierry Sengstag, Gareth Hughes, Lyn A. Elliston, et al. 2006. “Regional and Cellular Gene Expression Changes in Human Huntington’s Disease Brain.” *Human Molecular Genetics* 15 (6): 965–77.

Holbert, Sébastien, Alpaslan Dedeoglu, Sandrine Humbert, Frédéric Saudou, Robert J. Ferrante, and Christian Néri. 2003. “Cdc42-Interacting Protein 4 Binds to Huntingtin: Neuropathologic and Biological Evidence for a Role in Huntington’s Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (5): 2712–17.

Hoss, Andrew G., Vinay K. Kartha, Xianjun Dong, Jeanne C. Latourelle, Alexandra Dumitriu, Tiffany C. Hadzi, Marcy E. MacDonald, et al. 2014. “MicroRNAs Located in the Hox Gene Clusters Are Implicated in Huntington’s Disease Pathogenesis.” *PLoS Genetics* 10 (2): e1004188.

Hoss, Andrew G., Adam Labadorf, Thomas G. Beach, Jeanne C. Latourelle, and Richard H. Myers. 2016. “microRNA Profiles in Parkinson’s Disease Prefrontal Cortex.” *Frontiers in Aging Neuroscience* 8 (March): 36.

Hoss, Andrew G., Adam Labadorf, Jeanne C. Latourelle, Vinay K. Kartha, Tiffany C. Hadzi, James F. Gusella, Marcy E. MacDonald, et al. 2015. “miR-10b-5p Expression in Huntington’s Disease Brain Relates to Age of Onset and the Extent of Striatal Involvement.” *BMC Medical Genomics* 8 (March): 10.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009-1. “Bioinformatics Enrichment Tools: Paths toward the Comprehensive

Functional Analysis of Large Gene Lists." *Nucleic Acids Research* 37 (1): 1–13.

———. 2008. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." *Nature Protocols* 4 (1): 44–57.

Hunter, J. D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science Engineering* 9 (3): 90–95.

Huntington G. 1872. "On Chorea." *Medical and Surgical Reporter* 26 (15): 317–21.

Jackson, Brian C., Christopher Carpenter, Daniel W. Nebert, and Vasilis Vasiliou. 2010. "Update of Human and Mouse Forkhead Box (FOX) Gene Families." *Human Genomics* 4 (5): 345.

Jenner, Peter. 2003. "Oxidative Stress in Parkinson's Disease." *Annals of Neurology* 53 Suppl 3: S26–36; discussion S36–38.

Jones, Steve E., and Catherine Jomary. 2002. "Clusterin." *International Journal of Biochemistry & Cell Biology* 34 (5): 427–31.

Joshi NA, Fass NJ. 2011. *Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files* (version 1.33). https://github.com/najoshi/sickle.

Joung, Je-Gun, and Zhangjun Fei. 2009. "Identification of microRNA Regulatory Modules in Arabidopsis via a Probabilistic Graphical Model." *Bioinformatics* 25 (3): 387–93.

Joung, Je-Gun, Kyu-Baek Hwang, Jin-Wu Nam, Soo-Jin Kim, and Byoung-Tak Zhang. 2007. “Discovery of microRNA–mRNA Modules via Population-Based Probabilistic Learning.” *Bioinformatics* 23 (9): 1141–47.

Kaltschmidt, C., B. Kaltschmidt, H. Neumann, H. Wekerle, and P. A. Baeuerle. 1994-6. “Constitutive NF-Kappa B Activity in Neurons.” *Molecular and Cellular Biology* 14 (6): 3981–92.

Kanehisa, Minoru, and Susumu Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1): 27–30.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. 2013. “TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions.” *Genome Biology* 14 (4): R36.

Kim, Y. J., Y. Yi, E. Sapp, Y. Wang, B. Cuiffo, K. B. Kegel, Z. H. Qin, N. Aronin, and M. DiFiglia. 2001. “Caspase 3-Cleaved N-Terminal Fragments of Wild-Type and Mutant Huntingtin Are Present in Normal and Huntington’s Disease Brains, Associate with Membranes, and Undergo Calpain-Dependent Proteolysis.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (22): 12784–89.

Kinsella, Rhoda J., Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, et al. 2011. “Ensembl BioMarts: A Hub for Data Retrieval across Taxonomic Space.” *Database: The Journal of Biological Databases and Curation* 2011 (July): bar030.

Kolde, Raivo, Sven Laur, Priit Adler, and Jaak Vilo. 2012. “Robust Rank Aggregation for Gene List Integration and Meta-Analysis.” *Bioinformatics* 28 (4): 573–80.

Kozomara, Ana, and Sam Griffiths-Jones. 2011. “miRBase: Integrating microRNA Annotation and Deep-Sequencing Data.” *Nucleic Acids Research* 39 (Database issue): D152–57.

———. 2014. “miRBase: Annotating High Confidence microRNAs Using Deep Sequencing Data.” *Nucleic Acids Research* 42 (Database issue): D68–73.

