

12-15-2017

Understanding Huntington's disease using Machine Learning Approaches

Sonali Lokhande

Keck Graduate Institute

Recommended Citation

Lokhande, Sonali. (2017). *Understanding Huntington's disease using Machine Learning Approaches*. KGI Theses and Dissertations, 4.
https://scholarship.claremont.edu/kgi_theses/4. doi: 10.5642/kgitd/4

This Restricted to Claremont Colleges Dissertation is brought to you for free and open access by the KGI Student Scholarship at Scholarship @ Claremont. It has been accepted for inclusion in KGI Theses and Dissertations by an authorized administrator of Scholarship @ Claremont. For more information, please contact scholarship@cuc.claremont.edu.

UNDERSTANDING HUNTINGTON'S DISEASE USING
MACHINE LEARNING APPROACHES

BY

SONALI JAYANT LOKHANDE

A Dissertation submitted to the Faculty of Keck Graduate Institute of Applied
Life Sciences in partial fulfillment of the requirements for the degree of Doctor of
Philosophy in Applied Life Sciences

Claremont, California
2017

Approved by:


(Prof. Animesh Ray)

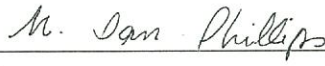
Copyright by Sonali Jayant Lokhande 2017
All rights Reserved

We, the undersigned, certify that we have read this dissertation of Sonali Jayant Lokhande and approve it as adequate in scope and quality for the degree of Doctor of Philosophy.

Dissertation Committee:



(Prof. Animesh Ray), Chair



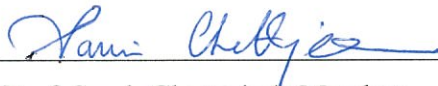
(Prof. Ian Phillips), Member



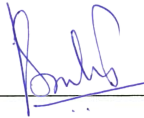
(Dr. Seongjoon Koo), Member



(Prof. Amarnath Gupta), Member



(Prof. Samir Chatterjee), Member



(Prof. Srikanth Kolluru), PhD Program Director

Abstract of the Dissertation

Understanding Huntington's Disease using Machine Learning

Approaches

By

Sonali Jayant Lokhande

Keck Graduate Institute of Applied Life Sciences: 2017

Huntington's disease (HD) is a debilitating neurodegenerative disorder with a complex pathophysiology. Despite extensive studies to study the disease, the sequence of events through which mutant Huntingtin (mHtt) protein executes its action still remains elusive. The phenotype of HD is an outcome of numerous processes initiated by the mHtt protein along with other proteins that act as either suppressors or enhancers of the effects of mHtt protein and PolyQ aggregates. Utilizing an integrative systems biology approach, I construct and analyze a Huntington's disease integrome using human orthologs of protein interactors of wild type and mHtt protein. Analysis of this integrome using unsupervised machine learning methods reveals a novel connection linking mHtt protein with chromosome condensation and DNA repair. I generate a list of candidate genes that upon validation in a yeast and drosophila model of HD are shown to affect the mHtt

phenotype and provide an *in-vivo* evidence of our hypothesis. A separate supervised machine learning approach is applied to build a classifier model that predicts protein interactors of wild type and mHtt protein. Both the machine learning models that I employ, have important applications for Huntington's disease in predicting both protein and genetic interactions of huntingtin protein and can be easily extended to other PolyQ and neurodegenerative disorders such as Alzheimer's and Parkinson's disease.

Dedication

Dedicated to

my Teachers, my Family

and

my Husband

Acknowledgements

The research presented in this thesis was funded by Keck Graduate Institute of Life Sciences, Claremont CA and by the CHDI foundation. I would like to express my sincere thanks for the facilities and funding provided during my research.

I would like to thank Gaelle-Sop-Kamga, Susan Ye, Dr. Biranchi Patra, Rishov Chatterjee, Mike De LaCruz who were collaborators for the unsupervised machine learning experiments. I am grateful to Dr. Tamas Nepusz for providing critical feedback and expert advice and to Dr. Mahidhar Tatineni for providing support at the San Diego Supercomputing Center (SDSC). A special thanks to Dr Biranchi Patra and (Late) Susan Lindquist who were collaborators on the yeast validation experiments. and to Assoc. Prof. Katerina Venderova for collaborating on the drosophila validation experiments.

In particular, I would like to thank Prof Animesh Ray, Dr. Seongjoon Koo, Prof. Amarnath Gupta, Prof. Samir Chatterjee and Prof. Ian Phillips for providing extensive guidance and support throughout the research. Special thanks to all my colleagues at the Keck Graduate Institute, all of whom made working at KGI a memorable and wonderful experience.

TABLE OF CONTENTS

1	Huntington's disease and its links to chromatin condensation mechanisms.....	1
1.1	Introduction:	1
1.2	The Genetic Basis	2
1.3	Clinical Presentation	4
1.4	Cell/Animal Models for HD.....	6
1.5	Classical Views of Molecular Pathogenesis in HD	9
1.6	Mitochondrial dysfunction and oxidative stress in HD	11
1.7	A possible mechanism behind the gradual onset of mHtt toxicity	12
1.8	Transcriptional regulation, Chromatin dynamics and the role of SIRT1 in HD. 16	
1.9	DNA damage and chromatin condensation defects	20
1.10	rDNA condensation defects in HD	23
1.11	Conclusion	26
1.12	Tables and Figures	27
2	Constructing the Huntington's disease Integrome	30
2.1	Introduction	30
2.2	Method	39
2.2.1	Interaction data sets.....	39
2.2.2	Ortholog detection and construction of orthologous HD interactome	40
2.2.3	Network construction and analysis	40
2.2.4	Gene Ontology (GO) enrichment.....	41
2.2.5	Yeast strains, media and plasmids	42
2.3	Results.....	43
2.3.1	Construction of HD protein interactome	43
2.3.2	Network properties of Huntington's disease integrome	44
2.4	Tables and Figures.....	46
3	Analysis of Huntington's disease integrome and candidate gene validation using unsupervised machine learning.....	49
3.1	Pre-requisites to choose a clustering algorithm for the HDI.	50

3.2	ClusterONE Algorithm	51
3.2.1	Cohesiveness of nodes:	53
3.2.2	Evaluating the quality of detected complexes:	55
3.2.3	Results using ClusterONE	56
3.2.4	Scalability of ClusterONE	59
3.2.5	Robustness of ClusterONE	60
3.2.6	Reproducibility of results	62
3.2.7	Jerarca algorithm.....	63
3.2.8	Results using Jerarca.....	66
3.3	HD integrome reveals novel genes that link chromosome condensation defects with Htt toxicity	67
3.4	Validation in a yeast HD model.	67
3.5	Identification of candidate genes	69
3.6	Candidate gene validation in a Drosophila HD model	70
3.7	Conclusion	72
3.8	Tables and Figures.....	73
4	Predicting physical interactors of the Huntingtin protein using Supervised Machine Learning methods.	91
4.1	Data.....	92
4.1.1	Model development dataset:	92
4.1.2	Input features	93
4.1.3	Dataset formatting.....	95
4.1.4	Classification target	96
4.2	Methods	97
4.2.1	Variable and dimension reduction methods	97
4.2.2	Logistic regression with regularization	98
4.2.3	Random forest.....	100
4.2.4	Gradient Boosting Machine (GBM).....	101
4.3	Results.....	102
4.3.1	Variable and dimension reduction	102
4.3.2	Logistic regression with regularization	103

4.3.3	Lasso with data segmentation	105
4.3.4	Random forest.....	106
4.3.5	Gradient Boosting Machine (GBM).....	107
4.3.6	GBM with data segmentation	108
4.3.7	Important Predictor variables	110
4.4	Conclusion	113
4.5	Tables and Figures.....	115
5	Summary.....	131
6	Appendices	134
6.1	Network Properties of HDI	134
6.1.1	Network heterogeneity	134
6.1.2	Average number of neighbors	135
6.1.3	Network density.....	135
6.1.4	Network diameter	135
6.1.5	Clustering coefficient.....	136
6.1.6	Average shortest path length	136
6.2	Rscripts	137
6.2.1	IV calculation and Lasso Regression – Experiment 2.....	137
6.2.2	Random forest – Experiment 2	143
6.2.3	Gradient Boosting Machine – Experiment 2	144
6.2.4	Gradient Boosting Machine – with Data segmentation.....	146
7	Bibliography	149

Abbreviations

AC	Affinity Coefficient
ANAPC7	Anaphase Promoting Complex Subunit 7
AP-MS	Affinity Purification Mass Spectrometry
AUC	Area Under Curve
BDNF	Brain-Derived Neurotrophic Factor
<i>BNA5</i>	Kynureninase
<i>C.elegans</i>	Caenorhabditis Elegans
COPS2	Cop9 Signalosome Subunit 2
CPU	Core Performance Unit
CREB	Camp Responsive Element Binding Protein
CUL2	Cullin 2
DDX4	Dead-Box Helicase 4
DRPLA	Dentatorubral-Pallidoluysian Atrophy
ERAD	Endoplasmic-Reticulum Associated Degradation
ERCC3	Ercc Excision Repair 3, Tfiif Core Complex Helicase Subunit
ESET	Erg-Associated Protein With Set Domain
ESPL1	Extra Spindle Pole Bodies Like 1, Separase
FANCI	Fanconi Anemia Complementation Group I
FDR	False Discovery Rate
fl-Htt	Full Length Htt
fMRI	Functional Magnetic Resonance Imaging
FOXO3a	Forkhead Box O3A
GBM	Gradient Boosting Machine
GG-NER	Global Genome Nucleotide-Excision Repair
GO	Gene Ontology
HD	Huntington's Disease
HDI	Huntington's Disease Integrome
HDL1	Huntington's Disease -Like 1
HDL2	Huntington's Disease -Like 2

HEK293T	Human Embryonic Kidney
HIPPIE	Human Integrated Protein-Protein Interaction Reference
HIST1H1	Histone Cluster 1 H1
HNRNPQ	Synaptotagmin Binding Cytoplasmic Rna Interacting Protein
HSF1	Heat Shock Transcription Factor 1
HSP90	Heat Shock Protein 90
<i>HTT</i>	Huntingtin
INNs	Intraneuronal Nuclear Inclusions
IV	Information Value
LMNB1	Lamin B1
Mdm2	Mdm2 Proto-Oncogene
mHtt	Mutant Huntingtin
MORF	Movable Open Reading Frame
MPT	Mitochondrial Permeability Transition
MRN	Mre11-Rad50-Nbs1
MS	Mass Spectrometry
MSH4	Muts Homolog 4
mtTFA	Mitochondrial Transcription Factor A
NADH	Nicotinamide Adenine Dinucleotide
NBS1	Nijmegen Breakage Syndrome 1 (Nibrin)
NDC80	Ndc80, Kinetochore Complex Component
NLRC4	Nlr Family Card Domain Containing 4
NMD3	Nmd3 Ribosome Export Adaptor
NPLOC4	Npl4 Homolog, Ubiquitin Recognition Factor
NRF	Nuclear Respiratory Factor
NuMA	Nuclear Mitotic Apparatus
p53	Tumor Protein P53
PC	Principal Component
PCA	Principal Component Analysis
PGC-1α	Ppar γ Coactivator-1 α
PPI	Protein-Protein Interactions
PPP6C	Protein Phosphatase 6 Catalytic Subunit

PSMC	Proteasome 26S Subunit, Atpase
RAD23A	Rad23 Homolog A, Nucleotide Excision Repair Protein
RAD23B	Rad23 Homolog B, Nucleotide Excision Repair Protein
rDNA	Ribosomal DNA
ROS	Reactive Oxygen Species
SHFM1	Sem1, 26s Proteasome Complex Subunit
SIRT	Sirtuin 1
SMAX1/SBMA	Spinal And Bulbar Muscular Atrophy, X-Linked 1
SMC	Structural Maintenance Of Chromosomes
SPC24	Spc24, Ndc80 Kinetochore Complex Component
TAP	Tandem Affinity Purification
TRIP12	Thyroid Hormone Receptor Interactor 12
TUBGCP2	Tubulin Gamma Complex Associated Protein 2
UBC	Ubiquitin C
UBTF	Upstream Binding Transcription Factor, Rna Polymerase I
UCHL5	Ubiquitin C-Terminal Hydrolase L5
UFD1L	Ubiquitin Recognition Factor In Er Associated Degradation 1
UME1	Transcriptional Regulatory Protein
UPGMA	Unweighted Pair-Group Method With Arithmetic Mean
UPR	Unfolded Protein Response
URA3	Orotidine 5'-Phosphate Decarboxylase
VCP	Valosin Containing Protein
WAPAL	Wapl Cohesin Release Factor
Y2H	Yeast Two-Hybrid
YWHAZ	Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Zeta

List of Tables

Table 2.1 Network properties of Huntington's disease integrome (HDI)	46
Table 3.1 Members of the first 10 statistically significant complexes detected in HDI by ClusterONE	73
Table 3.2 List of GO terms related to DNA repair, DNA damage and chromosome condensation found to be enriched in PPI network complexes	76
Table 3.3 List of pathways found to be enriched in HDI network complexes	79
Table 3.4 - Scalability analysis of ClusterONE algorithm	81
Table 3.5 Robustness analysis of ClusterONE	82
Table 3.6 - List of 27 candidate genes with a viable and visible phenotype chosen for validation in a Drosophila model of HD	83
Table 4.1 - Characteristics of data fed to the classifiers	115
Table 4.2 – Area under curve for regularized logistic regression using Lasso, Ridge and Elastic-net models from 10-fold cross-validation experiments.	116
Table 4.3 Area under curve for data segments of Experiment 2 obtained using the Lasso regression model.	117
Table 4.4 - Area under curve for random forest model from 10-fold cross-validation experiments.	118
Table 4.5 - Area under curve for GBM model from 10-fold cross-validation experiments.	119
Table 4.6 – Experiment 2 – Parameter tuning for GBM	119

Table 4.7 - Area under curve for GBM with data segmentation using 10-fold cross-validation.....	120
Table 4.8 – Overall importance of top 5 variables in predicting proteins interacting with mutant Htt protein using GBM model.	120
Table 4.9 Genes and their encoded proteins containing motifs of importance for GBM Segment Experiment 1	121
Table 4.10 – Overall importance of top 10 variables in predicting proteins interacting with mutant Htt protein using Lasso Regression model.	123
Table 4.11 Genes and their encoded proteins containing motifs and domains of importance found using Lasso regression for Experiment 2.....	124

List of Figures

Figure 1.1 Clinical progression of Huntington's disease as a function of accumulation of mHtt protein.	27
Figure 1.2 Models of mechanisms of action of mHtt.	28
Figure 1.3 Chromosome condensation defects in Huntington's disease.	29
Figure 2.1 Workflow to generate and identify candidate genes from the Huntington's disease interactome (HDI).....	48
Figure 2.2 Degree distribution of the Huntington's disease interactome (HDI). ..	48
Figure 3.1 An example graph depicting internal and external vertices and edges along with the boundary vertex.....	84
Figure 3.2 - Protein complexes within HDI that are enriched for DNA repair and chromosome condensation related functions.	85
Figure 3.3 Robustness of ClusterONE to edge perturbation.....	86
Figure 3.4 An adapted workflow layout of the Jerarca	87
Figure 3.5 Cladogram of HDI hierarchy.....	88
Figure 3.6 Suppressors of smc2-8 and condensin-cohesin genes suppress the mHtt toxicity (103Q) in yeast.	89
Figure 3.7 Hypothesis model of suppressors of smc2-8 mutation suppressing mHtt toxicity	90
Figure 4.1- Information Value of Motif and domain variables.....	126

Figure 4.2 - Full scree plot of variance explained by the top 150 principal components of motif and domain variables.	127
Figure 4.3- Receiver Operating Curves (ROC) of data segments for Experiment 2 using Lasso regression.....	128
Figure 4.4 - Variable importance of proteins interacting with mutant Htt as shown by the Random Forest.	129
Figure 4.5 – Receiver Operating Curves (ROC) for data segments of the master dataset using GBM model.....	130

Chapter 1

1 Huntington's disease and its links to chromatin condensation mechanisms

1.1 Introduction:

Huntington disease (HD) is an autosomal dominant neurodegenerative disorder caused by the expansion of CAG triplet repeats in the first exon of the Huntingtin (*HTT*) gene. Despite this apparently simple genetic basis, identified nearly three decades ago, the molecular pathophysiology of this inherited disorder that affects a significant number of people in the prime of their youth remains intractable, and with limited treatment options beyond managing the symptoms. This review summarizes the genetic basis of HD, lists the genetic models currently in use to study its molecular pathology, and critically examines the involvement of mitochondrial dysfunction, DNA damage repair pathways, and chromatin dynamics in HD progression. Specifically, this review summarizes recent studies that indicate a role for chromosome condensation defects, especially at the ribosomal DNA (rDNA), in molecular pathogenesis of HD. We hope to stimulate interest of HD researchers in further examining the novel associations among these apparently disparate cellular processes.

1.2 The Genetic Basis

The *HTT* gene encodes an expanded polyglutamine stretch in the *huntingtin* (Htt) protein. Despite being inherited in an autosomal dominant manner, the prevalence of this disease varies from just 5 to 6 per 100,000 people in North America, Europe and Australia with an incidence of 0.38 per 100,000 year.(Pringsheim et al., 2012) HD has been classified as a 'rare' or 'orphan disease'. The low level of occurrence of HD can be explained by reduced penetrance of mutant *HTT* alleles in individuals with 36-39 copies of CAG repeats.(Huntington Study Group COHORT Investigators and Dorsey, 2012; McNeil et al., 1997; Panegyres and Goh, 2011; Quarrell et al., 2007; Sequeiros et al., 2010) Moreover, inheriting a contracted CAG repeat (36-39 copies) from its parent could reduce the risk of the individual developing the disease.(Nahhas et al., 2009) On average, the age of disease onset is inversely correlated with the number of CAG repeats; however, the repeat number explains approximately 50-70 % of the age of onset data. The remaining variability in age of onset is attributed to various genetic and environmental factors that act in conjunction with the *HTT* gene.(Djousse et al., 2003; Project* and Wexler, 2004) Individuals with shorter CAG repeats exhibit a gradual decline in clinical progression than individuals with larger CAG repeats. Hence CAG repeat length is an important determinant of clinical progression of HD. This suggests that aging itself might influence clinical outcomes in Huntington's disease.(Rosenblatt et al., 2012) A peculiar observation is that the expanded *HTT* allele when inherited

through the male germ line, often leads to a more severe clinical course than when inherited through the female germ line. Children of HD sufferers experience HD symptoms 8 years earlier than that of their respective fathers.(Ranen et al., 1995) These observations suggest that the CAG expansion rate might be higher in the male germline(Duyao et al., 1993), the possibility of imprinting, which might affect the expression of genes involved in molecular pathogenesis(Farrer et al., 1992; Reik, 1988; Ridley et al., 1991), or a combination of these two factors.

The HD phenotype is a product of various aberrant genetic and epigenetic mechanisms triggered by the mutant *HTT* allele (which are reviewed in detail in the article by Valor Guiretti et al)(Valor and Guiretti, 2014) and also through various other proteins that are either suppressors or enhancers of the mutant Htt (mHtt) protein and the PolyQ aggregates. Quantitative interaction proteomic studies identify proteins interacting with both wild-type and mHtt protein.(Culver et al., 2012; Hosp et al., 2015; Ratovitski et al., 2012; Shirasaki et al., 2012) Conclusions from quantitative proteomics have been validated by suppressor/enhancer studies on a model organism (fly). Potential evidences of candidate gene involvement have been verified by the presence of SNPs within candidate human genes in patients.²⁰ The study of transcriptional dysfunction in HD through GWAS has enabled the charting of various genomic loci of Htt interacting proteins such as REST, PGC-1alpha, HSF1, and Foxp1(Lucas et al., 2012; Riva et al., 2012; Strand et al., 2007; Valor, 2014) and the Htt protein itself.(Benn et al., 2008) A combination of Genome

wide correlation of histone acetylation(McFarland et al., 2012) and gene expression defects in a HD mouse models identified aberrant deacetylation of H3K9/14ac and H4K12ac(McFarland et al., 2012) along with transcriptional dysregulation of other Htt interacting proteins.(Valor et al., 2013)

1.3 Clinical Presentation

Clinical presentation of HD has been observed to occur in individuals at any time between the ages of 1 and 80. The symptomatic phase of the disease (Table 1) is preceded by a prediagnostic phase in which patients show subtle changes in cognition and motor control, which mostly go unnoticed.(Walker, 2007) The prediagnostic phase leads to the diagnostic phase in which patients begin showing distinct chorea, motor incoordination and impersistence along with slow saccadic eye movements.(Watts and Koller, 1997; Weiner and Lang, 1989) Cognitive dysfunction affects executive functions and delays new motor skills(Craufurd and Snowden, 2002), which worsens with time (Figure 1.1). Depression and suicidal behavior are common, along with symptoms of psychosis and mania.(Walker, 2007) A question remains as to whether early cognitive dysfunction can be correlated with CAG repeat length in HD patients. A recent study using Functional Magnetic Resonance Imaging (fMRI) and structural-MRI, indicated that there is a significant correlation between executive function performance levels with disease

progression. No studies have been reported to our knowledge that explores cognitive dysfunction as a correlate of CAG repeat length.

Diagnosis of HD is usually definite when typical clinical symptoms start to develop coupled with a positive family history. However, in other patients, and those with early onset of the disease, the symptoms might resemble other disorders like the dentatorubropallidoluysian atrophy (DRPLA), Huntington's disease-like 2 (HDL2) and Spinocerebellar ataxia (SCA). (Margolis et al., 2004; T et al., 1985; Toyoshima et al., 2004; Walker, 2007)

Routine CET and MRI scans help assess the severity and progress of HD but both are usually not useful for early diagnosis of the disease. PET and functional MRI scans can show atrophy of the caudate nucleus and the putamen almost 9-11 years earlier than the disease onset. (Aylward et al., 2004; Küinig et al., 2000; Lawrence et al., 1998; Paulsen et al., 2004; Rosas et al., 2003) The PREDICT-HD study in 2008, also found striatal volume to be the strongest biomarker predicting the onset of disease when related to CAG repeat length. (Paulsen et al., 2008)

Genetic testing of HD patients and their families is a definitive diagnostic method that confirms the presence of the disease. These genetic tests can be administered as a predictive, pre-natal or a diagnostic assay depending on how the patient presents clinically. However, the uptake of the test has been minimal considering

the paucity of effective treatment(Lacone et al., 1999) and comes with its own set of ethical and psychosocial challenges that need to be addressed before and after the test has been conducted.(Tassicker et al., 2006)

Almost all treatment options currently available for the disease are symptomatic, and focus on addressing motor symptoms such as chorea and dystonia. Absence of reliable diagnostic biomarkers for early stages of the disease, and a lack of curative strategies call for further research in understanding this complex neurological disorder. Cell and various animal models have therefore been developed to better explain the molecular mechanisms underlying HD. While cell models help clarify the central apparatus of the disease, animal models help in recreating the HD genotype and phenotype to screen for therapeutic compounds and targets. The next section will discuss more about the currently available cell and animal models in HD.

