mRNA-Seq analysis of affected brain regions in manifest and asymptomatic Huntington’s Disease

## Background

Huntington’s Disease is a genetic neurodegenerative disorder that typically manifests in mid- to late life with progressive loss of motor and cognitive function, resulting in death. The disease is caused by an expanded trinucleotide CAG repeat in the first exon of the Huntingtin (*HTT*) gene and has nearly 100% penetrance. Individuals with 40 or more repeats will go on to develop HD, where longer repeats result in an earlier development of symptoms, and the most extreme mutations of 70 or more repeats cause juvenile onset. Though the presence of the mutation is sufficient to predict the disease, there is substantial, heritable variance in the age of onset of symptoms for individuals with the same repeat length. Though a genetic locus has recently been associated with modifying age of onset after adjusting for repeat length, there is presently no mechanistic insight into the underpinnings of when symptoms develop and why. Also, due to limited availability of post-mortem tissues from individuals with HD, very little is known about the disease process in asymptomatic HD brain.

The primarily affected region in the HD brain is the striatum, where medium spiny interneurons selectively degenerate. In individuals who die with late-stage HD, the striatal tissues are massively degenerated, and cellular composition is altered such that it is difficult if not impossible to study any neuron-related processes in the region. Other parts of the brain, including the afrontal cortex and cerebellum, are altered in HD but to a lesser extent than striatum by time of death. To date, the Myers lab has studied exclusively afrontal cortex (Brodmann Area 9, BA9) brain tissue from post-mortem HD individuals for this reason. However, asymptomatic HD individuals do not exhibit the same degree of degeneration in the striatum, making such tissues attractive for analysis. Post-mortem brain samples from asymptomatic HD individuals are extremely rare, but the Myers lab has obtained samples from two such brains.

The Myers lab has generated mRNA-Seq data from post-mortem BA9 of 29 individuals with HD, including two asymptomatic individuals, and 49 neurologically normal controls. Nine of the HD samples have not been previously studied or published. For the two asymptomatic individuals, mRNA was sequenced from striatal tissues as well, creating the exciting opportunity to compare the molecular signature of these two brain regions within the same individuals. Identifying which transcriptional changes in the cortex correlate with those in the striatum may lead to a better understanding of the active, as opposed to reactive, disease processes, and help to better interpret the signals observed in the brains of symptomatic HD brains.

This study has three aims:

1. Perform a comprehensive mRNA-Seq differential expression analysis using all BA9 HD and control samples.
2. Identify genes or groups of genes whose abundance correlates with CAG repeat length in HD BA9.
3. Identify genes or groups of genes that correlate well between BA9 and striatum in asymptomatic individuals.

## Hypothesis

While it is known that dramatic gene expression differences exist between HD and control brains, the increase sample size and improved methodology from previous studies will help identify changes with greater resolution and fidelity. Small sample size has previously hindered attempts at finding statistically significant genes that correlate with CAG repeat length. We hypothesize that the greater sample size in this study and focusing on groups of genes will identify factors related to disease severity. We also hypothesize that some gene expression signals in BA9 correlate with those in the striatum of asymptomatic individuals.

## Study Design

### Aim 1. HD vs Control Differential Expression

Sequencing data will be processed using the Bioinformatics Hub mRNA-Seq analytical pipeline. Reads from each sample will be assess for quality control with FastQC (FastQC 2016) and aligned to the hg38 human reference genome using the STAR aligner (Dobin 2013). Multimapping reads will be resolved with the MMR program (Kahles 2016). The datasets will be quantified against the GENCODE v25 transcriptome (Harrow 2012) using salmon (Patro 2015) and transcripts from the same gene will be collapsed to the gene level using the tximport bioconductor package (Soneson 2015). Alignment, library, and quantification quality will be assessed using RSeQC (REF) and multiqc (REF) tools.

Two analytic approaches will be applied to identify consistent mRNA abundance differences between HD and control. First, differential expression (DE) between HD and control samples will be assessed using Firth's logistic regression (Choi 2016). The resulting DE gene list will be analyzed with GSEA against the MsigDB gene set database v5 (Subramanian 2005). Second, the ASSIGN algorithm (Shen 2015) will be applied to the dataset to find pathways or gene sets that are consistently perturbed between HD and control.

### Aim 2. Association with CAG repeat length

The HD salmon mRNA abundance estimates from Aim 1 will be used to identify genes that are statistically associated with CAG repeat length. First the gene-wise salmon abundances will be transformed to follow a normal distribution using the Variable Stabilizing Transform from the DESeq2 package. The resulting matrix will then be analyzed with ComBat to assess the presence of latent confounders, covariates including age at death, and sequencing batch effects, producing adjusted normalized abundance estimates. Associations between each gene abundance estimate and CAG repeat length will be performed with limma.

A second exploratory analysis may be performed using WGCNA. WGCNA detects blocks of highly correlated genes and can perform statistical association analysis for a given variable on a summarized profile (i.e. meta-gene) of expression for each block. The algorithm will be applied to identify sets of correlated genes that are associated with CAG length.

### Aim 3. Comparative analysis of BA9 and Striatum

Since only two samples of paired BA9/striatal tissue are available, Gtex RNA-Seq datasets from BA9 and striatum (i.e. nucleus accumbens and putamen) will be downloaded from dbGAP and used in combination with the BA9 control samples in this study to form a basis comparison for cortex vs striatum gene expression. The Gtex datasets will be subjected to the same sequence analysis pipeline as the samples from Aim 1. The normalized, transformed salmon mRNA abundance estimates from Aim 2 will then be used to determine genes that are differentially expressed between the cortical and striatal tissues using DESeq2. The statistically differentially expressed genes from this analysis are genes expected to be different in the asymptomatic samples. The genes that are not differentially expressed, but whose log 2 fold change estimates have narrow confidence intervals are genes that we expect to be concordant in the asymptomatic samples.

Log 2 fold changes for each gene will be computed by taking the log 2 of the ratio of the BA9 vs striatal abundance estimates within each asymptomatic individual. Using the fold change distributions for each gene calculated using the control sample set, p-values will be calculated for these fold change point estimates for each sample. Genes found to be concordant (i.e. in the same direction) and significant by this test in both samples are potential candidates for genes involved in the disease process. These genes will be cross referenced with the results from the BA9 vs control study of Aim 1 to assess agreement.

## Scope and Responsibilities

TODO

## Deliverables

The following are deliverables for this project:

* Normalized mRNA abundance estimate matrix from salmon + tximport of all samples
* Firth logistic regression DE gene list for BA9 vs control
* Enriched MsigDB genesets as determined by GSEA
* Active genesets as determined by ASSIGN
* Genes associated with CAG repeat length using VST transformed, ComBat adjusted abundance estimates
* Normalized mRNA abundance estimate matrix from salmon + tximport of BA9 and striatal mRNA-Seq Gtex samples
* Differentially expressed genes between BA9 and striatal mRNA-Seq samples from Gtex and controls
* Genes whose fold change estimates significantly differ from the base fold change estimates of BA9 vs striatal samples

## Milestones

TODO

## Estimated Effort

This project is estimated to require 400 hours of analyst time. The Hub analyst will be paid an hourly rate of $16/hr. Adam’s effort on this project is provided at no cost.

## References

Shen, Ying, Mumtahena Rahman, Stephen R. Piccolo, Daniel Gusenleitner, Nader N. El-Chaar, Luis Cheng, Stefano Monti, Andrea H. Bild, and W. Evan Johnson. 2015. “ASSIGN: Context-Specific Genomic Profiling of Multiple Heterogeneous Biological Pathways.” Bioinformatics 31 (11): 1745–53.