Kwan, Wanda, Ulrike Träger, Dimitrios Davalos, Austin Chou, Jill Bouchard, Ralph Andre, Aaron Miller, et al. 2012. “Mutant Huntingtin Impairs Immune Cell Migration in Huntington Disease.” *The Journal of Clinical Investigation* 122 (12): 4737–47.

Labadorf, Adam, Andrew G. Hoss, Valentina Lagomarsino, Jeanne C. Latourelle, Tiffany C. Hadzi, Joli Bregu, Marcy E. MacDonald, et al. 2015. “RNA Sequence Analysis of Human Huntington Disease Brain Reveals an Extensive Increase in Inflammatory and Developmental Gene Expression.” *PLoS One* 10 (12): e0143563.

Labadorf, Adam T., and Richard H. Myers. 2015. “Evidence of Extensive Alternative Splicing in Post Mortem Human Brain HTT Transcription by mRNA Sequencing.” *PLoS One* 10 (10): e0141298.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, et al. 2001. “Initial Sequencing and

Analysis of the Human Genome.” *Nature* 409 (6822): 860–921.

Langfelder, Peter, and Steve Horvath. 2008. “WGCNA: An R Package for Weighted Correlation Network Analysis.” *BMC Bioinformatics* 9 (December): 559.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3): R25.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. “Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts.” *Genome Biology* 15 (2): R29.

Leek, Jeffrey T. 2014. “Svaseq: Removing Batch Effects and Other Unwanted Noise from Sequencing Data.” *Nucleic Acids Research* 42 (21). doi:10.1093/nar/gku864.

Lesage, Suzanne, and Alexis Brice. 2012. “Role of Mendelian Genes in ‘sporadic’ Parkinson’s Disease.” *Parkinsonism & Related Disorders* 18, Supplement 1 (January): S66–70.

Le, Thuc Duy, Junpeng Zhang, Lin Liu, and Jiuyong Li. 2015. “Ensemble Methods for MiRNA Target Prediction from Expression Data.” *PLoS One* 10 (6): e0131627.

Lewis, Benjamin P., Christopher B. Burge, and David P. Bartel. 2005a. “Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are microRNA Targets.” *Cell* 120 (1): 15–20.

———. 2005b. “Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are MicroRNA Targets.” *Cell* 120 (1): 15–20.

Li, Jun, and Robert Tibshirani. 2013. “Finding Consistent Patterns: A Nonparametric Approach for Identifying Differential Expression in RNA-Seq Data.” *Statistical Methods in Medical Research* 22 (5): 519–36.

Lin, Bingqing, Li-Feng Zhang, and Xin Chen. 2014. “LFCseq: A Nonparametric Approach for Differential Expression Analysis of RNA-Seq Data.” *BMC Genomics* 15 Suppl 10 (December): S7.

Linder, S., K. Hüfner, U. Wintergerst, and M. Aepfelbacher. 2000. “Microtubule-Dependent Formation of Podosomal Adhesion Structures in Primary Human Macrophages.” *Journal of Cell Science* 113 Pt 23 (December): 4165–76.

Ling, Hui, Muller Fabbri, and George A. Calin. 2013. “MicroRNAs and Other Non-Coding RNAs as Targets for Anticancer Drug Development.” *Nature Reviews. Drug Discovery* 12 (11): 847–65.

Li, Shi-Hua, Anna L. Cheng, Hui Zhou, Suzanne Lam, Manjula Rao, He Li, and Xiao-Jiang Li. 2002. “Interaction of Huntington Disease Protein with Transcriptional Activator Sp1.” *Molecular and Cellular Biology* 22 (5): 1277–87.

Liu, Bing, Jiuyong Li, and Anna Tsykin. 2009. “Discovery of Functional miRNA–mRNA Regulatory Modules with Computational Methods.” *Journal of Biomedical Informatics* 42 (4): 685–91.

Liu, Bing, Lin Liu, Anna Tsykin, Gregory J. Goodall, Jeffrey E. Green, Min Zhu, Chang Hee Kim, and Jiuyong Li. 2010. “Identifying Functional miRNA–mRNA Regulatory Modules with Correspondence Latent Dirichlet Allocation.” *Bioinformatics* 26 (24): 3105–11.

Liu, Zhaowen, Junying Zhang, Xiguo Yuan, Baobao Liu, Yajun Liu, Aimin Li, Yuanyuan Zhang, Xiaohan Sun, and Shouheng Tuo. 2015. “Detecting Pan-Cancer Conserved microRNA Modules from microRNA Expression Profiles across Multiple Cancers.” *Molecular bioSystems* 11 (8): 2227–37.

Li, Yue, Cheng Liang, Ka-Chun Wong, Jiawei Luo, and Zhaolei Zhang. 2014. “Mirsynergy: Detecting Synergistic miRNA Regulatory Modules by Overlapping Neighbourhood Expansion.” *Bioinformatics* 30 (18): 2627–35.