1.4 Cell/Animal Models for HD.

Several *in vitro* and *in vivo* genetic models developed for HD (summarized in Table 2), help in elucidating and dissecting various molecular pathways affected in this disease.

Cell lines and primary cultures such as the human embryonic kidney (HEK293T) cell line, are effective instruments in understanding the basic molecular processes contributing to neural degeneration and death.(Cisbani and Cicchetti, 2012)

Inducible cell systems on the other hand help experimentally modulate and allow assessment of the spatial and temporal activation of genes and proteins. Such systems therefore are useful for studying effects of gene and protein expression in a diseased state. *In vitro* models have been useful to characterize the cleavage mechanism of mHtt protein(Johri and Beal, 2010; Kim et al., 2001; Miller et al., 2010), and the factors influencing the process(Johri and Beal, 2010; Martindale et al., 1998), and how polyQ aggregates lead to cell death in HD.(Wytenbach et al., 2001; Zala et al., 2005)

Yeast models of HD have helped identify genes that modify mHtt toxicity and helped provide targets for validation in higher organisms. For example, a high throughput screening assay in yeast HD model (Htt-103Q) identified a small molecule inhibitor (C2-8) of polyQ aggregation which was validated in a *Drosophila* HD model to show suppression of neurodegeneration.(Zhang et al., 2005) A proof-of-concept study carried out in R6/2 mouse model of HD showed that mice treated with C2-8 improved motor performance and reduced neuronal atrophy with smaller huntingtin aggregates.(Chopra et al., 2007) A follow-up preclinical study of C2-8 in R6/2 also found evidence supporting its role in reducing the size of mHtt aggregates but did not find a significant role in improving behavioral deficits in this mouse model of HD.(Wang et al., 2013)

In vivo models of HD include organisms such as *Caenorhabditis elegans* (*C. elegans*), *Drosophila melanogaster*, mice and rats. *C. elegans* models are mostly generated by expression of N-terminal fragments of mHtt, ranging from 57 to 171 amino acids. Transgenic expression of mHtt in *C. elegans* can result in age-dependent mechanosensory defects, neuronal dysfunction and neurodegeneration.(Faber et al., 1999; Parker et al., 2001)

Drosophila models of HD mostly use an inducible gene expression system, UAS-GAL4, to express full-length or N-terminal Htt fragments ranging from 65 to 548 amino acids of the expanded repeat mutant genes. Such models recapitulate a progressive neurodegeneration phenotype of HD along with motor dysfunction and reduced survival.(Marsh et al., 2003; Robinow and White, 1988)

Transgenic mouse models for HD express N-terminal fragments of human *HTT* of various sizes. R6/1 and R6/2 mice express exon 1 of human *HTT* with 116 and 144 CAG repeats respectively and show somatic instability of the CAG repeat tract.(Mangiarini et al., 1996) Truncated N-terminal mouse models exhibit an accelerated degenerative phenotype including motor, cognitive and behavioral aberrations along with increased mortality.(Schilling et al., 1999)

Knock-in mouse models are created by inserting an extended CAG tract with CAG repeat sizes ranging from 50 to 200 into an endogenous mouse *HTT* gene.(Dougherty et al., 2013; Heng et al., 2007; Lin et al., 2001; Menalled et al.,

2002, 2003; Wheeler et al., 2002; White et al., 1997) Out of the many such knock-in mouse models, CAG140 (119/140 CAG repeats), the HdhQ111 (111 CAG repeats), and HdhQ150 (150 CAG repeats) mice are the most genetically appropriate to HD in terms of expression of mHtt (Pouladi et al., 2013) and their ability to generate neurological and neurodegenerative symptoms. (Hickey et al., 2008; Lerner et al., 2012; Menalled et al., 2003; Wheeler et al., 2002)

Some transgenic models such as the BACHD (97 CAG repeats) and YAC128 (128 CAG repeats) mice express the entire mutant *HTT* gene and show a comparatively milder and more progressive phenotype with cognitive disturbances along with striatal and cortical atrophy. (Ehrnhoefer et al., 2009; Gray et al., 2008; Raamsdonk, 2005)

1.5 Classical Views of Molecular Pathogenesis in HD

The mutant *HTT* gene confers a toxic gain of function and leads to protein misfolding and aggregation. (Duyao et al., 1995; Mangiarini et al., 1996; Nasir et al., 1995; Zeitlin et al., 1995) The toxic gain of function by the mutant *HTT* gene may be due to the short N-terminal fragments of mHtt protein produced through cleavage by various proteases. (Bizat et al., 2003; Cowan and Raymond, 2006; Gafni and Ellerby, 2002; Gafni et al., 2004; Goffredo et al., 2002; Graham et al., 2006; Kim et al., 2001; Lee and Kim, 2006; Wellington et al., 1998) Reduction in caspase 6 –dependent cleavage of full length mHtt significantly slows the

progression of neurodegeneration in YAC mouse model for HD.(Graham et al., 2006) Additional studies in neurons and mice models implicate that even a loss of potentially beneficial function of wild-type Htt contribute to HD disease phenotype.(Cattaneo et al., 2001; Faber et al., 1998; Hackam et al., 2000; Kalchman et al., 1997) Though, wild-type Htt is known to up-regulate of BDNF (Brain-derived neurotrophic factor) transcription gene, its function is lost in the cortex of YAC72 HD mice underlining the concept of loss of Htt-mediated function(Cattaneo et al., 2001) (Figure 1.2). Despite extensive experimental studies carried out in both gene knock-out and gene knock-in HD models, the multifaceted functions of the normal Htt protein or that of the mHtt protein remain elusive. The mHtt protein is known to physically interact with numerous other proteins, some of which do or do not interact with the wild-type Htt protein. The complexity of the molecular patho-physiology arises largely because of the proclivity and promiscuity of these interactions. Some of the interaction partners are involved in vesicle transport and gene transcription functions. These and many other functions are associated with the energy producing apparatus of the cell, such as the endoplasmic reticulum (ER) and mitochondria.

Studies with knock-in HD mouse model (homozygous for mHtt-Q111) showed that mHtt is associated directly with the outer mitochondrial membrane, and increases its susceptibility to calcium-induced permeability transition leading to cytochrome C release and apoptosis.(Choo et al., 2004; Panov et al., 2002; Petrasch-Parwez et

al., 2007) Furthermore, it has been suggested that the accumulation of abnormal misfolded proteins in the cells expressing mHtt protein overwhelms the ER apparatus responsible for quality control of proteins, which causes ER stress and trigger cell death.(Rao and Bredesen, 2004)

1.6 Mitochondrial dysfunction and oxidative stress in HD

The association of mHtt with mitochondria occurs both directly and indirectly and has been studied extensively over the past years. A decrease in the mitochondrial membrane potential due to a rise in proton conductivity(Panov et al., 2005; Sawa et al., 1999) and the opening of the mitochondrial permeability transition (MPT) pore (Choo et al., 2004), are a few well known mechanisms through which the mHtt protein affects mitochondrial function directly. Biochemical analysis of R6/2 mouse striatum demonstrate deficits in mitochondrial complex IV and aconitase activities, along with decrease in complex I/II and IV. (Hausladen and Fridovich, 1994; Tabrizi et al., 1999, 2000) Mitochondria in cells expressing mHtt are thus particularly susceptible to oxidative stress and cell death. Inhibition of the mitochondrial respiratory chain leads to increased production of reactive oxygen species (ROS), such as the superoxide ($O_2^{\cdot-}$) radical, hydrogen peroxide (H_2O_2) and peroxynitrite ($ONOO^-$), followed by fall in ATP levels.

Mitochondrial trafficking is indirectly affected by mHtt due to sequestration of trafficking components that are required for efficient axonal transport.(Trushina et

al., 2004) Specific N-terminal mHtt fragments accumulate as aggregates and physically block the movement of mitochondria within neurons (Chang et al., 2006; Orr et al., 2008) and also affect the mobility of Mitochondria trapped in the vicinity of the mHtt aggregates. (Chang et al., 2006) Fragmented mitochondria are a significant feature of HD cells, and exhibit alterations in cristae structure reflecting a disruption of electronic transport mechanisms (Solans et al., 2006), release of cytochrome C and eventual destruction of mitochondria (Costa et al., 2010) (Figure 1.2). Readers are referred to some excellent reviews on this subject (Chakraborty et al., 2014; Damiano et al., 2010; Johri et al., 2013) for detailed information.

1.7 A possible mechanism behind the gradual onset of mHtt toxicity

It appears reasonable to postulate that the gradual onset of the severity of HD phenotype is related to a cumulative process that is proportional to the polyQ repeat length as well as time during which the toxicity of polyQ is expressed. We here examine the evidences that point to several such cumulative processes in cellular physiology, which are thought to be affected by proteins containing polyQ tracts, especially by the pathogenic mHtt proteins.

Pathogenic polyQ repeats similar to those produced by the expanded CAG repeats in mutant *HTT* genes have been found to affect several cellular processes through alteration of protein conformation causing aberrant protein-protein

interactions,(Schaffar et al., 2004) leading to depletion of tRNA and translational frameshifting.(Girstmair et al., 2013) Q-rich tandem repeats (TR) are often present in transcriptional regulators(Gemayel et al., 2010; Legendre et al., 2007), a property that appears to be evolutionarily conserved.(Schaper et al., 2014) A comparative genomic study by Gemayel et al carried out in repeat variants of the yeast transcriptional regulatory protein Ssn6p (Cyc8p) demonstrated that an alteration of the repeat length leads to altered gene expression and phenotypic variation.(Gemayel et al., 2015) This study found that targets of Q-rich regulators exhibit increased gene expression across various time scales. Htt protein appears to be a transcription factor itself, and variation in its PolyQ length might therefore cause changes in its transcriptional activity.(Benn et al., 2008) Does altered gene expression in the presence of mHtt protein gradually lead to neuronal death in striatal cells? Below we examine the effects of mHtt on specific transcriptional regulators, which could potentially accumulate over time.

Indirect evidence links mHtt to mitochondrial dysfunction through its interaction with Tumor Protein 53 (p53). Studies with HD patient lymphoblasts stably expressing Htt N63-148Q, and in 293T (derivative of human embryonic kidney 293 cells, containing the SV40 T-antigen) cells transfected with a gene encoding the 171 amino acid long N-terminal-fragment of Htt containing either 23 or 148 glutamine repeats (N171-23Q or N171-148Q), respectively, or the full-length (FL) version of Htt (FL-23Q or FL-82Q) show that the mHtt binds selectively to nuclear

p53 and stabilizes the latter to increase transcriptional activity of p53 in the nucleus(Bae et al., 2005) (Figure 1.2). It has also been speculated that the interaction of mHtt with p53 might interfere with p53's interaction with its negative regulator, E3 Ubiquitin Protein Ligase (Mdm2) in the cytoplasm, thus stabilizing p53 further(Bae et al., 2005) (Figure 1.2). Over-expression of p53 was shown to increase the expression of various mitochondria-related pro-apoptotic proteins such as BCL2-Associated X Protein (Bax) and BCL2 Binding Component 3 (BBC3/PUMA) responsible for regulating mitochondrial depolarization and ROS levels(Bae et al., 2005; Toshiyuki and Reed, 1995; Yu et al., 2001) (Figure 1.2). Grison et al.(Grison et al., 2011) demonstrated that mHtt expression, which in turn causes p53 stabilization, leads to increased phosphorylation of p53 on Ser46. This causes p53 to interact with phosphorylation-dependent prolyl isomerase Pin1, and induce the expression of pro-apoptotic genes – Bax and PUMA.(Grison et al., 2011)

Other mitochondrial proteins known to be activated by p53 include Protein Noxa (NOXA) and Tumor Protein P53 Regulated Apoptosis Inducing Protein 1 (P53AIP1), all of which induce apoptosis when overexpressed.(Oda et al., 2000a, 2000b) Considering the above factors, it might be speculated that p53 is in a hyperactive state in neurons containing the mHtt protein in contrast to its activity in normal neuronal cells. p53 protein would then be expected to trigger apoptosis when a threshold level of oxidative damage has occurred in the mitochondria. Until then, the defective mitochondria may continue to divide and compromise the

cellular energetics that manifest itself phenotypically through various clinical signs and symptoms of HD.

The mHtt protein could bring about mitochondrial dysfunction through additional transcriptional mechanisms not involving p53. mHtt interacts with PPAR γ coactivator-1 α (PGC-1 α). In addition to playing important roles in glucose metabolism and adaptive thermogenesis, PGC-1 α is also required for the expression of nuclear respiratory factors such as NRF1 and NRF2 (Wu et al., 1999) along with other mitochondrial genes (*e.g.*, cytochrome C, (Andersson and Scarpulla, 2001) mitochondrial transcription factor A (mtTFA) (Wu et al., 1999) and respiratory complexes I-IV. (Kelly and Scarpulla, 2004) PGC-1 α is thus an important transcription factor controlling mitochondrial biogenesis (Wu et al., 1999) and for the production of mitochondrial –ROS detoxifying enzymes (Kukidome et al., 2006; St-Pierre et al., 2003; Valle et al., 2005) (Figure 1.2). mHtt inhibits the transcription of PGC-1 α by obstructing the promoter-binding activity of CREB/TAF4 (cAMP responsive element-binding/TATA-binding protein-associated factor 4) in mouse striatal cells expressing mHtt 111Q. (Cui et al., 2006) Indeed, lentivirus mediated overexpression of PGC-1 α in the brain striatum of transgenic (R6/1) mice reverses and rescues the mitochondrial dysfunction as well as neuronal degeneration. (Cui et al., 2006)

Taken together, the accumulation of molecular lesions related to defective transcription factor activity leading to increasing mitochondrial dysfunction can be reconciled with the observed gradual onset of the severity of neuronal dysfunction in HD affected brains. These, however, are unlikely to be exclusive mechanisms of the gradual onset of HD severity. Accumulation of unfolded or misfolded proteins due to defective chaperone activity in HD neurons may also be a major contributor.(Chai et al., 1999; Guzhova et al., 2011; Tagawa et al., 2007; Wacker et al., 2009) Nonetheless, misfolded proteins appear to accumulate in nearly all cells expressing mHtt while only the striatal neurons appear to be the most vulnerable and cause the early disease phenotype. This aspect of the disease mechanism has been reviewed elsewhere.(Cowan and Raymond, 2006)

1.8 Transcriptional regulation, Chromatin dynamics and the role of SIRT1 in HD.

Recent studies have increasingly highlighted the role of epigenetic mechanisms involved in chromosome dynamics and cell death in HD. This section will summarize the most recent findings and will attempt to define future questions and directions.

Since only certain neuronal cells of the striatal cortex undergo preferential early death in HD patients whereas mHtt protein is ubiquitously expressed, a major interest lies in understanding what makes these neuronal cell types more vulnerable

than others, and how to explain the delayed and progressive effects of mHtt toxicity on these cells. One rationale for addressing these questions is epigenetics, because it is the epigenetic modification of the cell's gene expression states that define cell-type specificities. Furthermore, molecular mechanisms of epigenetic processes, such as DNA methylation levels across genomic landscape are often progressive and cumulative, thus providing a framework to explain the progressive accumulation of mHtt toxic effects on the brain. Note also that signals from other cells, such as input connections from other neurons and their activities (Borrelli et al., 2008; Meadows et al., 2015; Singh-Taylor et al., 2015), also might affect the epigenetic states of neurons, thus providing the second rationale framework for the observation that excitotoxicity is important for susceptibility to mHtt toxicity. (Fan and Raymond, 2007; Young et al., 1988) Here we approach these problems primarily through the viewpoint that mHtt's interaction with mediators of epigenetic programming might explain both these aspects of HD.

First, we consider possible epigenetic players known to directly interact with mHtt. Sirtuins (yeast Sir2 homologous proteins) represent a family of epigenetic regulatory proteins that are highly conserved in evolution from prokaryotes to higher eukaryotes, including humans (Imai et al., 2000), and at least one member of this family is known to directly interact with the mHtt protein. There are seven recognizable Sirtuin genes in human: *SIRT1* to *SIRT7*. The functions of human *SIRT* genes are unknown. Based on homology with genes in yeast and other

organisms, these are thought to be important in epigenetic gene regulation, chromatin silencing and suppression of recombination within ribosomal DNA (rDNA) repeats. The encoded proteins have ADP-mono-ribosyltransferase activity. The seven members of the mammalian Sirtuin family occupy different subcellular compartments: *SIRT6* and *SIRT7* are nuclear proteins, *SIRT2*, *SIRT3*, and *SIRT4* are located in the mitochondria, while *SIRT1* and *SIRT5* are found both in the nucleus and in the mitochondria, and their relative expression levels vary somewhat based on the cell and tissue type. (Michishita et al., 2005) These proteins are collectively thought to transduce information on the energetic state of the cell (through sensing NAD/NADH ratio) to epigenetic processes (through their protein deacetylase activity). (Imai et al., 2000) Might these functions of Sirtuins be the key to imposing the specificity of early cell death, mainly the result of excitotoxicity, of striatal neurons in HD, by dint of a possibly unique metabolic state needed for the normal survival of these neurons? Sirt1 protein likely functions as a neuroprotective molecule through several mechanisms via its enzymatic function as a protein deacetylase on a number of distinct protein substrates. One such well-studied substrate is the forkhead box O3A (FOXO3a) transcription factor, which is highly expressed in adult brain. (Kops et al., 2002; Mojsilovic-Petrovic et al., 2009; Peng et al., 2010) mHtt is known to directly interact with Sirt1 and to inhibit the protein deacetylase activity of Sirt1, leading to hyperacetylation of Foxo3a (Figure 1.2). This is correlated with reduced neuronal survival in HD cell models. (Jiang et al.,

2012) Additional indirect evidence for interaction between mHtt protein and Sirt1 comes from experiments with HEK293 t/17 cells (having partial neuronal characteristics) containing mHtt (68Q), which showed increased acetylation of p53 –Sirt1 substrate, compared to cells containing wild-type Htt (17Q). This indicates that mHtt protein might interfere with p53 activity through modulation of Sirt1 deacetylase activity(Jiang et al., 2012) (Figure 1.2). Hyperacetylation of p53 is directly responsible for triggering DNA damage response of the cell (Figure 1.2). Similarly, PGC-1alpha is also subject to SIRT1 deacetylase activity.(Nemoto et al., 2005) Studies on PC12 cells (derived from a transplantable rat pheochromocytoma, a neuro-endocrine tumor), demonstrated a direct molecular interaction between SIRT1 and PGC1-alpha causing deacetylation of the latter protein. The resultant inhibition of transcriptional activity of PGC1-alpha leads to a complex chain of abnormal downstream interactions which affect both the mitochondrial function and the respiratory chain complex.(Nemoto et al., 2005) PGC1-alpha is therefore subject to regulation by various proteins including mHtt and acts in a manner similar to a master switch that controls mitochondrial function and cellular respiration. Overexpressed SIRT1 has been shown to attenuate brain atrophy and improve motor functions in both N-terminal fragment (N171-82Q) and full-length Htt (BACHD) mice models(Jeong et al., 2012; Jiang et al., 2012) by maintaining optimal levels of DARPP32 needed for dopamine signaling.(Fienberg et al., 1998;

Greengard et al., 1999) These studies also demonstrated a neuroprotective effect of SIRT1 against mHtt induced striatal atrophy.

It is interesting to note that *SIRT1* gene overexpression in N171-82Q Huntington disease mice model also leads to improved glucose tolerance and insulin sensitivity in these mice along with restoration of normal BDNF levels that are vital for controlling both glucose metabolism and DARPP32 expression. These findings suggest a broader metabolic and neuroprotective role of SIRT1 in Huntington's disease.(Jiang et al., 2012) In light of the above, mHtt could possibly be one of the key players connecting the glucose metabolism and neuronal survival pathways through its interaction with SIRT1—a possibility worthy of further investigations.

1.9 DNA damage and chromatin condensation defects

There is a second pathway in which SIRT1 might be important in HD: through SIRT1's role in chromatin condensation. Mice embryos with homozygous null mutation in *SIRT1* gene being to die at E9.5 and no homozygous live animals are produced. Some mice heterozygous for null *SIRT1* mutation exhibit brain development defects including exencephaly. Complete loss of *SIRT1* causes arrest of cell division in some cells in the early mitotic phase, and the arrested cells exhibit abnormal chromosome condensation, loss of DNA-damage induced G2/M checkpoint arrest, aneuploidy, apoptosis, higher frequency of spontaneous DNA double strand breaks, and the presence of hyper-acetylated lysine-16 (K16) of

histone H4 and lysine-9 (K9) of histone H3.(Wang et al., 2008) These findings suggest SIRT1 is important for DNA damage repair in mitotic cells. Given this function of SIRT1, it is anticipated that mHtt, which inhibits SIRT1 protein function, should adversely affect DNA-damage repair and cause abnormal chromatin condensation. Consistently, deficiency of SIRT1 protein (caused by siRNA) in HEK293T cells cause abnormal loading of histone 1 and the condensin I protein complex on to the mitotic chromosomes.(Fatoba and Okorokov, 2011) While it is possible that these findings might indicate higher levels of DNA damage accumulation in neural cells with depleted SIRT1 activity due to mHtt binding, these results should be interpreted with caution because condensin I is cytoplasmic, and its loading on to the chromatin occurs only during mitotic metaphase during which the nuclear membrane does not exist. Similar adverse effects of SIRT1 depletion on prophase-specific condensin II was not observed in HEK293T cells. However, SIRT1 is also known to associate with the MRN (MRE11-RAD50-NBS1, a protein complex that processes broken DNA) complex.(Tauchi et al., 2002) A deacetylated NBS1 protein enables the MRN complex in detecting DNA damage; therefore, hyperacetylation of NBS1 by the inhibition of SIRT1 activity by mHtt is expected to interfere with DNA damage repair. In fact, hyperacetylated NBS1 was shown to negatively affect intra-S phase checkpoint of the cell cycle(Yuan et al., 2007), suggesting that in addition to affecting G2/M DNA-damage checkpoint arrest, the loss of SIRT1 activity might cause increased

persistence of DNA damage lesions. If the adverse effects of mHtt on DNA damage repair is indeed relevant for excitotoxicity in striatal neurons, the question of cell-type specificity remains unanswered unless one postulates an increased DNA damage load in cells with high oxidative metabolism due to electrical activities of this group of neurons. While DNA damage by free radicals are generated by oxidative metabolism and increased mitochondrial respiration, a clear cause-and-effect relationship between the metabolic state and mHtt mediated cell death has not yet been established. Measurement of region-specific metabolomes in mouse brain indicated that no single metabolite but a complex of metabolites (metabolite signatures) correlated well with the sensitivity of brain regions to excitotoxicity.(Jaeger et al., 2015) Specifically, striatal neurons have a unique metabolite signature. Future studies of region-specific metabolomes of HD mouse brains, coupled with measurements of intracellular DNA lesions and epigenetic modification of these regional neurons, should be valuable for understanding how metabolites might affect neuronal specificity in HD pathogenesis.