López-Gómez, Carlos, Oscar Fernández, Juan Antonio García-León, María Jesús Pinto-Medel, Begoña Oliver-Martos, Jesús Ortega-Pinazo, Margarita Suardíaz, et al. 2011. “TRAIL/TRAIL Receptor System and Susceptibility to Multiple Sclerosis.” *PLoS One* 6 (7): e21766.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *bioRxiv*, May, 002832.

Lumpkins, Kimberly M., Grant V. Bochicchio, Kaspar Keledjian, J. Marc Simard, Maureen McCunn, and Thomas Scalea. 2008. “Glial Fibrillary Acidic Protein Is Highly Correlated With Brain Injury.” *Journal of Trauma-Injury Infection* 65 (4): 778–84.

Lu, Yiming, Yang Zhou, Wubin Qu, Minghua Deng, and Chenggang Zhang. 2011. “A Lasso Regression Model for the Construction of microRNA-Target Regulatory Networks.” *Bioinformatics* 27 (17): 2406–13.

MacDonald, Marcy E., Christine M. Ambrose, Mabel P. Duyao, Richard H. Myers, Carol Lin, Lakshmi Srinidhi, Glenn Barnes, et al. 1993. “A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington’s Disease Chromosomes.” *Cell* 72 (6): 971–83.

Malagelada, Cristina, Zong Hao Jin, Vernice Jackson-Lewis, Serge Przedborski, and Lloyd A. Greene. 2010. “Rapamycin Protects against Neuron Death in in Vitro and in Vivo Models of Parkinson’s Disease.” *Journal of Neuroscience* 30 (3): 1166–75.

Malagelada, Cristina, Miguel Angel López-Toledano, Ryan T. Willett, Zong Hao Jin, Michael L. Shelanski, and Lloyd A. Greene. 2011. “RTP801/REDD1 Regulates the Timing of Cortical Neurogenesis and Neuron Migration.” *Journal of Neuroscience* 31 (9): 3186–96.

Mantamadiotis, Theo, Thomas Lemberger, Susanne C. Bleckmann, Heidrun Kern, Oliver Kretz, Ana Martin Villalba, François Tronche, et al. 2002. “Disruption of CREB Function in Brain Leads to Neurodegeneration.” *Nature Genetics* 31 (1): 47–54.

Marcora, Edoardo, and Mary B. Kennedy. 2010. “The Huntington’s Disease Mutation Impairs Huntingtin's Role in the Transport of NF-?B from the Synapse to the Nucleus.” *Human Molecular Genetics* 19 (22): 4373–84.

Martín-Flores, Núria, Joan Romaní-Aumedes, Laura Rué, Mercè Canal, Phil Sanders, Marco Straccia, Nicholas D. Allen, et al. 2015. “RTP801 Is Involved in Mutant Huntingtin-Induced Cell Death.” *Molecular Neurobiology*, April. doi:10.1007/s12035-015-9166-6.

McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” In *Proceedings of the 9th Python in Science Conference*, 51–56.

McNaught, K. S., and P. Jenner. 2001. “Proteasomal Function Is Impaired in Substantia Nigra in Parkinson’s Disease.” *Neuroscience Letters* 297 (3): 191–94.

Meffert, Mollie K., Jolene M. Chang, Brian J. Wiltgen, Michael S. Fanselow, and David Baltimore. 2003. “NF-κB Functions in Synaptic Signaling and Behavior.” *Nature Neuroscience* 6 (10): 1072–78.

Milacic, Marija, Robin Haw, Karen Rothfels, Guanming Wu, David Croft, Henning Hermjakob, Peter D’Eustachio, and Lincoln Stein. 2012. “Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome.” *Cancers* 4 (4): 1180–1211.

Mizukoshi, Eishiro, Kazumi Fushimi, Kuniaki Arai, Tatsuya Yamashita, Masao Honda, and Shuichi Kaneko. 2012. “Expression of Chondroitin-Glucuronate C5-Epimerase and Cellular Immune Responses in Patients with Hepatocellular Carcinoma.” *Liver International* 32 (10): 1516–26.

Moreira, Paula I., Cristina Carvalho, Xiongwei Zhu, Mark A. Smith, and George Perry. 2010. “Mitochondrial Dysfunction Is a Trigger of Alzheimer’s Disease

Pathophysiology.” *Biochimica et Biophysica Acta* 1802 (1): 2–10.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28.

Morton, A. Jennifer. 2013. “Circadian and Sleep Disorder in Huntington’s Disease.” *Experimental Neurology* 243 (May): 34–44.

Müller, Sarina K., Andreas Bender, Christoph Laub, Tobias Högen, Falk Schlaudraff, Birgit Liss, Thomas Klopstock, and Matthias Elstner. 2013. “Lewy Body Pathology Is Associated with Mitochondrial DNA Damage in Parkinson’s Disease.” *Neurobiology of Aging* 34 (9): 2231–33.