Second, we approach the problem from the idea that DNA damage might not be directly related to the metabolic activity of the striatal neurons, but is the result of specific neuronal functions. Huntingtin co-localizes with microtubule organizing bodies, and is thought to facilitate the dynein/dynactin-mediated transport of organelles including mitochondria along microtubules in neuronal cells(Godin et al., 2010); siRNA against normal Htt causes mislocalization of *p150^{Glued}* (subunit

of dynein), dynein, and the large nuclear mitotic apparatus (NuMA) protein – which is essential for microtubule assembly and maintenance.(Radulescu and Cleveland, 2010) It is possible that mHtt might cause problems with organelle transport in striatal neurons, which could indirectly influence chromosome integrity. However, it is hard to see how microtubule malfunction could lead to DNA damage because neurons do not divide, and therefore microtubule motors are not expected to be involved in chromosome dynamics in G0-arrested neurons. Nevertheless, DNA damage might also be directly caused by mHtt interaction with proteins that modulate chromosome dynamics in a G0 neuron, as discussed in the next section.

1.10 rDNA condensation defects in HD

Abnormal nucleolar ribosomal DNA (rDNA) condensation, during which condensin I and II subunits, including Smc2p and Smc4p, are recruited to the rDNA loci, is related to apoptotic death and DNA damage in mammalian cells(Blank et al., 2006) (Figure 1.3). Repeat-containing RNA has been demonstrated to associate with nucleolar protein complexes leading to “nucleolar stress”, which is thought to trigger cell death through activating the p53 mediated pathway.(Kreiner et al., 2013) This phenomenon has been shown to be relevant for certain degenerative disorders of repeat sequences, including HD (Tsoi et al., 2012), ALS and frontotemporal dementia.(Haeusler et al., 2014) Interestingly, results from our

laboratory have recently demonstrated that mHtt toxicity in yeast can be suppressed by several rDNA genes (Chatterjee et al., 2013), most notably by the gene encoding L12p, a member of the L11p subgroup of ribosomal proteins. More recently it has been shown that L11p which is recruited to the nucleolus within a poorly-defined complex associated with repeat-containing noncoding RNA and chromatin proteins, is directly responsible for triggering apoptosis through activating p53 (Maehama et al., 2014) (Figure 1.3).

An emerging model is that premature or abnormal condensation of the chromatin at rDNA loci, facilitated by L11p-like proteins and condensins, such as Smc2p, sequesters an inhibitor of p53 (Mdm2) at the chromatin during S-phase, with the resulting activation of p53-mediated apoptosis (Figure 1.3). However, there is yet no direct evidence that mHtt-mediated apoptosis in human cells can be triggered by abnormal chromatin condensation. Consistent with these is the finding that an RNA Pol II associated protein UBTF is tri-methylated by ESET, a H3K9 methyl transferase, and this trimethylation deregulates rDNA condensation in a striatal Q111 knock-in cell line (relative to Q7 striatal cell line) model of HD (Hwang et al., 2014) (Figure 1.3). Reduced acetylation of UBTF at K352 by an siRNA against the CREB protein also reduces rDNA transcription in striatal Q111 expressing neurons but not in Q7-expressing neurons. (Lee et al., 2011) A genetic evidence connecting

rDNA expression with mHtt toxicity comes from recent experiments with the yeast HD model (chromosomal 103Q N-terminal fragment expressed from GAL promoter in yeast)(Chatterjee et al., 2013) in which several strong rDNA suppressors of HD toxicity were discovered. In a genome-wide gain-of-function suppressor screen for a defect in chromosome condensation by mutation in the condensin-encoding gene *smc2*, we identified *UME1* and *BNA5* as suppressors.(Patra et al., 2013) *UME1* encodes a member of the histone deacetylase complex, underscoring the importance of epigenetic processes in mHtt toxicity, and *BNA4* and *BNA5* encode two successive enzymes in the biosynthesis of NAD from kynurenine. The latter gene, *BNA5*, is a current target of HD drug development.(Beconi et al., 2012; Harris et al., 1998; Santamaría et al., 1996; Zwilling et al., 2011) The Bna4p enzyme localizes to the mitochondrial outer membrane, whereas Bna5p localizes both to the nucleus and the cytoplasm (Figure 1.3), suggesting that these two enzymes might form a link between the chromatin, p53, and mitochondrial abnormality in cells expressing mHtt. These results are in general accordance with the idea that a defect in rDNA condensation, brought about by mHtt, could directly lead to DNA damage in the striatal neurons, thus triggering apoptosis.

1.11 Conclusion

We have summarized here the current experimental models of HD, their relative usefulness, and have discussed molecular mechanism of HD pathogenesis. Evidence point to the importance of epigenetic mechanisms related to mHtt's (a) direct interaction with epigenetic regulators, (b) indirect interaction with regulators of metabolic states of neurons leading to DNA damage and its persistence, (c) direct interaction with proteins important for chromosome condensation at the rDNA. Synergistic genetic interactions between mHtt and p53 or PGC1- α or both appear to be amplifier mechanisms of HD progression. A better understanding of the mechanisms of molecular pathogenesis of HD should be possible in the context of integrated networks of genetic modifiers. We have suggested areas of future experimental approaches to better understand the molecular pathogenesis of HD, such that this crippling disease becomes amenable to rational drug development.

1.12 Tables and Figures

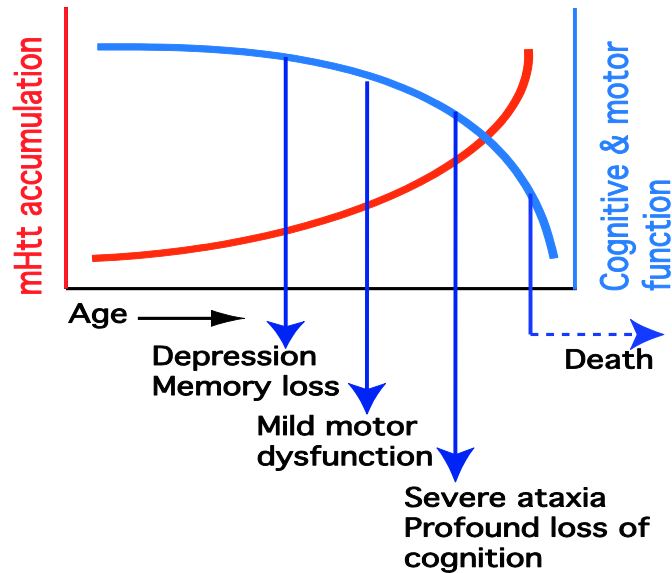


Figure 1.1 Clinical progression of Huntington's disease as a function of accumulation of mHtt protein.

The Y-axis on the left represents mHtt accumulation in the brain; y-axis on the right represents cognitive, motor activity. Axes are in arbitrary units. There is a substantial heterogeneity among patients as to when salient markers become visible, the age of onset of clinical presentations, the rate of progression of the disease and the time of death since diagnosis. While CAG repeat length at HTT is a strong determinant of the age of onset, there is still much variability among individual patients, indicating the influence of genetic or epigenetic modifiers as well as environmental effects

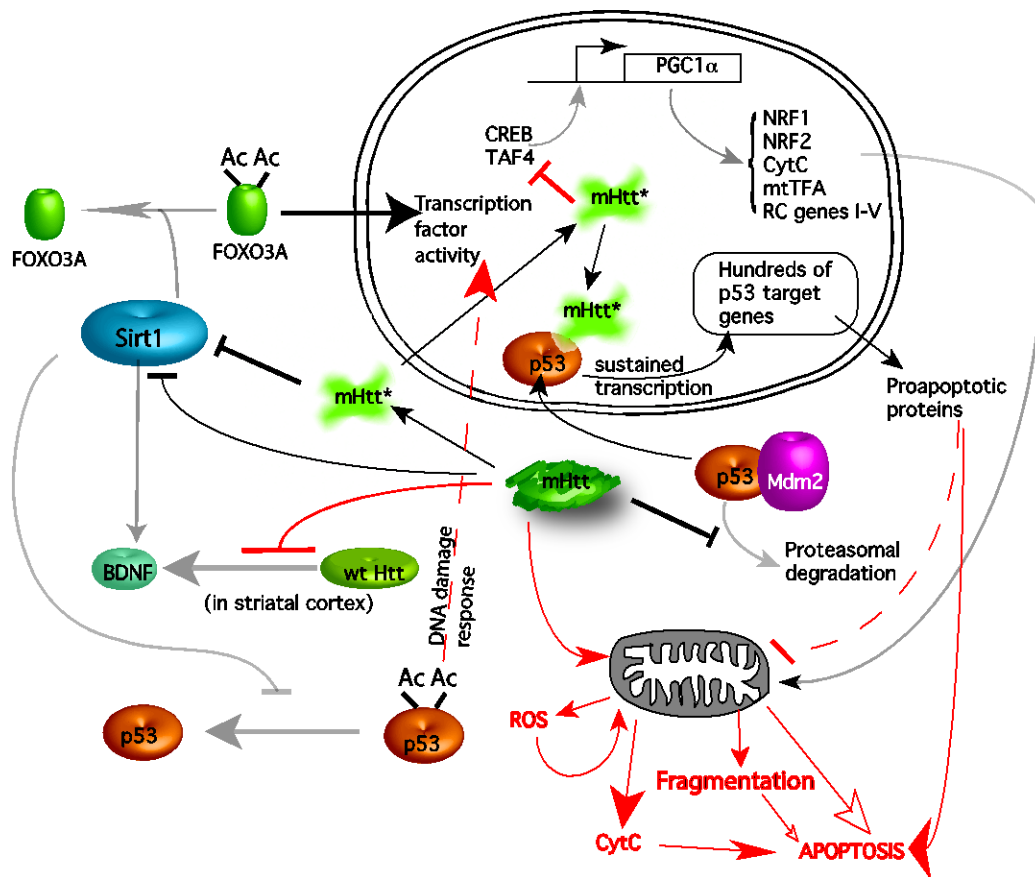


Figure 1.2 Models of mechanisms of action of mHtt.

Full length mutant Huntingtin protein (mHtt) interferes with the interaction of p53 with its regulator Mdm2, stabilizing p53 and inhibiting the latter's downstream proteasomal degradation. N-terminal fragments of mHtt (mHtt*) enter the nucleus leading to persistent transcription of many genes by p53, triggering the synthesis of pro-apoptotic proteins, such as Bax and Puma, eventually causing mitochondrial destruction. mHtt* also inhibits CREB and TAF4 to decrease the expression of transcriptional co-activator PGC1-alpha, leading to increased levels of reactive oxygen species (ROS) and cytochrome C release from the mitochondria, and also defects in the synthesis of electron transport chain components. These subsequently lead to mitochondrial fragmentation, and cell death. mHtt (mHtt*) also inhibits protein deacetylase activity of SIRT1 which leads to abnormal accumulation of hyperacetylated FOXO3a and p53 proteins. Hyperacetylated FOXO3a and p53 participate in abnormal transcriptional factor activity, the latter specifically involving DNA damage response genes.

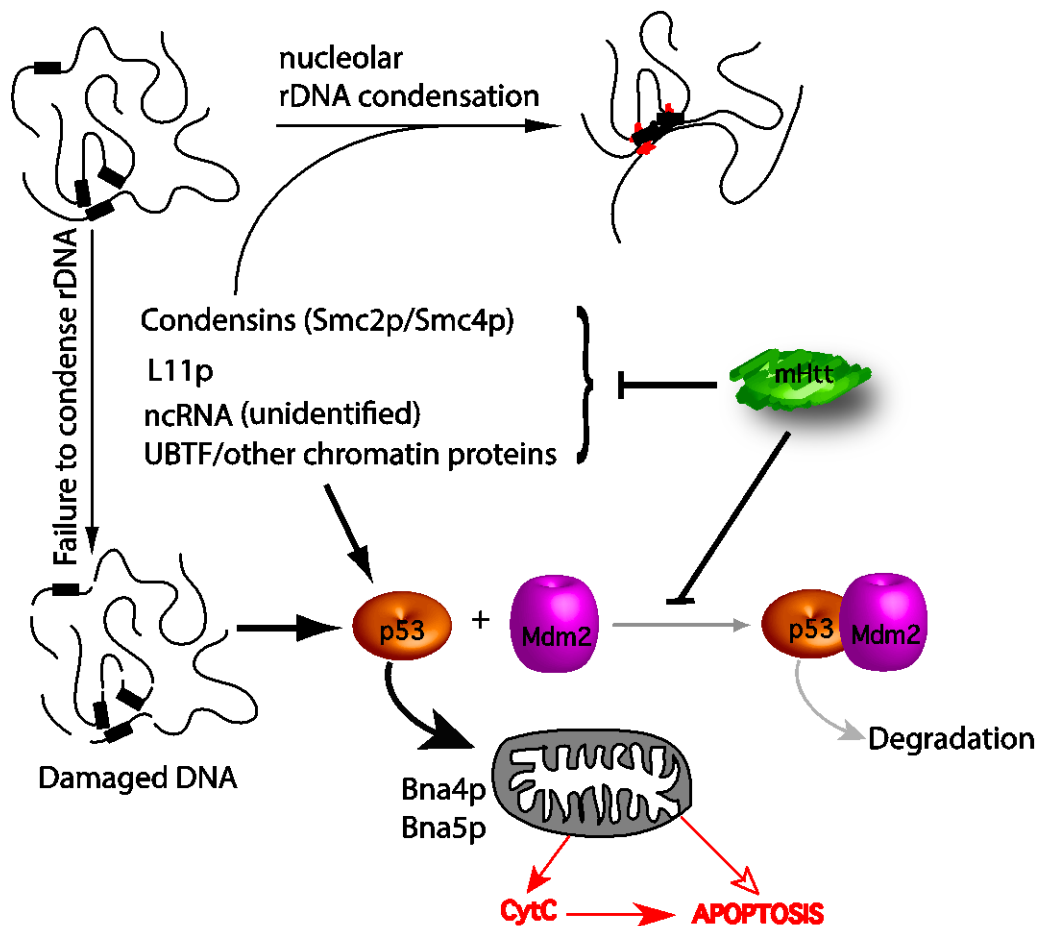


Figure 1.3 Chromosome condensation defects in Huntington's disease.

mHtt inhibits the function of ribosomal protein L11p, condensin proteins Smc2p/Smc4p, UBTF and other chromatin proteins, which are responsible for nucleolar rDNA condensation. These may lead to fragmentation of nuclear DNA, initiating DNA damage response, also involving p53, leading to triggering of apoptosis. Subsequent mitochondrial dysfunction may also contribute to apoptosis as described in Figure.1.2. Black boxes represent rDNA loci; red dots on chromosomes represent condensed rDNA regions including rRNA transcripts in the nucleolus.

Chapter 2

2 Constructing the Huntington's disease

Integrome

2.1 Introduction

Huntington disease (HD) is a rare neurodegenerative disorder inherited in an autosomal dominant manner which is thought to be caused by a CAG triplet repeat expansion in exon1 of the Huntingtin (*HTT*) gene. HD is one among the group of PolyQ repeat disorders that include spinocerebellar ataxias, dentatorubral-pallidoluysian atrophy (DRPLA) and spinal and bulbar muscular atrophy, X-linked 1 (SMA1/SBMA). (Fan et al., 2014) Other disorders such as Huntington's disease-like 1 (HDL1) and Huntington's disease-like 2 (HDL2) show close resemblance with the neuropathological and clinical presentation of HD. (Schneider and Bird, 2016) On an average, the age of onset of disease is inversely proportional to the number of CAG repeats. While these repeat numbers account for approximately 50-70% of the age of onset of data, various genetic and environmental factors explain the remaining variability in the age of onset. (Djousse et al., 2003; Project* and Wexler, 2004) The number of CAG repeats is also an important determinant of the rate of clinical progression. Individuals with shorter CAG repeats display a gradual increase in clinical progression than patients with longer CAG repeats. Clinical features of HD range from subtle changes in cognition and motor control

(Walker, 2007) to distinct chorea, and profound motor incoordination (Watts and Koller, 1997; Weiner and Lang, 1989), cognitive dysfunction (Craufurd and Snowden, 2002), depression and suicidal behavior. (Walker, 2007) The patient generally succumbs to the disease. Although extensive studies reveal a complex pathophysiology of this severely crippling disorder, the sequence of events through which the mutant Huntingtin (mHtt) protein executes its action still remains elusive.

The mHtt protein is thought to exert its effects mostly through a toxic gain of function via the short N-terminal fragments of mHtt that are produced as a result of proteolytic cleavage. (Bizat et al., 2003; Cowan and Raymond, 2006; Gafni and Ellerby, 2002; Gafni et al., 2004; Goffredo et al., 2002; Graham et al., 2006; Kim et al., 2001; Lee and Kim, 2006; Wellington et al., 1998) Neuronal studies and mice models of HD also point to a loss of beneficial function of wild-type Htt protein. (Cattaneo et al., 2001; Faber et al., 1998; Hackam et al., 2000; Kalchman et al., 1997) Accumulation of PolyQ aggregates also called as intraneuronal nuclear inclusions (INNs) in the neuronal cells (Davies et al., 1997) due to protein misfolding is one of the hallmarks of the HD. The overall burden of INNs correlates with the severity of the clinical symptoms of HD. The cellular stress that builds up due to increasing levels of INNs unleashes a sustained unfolded protein response (UPR) and eventual neuronal apoptosis. (Soto, 2003) The normal *HTT* gene is

essential for survival in early embryogenesis, but its function, though ubiquitously expressed in later stages of development, is not well understood. The complexity of the pathophysiology of HD can be attributed to the tendency of mHtt to abnormally interact with various other proteins that either do or do not interact with the wild-type Htt protein in normal conditions. This is compounded by the presence of the Htt protein at various subcellular locations where it is proposed to participate in various signaling pathways and/or associate with numerous other protein partners during its normal course of action.(Cattaneo et al., 2005; MacDonald, 2003; Marcora et al., 2003) Among the several molecular and cellular functions affected in HD, some important ones include transcriptional activity(Benn et al., 2005; Luthi-Carter et al., 2000; Schilling et al., 2004), vesicle transport(DiFiglia et al., 1995; Gauthier et al., 2004; Velier et al., 1998), synaptic transmission.(Gutekunst et al., 1999; Li et al., 2000; Luthi-Carter et al., 2000; Sapp et al., 1999; Trettel et al., 2000; Velier et al., 1998), and mitochondrial functions.(Benchoua et al., 2006; Bezprozvanny and Hayden, 2004; Panov et al., 2002)

Curiously, although differentiated cortical neurons affected by HD do not divide, recent studies have highlighted the role of chromosome dynamics in the pathophysiology of HD and how epigenetic mechanisms might contribute to DNA

damage and cell death. SIRT1, whose protein deacetylase activity is inhibited by mHtt (Jiang et al., 2012), is one such epigenetic player. Several studies have shown SIRT1 to be important for DNA damage repair in mitotic cells (Tauchi et al., 2002; Wang et al., 2008) by helping with loading of histone 1 and condensin 1 complexes on to the mitotic chromosomes.(Fatoba and Okorokov, 2011) Association of SIRT1 with the MRE11-RAD50-NBS1 (MRN) complex, leads to a deacetylated NBS1 that helps in DNA damage detection.(Tauchi et al., 2002) Inhibition of SIRT1 by mHtt has therefore been thought to interrupt the process of DNA damage repair(Yuan et al., 2007), further affecting the processes of G2/M DNA-damage checkpoint arrest, and intra-S phase checkpoint of cell cycle.(Yuan et al., 2007) SIRT1 is known to be important for chromatin condensation during normal cell division.(Wang et al., 2008) It functions as a neuroprotective agent through its protein deacetylase activity on a number of protein substrates such as FOXO3a (Jiang et al., 2012; Kops et al., 2002; Mojsilovic-Petrovic et al., 2009; Peng et al., 2010), p53(Bae et al., 2005; Grison et al., 2011; Toshiyuki and Reed, 1995; Yu et al., 2001), PGC1-alpha (Nemoto et al., 2005) and BDNF(Cattaneo et al., 2001) to regulate neuronal function including their metabolic states. Such indirect actions of mHtt through SIRT1 may lead to continuing accumulation of DNA damage.

Recent studies from our laboratory demonstrated that several rDNA genes encoding ribosomal proteins, specifically L12p suppresses mHtt toxicity in yeast (Chatterjee et al., 2013); Yeast L12p is equivalent to the human L11p, and L11p is known to be directly implicated in triggering apoptosis through the p53 pathway.(Maehama et al., 2014) An additional connection between mHtt and chromosome condensation defects exists which is implicated in DNA damage and apoptotic cell death in mammalian cells.(Blank et al., 2006) Abnormal association of repeat-containing RNA with nucleolar protein complexes is known to cause “nucleolar stress” and activates the p53 pathway to trigger cell death(Kreiner et al., 2013), a process shown to be important for HD.(Tsoi et al., 2012) Thus, while neurons do not undergo mitosis, their death might indeed be triggered by rDNA condensation defects through the same molecular pathways that are used in mitotic DNA repair.