Myers, R. H., K. Marans, and M. E. MacDonald. 1998. “Huntington’s Disease.” In *Genetic Instabilities and Hereditary Neurological Diseases*, 301–23. Academic Press.

Myers, Richard H. 2004-4. “Huntington’s Disease Genetics.” *NeuroRx* 1 (2): 255–62.

Nalls, Mike A., Nathan Pankratz, Christina M. Lill, Chuong B. Do, Dena G. Hernandez, Mohamad Saad, Anita L. DeStefano, et al. 2014. “Large-Scale Meta-Analysis of Genome-Wide Association Data Identifies Six New Risk Loci for Parkinson’s Disease.” *Nature Genetics* 46 (9): 989–93.

Neef, Daniel W., Michelle L. Turski, and Dennis J. Thiele. 2010. “Modulation of Heat Shock Transcription Factor 1 as a Therapeutic Target for Small Molecule Intervention in Neurodegenerative Disease.” *PLoS Biology* 8 (1):

e1000291.

Obrietan, Karl, and Kari R. Hoyt. 2004. “CRE-Mediated Transcription Is Increased in Huntington’s Disease Transgenic Mice.” *Journal of Neuroscience* 24 (4): 791–96.

O’Neill, L. A. J., and C. Kaltschmidt. 1997. “NF-kB: A Crucial Transcription Factor for Glial and Neuronal Cell Function.” *Trends in Neurosciences* 20 (6): 252–58.

Ota, Kristie T., Rong-Jian Liu, Bhavya Voleti, Jaime G. Maldonado-Aviles, Vanja Duric, Masaaki Iwata, Sophie Dutheil, et al. 2014. “REDD1 Is Essential for Stress-Induced Synaptic Loss and Depressive Behavior.” *Nature Medicine* 20 (5): 531–35.

Park, Sang Gyu, Paul Schimmel, and Sunghoon Kim. 2008. “Aminoacyl tRNA Synthetases and Their Connections to Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (32): 11043–49.

Peng, Xinxia, Yu Li, Kathie-Anne Walters, Elizabeth R. Rosenzweig, Sharon L. Lederer, Lauri D. Aicher, Sean Proll, and Michael G. Katze. 2009. “Computational Identification of Hepatitis C Virus Associated microRNA-mRNA Regulatory Modules in Human Livers.” *BMC Genomics* 10 (1): 373.

Pérez, Fernando, and Brian E. Granger. 2007. “IPython: A System for Interactive Scientific Computing.” *Computing in Science & Engineering* 9 (3): 21–29.

Pizzorusso, T., G. M. Ratto, E. Putignano, and L. Maffei. 2000. “Brain-Derived Neurotrophic Factor Causes cAMP Response Element-Binding Protein Phosphorylation in Absence of Calcium Increases in Slices and Cultured Neurons from Rat Visual Cortex.” *Journal of Neuroscience* 20 (8): 2809–16.

Polito, Letizia, Antonio Greco, and Davide Seripa. 2016. “Genetic Profile, Environmental Exposure, and Their Interaction in Parkinson’s Disease.” *Parkinson’s Disease* 2016 (January). doi:10.1155/2016/6465793.

Ponomarev, Eugene D., Leah P. Shriver, and Bonnie N. Dittel. 2006. “CD40 Expression by Microglial Cells Is Required for Their Completion of a Two-Step Activation Process during Central Nervous System Autoimmune Inflammation.” *Journal of Immunology* 176 (3): 1402–10.

Qi, Lin, and Xing-Ding Zhang. 2013. “Role of Chaperone-Mediated Autophagy in Degrading Huntington’s Disease-Associated Huntingtin Protein.” *Acta Biochimica et Biophysica Sinica*, December. doi:10.1093/abbs/gmt133.

Razzell, William, Iwan Robert Evans, Paul Martin, and Will Wood. 2013. “Calcium Flashes Orchestrate the Wound Inflammatory Response through DUOX Activation and Hydrogen Peroxide Release.” *Current Biology* 23 (5): 424–29.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rehmsmeier, Marc, Peter Steffen, Matthias Hochsmann, and Robert Giegerich. 2004. “Fast and Effective Prediction of microRNA/target Duplexes.” *RNA* 10

(10): 1507–17.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40.

Romaní-Aumedes, J., M. Canal, N. Martín-Flores, X. Sun, V. Pérez-Fernández, S. Wewering, R. Fernández-Santiago, et al. 2014. “Parkin Loss of Function Contributes to RTP801 Elevation and Neurodegeneration in Parkinson’s Disease.” *Cell Death & Disease* 5 (August): e1364.

Rubinsztein, David C., and Jenny Carmichael. 2003. “Huntington’s Disease: Molecular Basis of Neurodegeneration.” *Expert Reviews in Molecular Medicine* 5 (20): 1–21.

Russo, Cinzia V., Elena Salvatore, Francesco Saccà, Tecla Tucci, Carlo Rinaldi, Pierpaolo Sorrentino, Marco Massarelli, et al. 2013. “Insulin Sensitivity and Early-Phase Insulin Secretion in Normoglycemic Huntington’s Disease Patients.” *Journal of Huntington’s Disease* 2 (4): 501–7.

Sanders, Laurie H., Jennifer McCoy, Xiaoping Hu, Pier G. Mastroberardino, Bryan C. Dickinson, Christopher J. Chang, Charleen T. Chu, Bennett Van Houten, and J. T. Greenamyre. 2014. “Mitochondrial DNA Damage: Molecular Marker of Vulnerable Nigral Neurons in Parkinson’s Disease.” *Neurobiology of Disease* 70 (October): 214–23.

Scaramuzzino, Chiara, Ian Casci, Sara Parodi, Patricia M. J. Lievens, Maria J. Polanco, Carmelo Milioto, Mathilde Chivet, et al. 2015. “Protein Arginine

Methyltransferase 6 Enhances Polyglutamine-Expanded Androgen Receptor Function and Toxicity in Spinal and Bulbar Muscular Atrophy." *Neuron* 85 (1): 88–100.

Schaefer, Carl F., Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. 2009-1. "PID: The Pathway Interaction Database." *Nucleic Acids Research* 37 (Database issue): D674–79.

Schapira, Anthony H. V. 2008. "Mitochondria in the Aetiology and Pathogenesis of Parkinson's Disease." *Lancet Neurology* 7 (1): 97–109.

Schapira, Anthony H. V., C. Warren Olanow, J. Timothy Greenamyre, and Erwan Bezard. 08/2014. "Slowing of Neurodegeneration in Parkinson's Disease and Huntington's Disease: Future Therapeutic Perspectives." *The Lancet* 384 (9942): 545–55.

Schultz, Wolfram. 2007. "Multiple Dopamine Functions at Different Time Courses." *Annual Review of Neuroscience* 30: 259–88.

Selemon, Lynn D., Grazyna Rajkowska, and Patricia S. Goldman-Rakic. 2004. "Evidence for Progression in Frontal Cortical Pathology in Late-Stage Huntington's Disease." *Journal of Comparative Neurology* 468 (2): 190–204.

Seok, Junhee, H. Shaw Warren, Alex G. Cuenca, Michael N. Mindrinos, Henry V. Baker, Weihong Xu, Daniel R. Richards, et al. 2013. "Genomic Responses in Mouse Models Poorly Mimic Human Inflammatory Diseases." *Proceedings of the National Academy of Sciences of the United States of America* 110

(9): 3507–12.

Setty, Manu, Karim Helmy, Aly A. Khan, Joachim Silber, Aaron Arvey, Frank Neezen, Phaedra Agius, Jason T. Huse, Eric C. Holland, and Christina S. Leslie. 2012. “Inferring Transcriptional and microRNA-Mediated Regulatory Programs in Glioblastoma.” *Molecular Systems Biology* 8 (1). doi:10.1038/msb.2012.37.

Sharma, Manu, Demetrius M. Maraganore, John P. A. Ioannidis, Olaf Riess, Jan O. Aasly, Grazia Annesi, Nadine Abahuni, et al. 2011. “Role of Sepiapterin Reductase Gene at the PARK3 Locus in Parkinson’s Disease.” *Neurobiology of Aging* 32 (11): 2108.e1–5.

Shenoy, Archana, and Robert H. Blelloch. 2014. “Regulation of microRNA Function in Somatic Stem Cell Proliferation and Differentiation.” *Nature Reviews. Molecular Cell Biology* 15 (9): 565–76.

Shi, Yang, Arul M. Chinnaiyan, and Hui Jiang. 2015. “rSeqNP: A Non-Parametric Approach for Detecting Differential Expression and Splicing from RNA-Seq Data.” *Bioinformatics* 31 (13): 2222–24.

Shulman, Joshua M., Philip L. De Jager, and Mel B. Feany. 2011. “Parkinson’s Disease: Genetics and Pathogenesis.” *Annual Review of Pathology: Mechanisms of Disease* 6 (1): 193–222.

Silvestroni, Aurelio, Richard L. M. Faull, Andrew D. Strand, and Thomas A. Moller. 2009. “Distinct Neuroinflammatory Profile in Post-Mortem Human Huntington’s Disease. [Miscellaneous Article].” *Neuroreport* 20 (12): 1098–

1103.

Soneson, Charlotte, and Mauro Delorenzi. 2013. “A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data.” *BMC Bioinformatics* 14 (1): 91.

Sotrel, A., P. A. Paskevich, D. K. Kiely, E. D. Bird, R. S. Williams, and R. H. Myers. 1991. “Morphometric Analysis of the Prefrontal Cortex in Huntington’s Disease.” *Neurology* 41 (7): 1117–1117.

Sotrel, A., R. S. Williams, W. E. Kaufmann, and R. H. Myers. 1993. “Evidence for Neuronal Degeneration and Dendritic Plasticity in Cortical Pyramidal Neurons of Huntington’s Disease: A Quantitative Golgi Study.” *Neurology* 43 (10): 2088–96.