Considering the above, it is now increasingly evident that the phenotype of HD is an outcome of numerous processes initiated by the mHtt protein (Culver et al., 2012) along with other proteins that act as either suppressors or enhancers of the effects of mHtt protein and PolyQ aggregates. Detection and analysis of proteins that physically interact with wild-type and mHtt proteins have provided valuable information on various molecular and cellular processes affected in the mutant cells. Physical interactors of Htt proteins have been discovered using yeast two-

hybrid (Y2H) and in vitro affinity pull-down experiments.(Faber et al., 1998; Goehler et al., 2004; Holbert et al., 2001; Li et al., 1995; Passani, 2000; Savas et al., 2008; Wanker et al., 1997; Yamamoto et al., 2006) Such proteomic studies have identified high-confidence Htt-associated proteins by using both wild-type and mutant full length Htt (fl-Htt) proteins.(Culver et al., 2012; Shirasaki et al., 2012) Proteomic analysis of human HD brain specimens reveal several differentially expressed proteins in substantia nigra (Chen et al., 2012), cortex (Schönberger et al., 2013) and striatum.(Sorolla et al., 2008) A recent proteomic study of HD and HDL2 disease brains uncovered several concomitantly affected pathways such as Rho-mediated signaling, axonal guidance and DNA/RNA processing.(Ratovitski et al., 2016) While these recent analytical studies have generated long lists of protein interactions that are affected in HD, the potential of using them to reveal disease mechanisms, poses interesting challenges considering the high volume data involved.

Network-based analysis of disease proteomes have provided important clues to mechanisms of molecular pathogenesis in HD. Such studies have so far been largely conducted on data derived from yeast-two hybrid protein interaction experiments.(Goehler et al., 2004; Riechers et al., 2016; Tourette et al., 2014) These studies have revealed important roles of Rho GTPase pathway (Tourette et al.,

2014) and, caspase-6 interactors (Riechers et al., 2016) in HD pathogenesis. A spatiotemporal proteomic study of HD and wild-type mouse brains performed using Affinity Purification Mass Spectrometry (AP-MS), identified candidate proteins found in complex with Htt protein and uncovered a modular network of Htt-interacting proteins enriched in functions such as proteostasis, microtubule-based transport and 14-3-3 signaling.(Shirasaki et al., 2012) Construction of separate protein-protein interaction (PPI) networks for wild-type and mHtt protein have helped identify new interacting partners of the mHtt protein acquired in the diseased state.(Basu et al., 2013) Such protein-protein networks alone are of limited value without additional consideration, such as the functions of groups of these proteins, which might be affected by the disease.(Kelley and Ideker, 2005)

Building network models by integrating genetic and physical interactions have been invaluable in understanding the organization and functions of disease pathways.(Kelley and Ideker, 2005) For example, integrated analyses of multiple datasets such as genome-wide linkage studies, genome-wide association studies and genome-wide expression profiling followed by PPI network modeling has prioritized candidate genes for Alzheimer's disease.(Talwar et al., 2014) An interesting approach has been the integration of gene expression and protein interaction data from HD patients and progressively filtering the resultant HD

interaction network to predict brain-specific interactors of Huntingtin protein (Stroedicke et al., 2015).

While protein interactions are valuable tools to explore still unknown molecular processes and functions, it has been shown that the ability of PPI data alone to predict interactions and molecular pathways improves with the inclusion of additional information. Such information may include tissue and cell type-specific gene expression data and further evidence about highly interacting proteins (Lopes et al., 2011). To enable functional linkages between PPI networks and biological functions, studies on model organisms have been of outstanding promise.

To enable functional linkages between PPI networks and biological function, studies on model organisms have been of outstanding promise. *C.elegans*, *D.melanogaster* and *S.cerevesiae* models of HD toxicity have enabled the identification of genetic modifiers of HD (Chatterjee et al., 2013; Faber et al., 2002; Giorgini et al., 2005; Imamura et al., 2016; Jimenez-Sanchez et al., 2015; Mason and Giorgini, 2011; Mason et al., 2013; Parker et al., 2004; Silva et al., 2011; Willingham, 2003; Yamamoto et al., 2006; Zhang et al., 2010). However, the complexity of the mechanisms of action of these genetic modifiers on their targets pose a serious challenge for deciphering the causative mechanisms behind the

disease. It may be speculated that an integrative systems approach that includes both physical and genetic interactions could be a powerful avenue to decipher the complexities of this multi-system disorder and simultaneously obtain a comprehensive depiction of the perturbed molecular processes in HD.

In this study, we construct an orthologous human HD interactome, using human orthologs of protein interactors of wild-type and mHtt in a mouse HD model (Shirasaki et al., 2012) combined with genetic modifiers of mHtt toxicity found in yeast HD models.(Chatterjee et al., 2013; Giorgini et al., 2005; Mason and Giorgini, 2011; Mason et al., 2013; Willingham, 2003) Computational analysis of the orthologous interactome revealed a modular structure functionally enriched for DNA damage response, regulation of chromatid cohesion and chromosome organization, suggesting the hypothesis that these processes might be abnormal in cells expressing mHtt. We tested this hypothesis by combining a series of gene over-expression constructs with genes encoding the normal or mutant versions of the human Htt N-terminal fragments, and observing that 24 yeast genes related to chromosome dynamics, when over-expressed from multicopy plasmids, suppress mHtt-mediated toxicity. The computational predictions with the human HD interactome were confirmed further by an independent computational technique,

which allowed the identification of 27 candidate human genes as possible genetic modifiers of HD. We have validated 3 of these genes in a *Drosophila* model of HD

The novelty of our approach is an iterative process: by integrating genetic interaction and protein interaction data from two different model organisms, we generate an orthologous human interactome, analyze that orthologous interactome for the presence of molecular functions enriched within topologically defined modules by unsupervised machine learning, and validate gene candidates as modifiers of HD in model organisms.

2.2 Method

2.2.1 Interaction data sets

The huntingtin protein interactome was built using published datasets. Five sets of primary interactors of mHtt protein were retrieved from experiments performed in yeast (Chatterjee et al., 2013; Giorgini et al., 2005; Mason and Giorgini, 2011; Mason et al., 2013; Willingham, 2003) and mouse models (Shirasaki et al., 2012) of HD. Studies performed in yeast HD model (Chatterjee et al., 2013; Giorgini et al., 2005; Mason and Giorgini, 2011; Mason et al., 2013; Willingham, 2003) comprise of genetic suppressors and enhancers of mHtt protein, while Shirasaki et al. is a spatiotemporal set of protein interactors identified using AP-MS that form

complexes with Htt in both wild-type and BACHD mouse brains (Shirasaki et al., 2012).

2.2.2 Ortholog detection and construction of orthologous HD interactome

Human orthologs ($n = 601$) of the protein interactors were found using the Inparanoid algorithm using a threshold of 1 for detecting orthologs (O'Brien et al., 2005; Remm et al., 2001) (R packages – hom.Hs.inp.db_3.0.0, hom.Mm.inp.db_3.0.0 and hom.Sc.inp.db_3.0.0). The first degree protein interactions for these human orthologs were obtained by querying the Human Integrated Protein-Protein Interaction rEference (HIPPIE) database – a human PPI database which contains an integration of multiple experimental PPI datasets normalized using a confidence score ranging from 0 to 1 for each protein interaction. (Schaefer et al., 2012) Protein interactions with a confidence score larger than 0.3 were chosen for network construction and analysis (Figure S1 in Text S1).

2.2.3 Network construction and analysis

HD interactome was built using the web-based Cytoscape tool. (Shannon et al., 2003) Topological network analysis was performed using NetworkAnalyzer. (Assenov et al., 2008) ClusterONE (Nepusz et al., 2012) was

used to identify overlapping protein complexes within the HD interactome. Hierarchical network structure was obtained using Jerarca.(Aldecoa and Marín, 2010) Additional information about ClusterONE and Jerarca algorithms is given in the forthcoming sections Dendrogram was visualized using the Phylowidget tool.(Jordan and Piel, 2008)

2.2.4 Gene Ontology (GO) enrichment

GO enrichment of the protein complexes was performed using the g:GOST tool from the g:Profiler package.(Reimand et al., 2016) Significant GO enriched processes and pathways were estimated using the hypergeometric probability distribution given by the following equation:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}$$

where, n is the number of genes of interest, k is the number of genes within that list that are annotated with a GO term, N is the total number of background genes in the distribution, M is the number of genes within that distribution that are annotated directly or indirectly with the GO term of interest.(Boyle et al., 2004) False Discovery Rate (FDR) was controlled using the Benjamini-Hochberg method for multiple correction.(Benjamini and Hochberg, 1995)

2.2.5 Yeast strains, media and plasmids

The yeast strains for studying mHtt toxicity study were generous gifts from S. Lindquist (MIT). The two different versions of the N-terminal Htt fragment from exon 1 (25Q(normal) and 103Q(toxic)) are integrated into the chromosome of W303 yeast strain background for expression of poly(Q). W303 is a haploid yeast strain expressing FLAG-htt-poly(Q)-CFP under GAL1 promoter control in *his3* locus and the two strains used in this study are *can1-100, his3-11,15::FLAGhtt103Q-CFP, leu2-3,112, trp1-1, ura3-1, ade2-1*; and *can1-100, his3-11,15::FLAGhtt25Q-CFP, leu2-3,112, trp1-1, ura3-1, ade2-1*. Suppressors genes of *smc2-8* were selected from Patra et al 2017.(Patra et al., 2017). They were individually selected from the MORF (Movable open reading frame) library.(Gelperin, 2005; Patra et al., 2017) The MORF library contains 5,871 ORFs in 2 μ plasmids with galactose inducible promoter and a URA3 selectable marker. The plasmid DNAs were isolated separately for further transformation into the above two yeast strains. The yeast strains having 25Q and 103Q repeats, were grown in yeast complete media containing 1% raffinose followed by transformation of each strain with 1 μ g of individual gene. The transformants were selected on synthetic defined medium lacking uracil containing 1% raffinose and streaked for single colonies. A single colony corresponding to each candidate suppressor gene in each of the two different yeast strains was suspended in liquid broth for serial 10X dilution and titration spotted on synthetic media containing either 2% glucose

(repression) or 2% galactose (induction). The wild-type strain containing Htt25Q can grow normally when GAL1-25Q is overexpressed in presence of galactose, whereas the mHtt strain having Htt103Q should die or show reduced growth in presence of galactose.

2.3 Results

2.3.1 Construction of HD protein interactome

We compiled a set of 1188 physical and genetic interactors of the normal Htt and mHtt proteins from published data (Chatterjee et al., 2013; Giorgini et al., 2005; Mason and Giorgini, 2011; Mason et al., 2013; Shirasaki et al., 2012; Willingham, 2003) (Table S1) and obtained 601 human orthologs that were designated as primary interactors of Htt or mHtt protein. Next, we acquired the interacting protein partners of the above 601 human orthologs from HIPPIE (version 1.7), with a confidence score assigned to each interaction. (Schaefer et al., 2012) A total of 32365 interactions were obtained, which comprised of both direct and indirect interacting partners of Htt protein. The interaction list was further narrowed down to 32243 interactions by selecting only those interactions that had a confidence score of more than 0.45 in the HIPPIE database. An integrated HD interactome was constructed from this final list of interactions and will be referred to as the Huntington's disease interactome (HDI) (Figure 2.1). The HDI is densely connected

containing 7418 nodes that represent first and second-degree protein interactors of wild-type and mHtt protein and 31185 edges (self-loops and duplicate edges removed) that represent total interactions among the protein pairs. Note these lists of interactions contains two kinds of information (1) orthologous PPI pairs of mouse proteins mapped to the human proteome, (2) orthologous genetic interaction pairs of yeast genes mapped to the human proteome. Thus, these lists encapsulate heterogeneous properties of proteins and genes.

2.3.2 Network properties of Huntington's disease integrome

The network properties of the HDI have been summarized in Table 2.1 and discussed in brief in the Appendix. The HDI follows a power-law degree distribution (Figure 2.2), with a clustering coefficient of 0.140 and a network density of 0.001 indicating a large sparse network of interactions characteristic of most biological networks.(Jeong et al., 2000, 2001; Watts and Strogatz, 1998) Out of the 601 human orthologs identified as primary interactors of the Htt protein, 578 genes are directly connected to each other. The 5 most highly connected nodes with a degree > 350 are HSP90AA1, YWHAZ, HSP90AB1, VCP and CUL2 which act as network 'hubs' indicating their probable role in HD processes. Indeed, the binding activity of HSF1 – a master regulator of heat shock proteins such as HSP90AA1 and HSP90AB1 is known to be altered in Poly-Q cells (ST Hdh^{Q111})

expressing full-length Htt.(Riva et al., 2012) Both wild type and mHtt are known to interact with HSP90; a pharmacological inhibition of this interaction leads to increased clearance of mHtt and degradation through the ubiquitin-proteasome systems.(Baldo et al., 2012) HSP90AA1 has also been identified has a potential gene of interest in HD in another study involving microarray analysis of post-mortem human brain samples.(Chandrasekaran and Bonchev, 2016) VCP is known to localize chiefly in the nucleus of the adult neurons and interacts with both wild and type mHtt protein. Its interaction with mHtt alters the recruitment of VCP to DNA damage foci causing inhibition of DNA repair (Fujita et al., 2013) and modulates neurodegeneration. Loss-of-function mutants of *ter94* gene in *Drosophila* encoding for VCP act as genetic modifiers of neurodegeneration.(Higashiyama et al., 2002) Overexpression of *cul-2*, a *Drosophila* ortholog for the human gene CUL2, is known to suppress an expanded (128Q) Htt- fragment induced neurodegeneration in the *Drosophila* eye.(Kaltenbach et al., 2007) Taken together, these findings present evidence that our orthologous HDI displays a ‘network-hub’ driven structure characteristic of a scale-free topology associated with robust biological interaction networks, where the major hubs represent genes/proteins known to be important in HD pathogenesis.

Chapter 3 covers the analysis of the integrome using an unsupervised machine learning approach and the results of validation.

2.4 Tables and Figures

Table 2.1 Network properties of Huntington's disease integrome (HDI)

Network Properties	
Number of nodes	7418
Number of edges	31185
Connected components	12
Network density	0.001
Network heterogeneity	3.405
Network diameter	9
Clustering Coefficient	0.14
Average shortest path length	3.545
Average number of neighbors (mean degree)	8.408

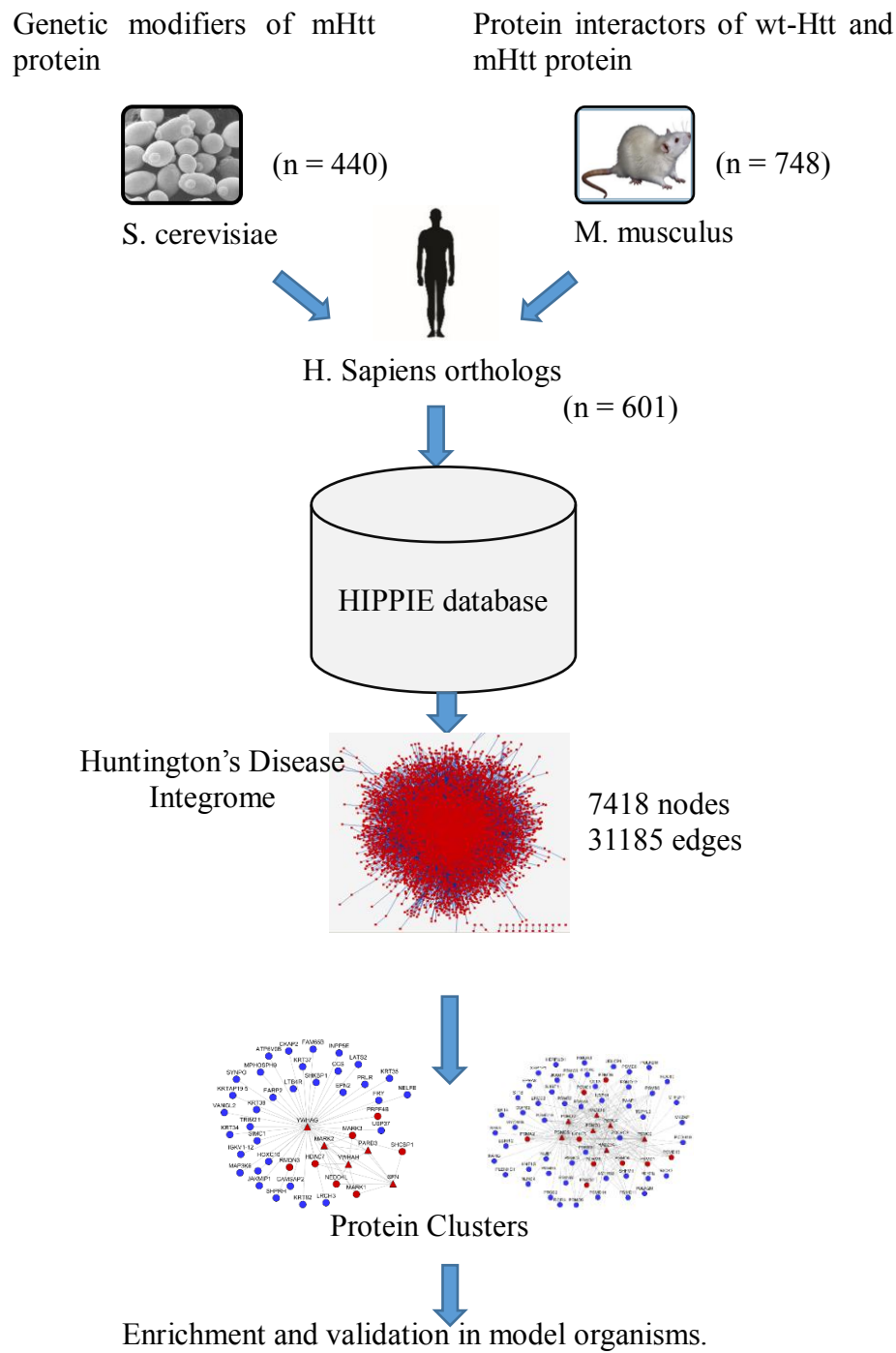


Figure 2.1 Workflow to generate and identify candidate genes from the Huntington's disease interactome (HDI).

The workflow for the interactome includes integrating physical and genetic interactors of wild-type and mHtt protein from various HD model organisms finding the human orthologs ($n = 601$) for these interactors, and mining a protein interaction database for secondary interactions. The integrated network thus constructed contains first and second-degree interactors of the wild-type and mutant Htt protein.

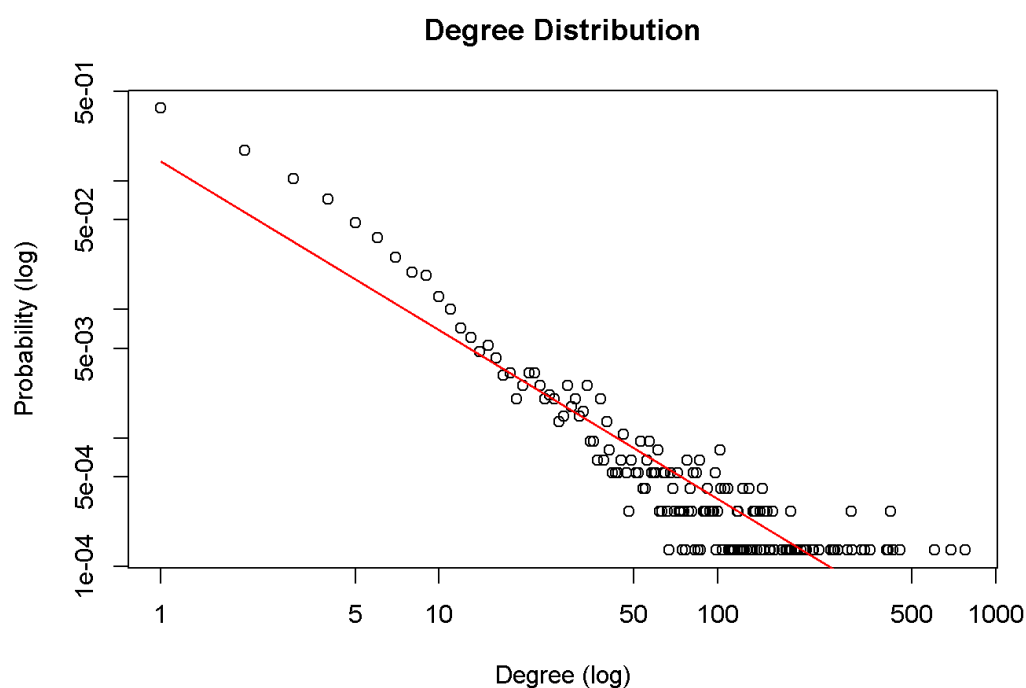


Figure 2.2 Degree distribution of the Huntington's disease interactome (HDI).

The HDI follows a power-law distribution fit in the form of $y = ax^{-b}$ ($a = 1053.7$, $b = 1.313$ and $R^2 = 0.857$)

Chapter 3

3 Analysis of Huntington's disease interactome and candidate gene validation using unsupervised machine learning

An important approach to understand the biological significance of a large interaction network is to apply unsupervised machine learning algorithm to reduce the dimensionality of the data. When the interaction data are reduced to smaller numbers of subclasses according unbiased methods of dimensionality reduction, one can then ask whether such classes have biological significance. One such method of unsupervised machine learning method for dimensionality reduction is clustering the network members on the basis of some graph-theoretic properties. We thought that within a large interactome with potentially many false positives, nodes (genes/proteins) that are connected topologically to other genes/proteins, especially those within overlapping clusters where connections to the same members are repeated many times, could provide clues to important biological functions in common among those proteins belonging to overlapping clusters.

With the aim of detecting highly connected overlapping clusters of proteins within the HDI,

3.1 Pre-requisites to choose a clustering algorithm for the HDI.

Considering the heterogenous nature of the information contained in the HDI network, we devised a list of pre-requisites that needed to be fulfilled that would help extract biological meaningful information from the graph. Since proteins have multiple biological functions, and operate in various cellular compartments over time, a clustering method that captures these overlapping relationships of functions with each other and also with other proteins was an essential pre-requisite. Computationally efficiency of the clustering approach and the ability of the algorithm to scale up to larger networks was also a necessary requisite in choosing a method. Ability of the algorithm to be robust and produce consistent results, albeit with some performance degradation, even when a certain level of noise is introduced in the graph was also an important criterion. Since there is no gold standard dataset for HD for the algorithm to compare against, we believe, another essential criterion to be the ability of the algorithm to assess the quality and accuracy of the cluster output by using an internal validity measure that is built into the algorithm itself.