Squitieri, Ferdinando, Milena Cannella, Maria Simonelli, Jenny Sassone, Tiziana Martino, Eugenio Venditti, Andrea Ciammola, Claudio Colonnese, Luigi Frati, and Andrea Ciarmiello. 2009. “Distinct Brain Volume Changes Correlating with Clinical Stage, Disease Progression Rate, Mutation Size, and Age at Onset Prediction as Early Biomarkers of Brain Atrophy in Huntington’s Disease.” *CNS Neuroscience & Therapeutics* 15 (1): 1–11.

Stachtea, Xanthi N., Emil Tykesson, Toin H. van Kuppevelt, Ricardo Feinstein, Anders Malmström, Rogier M. Reijmers, and Marco Maccarana. 2015. “Dermatan Sulfate-Free Mice Display Embryological Defects and Are Neonatal Lethal Despite Normal Lymphoid and Non-Lymphoid Organogenesis.” *PLoS One* 10 (10): e0140279.

Suárez, I., G. Bodega, and B. Fernández. 2002. “Glutamine Synthetase in Brain: Effect of Ammonia.” *Neurochemistry International* 41 (2–3): 123–42.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.

Thrane, Alexander S., Phillip M. Rappold, Takumi Fujita, Arnulfo Torres, Lane K. Bekar, Takahiro Takano, Weiguo Peng, et al. 2011. “Critical Role of Aquaporin-4 (AQP4) in Astrocytic Ca2+ Signaling Events Elicited by Cerebral Edema.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (2): 846–51.

Tibshirani, Robert. 1994. “Regression Shrinkage and Selection Via the Lasso.” In *Journal of the Royal Statistical Society, Series B*. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574.

Tobin, Jennifer E., Jing Cui, Jemma B. Wilk, Jeanne C. Latourelle, Jason M. Laramie, Ann C. McKee, Mark Guttman, Samer Karamohamed, Anita L. DeStefano, and Richard H. Myers. 2007. “Sepiapterin Reductase Expression Is Increased in Parkinson’s Disease Brain Tissue.” *Brain Research* 1139 (March): 42–47.

Träger, Ulrike, Ralph Andre, Nayana Lahiri, Anna Magnusson-Lind, Andreas Weiss, Stephan Grueninger, Chris McKinnon, et al. 2014. “HTT-Lowering

Reverses Huntington’s Disease Immune Dysfunction Caused by NFκB Pathway Dysregulation.” *Brain* 137 (3): 819–33.

Tran, Dang Hung, Kenji Satou, and Tu Bao Ho. 2008. “Finding microRNA Regulatory Modules in Human Genome Using Rule Induction.” *BMC Bioinformatics* 9 (Suppl 12): S5.

Tusher, V. G., R. Tibshirani, and G. Chu. 2001. “Significance Analysis of Microarrays Applied to the Ionizing Radiation Response.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (9): 5116–21.

Vadysirisack, Douangsone D., Franziska Baenke, Benjamin Ory, Kui Lei, and Leif W. Ellisen. 2011. “Feedback Control of p53 Translation by REDD1 and mTORC1 Limits the p53-Dependent DNA Damage Response.” *Molecular and Cellular Biology* 31 (21): 4356–65.

van Roon-Mom, Willeke M. C., Suzanne J. Reid, A. Lesley Jones, Marcy E. MacDonald, Richard L. M. Faull, and Russell G. Snell. 2002. “Insoluble TATA-Binding Protein Accumulation in Huntington’s Disease Cortex.” *Brain Research. Molecular Brain Research* 109 (1–2): 1–10.

Vonsattel, J. P., R. H. Myers, T. J. Stevens, R. J. Ferrante, E. D. Bird, and E. P. Richardson. 1985. “Neuropathological Classification of Huntington’s Disease.” *Journal of Neuropathology and Experimental Neurology* 44 (6): 559–77.

Walker, Francis O. 2007. “Huntington’s Disease.” *Lancet* 369 (9557): 218–28.

Wijemanne, Subhashie, and Joseph Jankovic. 2015. “Dopa-Responsive Dystonia–Clinical and Genetic Heterogeneity.” *Nature Reviews. Neurology* 11 (7): 414–24.

Wojda, Urszula, Elzbieta Salinska, and Jacek Kuznicki. 2008. “Calcium Ions in Neuronal Degeneration.” *IUBMB Life* 60 (9): 575–90.

Wong, Nathan, and Xiaowei Wang. 2015. “miRDB: An Online Resource for microRNA Target Prediction and Functional Annotations.” *Nucleic Acids Research* 43 (Database issue): D146–52.

Yager, L. M., A. F. Garcia, A. M. Wunsch, and S. M. Ferguson. 2015. “The Ins and Outs of the Striatum: Role in Drug Addiction.” *Neuroscience* 301 (August): 529–41.