Based on the above pre-requisites we chose ClusterONE – an overlapping complex detection method as the primary clustering algorithm for HDI. While ClusterONE detects overlapping relationships between clusters, it also was found to be a

computationally efficient method that could be scaled up to support clustering of larger networks. Another method Jerarca – a non-overlapping hierarchical clustering algorithm was used to validate in-silico the findings obtained from ClusterONE. Jerarca though computationally more expensive – a typical feature of hierarchical clustering algorithm, compares reasonably well to other algorithms in the same category Both these algorithms have internal parameters built into them that ensure that the quality and accuracy of clusters compare favorably to those found in the literature. These two clustering methods will be overviewed in the forthcoming sections

3.2 ClusterONE Algorithm

The ClusterONE algorithm (Nepusz et al., 2012) detects potentially overlapping protein complexes in protein-interaction networks. Typically, a protein complex within a network is a group of nodes (proteins) that are densely connected to each other, as compared to the rest of the network. ClusterONE algorithm explores this inherent property of protein complexes to identify overlapping network clusters.

The algorithm implements a function called ‘cohesiveness’ of nodes to identify the quality of the nodes included in the complex. Some basic terminology used by the algorithm to determine cohesiveness of nodes is explained briefly as follows:

Consider the graph G in Figure 3.1 below with a group of vertices V_0 (the shaded gray region)

The vertices of V_0 are called *internal vertices*, while the vertices not included in V_0 are called *external vertices*. The edge between two internal vertices is called an *internal edge*, while an edge between an internal and external edge is called the *boundary edge*. The edge between two external vertices is called the *external edge*. An *internal boundary vertex* is an internal vertex that has at least one boundary edge incident on it. An *external boundary vertex* is an external vertex that has at least one boundary edge incident on it.

The ClusterONE algorithm proceeds through three distinct stages:

Step 1: Beginning from a single seed vertex – a protein with the highest degree, the algorithm greedily adds or removes vertices to find groups of nodes with high cohesiveness (See below).

Step 2: The next step, measures the extent of overlap between each pair of node groups and merges those groups that have an overlap score greater than a specified threshold (an overlap threshold score of 0.8 was used for our analysis). The overlap score between two protein sets A and B is given as follows (Bader and Hogue, 2003):

$$\omega(A, B) = \frac{|A \cap B|^2}{|A||B|}$$

Basically, ClusterONE calculates the overlap scores of each pair of protein set and builds an overlap graph in which each vertex represents a cohesive group and the two protein sets are connected by an edge if they have an overlap score more than the overlap threshold. Groups of proteins that are connected to each other directly through an edge or indirectly through a path of edges are merged as protein complex candidates.

Step 3: In this step, the algorithm discards those nodes from complexes that have size less than 3 proteins or whose density δ is less than a given threshold (0.3 in this case for weighted networks). The density of a complex with n proteins is defined as the total weight of its internal edges, divided by $n(n-1)/2$.

3.2.1 Cohesiveness of nodes:

The algorithm uses the concept of cohesiveness (Nepusz et al., 2012) to greedily build groups of proteins in the PPI network. Cohesiveness measures how likely the group of proteins can form a protein complexes.

Let $w^{in}(V)$ be the total weight of edges within a group of proteins V , and $w^{bound}(V)$ be the total weight of edges that connect the group with the rest of the network. Then the cohesiveness of V is given by

$$f(V) = \frac{w^{in}(V)}{w^{in}(V) + w^{bound}(V) + p|V|}$$

Where, $p|V|$ is a penalty term that models the uncertainty in the data by presuming the existence of yet undiscovered interactions in the PPI network. The penalty term is set to 2 for the current analysis. Cohesiveness is a simple and efficient way to evaluate how well a group of proteins aligns with respect to its w^{in} and w^{bound} property. An increase in cohesiveness indicates two kinds of subgraphs: a subgraph with many reliable edges and hence a high w^{in} or a well-separated subgraph with a low w^{bound}

The algorithm retains computational efficiency during its implementation by maintaining two variables (w_i^{in} and w_i^{out}) for every protein i in the network. If we consider V_t to denote the cohesive subgroup in step t , then,

w_i^{in} denotes the total edge weight that connect protein i with members of V_t , and

w_i^{out} denotes the total edge weight that connect protein i with non-members of V_t .

A *boundary-set* of V_t is also retained by the algorithm to calculate the cohesiveness measure. The cohesiveness after adding protein i , is calculated the above preserved current variables w_i^{in} and w_i^{out} of V_t and is given by:

$$f(V_t \cup \{i\}) = \frac{w^{in}(V_t) + w_i^{in}}{w^{in}(V_t) + w^{out}(V_t) + w_i^{out} + p(|V_t| + 1)}$$

3.2.2 Evaluating the quality of detected complexes:

Since the amount of information contained in disease proteomes is still incomplete, it is difficult to construct a gold standard to compare predicted complexes against. Therefore, in such cases, the quality of the predicted complexes can be assessed by determining the cellular localization of its protein candidates.(Jansen et al., 2003)

If the protein members within a complex are localized in the same cellular compartment, then it is highly likely that they are members of the same complex. ClusterONE uses a ‘co-localization score’(Friedel et al., 2009) using localization annotations of yeast proteins and a standard overrepresentation analysis of biological process, molecular function and cellular component terms from the Gene Ontology to evaluate the biological significance of predicted complexes. The significance levels of the p-values of this overrepresentation analysis were adjusted using the Benjamini-Hochberg method.(Benjamini and Hochberg, 1995)

The quality of clusters is also evaluated by calculating the p-value of a one-sided Mann-Whitney U test performed on the in-weights and out-weights of the vertices. A low p-value indicates that the in-weights are significantly larger than the out-weights, and hence it is more likely that the cluster is a valid finding and not the result of random fluctuations. While internal quality indices such as Dunn’s index or Silhouette index are used to measure quality of clusters, these methods are used to cluster points in a high-dimensional space where a sensible distance or similarity matrix can be defined. Since ClusterONE is a graph clustering algorithm, the notion

of ‘distance’ does not apply in our particular case. Hence, based on communication with the author of the algorithm, the p-value of the detected complexes was decided to be a suitable internal validation measure to evaluate the quality of a cluster in the graph,. (Nepusz, 2016)

3.2.3 Results using ClusterONE

An implementation of the ClusterONE algorithm on the Huntington’s disease integrome revealed 3065 overlapping protein complexes out of which 48 complexes were found to be statistically significant in terms of their p-value. (p-value < 0.05). 12 complexes out of 48 were further chosen for functional enrichment analysis. The members of these complexes are shown in Table 3.1

3.2.3.1 Gene Ontology Functional Enrichment

The g:GOST tool.(Ashburner et al., 2000; Reimand et al., 2016; The Gene Ontology Consortium, 2015) was chosen for detecting the cellular functions enriched in the selected 12 complexes.

The primary interactors within the HDI (n = 578) are directly connected to each other through biological and molecular processes representing various protein transport and mitochondrial functions

3.2.3.2 *DNA repair and chromosome condensation related functions enriched in HDI*

Among the 12 statistically significant complexes (p-value<0.05), we found four modules specifically enriched for functions related to DNA repair and chromosome condensation. We name these modules as VCP complex, PSMC complex, YWHAG complex and CCT complex for convenience (Figure 3.2). The members of these four complexes were functionally enriched for molecular processes such as “DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest” (GO:0006977, p-value = 9.7E-49), “G1 DNA damage checkpoint” (GO:0044783, p-value= 1.3E-46), “cellular response to DNA damage stimulus” (GO:0006974, p-value = 3.8E-28) and “DNA repair” (GO:0006281, p-value= 0.00549) (Table S5). More interestingly, biological functions closely related to chromatin cohesion and chromosome organization such as “negative regulation of sister chromatid cohesion” (GO:0045875, p-value = 0.00016), “chromosome organization” (GO:0051276, p-value =0.00582) were also found to be enriched within three of the four complexes. (Figure 3.2) (Table 3.2). A detailed list of the enriched GO terms and Reactome pathways and their corresponding gene lists is given in Table 3.2 and Table 3.3 respectively.

mHtt interrupts the processes of DNA damage repair including G2/M and intra-S phase cell cycle checkpoints (Yuan et al., 2007) and chromatin condensation (Fatoba and Okorokov, 2011) by inhibiting SIRT1 deacetylase activity. Ku70, a component of the DNA damage repair complex, is also targeted by mHtt for impairing its DNA-dependent protein kinase function (Enokido et al., 2010). Considering these and several additional information, we have recently proposed that mHtt could affect cell cycle checkpoint regulation related to chromatin condensation defects, chromosome breaks and abnormal DNA repair processes, including those involved in ribosomal DNA condensation pathways.(Lokhande et al., 2016) GO enrichment results described here are generally supportive of this proposal.

Interestingly, functional GO enrichment of the first significant protein complex (n = 105) revealed enrichment for various GO terms related to mitochondrial functions and processes. Some of the significantly enriched terms include - “mitochondrial translation (GO:0032543, p value = 1.35×10^{-54}), “mitochondrion organization” (GO: 0007005, p value = 1.7×10^{-47}), “mitochondrial electron transport, NADH to ubiquinone” (GO:0006120, p-value = 2.5×10^{-11}). These findings recapitulate observations from various animal model studies that have found the role of functions such as maintenance of electron transport chain (Hausladen and Fridovich, 1994; Tabrizi et al., 1999, 2000), mitochondrial organization, biogenesis

(Chakraborty et al., 2014; Damiano et al., 2010; Johri et al., 2013) and mobility (Chang et al., 2006; Orr et al., 2008; Trushina et al., 2004) in HD.

3.2.4 Scalability of ClusterONE

To ensure the algorithm we used for clustering HDI was computationally efficient for other datasets of comparable size or larger, we ventured to determine the scalability of ClusterONE to larger datasets.

Before we tested if the method was scalable, we attempted to find the Big O notation of the method. Considering that this algorithm forms initial clusters based on the cohesiveness measure of each node in the graph, we assumed that the initial step would require a total of n^2 operations. The next step involves measuring the extent of overlap between each pair of node groups formed in the first step. This step is dependent on the number of groups created in the previous stage and hence it becomes difficult to formally compute the complexity of the algorithm and assign a definite BigO notation to the method. This conclusion was drawn after communication with the author of the algorithm and indicates that the runtime of the algorithm depends on the exact structure of the network it is trying to cluster. While there might be artificially constructed graphs that might make ClusterONE run slowly, these graphs will most likely not be a representative of ‘real’ world datasets that the algorithm will be confronted with.

We therefore proceeded to analyze the run time of the algorithm by implementing it on the entire Human protein-protein interaction (PPI) network from BIOGRID. The run time on this network with 19634 unique proteins and 270,970 non-redundant interactions was found to be 2 mins 40 secs.

ClusterONE has a scaled-up version of its algorithm that was obtained from the author of this technique. This version parallelizes the cluster growth phase of the algorithm by using multiple CPU cores. We implemented this scaled version on two large datasets –

- The Amazon graph - 863 nodes and 925,872 edges
- The YouTube graph – 1,13,4890 nodes and 2,98,7624 edges

The run time of the scaled-up version of ClusterONE on the Amazon graph and the YouTube graph was 7.23 seconds and 2.3 minutes (141 seconds) respectively. Details of the graph properties and the run times are given in Table 3.4. The above scalability studies were performed using a Window PC laptop with the following processor:

Processor – Intel(R) Core(TM) i7 – 6500 CPU @ 2.50GHz 2.59 GHz, 2 Cores, 4 Logical Processors (x64-based processor)

3.2.5 Robustness of ClusterONE

To demonstrate the robustness of the algorithm used for clustering HDI, we subjected the HDI network to the following perturbations:

- Dropping edges weights below a threshold score of 0.5 (~10% edges), 0.6 (~20% edges) and 0.63 (~50% edges)
- Rewiring the edges and edge weights within a range of 20% - 100% while maintaining a constant degree distribution of the graph.

For each perturbation, we noted the number of statistically significant complexes detected, and calculated the significance of overlap (p-value) of the member of these complexes with the statistically significant complexes detected from HDI. A background genome size of 19000 was used for calculating the overlap significance. A Jacquard's index and Odds ratio was also determined for each perturbation. We find that the performance of the algorithm starts showing signs of degradation when we drop edges with a score of 0.6 and less (~ 20% of the edges). The network fragments into smaller connected components as the percentage of dropped edges increases to 50% which is also indicated by a drop in the clustering coefficient of the graph. (Figure 3.3) On the other hand, rewiring just 20% of the network edges and edge weights, lead to sharp drop in the Jacquard's index and Odds ratio which also indicates a significant departure from the clusters obtained from HDI. Such a performance degradation is however expected considering that the HDI graph is a weighted graph and contains carefully curated and scored interactions between nodes. A perturbation via rewiring essentially alters the biological information contained in the graph and inadvertently introduces a high number of false positives in the process leading to performance degradation.

Table 3.5 summarizes the results of robustness analysis. We conclude that while ClusterONE can handle a drop of ~20% of weighted interactions comfortably without affecting the connectedness of the graph, its performance significantly degrades after 20% or more of HDI edges and edge weights are rewired.

3.2.6 Reproducibility of results

To ensure the results obtained using ClusterONE are reproducible, the Rscripts and step by step code used to bring about this analysis has been documented in the form of Jupyter notebooks and have been uploaded on GitHub. This will enable users to run these scripts on the data themselves and verify the results.

The above clustering approach using ClusterONE can find multiple applications in other PolyQ disorders such as spinocerebellar Ataxia, dentatorubral-pallidoluysian atrophy (DRPLA) and spinal and bulbar muscular atrophy, X-linked 1 (SMA1/SBMA). Other neurodegenerative disorders such as Parkinson's disease and Alzheimer's disease can also find this algorithm useful for analysis. Since the algorithm creates clusters based on a cohesiveness measure, any biological data which depicts relationships between two or more proteins, genes, cellular components, functions or processes could serve as a potential source of input. We however believe that assigning weights to the edges and carefully curating the input data allows for elimination of noise to a reasonable extent which has positive implications on the algorithm output as evident from our HDI analysis.

3.2.7 Jerarca algorithm

We implemented a different clustering approach on the HD integrome, and performed an in-silico validation of the results obtained using ClusterONE before we proceeded find *in-vitro* and *in-vivo* support for any of the findings. We specifically looked at the hierarchical clustering algorithm Jerarca to understand the hierarchy of HDI.

3.2.7.1 Description of the Jerarca algorithm:

Jerarca (Aldecoa and Marín, 2010) is a suite of three hierarchical clustering algorithms that converts a PPI network into dendrograms through iterative hierarchical clustering. The suite consists of three algorithms namely UVCluster (a modified version), RCluster and SCluster. For purposes of analysis of the HDI, we have implemented the modified more efficient version of UVCluster to reveal the hierarchical structure of the network.

The suite implements the following steps to detect a hierarchy in the PPI network.

1. The program reads an input file to create an adjacency matrix A of the PPI graph; where $A_{ij} = 1$ if vertices i and j are connected and $A_{ij} = 0$ if they are not.
2. An iterative UVCluster algorithm (please refer below for details) is run depending upon the number of nodes in the network. The ideal number of

iterations suggested by authors is approximately 10 times the number of nodes in the network. We have used 70000 iterations for our analysis.

3. Step 3 - A matrix of secondary distances is calculated. For each pair of nodes, the algorithm saves, the number of iterations for which the nodes have been separately clustered. The secondary distances are then computed by finding the ratio of these values to the number of iterations.
4. Step 4 - A dendrogram is built using one of the two phylogenetic algorithms – UPGMA and Neighborhood-joining. We have used the UPGMA to build a dendrogram from the secondary distance matrix.
5. Step 5 – The resulting dendrogram is evaluated at each level by using two indices – modularity (Q) and the Surprise (H) score. The optimal partition of the tree is saved for both the indices.

The workflow of the Jerarca suite considering UVCluster is given in Figure 3.4

3.2.7.1.1 UVCluster Algorithm

The modified version of UVCluster algorithm(Arnau et al., 2005) iteratively uses agglomerative hierarchical clustering using the primary distance matrix of the PPI graph. Based on the number of iterations (N) specified by the user before the analysis starts, the algorithm generates N clustering solutions by randomly sampling elements of the dataset. The elements in the dataset are clustered using the average linkage method (Everitt et al 2001). Another parameter that is set by the user besides the number of iterations is the Affinity Coefficient (AC) which is

the global stopping parameter for the agglomerative process. The AC is given by the following equation:

$$AC = 100[(P_m - C_m) / (P_m - 1)] \text{ where,}$$

C_m is the cluster mean which is the average of distances for all the members in the clusters, and,

P_m is the partition mean which is the average of the distances for the entire set of selected proteins.

If $C_m = 1$, then $AC = 100$, which signifies that the proteins clustered together have a distance equal to 1.

A modified version of UVCluster fixes the AC value as 100 (maximally strict) for the Jerarca suite of algorithms. This value is suitable for PPI networks wherein many proteins are directly connected to each other and have shorter average path lengths. The Huntington's Disease Integrome (HDI) has an average path length of 3.54 and hence an AC value of 100 was deemed appropriate to find protein clusters in the network. A strict AC value essentially helps to find clique-like subgraphs within the network.

After obtaining N clustering solutions in the previous step, the algorithm proceeds to calculate a secondary distance matrix as described in the section above. Such a secondary distance matrix displays the strength of connection between all pairs of

elements in the dataset. This distance matrix serves as an input to build a dendrogram.

3.2.8 Results using Jerarca

The time complexity of the Jerarca algorithm (UV cluster implementation) with an AC of 100 is $O(n^2)$ (Aldecoa and Marín, 2010) Hence we ran this algorithm using a Trial allocation of 1000 SUs on the Comet cluster of the San Diego Supercomputing Center (SDSC). The run time of the algorithm was 4 hours and 41 minutes.

Jerarca partitioned the HDI network into 592 distinct non-overlapping clusters (Figure 3.5) with a modularity (Q) score of 0.31. Examination of the dendrogram output showed results that were coherent with the output of ClusterONE. Genes such as FANCI, WAPAL, ESPL1, TUBGCP2 and ANAPC7, which are implicated in chromosome condensation related functions, were seen to assemble in a single module (Figure 3.5). Similarly, genes such as ERCC3, UCHL5, SHFM1, NPLOC4, UFD1L and COPS2 that are related to DNA damage response and DNA repair were found to be clustered in a separate module. Consistent with results from ClusterONE, we also found several genes related to mitochondrial functions also clustered together in a separate module. Such functional segregation in modules is expected. considering that the hierarchical algorithm does not output overlapping clusters. These results using an entirely distinct approach recapitulate the

observations from ClusterONE and lend an additional in-silico support to the biological quality of data obtained.

We now proceeded to closely examine the clusters obtained using ClusterONE.

3.3 HD integrome reveals novel genes that link chromosome condensation defects with Htt toxicity

The HDI dataset that yielded a possible link between Huntington's disease protein and chromosome condensation/DNA damage repair pathways was only indirectly derived from yeast genetic interaction data. A fraction of these interactions contained suppressor and enhancer mutations of mHtt toxicity in yeast. In addition, they contained 41 ribosomal gene dosage suppressors of mHtt toxicity. Previous studies with these interaction datasets did not reveal a noteworthy enrichment of chromosome condensation or DNA damage response pathways among suppressors or enhancers of mHtt toxicity. We have recently generated a large collection of gene-dosage suppressors of lethal mutations in yeast (Patra et al., 2017), which presented an opportunity to ask whether there might be any indirect connection between suppressors of chromosome condensation defects and genes known to modify mHtt toxicity.

3.4 Validation in a yeast HD model.

We noticed four yeast homologs among two of the 12 statistically significant complexes within HDI, detected using ClusterONE, which are known to function

in chromosome dynamics. These four interactors are *SPC24*, *MSH4*, *NDC80*(*TID3*) and *NMD3*. Among these, Spc24p and Ndc80p are part of core kinetochore protein complex involved in chromosome segregation. This provoked us to hypothesize that *SPC24* and *NDC80* might genetically interact with mHtt in yeast. However, loss of function mutations in these two genes were not previously recovered as suppressors or enhancers of mHtt toxicity in previous studies. Therefore, we hypothesized that gain of function (overexpression) could potentially modify the phenotype of mHtt expression.

We tested this hypothesis by directly overexpressing *SPC24* gene under the control of a galactose-inducible promoter on a multicopy yeast plasmid (see Methods). This plasmid was introduced into a yeast strain that expresses a chromosomal copy of mHtt encoding the N-terminal (103Q), also under the control of a galactose-inducible promoter. If grown on glucose, these cells do not produce mHtt. However, in the presence of galactose, these cells produce both mHtt and Spc24p proteins. A control strain contains exactly the same constructs except that the N-terminal fragment of normal Htt is produced in the presence of galactose. Overexpression of *SPC24* was found to suppress mHtt toxicity (Figure 3.6).

Incidentally, *SPC24* was also found to be among the set of genes that suppresses smc2-8 chromosome condensation defective mutant.(Patra et al., 2017) These

findings led us to hypothesize whether overexpression of *smc2-8* suppressors could suppress mHtt toxicity (Figure 3.7). As a test of this hypothesis, we examined if a set of suppressors of *smc2-8* mutation (including *SMC2*), and a selected set of cohesin/condensin genes could suppress mHtt toxicity in yeast. We examined 36 suppressors of *smc2-8* mutant and seven condensin/cohesion related genes, which were not previously described as modifiers of *smc2-8*. Results showed that approximately 50 percent of these genes (23 of 43 tested), suppressed mHtt toxicity. The panel that tested for SPC24, is shown in Figure 3.6. Since these 23 genes are all known to function in chromosome condensation/cohesion processes, these results are consistent with the hypothesis that mHtt toxicity is related to these processes.

3.5 Identification of candidate genes

Emboldened by the results with yeast genes, we ventured to look for genes with similar functions in HDI because we thought chromatin condensation/dynamics related genes in humans might indeed play important roles in HD. We specifically chose a subset of 27 candidate genes (Table 3.6) within the first 12 statistically significant ClusterONE complexes with the additional criteria that they have orthologs in *Drosophila*. Among these genes, FANCI, WAPAL, ESPL1, TRIP12, ANAPC7, UBC, NDC80, TUBGCP2, PPP6C were functionally enriched for processes such as chromosome organization, regulation of chromosome

segregation, mitotic cell cycle and cellular response to DNA damage. These findings are significant considering their enrichment is represented across all the three aspects of GO classes (biological process, molecular function and cellular component) and the Reactome pathways (Table 3.2 and Table 3.3)

3.6 Candidate gene validation in a *Drosophila* HD model

The validation experiments in this section were carried out in collaboration with a different laboratory and is not a contribution by the author of this thesis. For purposes of validation, two fully balanced lines: one that stably expresses the pathogenic Htt variant (128-Q) in *Drosophila* eye (under the eye-specific promoter), and one that stably expresses the non-pathogenic Htt variant (16-Q) that serves as one of the controls was generated. Each of these lines was crossed with an RNAi, overexpression and/or deletion line from the list of 27 potential interacting genes. The phenotype of the double transgenic flies was assessed and compared to the driver-alone control and to the eye expressing the non-pathogenic Htt variant. Any suppression or enhancement of the phenotype strongly suggests a genetic interaction.

Using this approach, it was found that a loss of function of NPLOC4 (*Npl4*) gene in *drosophila* suppresses mHtt effects in the eye. A loss of function of TUBGCP2 (*Grip84*) gene was also found to suppress mHtt effects in the *drosophila* eye. On

the other hand, a loss of function of the NLRC4 (*Diap2*) gene was found to enhance the mHtt effects in the drosophila eye.

NPLOC4 is known to form a complex with VCP and UFD1L, which in turn complexes with FAF1 receptor to promote Endoplasmic-reticulum associated degradation (ERAD) of polyubiquitinated proteins. (Lee et al., 2013) Additionally, NPLOC4 is a crucial component of the Cdc48/p97–Ufd1–Npl4 complex which negatively regulates Aurora B early in the mitosis of human somatic cells. A depletion of the Ufd1–Npl4 by using siRNA is known to cause defects in chromosome alignment and in the anaphase. (Dobrynin et al., 2011) These findings lend support to our assumption that NPLOC4 could be an intermediate partner influencing the action of mHtt. This makes NPLOC4 an attractive therapeutic candidate for Huntington's disease. TUBGCP2 is component of the Gamma-tubulin complex and is necessary for microtubule nucleation at the centrosome (Murphy et al., 1998) However, there have been no studies reported so far documenting its interaction with the Htt protein and its downstream effects. NLRC4 encodes a member of the NLR family and contains the caspase recruitment domain. It is known to be essential in eliciting an innate immune response to a wide range of tissues and organisms, thus playing an important role in tissue damage and cellular stress. (Kitamura et al., 2014; Romberg et al., 2014; Thalappilly et al., 2006) While the role of caspases in apoptosis and mitochondrial dysfunction in HD has been studied, its role in causing an auto-inflammatory response leading to

cellular stress and damage remains to be fully explored. The above findings from validation experiments throw light on this portion of NLRC4 in causing apoptosis and neuronal cell death in HD.

3.7 Conclusion

In this study, we have integrated and analyzed a diverse set of human orthologs of Htt interacting proteins obtained using both physical and genetic studies. The resulting HDI captured a brief snapshot of the HD pathogenetic process and helped us identify clusters of genes that were overexpressed for functions related to chromosome condensation, DNA damage and DNA repair.

In particular, we have identified 27 candidate genes as potential targets that can alter or modify the progress of HD. Three of these candidate genes NPLOC4, TUBGCP2 and NLRC4 have been successfully validated in a drosophila model of HD and could be considered as potential therapeutic targets to alter the course of HD. Our results demonstrate and support our hypothesis that mHtt affects and alters the processes related to chromosome condensation and DNA repair eventually leading to cell death.

3.8 Tables and Figures

Table 3.1 Members of the first 10 statistically significant complexes detected in HDI by ClusterONE

Complex No.	Member of complex
1	MUT,ADCK3,PNPT1,ACADVL,SUPV3L1,C10orf2,FARS2,MTPAP,TFAM,CPT2,MCU,N DUFS1,PCCB,AFG3L2,TMLHE,NDUFS2,DBT,PYCR2,NME4,NDUFV1,YARS2,GLDC,N DUFS3,CLPX,AARS2,MTERF4,ACOT9,AASS,ALDH1B1,ALDH1L2,ATAD3B,C6orf203, CECR5,CHCHD1,CLPTM1,SLC25A10,ECH1,ECI1,FASTKD1,FASTKD2,GADD45GIP1,S LC25A18,GRSF1,GTPBP10,GRPEL1,MTIF2,ICT1,ISCA1,MALSU1,MMAB,METTL17,R NMTL1,MTERF3,PMPCB,MRM1,MTG1,NDUFS8,NDUF3F3,NDUFV2,NGRN,NDUFA2, NDUFA9,NDUFS7,NSUN4,NOA1,POLDIP2,PTCD1,PUSL1,RBFA,MRPL21,MRPL28,M RPL48,MRPL3,MRPL27,MRPL32,MRPL18,MRPL19,MRPL54,MRPL39,MRPS11,MRPS1 5,MRPS18A,MRPS22,MRPS6,MRPS23,MRPS25,DAP3,MRPS24,MRPS27,RPUSD3,MRP S7,MRPS17,MRPS18B,MRPS31,HARS2,PARS2,VAR2,SYNJ2BP,TARS2,RARS2,TFB1 M,THNSL1,TEFM,TRUB2,SLC25A1
2	VCP,RAD23A,RAD23B,PSMD4,UBC,ANXA5,ANAPC7,ARFGAP2,ARIH1,ATXN3,BAI AP2L1,CDC42EP1,BRAT1,UBQLN1,CAAP1,CDKN2AIP,C3orf17,CCDC132,CCDC134,C DK2AP1,CENPH,CIDEC,KIAA1524,CNOT10,COG5,COMMD6,COMT,HSP90B2P,ESPL 1,FANCI,G3BP2,TUBGCP2,GOLPH3L,HS1BP3,H2AFJ,HAUS1,HEATR1,HELLS,HOO K1,HSBP1,INF2,NPLOC4,LRIG1,MAP7D3,NGLY1,NMD3,DDIAS,PPP6C,UFD1L,PTCRA, UBE4B,POLR3C,SPAST,SPC24,TAF6L,TBC1D10B,TBC1D9B,SCD,CLN6,DCAF11,MSH 4,RHBDL3,UBE2J1,UBXN11,GTF3C3,GTF3C5,TMEM33,TRIP12,ULK3,VIL1,WAPAL
3	NAPA, SNAP25, STX1A, ANKRD35, STXBP1, STX1B, STX2, CCSER2, KIAA0319L, LAMA4, MYH7B, CPLX1, SLC6A5, SLC6A9, SCNN1A, STXBP2, SNAP23, UNC13B, VAMP8, STMN4, STX17, CAPG, CCDC93, FAM161B, FGB, FUBP3, SLC6A2, SCRT1,

Complex No.	Member of complex
	SNPH, STXBP5, TTC3, TXLNA, TXLNB, ZNF189, ZNF226, ZNF254, ZNF526, ZNF799, VAMP2, TRIM14, TSPAN7, TXLNG, UACA, VAPB
4	YWHAZ, YWHAH, HDAC4, HDAC7, MARK2, NEDD4L, PARD3, MST1R, MARK1, MARK3, PARD6G, PRKCI, CDC25B, PARD6A, SFN, BSPRY, EFNB3, KRT19, LDB1, PDE1A, SNX24, POTEKP, ACTN3, ADRA2C, ARL6IP1, ATL2, ATL3, CTH, CHTOP, CIC, DCAF7, TSFM, HIST2H2BF, EIF5B, HECTD4, CMPK1, WWC1, MEF2C, MYH3, NAPSA, NUFIP1, PGLYRP1, PRIM2, ADSSL1, REEP6, RPRD1A, SMC5, SUPT6H, SRGAP1, SRSF8, TNS1, SNRPD2P1
5	ADRM1, UCHL5, RAD23A, RAD23B, PSMD4, PSMA1, PSMC4, PSMC2, PSMD3, SHFM1, PSMD6, CIITA, PSMC5, PSMD2, COPS2, PSMD8, PSMA2, PSMD13, EPHA8, ERCC3, POLR2M, HERPUD1, PSMD7, NDC80, NLRC4, PCDH10, GTF2F1, USP14, ZCCHC8, PSMC6, PSMC1, PSMC3, HTR1E, ATG4C, PAAF1, PSMB2, PSMD10, PSMD5, RIOK3, POLR2M, MYZAP, PSMD14, ESRRG, HNF4G, MYO18B, PTGS2, PLEKHO1, PSMB9, PSMD1, RORA, PSMA8, PSMB6, PSMB8, PSMD11, PSMD12, HMOX1, JKAMP, NUB1, ACTR3B, TSPYL2, RARB, RARG, SUGT1, ST18, TEK4, UBLCP1, XBP1P1
6	CTNNB1, CTNNBIP1, CDH1, CTNNA1, JUP, CDH10, CDH17, CDH18, CDH5, CDH6, AJAP1, BOC, CDH11, CDH7, CDH8, CDH9, CDON, DLG5, JRK, LEF1, PCSK1, NEURL2, PROP1, TCF7L2, UHRF2, VEZT, CTNND1, FOXO4, SOX6, TAX1BP3, BCR/ABL
7	YWHAG, YWHAH, HDAC7, MARK2, NEDD4L, PARD3, MARK1, MARK3, SIMC1, FARP2, HOXC10, LATS2, MAP3K6, PRPF4B, RMDN3, SFN, CAMSAP2, CCS, CKAP2, EPN2, FAM65B, FRY, INPP5E, JAKMIP1, KRTAP195, KRT34, KRT35, KRT37, KRT38, KRT82, LRCH3, LTB4R, MPHOSPH9, NELFE, PRLR, TRIM21, SHCBP1, SHKBP1, SHPRH, SYNPO, USP37, VANGL2, ATP6V0B, IGKV1-12,170549

Complex No.	Member of complex
8	PPP2R2B, HDAC3, RFWD2, PPP2R2A, ATG16L1, CCT7, CCT8, CCT4, CCT5, CCT2, TCP1, TBK1, PPP2R2C, METTL20, METTL21B, RPAP1, PPP2R4, PPP4C, CCT6A, PPP2CA, BBS7, GCNT1, GPN1, MKKS, PACRG, IGBP1, KDM5A, MLST8, MED31, PPP2CB, MYBPC2, MOB4, STRN, STRN3, STRN4, PPP2R2D, BBS10, CTTNBP2, DOCK5, FAM86B2, GPR37, TRAF3IP3, THEG, CCT3, CCT6B, IMPA2, MLX, PARP4
9	GNB5, GNG2, PDCL, GNB2, GNGT1, GNG10, GNG13, GNG3, GNG4, GNG5, GNG7, GNGT2, RGS6, RASD2, GNG12, GNB3, GNB4, GNG8
10	GNB5, GNG2, GNB1, GNB2, KCNJ3, GNG10, GNG11, GNG13, GNG3, GNG4, GNG5, GNG7, GNGT2, GNG12, GNB3, GNB4
11	CTBP1, ACTL6B, HIC1, MECOM, LCOR, ZEB1
12	CTBP1, EHMT1, HIC1, MECOM, LCOR, ZEB1

Table 3.2 List of GO terms related to DNA repair, DNA damage and chromosome condensation found to be enriched in PPI network complexes

GO term ID	p-value	No. of genes	GO type	GO term description	List of genes
GO:0006977	9.70E-49	26	BP	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	PSMC4, PSMC5, PSMD5, PSMD8, PSMC6, PSMC1, PSMD10, PSMD7, PSMA2, PSMD3, PSMD11, PSMD14, PSMB2, PSMA1, PSMB6, PSMA8, PSMD4, PSMC2, PSMD6, PSMC3, PSMD1, PSMD2, PSMD13, PSMD12, PSMB8, PSMB9
GO:0072431	1.60E-48	26	BP	signal transduction involved in mitotic G1 DNA damage checkpoint	
GO:0044783	1.30E-46	26	BP	G1 DNA damage checkpoint	
GO:0006974	3.80E-28	33	BP	cellular response to DNA damage stimulus	PSMC4, PSMC5, PSMD5, PSMD8, HMOX1, PSMC6, PSMC1, PSMD10, PSMD7, PSMA2, PSMD3, PSMD11, PSMD14, UCHL5, RAD23B, PSMB2, SHFM1, PSMA1, PSMB6, PSMA8, PSMD4, PSMC2, ERCC3, PSMD6, PSMC3, COPS2, PSMD1, PSMD2, RAD23A, PSMD13, PSMD12, PSMB8, PSMB9
GO:0006974	5.90E-05	13	BP	cellular response to DNA damage stimulus	MSH4, ATXN3, UFD1L, BRAT1, RAD23B, FANCI, UBC, TRIP12,

GO term ID	p-value	No. of genes	GO type	GO term description	List of genes
					PSMD4, VCP, CDKN2AIP, RAD23A, NPLOC4
GO:0045875	0.00016	2	BP	negative regulation of sister chromatid cohesion	WAPAL, ESPL1
GO:0006281	0.00017	10	BP	DNA repair	MSH4, ATXN3, UFD1L, RAD23B, FANCI, UBC, TRIP12, VCP, RAD23A, NPLOC4
GO:0007084	0.00325	2	BP	mitotic nuclear envelope reassembly	PPP2CA, PPP2R2A
GO:0003684	0.00422	3	MF	damaged DNA binding	RAD23B, ERCC3, RAD23A
GO:0006281	0.00549	7	BP	DNA repair	PSMD14, UCHL5, RAD23B, SHFM1, ERCC3, COPS2, RAD23A
GO:0051276	0.00582	12	BP	chromosome organization	MSH4, WAPAL, ATXN3, RAD23B, HELLS, ESPL1, UBC, CENPH, TRIP12, TAF6L, ANAPC7, H2AFJ
GO:0051985	0.00922	3	BP	negative regulation of chromosome segregation	WAPAL, ESPL1, ANAPC7
GO:2001251	0.0101	4	BP	negative regulation of chromosome organization	WAPAL, ESPL1, TRIP12, ANAPC7
GO:0030472	0.0298	1	BP	mitotic spindle organization in nucleus	PPP2R4

GO term ID	p-value	No. of genes	GO type	GO term description	List of genes
GO:1903047	1.89E-23	30	BP	mitotic cell cycle process	PSMC4, NDC80, PSMC5, NLRC4, PSMD5, PSMD8, PSMC6, PSMC1, PSMD10, PSMD7, PSMA2, PSMD3, PSMD11, PSMD14, PSMB2, PSMA1, PSMB6, PSMA8, PSMD4, PSMC2, ERCC3, PSMD6, SUGT1, PSMC3, PSMD1, PSMD2, PSMD13, PSMD12, PSMB8, PSMB9
GO:0000922	0.00551	4	CC	spindle pole	FRY, CKAP2, RMDN3, LATS2

Table 3.3 List of pathways found to be enriched in HDI network complexes

Pathway ID	p-value	No of genes	GO term description	List of genes
REAC:69563 REAC:69580	1.9E-45	26	p53-Dependent G1 DNA Damage Response/ p53-Dependent G1 DNA damage checkpoint	PSMC4, PSMC5, PSMD5, PSMD8, PSMC6, PSMC1, PSMD10, PSMD7, PSMA2, PSMD3, PSMD11, PSMD14, PSMB2, PSMA1, PSMB6, PSMA8, PSMD4, PSMC2, PSMD6, PSMC3, PSMD1, PSMD2, PSMD13, PSMD12, PSMB8, PSMB9
REAC:24678 13	3.9E-35	27	Separation of Sister Chromatids	PSMC4, NDC80, PSMC5, PSMD5, PSMD8, PSMC6, PSMC1, PSMD10, PSMD7, PSMA2, PSMD3, PSMD11, PSMD14, PSMB2, PSMA1, PSMB6, PSMA8, PSMD4, PSMC2, PSMD6, PSMC3, PSMD1, PSMD2, PSMD13, PSMD12, PSMB8, PSMB9
REAC:74752	1.3E-25	26	Signaling by Insulin receptor	PSMC4, PSMC5, PSMD5, PSMD8, PSMC6, PSMC1, PSMD10, PSMD7, PSMA2, PSMD3, PSMD11, PSMD14, PSMB2, PSMA1, PSMB6, PSMA8, PSMD4, PSMC2, PSMD6, PSMC3, PSMD1, PSMD2, PSMD13, PSMD12, PSMB8, PSMB9
REAC:24659 10	1.3E-05	3	MASTL Facilitates Mitotic Progression	PPP2CB, PPP2CA, PPP2R2D
REAC:75035	1.67E-05	3	Chk1/Chk2(Cds1) mediated inactivation	YWHAH, YWHAG, SFN

Pathway ID	p-value	No of genes	GO term description	List of genes
			of Cyclin B: Cdk1 complex	
REAC:69278	0.000407	9	Cell Cycle, Mitotic	WAPAL, TUBGCP2, ESPL1, UBC, HAUS1, CENPH, PSMD4, SPC24, ANAPC7
REAC:73894	0.000818	7	DNA Repair	UFD1L, RAD23B, FANCI, UBC, VCP, RAD23A, NPLOC4
REAC:5696394	0.00163	3	DNA Damage Recognition in GG-NER	RAD23B, COPS2, RAD23A
REAC:69473	0.00983	3	G2/M DNA damage checkpoint	YWHAH, YWHAG, SFN

Table 3.4 - Scalability analysis of ClusterONE algorithm

Network	Network size	Minimum Complex Size	Node penalty	Minimum Density	Overlap threshold	Run time
Human PPI network – BIOGRID	Nodes: 19634 Edges: 270,970	5	2	0.5	0.8	2 mins 40 seconds
Scaled version of ClusterONE						
Human PPI network – BIOGRID	Nodes: 19634 Edges: 270,970	5	2	0.5	0.8	16.3 seconds
Amazon graph	Nodes: 334,863 Edges: 925,872	5	2	Auto	0.8	7.23 seconds
YouTube graph	Nodes: 1134890 Edges: 2987624	5	2	Auto	0.8	2.3 minutes (141 secs)

Table 3.5 Robustness analysis of ClusterONE

Network name	Perturbation	Edges	Nodes	Percent of interactions dropped	No. of significant complexes	Intersection size of members of significant complexes	Overlapping p-value (overlap significance) against a genome size of 19000	Jaccard index	Odds Ratio
ED network1	3628	27557	7418	11.6	53	$537 \cap 481 = 412$	0 e +00	0.68	891
ED network2	6736	24449	7418	21.6	71	$537 \cap 621 = 364$	0 e +00	0.5	148
ED network3	14888	16297	7418	47.7	98	$537 \cap 722 = 299$	1.5e -299	0.3	53.5
EWS network1	all edges and weights shuffled	31185	7418	0	20	$537 \cap 72 = 38$	6.20E-40	0.06	41.2
EWS network2	20% edges and weights shuffled	31185	7418	0	9	$537 \cap 207 = 38$	2.40E-20	0.05	8.24
EWS network3	40% edges and weights shuffled	31185	7418	0	8	$537 \cap 292 = 58$	1.70E-32	0.075	9.43

Table 3.6 - List of 27 candidate genes with a viable and visible phenotype chosen for validation in a *Drosophila* model of HD

Human Genes	Drosophila Orthologs
UFD1L	<i>Ufd1-like</i>
FANCI*	<i>FANCI</i> *
UBC	<i>Ubi-p63E</i>
TRIP12*	<i>Ctrip</i> *
NPLOC4	<i>Npl</i>
WAPAL*	<i>wpl</i> *
TUBGCP2*	<i>Grip84</i> *
ESPL1	<i>Sse</i>
ANAPC7	<i>APC7</i>
PPP6C	<i>PpV</i>
UCHL5	<i>uch-L5</i>
SHFM1	<i>sem1</i>
ERCC3*	<i>hay</i> *
COPS2	<i>alien</i>
NDC80*	<i>Ndc80</i> *
NLRC4	<i>Diap2</i>
SUGT1	<i>Sgt1</i>
YWHAH	<i>14-3-3zeta</i>
YWHAG	<i>14-3-3zeta</i>
SFN	<i>14-3-3zeta</i>
FRY	<i>fry</i>
RMDN3	<i>CG1575</i>
LATS2*	<i>Wts</i> *
PPP2CB	<i>mts</i>
PPP2CA*	<i>mts</i>
PPP2R2A	<i>tw</i>
PPP2R4	<i>Ptpa</i>

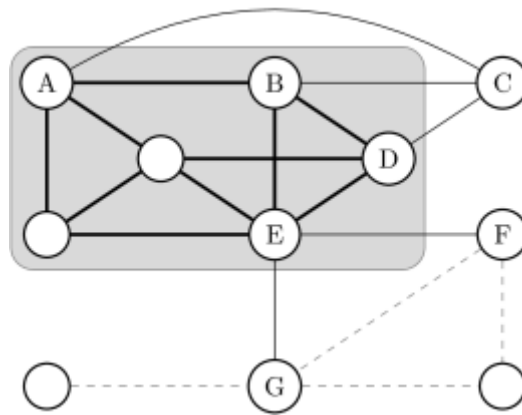


Figure 3.1 An example graph depicting internal and external vertices and edges along with the boundary vertex

The primary interactors within these complexes as colored as follows: genes interacting with wt-Htt (blue), with mHtt (red), with both wt and mHtt (purple).