Yoon, Sungroh, and Giovanni De Micheli. 2005. “Prediction of Regulatory Modules Comprising microRNAs and Target Genes.” *Bioinformatics* 21 (suppl 2): ii93–100.

Yu, Guangchuang, Li-Gen Wang, Guang-Rong Yan, and Qing-Yu He. 2015. “DOSE: An R/Bioconductor Package for Disease Ontology Semantic and Enrichment Analysis.” *Bioinformatics* 31 (4): 608–9.

Zhang, Junpeng, Bing Liu, Jianfeng He, Lei Ma, and Jiuyong Li. 2012. “Inferring Functional miRNA–mRNA Regulatory Modules in Epithelial–mesenchymal Transition with a Probabilistic Topic Model.” *Computers in Biology and Medicine* 42 (4): 428–37.

Zhang, Shihua, Qingjiao Li, Juan Liu, and Xianghong Jasmine Zhou. 2011. “A Novel Computational Framework for Simultaneous Integration of Multiple Types of Genomic Data to Identify microRNA-Gene Regulatory Modules.” *Bioinformatics* 27 (13): i401–9.

Zhang, Wensheng, Andrea Edwards, Wei Fan, Erik K. Flemington, and Kun Zhang. 2012. “miRNA-mRNA Correlation-Network Modules in Human Prostate Cancer and the Differences between Primary and Metastatic Tumor Subtypes.” *PLoS One* 7 (6): e40130.

Zuccato, Chiara, Manuela Marullo, Paola Conforti, Marcy E. MacDonald, Marzia Tartari, and Elena Cattaneo. 2008. “RESEARCH ARTICLE: Systematic Assessment of BDNF and Its Receptor Levels in Human Cortices Affected by Huntington’s Disease.” *Brain Pathology* 18 (2): 225–38.

CURRICULUM VITAE

Adam Thomas Labadorf

1298 Cambridge St 3L | Cambridge, MA 02139 | alabadorf@gmail.com

Education

- 2012 – 2016: Boston University, PhD Candidate, IGERT Fellow, Bioinformatics Program GPA: 4.0
- 2006 – 2010: Colorado State University, MS, Bioinformatics and Computer Science GPA: 3.87
- 1999 – 2003: Dickinson College, BS, Computer Science with Honors GPA: 3.4

Expertise Areas

- Biology: genetics and genomics, RNA biology, transcriptional regulation, epigenetics and epigenomics
- Bioinformatics: NGS analysis, differential gene expression analysis, genotype and GWA studies
- Computation: statistical modeling, high performance computing, machine learning, artificial intelligence
- Technical: *nix, python, R, web based technologies, AWS, data visualization, cluster computing, databases

Research Experience

- Research Assistant *Myers Lab, Department of Neurology, BU School of Medicine* 6/2013 – present. Analysis of **high-throughput genomic data** in post mortem human brain of **Huntington's** and **Parkinson's disease** patients, statistical methods for association of clinical covariates to molecular signatures, translation of computational results into biological hypotheses in collaboration with a diverse lab team, integration of mRNA, miRNA, and genotype data.
- Research Assistant *Colorado State University* 9/2007 – 4/2010. Conducted alternative splicing (AS) detection analysis, designed software pipelines for EST and high-throughput sequencing datasets from AS and ChIP-Seq experiments, gained close familiarity with common bioinformatics tools, designed and deployed cluster computation distribution systems specifically for bioinformatics tasks.

Professional Experience

- Neurodegenerative Disease Consultant *Immuneering Corporation, Cambridge MA,* 6/2015 – present. Providing insight into **neurodegenerative disease biology** and analytical methods pursuant of company objectives. Software development implementing published statistical methods. Data analysis of publicly available gene expression data.
- Summer Intern, *GNS Healthcare, Cambridge MA,* 6/2014–8/2014. Applied existing and original **network- based computational algorithms** to mRNA-Seq data from Huntington's Disease patients to identify causal gene expression relationships underlying HD pathogenesis.
- GSI NGS Workshop Facilitator *Genome Science Institute (GSI), BU School of Medicine,* 6/2014, 6/2015. Worked with vendor Globus Genomics to develop and deliver training on a Galaxy-based **sequencing analysis pipeline** to participants of an NGS workshop sponsored by GSI.
- Computational Technician *Fraenkel Lab, MIT,* 4/2010 – 6/2012. Worked with experimental and computational biologists to design and implement data analyses, developed custom **ChIP-Seq pipeline** and other computational tools for lab, assisted lab members with manuscript and figure preparation, built web-based applications and databases to support lab infrastructure, developed and supported lab computational resources.
- Illumina Data Analysis Specialist *BioMicroCenter, MIT,* 8/2011 – 6/2012. Designed and developed software to transform proprietary **Illumina data** to standardized formats with state of the art analysis tools. Implemented custom database integrations to organize and automate pipeline execution and communication with client labs. Worked with non-computational scientists to motivate design choices to produce data that is easily accessible and interpretable to labs without dedicated computational personnel and resources.
- Senior Information Technology Consultant *SMART, LLP, Devon, PA,* 8/2004 – 12/2007. Implemented and supported technology systems for several large legal departments, collaborated closely with clients in non-technical positions, translated business requirements into technical solutions, managed and supported database systems and client data, developed and provided training sessions for core clients and technologies.
- Web Application Developer *Universal Payment Solutions, Newtown PA (defunct),* 6/2003 – 8/2004.