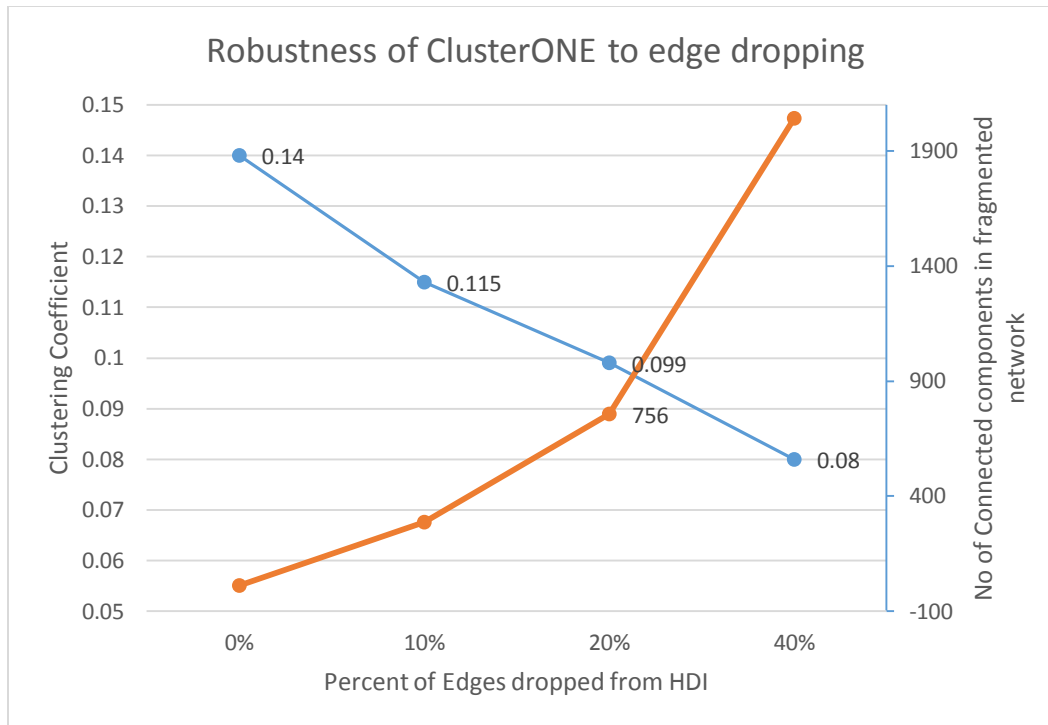


Figure 3.3 Robustness of ClusterONE to edge perturbation.

Dropping edges ranging from 10% to approximately 50% leads to a drop in the clustering coefficient and an increase in the number of connected components in the fragmented network.

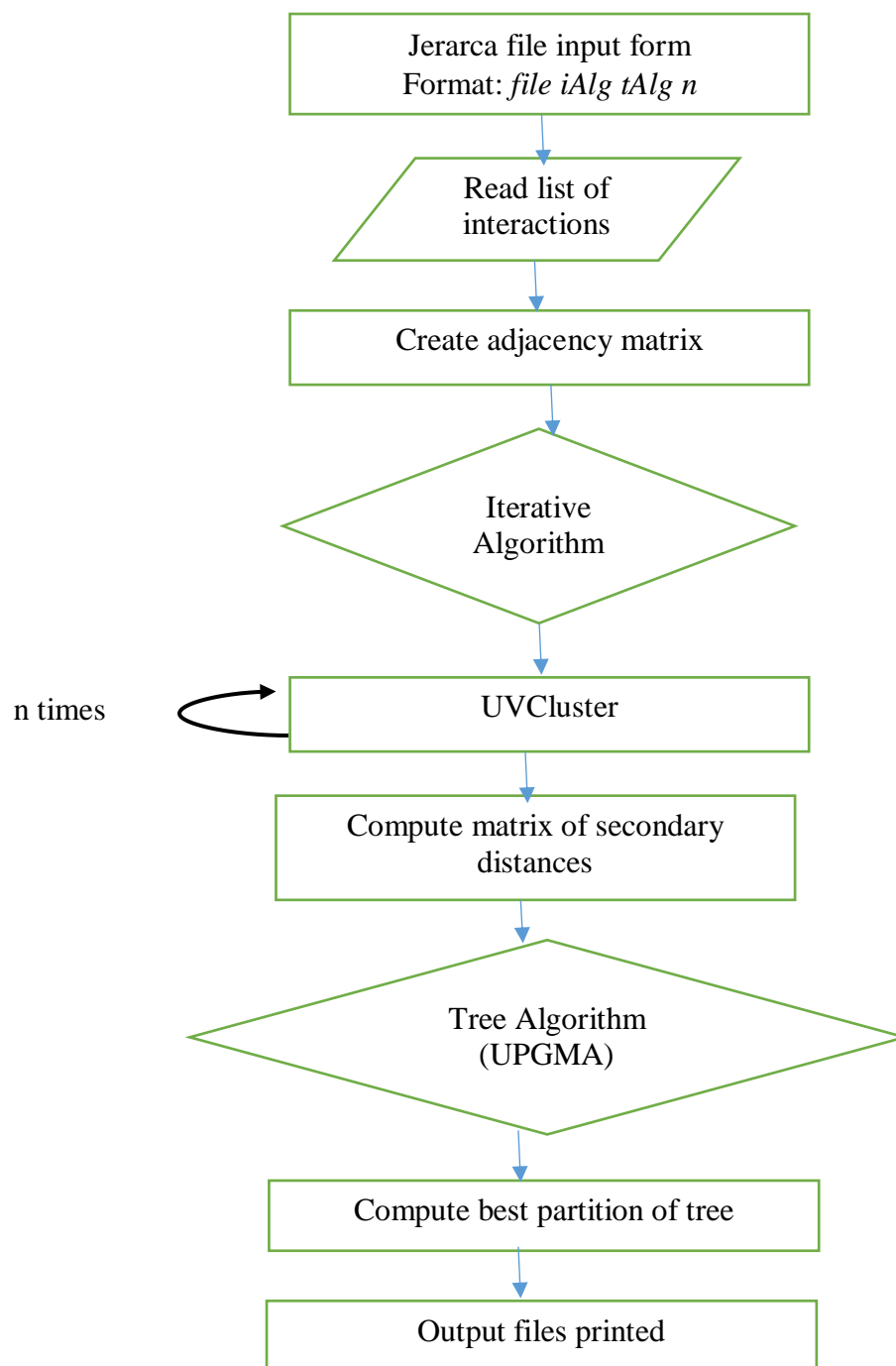


Figure 3.4 An adapted workflow layout of the Jerarca

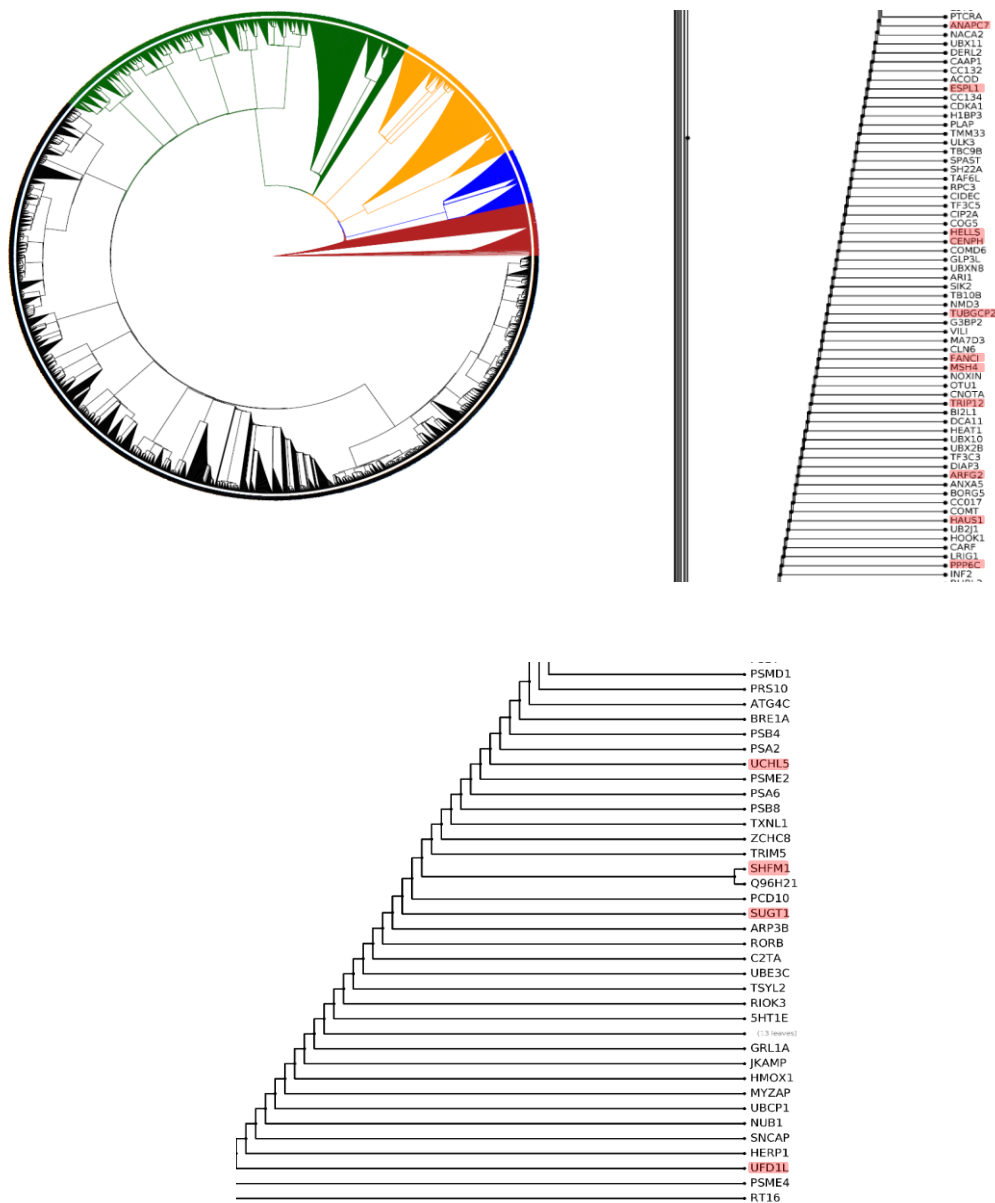


Figure 3.5 Cladogram of HDI hierarchy

The top left panel shows a hierarchical structure of HDI depicted in a circular cladogram. The top right panel shows a module containing candidate genes functionally enriched for chromosome condensation processes, while the bottom panel shows the module containing candidate genes functionally enriched for DNA damage and mitotic cell cycle processes.

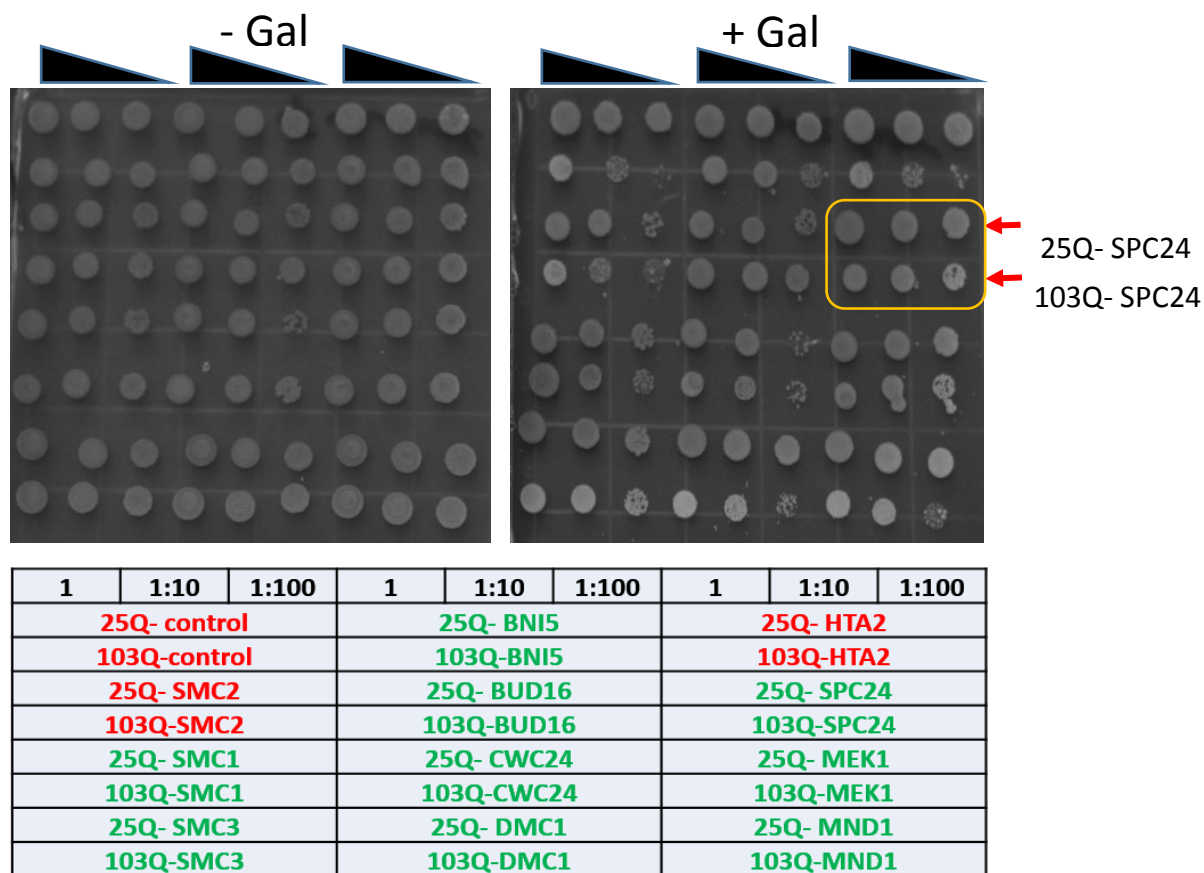


Figure 3.6 Suppressors of *smc2-8* and condensin-cohesin genes suppress the mHtt toxicity (103Q) in yeast.

A group of 36 *smc2-8* mutant suppressors and seven condensin/cohesion related genes were tested. 23 genes were found to suppress mHtt (103Q) toxicity in yeast

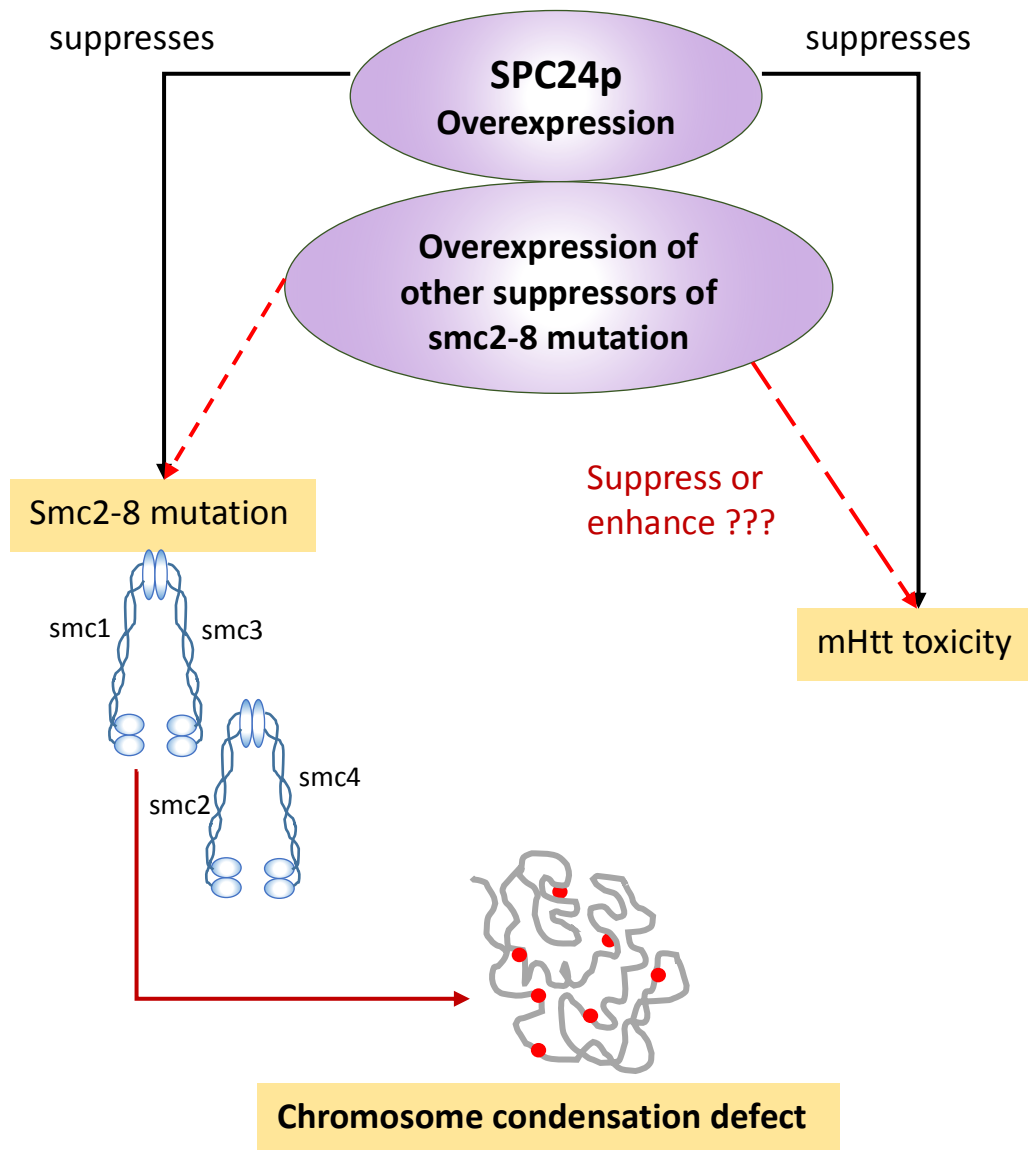


Figure 3.7 Hypothesis model of suppressors of smc2-8 mutation suppressing mHtt toxicity

Overexpression of SPC24, a yeast homolog from one of the significant protein complexes detected in HDI, suppresses smc2-8 mutation and also suppresses mHtt toxicity. Hence, we hypothesize that overexpression of a set of genes known to suppress smc2-8 mutant, can also suppress mHtt toxicity.

Chapter 4

4 Predicting physical interactors of the Huntingtin protein using Supervised Machine Learning methods.

Experimental approaches such as Y2H (yeast two-hybrid) mass spectrometry (MS), Tandem Affinity Purification (TAP) and protein microarrays have been the most widely adopted method to identify protein-protein interactions (PPIs). While Y2H is sensitive to detection of potential protein partners, it cannot detect interactions involving more than two protein partners. Additionally, these interactions are detected by virtue of their occurrence in the Y2H system and do not affirm their interaction in a physiological state.

The biological data generated using these experimental approaches though valuable, is subject to disadvantages; a high number of false positives being an important one of them. Machine learning approaches utilize the existing knowledge of protein interactors generated using these experimental approaches and help predict protein interactors. These methods use various protein features related to their structure, function or sequence to make PPI predictions to achieve better accuracy (A. Theofilatos et al., 2011). Other computational methods that integrate

various protein features into one predictor-classifier model have been able to accomplish higher accuracy (A. Theofilatos et al., 2011; Chen and Liu, 2005)

In this study, we integrate various protein features such as motifs, domains and their topological properties in a PPI network, to predict protein interactors of mutant Htt (mHtt) protein. We propose a Gradient Boosting Modeling (GBM) based classifier that helps to predict Htt-interacting proteins. This classifier examines the relationships between the topological characteristics of proteins within a protein-protein interaction network along with the structural and functional properties of the proteins to group them as interactors or non-interactors of mHtt protein. We study the extent of information captured by structural and topological aspects of proteins and investigate whether this information is sufficient enough to predict wt and/or mHtt protein interactors.

4.1 Data

4.1.1 Model development dataset:

The machine learning model was built using a set of primary interactors of Htt protein experimentally detected in wild-type and BACHD mouse brains. This dataset is spatiotemporal collection of 747 candidate proteins identified using AP-MS that form complexes with Htt in both wild-type and BACHD mouse brains (Shirasaki et al., 2012). This dataset was divided into 3 separate, non-overlapping groups as follows: Group 1 – containing proteins that interact with wt Htt protein

only, Group 2 – containing proteins that interact with mHtt only and Group 3 – containing proteins that interact with both wt and mHtt proteins.

4.1.2 Input features

4.1.2.1 Motif and domain properties

Motif and domain information related to each of the three groups in the dataset were obtained from the Uniprot database (Magrane and Consortium, 2011) and used as features for the input data.

4.1.2.2 PIN graph-theoretic properties

Additionally, graph properties were computed for each protein in the dataset and used as feature inputs to the machine learning classifier. To compute these network properties, we used the protein-protein interactions in mouse, curated by the BIOGRID database. The mouse PPI network obtained from BIOGRID consists of 8629 proteins and 19828 interactions. The following graph properties were calculated for candidate proteins in the input set:

- (a) Average Shortest Path Length: also, known as the characteristic path length. It measures the expected distance between two connected nodes in a network.
(Assenov et al., 2008)
- (b) Betweenness Centrality: If $\sigma_{p,q}$ is the number of shortest paths between proteins p and q , and $\sigma_{p,q}(r)$ is the number of shortest paths between p and q that pass through protein r in a protein interaction network, then betweenness

- centrality of the protein r is defined as $\Sigma \sigma_{p,q}(r) / \sigma_{p,q}$, where the sum is taken over all distinct pairs p and q . The betweenness value for each node r is normalized by dividing by the number of node pairs excluding r : (Freeman, 1977)
- (c) Closeness Centrality: it measures the extent to which a protein r is close to all the proteins in the network. If $d(r, s)$ is the shortest distance between proteins r and s in a protein network, then the closeness centrality of protein r is defined as $(n - 1) / \Sigma_q d(r, s)$, where n is the total number of proteins in the network (Beauchamp, 1965).
- (d) Clustering Coefficient: it is the fraction of the total possible interactions among direct neighbors of a protein in a protein interaction network. It is always a number between 0 and 1 (Watts and Strogatz, 1998).
- (e) Degree: is the number of edges connected to a node.
- (f) Eccentricity: it the maximum (non-infinite) length of a shortest path between r and another node in the network. If r is an isolated node, the value of this attribute is zero.
- (g) Neighborhood Connectivity: The neighborhood connectivity of a node r is defined as the average connectivity of all neighbors of r . The neighborhood connectivity distribution gives the average of the neighborhood connectivities of all nodes r with k neighbors for $k = 0, 1, \dots$. Therefore, if the neighborhood connectivity distribution is a decreasing function of k , then the network

displays edges between low connected and highly connected nodes (Maslov and Sneppen, 2002)

(h) Radiality: it is an index computed as follows:

(Diameter of the connected component of node r) – (Average shortest path length of a node r) + 1.