Teaching and Leadership Experience

- Instructor: BF591 - Applications in Translational Bioinformatics *Boston University*. Fall 2014–Spring 2015. Conceived, designed, and delivered a graduate level course teaching hands-on analysis techniques for genomic datasets with two colleagues. Topics include **microarray**, **mRNA-Seq**, and **genotype analysis**, **gene set enrichment analysis techniques**, **biomarkers**, microbiome studies, systems biology, and pharmacogenetics
- Allele Specific Expression Challenge Project group leader *Boston University.* Fall 2014–Spring 2015. Lead a group of BU Bioinformatics first year students in the design and implementation of an analysis using mRNA-Seq and SNP data to identify **allele specific expression** in post mortem Parkinson's disease brain samples. Advised students on the implementation of a database-driven web application that presents the results of the analysis
- Graduate Teaching Assistant *Colorado State University.* Operating Systems, Fundamentals of Prog. Languages
- President, CS Department Graduate Student Association *Colorado State University.*
- Treasurer, Upsilon Pi Epsilon Colorado State University chapter, 2009–2010.
- Upsilon Pi Epsilon inductee Dickinson College chapter, 2003.

Publications

- Hoss AG, **Labadorf A**, Beach TG, Latourelle JC, Myers RH. *microRNA profiles in Parkinson's disease prefrontal cortex.* Frontiers in Aging Neuroscience 2016 (in press).
- Dumitriu A, Golji J, **Labadorf A**, Gao B, Beach TG, Myers RH, Longo KA, Latourelle JC. *Integrative analyses of proteomics and RNA transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci in Parkinson disease.* BMC Med Genomics. 2016 Jan 21;9(1):5.
- **Labadorf A**, Hoss A, Lagomarsino V, Latourelle J, Hadzi T, MacDonald M, Gusella J, Chen J, Akbarian S, Weng Z, Myers RH. *RNA sequence analysis of human Huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression.* PLoS One. 2015 Dec 4;10(12):e0143563.
- Dong X, Tsuji J, **Labadorf A**, Roussos P, Chen JF, Myers RH, Akbarian S, Weng Z. *The Role of H3K4me3 in Transcriptional Regulation Is Altered in Huntington's Disease.* PLoS One. 2015 Dec 4;10(12):e0144398
- **Labadorf A,** Myers RH. *Evidence of Extensive Alternative Splicing in Post Mortem Human Brain HTT Transcription by mRNA Sequencing.* PLoS One. 2015 Oct 23;10(10):e0141298.

- **Labadorf A**, Hoss A, Myers RH. *Neuroimmune Response and Inflammation in Huntington's disease.* Ikezu, T. Neuroimmune Pharmacology 2nd Edition (in press). Springer, US.
- Hoss A, **Labadorf A**, Latourelle J, Kartha V, Hadzi T, Gusella J, MacDonald M, Chen J, Akbarian S, Weng Z, Vonsattel J, Myers RH. *miR-10b-5p expression in Huntington's disease brain relates to age of onset and the extent of striatal involvement.* BMC Medical Genomics 2015, 8:10 doi:10.1186/s12920-015-0083-3.
- Lo KA, **Labadorf A**, Kennedy NJ, Han MS, Yap YS, Matthews B, Xin X, Sun L, Davis RJ, Lodish HF, Fraenkel E. *Analysis of In Vitro Insulin-Resistance Models and Their Physiological Relevance to In Vivo Diet-Induced Adipose Insulin Resistance.* Cell Rep. 2013 Oct 2. doi:pii: S2211-1247(13)00480-4. 10.1016/j.celrep.2013.08.039.
- Huang SS, Clarke DC, Gosline SJ, **Labadorf A**, Chouinard CR, Gordon W, Lauffenburger DA, Fraenkel E. *Linking proteomic and transcriptional data through the interactome and epigenome reveals a map of oncogene-induced signaling.* PLoS Comput Biol. 2013;9(2):e1002887. doi: 10.1371/journal.pcbi.1002887. Epub 2013 Feb 7.
- Carlson S, Chouinard C, **Labadorf A**, Lam C, Schmelzle K, Fraenkel E, White F. *Large-Scale Discovery of ERK2 Substrates Identifies ERK-Mediated Transcriptional Regulation by ETV3.* Sci. Signal. 4 (196), rs11. [DOI: 10.1126/scisignal.2002010]
- **Labadorf A**, Link A, Rogers M, Thomas J, Reddy ASN, Ben-Hur A. *Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii.* BMC Genomics 2010, 11:114 doi:10.1186/ 1471-2164-11-114