It is a number between 0 and 1.

(i) Stress Centrality: is the number of shortest paths passing through a node.

(j) Topological Coefficient: this is a measure attributed to those protein in the network that are not necessarily directly connected to each other. The measure is given by $TC_p = \text{average}(J(p,j)/k_p)$, where $J(p, j)$ denotes the number of nodes to which both p and j are linked, plus 1 if there is a direct link between p and j and k_p is the number of links of node p (Stelzl et al., 2005).

The graph properties of the proteins were calculated using the Network Analyzer application in Cytoscape (Assenov et al., 2008; Shannon et al., 2003)

4.1.3 Dataset formatting

Variable names for motif and domain information were coded, instead of their long raw names, with numerical identifiers for classifier models. Additionally, presence of motif or domain for a certain protein was denoted as '1' while absence of a motif was denoted as '0'. The resultant master dataset had 554 proteins as rows/observations and motifs, domains and graphical properties (n=779) as columns/dimensions. Detailed characteristics of the master dataset are given in the

Table 4.1. Evidently, the master dataset is sparsely populated and has a higher number of variables than the number of observations. This requires variable reduction and dimension reduction methods, which is explained in Section 4.2.1

4.1.4 Classification target

For classification, a multivariate prediction approach was initially used to accommodate the three response variables. However, we later moved on to a binomial prediction approach for better prediction power and simply focused on two response variables/groups of proteins viz. (a) proteins that interact with wHtt only ($n = 116$) (group1) and (b) proteins that interact with mHtt ($n = 438$) (group 2 ($n = 108$) + group 3 ($n = 330$)). This binary approach to analysis showed an improvement in the model's predictive power. The three classifiers used for model development are addressed in the next section.

4.2 Methods

4.2.1 Variable and dimension reduction methods

The set of variables that best capture the relationship between the response variable and the predictor variables was determined by calculating the Information Value (IV) of the predictor variables (Shannon, 1948). Information value helps in variable selection during model building. Information value of x for measuring y is a number that attempts to quantify the predictive power of x in capturing its relationship with y . Assuming that the target variable y is binary in nature, IV is defined as,

$$IV = \sum_{i=1}^{10} (bad_i - good_i) \ln \frac{bad_i}{good_i}$$

where,

i ranges from 1 to 10, in which the data is divided, bad_i is the proportion of bad accounts captured in the i^{th} division out of all bad accounts in the population and $good_i$ is the proportion of good accounts in the i^{th} division.

Additionally, Principal Component Analysis (PCA) of motif and domain variables was used for dimension reduction (Hotelling, 1933; Pearson, 1901). PCA converts a set of observations into a smaller set of linearly uncorrelated variables called as principal components through orthogonal transformation thus leading to variable reduction. Each principal component depicts variability in the data, in a descending order of magnitude, with the first component capturing the maximum variability.

4.2.2 Logistic regression with regularization

A logistic regression model depicts the relationship between the categorical dependent variable (response variable) and the independent feature variables (predictor variables) by estimating probabilities through a cumulative logistic distribution function (Wedderburn, 1974). Considering that the response variable for our dataset is binary in nature, the logistic regression model takes the form of a Generalized Logistic Model (GLM) through the following equation:

$$\log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta x$$

The output of a logistic regression model is the probability that a given protein is an interactor of either wt Htt, mHtt or both wt and mtHtt protein.

Logistic regression with regularization was used to obtain stable fit to the sparse data in this study. Regularization methods work by penalizing the coefficients of the features and minimize the error between the predicted and actual observations either through L2 regularization (Ridge regression) or through L1 regularization (Lasso regression) (Tibshirani, 1996). The cost function that needs to be minimized is also called as RSS (Residual Sum of Squares) and is given by the equation:

$$\text{Cost}(W) = \text{RSS}(W) = \sum_{i=1}^N \{y_i - \hat{y}_i\}^2 = \sum_{i=1}^N \left\{y_i - \sum_{j=0}^M w_j x_{ij}\right\}^2$$

where \mathbf{X} is the matrix of input features, \mathbf{y} is the actual outcome variable, $\hat{\mathbf{y}}$ is the predicted value of \mathbf{y} , \mathbf{w} is the weights or the coefficients, N is the total number of data points available, and M is the total number of features.

Penalizing the coefficients with a regularization parameter helps to avoid a large emphasis on any one particular feature and also helps to reduce the model complexity. Lasso regression performs L1 regularization by adding a penalty equal to the absolute value of the magnitude of the coefficients and is given by the following equation:

$$\text{Cost}(\mathbf{W}) = \text{RSS}(\mathbf{W}) + \lambda * (\text{sum of absolute value of weights})$$

$$= \sum_{i=1}^N \left\{ y_i - \sum_{j=0}^M w_j x_{ij} \right\}^2 + \lambda \sum_{j=0}^M |w_j|$$

Ridge regression performs L2 regularization by adding a penalty equal to the square of the magnitude of the coefficients and is given by the following equation:

$$\text{Cost}(\mathbf{W}) = \text{RSS}(\mathbf{W}) + \lambda * [\text{sum of square of weights}]$$

$$= \sum_{i=1}^N \left\{ y_i - \sum_{j=0}^M w_j x_{ij} \right\}^2 + \lambda \sum_{j=0}^M |w_j|^2$$

4.2.3 Random forest

Random Forest is an ensemble decision tree-based machine learning method that combines groups of weak tree models to result in a stronger model. This model grows multiple decision trees, wherein, each tree ‘votes’ for a class based on the attributes (predictor variables) the tree was built on. The model predicts new data by choosing the classification that receives the most votes over all the trees (Breiman, 2001). Random forest model thus reduces the variance of prediction while retaining a low bias. A lower bias and variance translates to a reduction in the prediction error and also avoids the issue of over-fitting the model to the training data.

4.2.3.1 *Implementation and Parameters used:*

We used the random forest package in R for analysis (Breiman, 2001; Liaw and Wiener, 2002). All the three types of predictor variables – motifs, domains and graphical properties were used as input for the Random Forest model. The data was scaled and centered prior to employing the model. The parameters used to run the model were as follows:

- `ntree` – number of trees to grow. Higher number of trees gives a better performance.,
- `mtry` – number of variables randomly sampled as candidates at each split.

- Nodesize – minimum size of terminal node/leaf of the decision tree. A smaller node size causes the model to capture more noise in the train data.

4.2.4 Gradient Boosting Machine (GBM)

Gradient Boosting is also a process that generates an ensemble of trees. However, the main premise of this model is the concept of ‘boosting’ that serially adds new prediction models to the ensemble. A new weak, base-learner model is trained at every iteration based on the negative gradient of the loss function of the entire ensemble obtained till that point (Friedman, 2001; Natekin and Knoll, 2013). Since our response variable is binary in nature, we used a ‘binomial’ distribution of the variable to calculate the loss of function gradient. The model complexity is controlled by using a shrinkage factor that reduces the impact of each base-learner model added to the ensemble. Shrinkage penalizes the magnitude of each iteration and reduces the size of additional steps. Such a method helps to improve the model accuracy through a series of smaller steps rather than a few large steps (Natekin and Knoll, 2013) The parameters used to run the model were as follows:

- n.trees – the total number of trees to fit which is equal to the number of iterations
- cv.folds – number of cross-validations to perform
- interaction depth – the maximum depth of variable interactions

- `n.minobsinnode` – minimum number of observations in the terminal nodes of the trees.
- `shrinkage` – also known as learning rate or step-size reduction parameter for the model that is applied to each tree during expansion.

4.3 Results

The master dataset was prepared before model training and testing, first, by imputing the missing values in the dataset. The missing values in the motif and domain predictor variables were replaced with “-1” while the missing values in the topology/graphical predictor variables were imputed with the mean of their respective column data. The 769 predictor variables in the master data, which consist of motif, domain, and graph-theoretic properties, were then reduced dimensionally using two approaches – (a) Information Value (IV) and (b) Principal Component Analysis (PCA)

4.3.1 Variable and dimension reduction

Figure 4.1 shows the information values of the predictors in descending order. The IV of the motif, domain, graph-theoretic variables ranged from 0.2256 to 0.0108. IV cutoffs of 0.1, 0.056 and 0.055 were initially selected for variable reduction. Among these, an IV cutoff ≥ 0.056 was chosen for variable reduction because of the large plateau, corresponding to 0.055, as seen in Figure 4.1. The accuracy of a lasso regression model using this cutoff value was found to be the best as well.

To reduce the dimension of the input variables, we also tried PCA analysis, applying only on motif and domain variables since they account for a vast majority of the input variables.

PCA on the motif and domain variables ($n = 769$) revealed 554 principal components (PCs). The three leading PCs captured most of the variance in the input data viz. 33.5 %, 7.9 % and 5.3 % respectively (Figure 4.2). These top three PCs were then combined with 10 graph-theoretic variables to form a development dataset for further testing.

4.3.2 Logistic regression with regularization

We have performed a series of experiments using the logistic regression with Lasso regularization to determine the best variable/dimension reduction approach. The following set of experiments were considered with various configurations of input data:

1. Experiment 1 – Raw input of master dataset with imputed missing values:
779 predictor variables
2. Experiment 2 – [Variable selection of motif and domain variables using an IV cutoff of ≥ 0.056]. + [Topology/Graphical predictors]: 157 predictor variables
3. Experiment 3 – [Variable reduction of motif and domain variables using PCA] + [Topology/Graphical predictors]: 13 predictor variables

Because of the small sample size, each experiment is evaluated using 10-fold cross-validation. 10-fold validation splits the original dataset randomly into 10 samples of equal size. Out of these 10 samples, one is retained as a validation dataset, while the remaining 9 samples are used as a training set. This cross-validation process is repeated k times ($k = 10$ in this experiment), with each of the 10 samples used exactly once as a validation dataset. This method therefore makes sure that all the observations are used for both training and validation while each observation gets to be used exactly once as a validation sample.

The AUC values of regularized logistic regression using the lasso, elastic-net and ridge regularization are given in the Table 4.2. AUC values for experiment 1 with 10-fold cross-validation ranged from 0.611 (Lasso and Elastic-net) to 0.584 (Ridge). For experiment 2, they ranged from 0.621 (Lasso and Elastic-net) to 0.619 (Ridge). For experiment 3, AUC values were 0.615 (Lasso), 0.613 (Elastic-net) and 0.599 (Ridge). Among the 3 different regularization methods, Lasso performed the best in all 3 experiments although the Elastic net results were very close to those of Lasso.

Among the three experiments, experiment 2 with variable selection by IV gave the best prediction accuracy (AUC = 0.621) in 10-fold cross-validation. Note that the performance of all three experiments with Lasso is very close, with AUC values

ranging between 0.611 and 0.621. Such minimal differences can be explained by the built-in variable selection methods within Lasso itself.

4.3.3 Lasso with data segmentation

We further adopted a segmentation approach to achieve better prediction accuracy with the Lasso model by dividing the Experiment 2 dataset into four segments as follows:

1. Lasso Segment 1- Motif-Topology segment - containing proteins with only motif and topological properties as the predictor variables (48 proteins, 157 predictor variables)
2. Lasso Segment 2- Domain-Topology segment – containing proteins with only domains and graphical properties as the predictor variables, (231 proteins, 157 predictor variables) and
3. Lasso Segment 3- Motifs and Domain-Topology segment- containing proteins with motifs, domains and graphical properties as the predictor variables (35 proteins, 157 predictor variables)
4. Lasso Segment 4 – Only Topology segment – containing proteins with only graphical properties as the predictor variables (240 proteins, 157 predictor variables).

Considering the low sample size of Segments 1 and 3, in-sample predictions were obtained for these segments using LOOCV (Leave one out cross-validation). The

AUC values for all the four data segments range from 0.6 to 0.85 (Table 4.3 and Figure 4.3). Though these values suggest a better prediction accuracy, it should be noted that in-sample predictions usually depict an optimistic picture of the model fit and the model accuracy more than often drops after using repeated cross-validation.

Nonetheless, these AUC values, encouraged us to believe that a segmentation approach using other models could help us reach better prediction accuracies.

4.3.4 Random forest

Random forest is a decision-tree based ensemble model known to reduce the variance and also retain a low bias in its model predictions, thus balancing accuracy and complexity of the model. A series of experiments were conducted by varying two main hyper-parameters used in Random forest, the number of trees (ntree) and the number of variables available for splitting at each node (mtry). The 10-fold cross-validated AUC values from these experiments are given in Table 4.4.

The AUC values range from 0.54 to 0.594, which are lower than the logistic regression with Lasso regularization. Experiment 2 with variable selection by IV again revealed the best prediction accuracy (AUC = 0.594, ntree = 500, mtry = 12) in 10-fold cross-validation among all the parameter tuning experiments. The variable importance for proteins interacting with mutant Htt in experiment 2 is

shown in Figure 4.4. It was found that Random forest relies heavily on the graphical properties of the proteins while fitting the model.

One of the explanations for a lower predictive power of the Random forest model could be the small sample size in our data. Random forest works well by intentionally overfitting the data with deep bushy trees and averaging out these overfit and diverse trees. The small samples size in our data makes it difficult to create diverse overfit trees, thus hindering Random forest's performance.

4.3.5 Gradient Boosting Machine (GBM)

GBM is another tree-based ensemble model by reducing bias step-by-step using shallow trees. A GBM model was used to fit the input data for all the three experiments, which examine the effect of variable and dimension reduction. For these initial experiments, the following parameters were used:

- 5000 trees with 10-fold cross-validation to determine the optimal number of trees
- interaction depth of 1
- number of minimum observation in each node equals to 1
- shrinkage of 0.001

Initial implementation of the GBM algorithm on all the three experiments showed that the AUC ranged from 0.584 to 0.6, with experiment 2 obtaining the highest AUC (0.6) among the three experiments (Table 4.5). Experiment 2 selects variable

by using Information value ≥ 0.056 and this turns out to be consistently the best method across all machine learning algorithms we tested.

Since GBM has more hyper-parameters to tune than the other machine learning methods we tested so far, we decided to conduct additional experiments to check the sensitivity to different hyper parameters. The hyper-parameter tuning experiments were conducted after the variable selection by IV (Experiment 2). The experimental design and the AUC values from 10-fold cross-validations are shown in Table 4.6. In these experiments, the interaction depth were set to 1 to avoid overfitting and 5000 trees with 10-fold cross validation was used within the fitting process to determine the optimal number of trees. Two key parameters, shrinkage and the minimum number of observation in each node, were varied.

It was found that a shrinkage factor of 0.001 and a minobsnode of 10 gave the highest AUC of 0.61 for experiment 2. This result is better than that of Random forest (AUC = 0.594) but is not as good as the logistic regression with Lasso (0.621).

4.3.6 GBM with data segmentation

Encouraged by the prospect of better prediction accuracy using data segments, we adopted a segmentation approach with the GBM model and divided the master dataset into three segments as follows:

5. Segment 1- Motif-Topology segment - containing proteins with only motif and topological properties as the predictor variables (48 proteins, 60 predictor variables)
6. Segment 2- Domain-Topology segment – containing proteins with only domains and graphical properties as the predictor variables, (231 proteins, 596 predictor variables) and
7. Segment 3- Motifs and Domain-Topology segment- containing proteins with motifs, domains and graphical properties as the predictor variables (35 proteins, 143 predictor variables)

Note that the number of input predictor variables vary for each segment since the set of proteins in each segment contains a different number of motifs and/or domains.

PCA analysis was used for motif and domain variable reduction for the above segments. The following set of GBM experiments were considered with various configurations of input data:

1. GBM Segment Experiment 1 – [Variable selection of motif variables using PCA]. + [Topology/Graphical predictors]: 22 predictor variables
2. GBM Segment Experiment 2 – [Variable selection of domain variables using PCA]. + [Topology/Graphical predictors]: 95 predictor variables

3. GBM Segment Experiment 3 – [Variable reduction of motif and domain variables using PCA] + [Topology/Graphical predictors]: 23 predictor variables

For these experiments, the following parameters were used:

- 5000 trees with 10-fold cross-validation to determine the optimal number of trees
- interaction depth of 1
- number of minimum observation in each node equals to 1
- shrinkage of 0.001

The 10-fold cross-validated AUC values from these experiments are given in Table 4.7. AUC values for all the three experiments range from 0.55 to 0.88. GBM Segment Experiment 1 with motifs and topology as predictor variables revealed the best prediction accuracy (AUC = 0.88, ntree = 5000, shrinkage factor = 0.001, n.minobsnode = 1, interaction depth = 1).

4.3.7 Important Predictor variables

The above results demonstrate that logistic regression with Lasso gives better prediction accuracy for experiment 2 among the three experiments that use IV for variable reduction. However, data segmentation allows us to achieve much better prediction accuracy using the GBM model, with GBM Segment Experiment 1 revealing the best AUC among the three data segment models (Figure 4.5).

We therefore proceeded to examine the variables of importance found from the GBM Segment Experiment 1. Table 4.8 shows the top 5 important variables for GBM Segment Experiment 1 in predicting proteins that interact with mutant Htt protein. Among the graphical properties of proteins, degree, average shortest path length, betweenness centrality and neighborhood connectivity were found to be the most important predictor variables. This is indeed true to imagine intuitively as a protein with numerous interacting proteins is more likely to interact with Huntingtin protein. Next, we examined the motif variables that contributed to the PC4, PC6 and PC10. Important motifs in the list were found to encode for an amino acid sequence relating to nuclear localization signals in proteins (Table 4.9). These specific proteins are encoded by genes such as RAB3D, RAB3A and RAB3B which are known to function in GTPase mediated signal transduction pathways and vesicle mediated transport. We also find the gene NPM1 that encodes for a protein that is essential for ribosome biogenesis, centrosome duplication, histone assembly and suppression of p53/TP53. Another set of proteins SLC25A4p and SLC25A5p are involved in chromosome segregation and in catalyzing exchange of ADP with mitochondrial ATP across the inner mitochondrial membrane. The above findings recapitulate the observations made in various animal and cell models of HD and therefore lend support to the results obtained by the GBM model.

We also examined motifs and domains of relative importance as found by the Lasso regression model (Table 4.10). Important motifs in the list were found to encode

for an amino acid sequence relating to nuclear localization signals in proteins (Table 4.11). Other motifs were found to encode a SUMO paralogue-specific binding sequence in proteins. Proteins containing these motifs are encoded by genes such as HNRNPQ, HNRNPA2B1 and LMNB1. HNRNPQ, HNRNPA2B1 are nuclear ribonucleoproteins involved in pre-mRNA processing in the nucleus, mRNA processing, RNA binding and splicing. HNRNPA2B1 has been shown to bind to telomeric DNA sequences thus protecting telomeric DNA from digestion. It is also involved in chromatin regulation and telomere extension. LMNB1 is a component of the inner nuclear membrane and is thought to interact with chromatin. DDX4p, a protein encoded by the gene DDX4 has ATP-dependent helicase activity and is involved in translational control and gene silencing processes by RNA in the mitotic cell cycle phase.

Similarly, proteins containing domains of importance are encoded by genes such as HIST1H1C, HIST1H1E and HIST1H1B that are members of the histone family (Table 4.11). Histones are required for condensation of nuclear chromatin and are known to regulate gene transcription to chromatin remodeling and DNA methylation. Histones are also involved in cellular response to stress. Another set of proteins containing domains of importance are encoded by the genes such as RAD23A and RAD23B that are known to play an important role in DNA nucleotide excision repair and in generating a cellular response to DNA damage. RAD23B is

specifically involved in global genome nucleotide-excision repair (GG-NER) and in modulating proteasomal degradation of ubiquinated proteins.

4.4 Conclusion

Our results demonstrate the informative value of motifs, domains of proteins in predicting interactors of mHtt. We show that graph theoretic properties of these protein interactors also help to determine a possible existence of interaction with Htt. Considering the sparse nature of predictor variables, we show that while using Information Value (IV) for variable reduction to provide us with better prediction accuracy, a segmentation approach using the GBM model coupled with PCA for dimension reduction, enables us to reach a higher prediction accuracy. The GBM model specifically reveals the importance of motifs and topology variables in predicting protein interactors of mutant Htt. The protein motifs of relative importance detected using this approach are known to annotated with functions such as vesicular transport, mitochondrial permeability and GTPase activity; all of which are established cellular processes known to be affected in HD. Additionally, we show that motifs and domains of importance required to predict proteins interacting with mHtt, as found by the Lasso model are annotated with functions such as condensation of nuclear chromatin, DNA nucleotide-excision repair, DNA and chromatin binding and cellular response to stress. These findings, support our assumption that mHtt interferes with chromosome condensation and DNA repair

processes and leads to accumulation of DNA damage in neuronal cells eventually leading to apoptosis.

4.5 Tables and Figures

Table 4.1 - Characteristics of data fed to the classifiers

Proteins	M1 ... Mxx	Myy- Mzz	D1 ... Dxx	Dyy ... Dzz	Topology	Response variables
P1 – Px	Only motif information				Topology	Group 1, Group 2, Group 3
Px – Py		Both motif and domain information			Topology	
Py - Pz				Only domain information	Topology	
Pz – P554					Only Topology information	

* Number of rows (proteins) = 554

* Number of predictors (motifs, domains and topology) = 779

* Response variables (Group 1, Group 2 and Group 3) are binary in nature.

Table 4.2 – Area under curve for regularized logistic regression using Lasso, Ridge and Elastic-net models from 10-fold cross-validation experiments.

The predictor variables include motifs, domains and graphical properties of the proteins.

Regularized Regression - AUC values (Predictor variables include motifs, domains and graphical properties)			
Experiment 1 – Raw input of master dataset with imputed missing values			
	Lasso ($\alpha = 1$)	Ridge ($\alpha = 0$)	Elastic net ($\alpha = 0.5$)
10-fold CV	0.611	0.584	0.61
Experiment 2 – [Variable selection with an IV cutoff ≥ 0.056]			
	Lasso ($\alpha = 1$)	Ridge ($\alpha = 0$)	Elastic net ($\alpha = 0.5$)
10-fold CV	0.621	0.619	0.62
Experiment 3 - [Top 3-PCs on all motif/domain information without IV filtering] + [Topology/Graphical predictors]			
	Lasso ($\alpha = 1$)	Ridge ($\alpha = 0$)	Elastic net ($\alpha = 0.5$)
10-fold CV	0.615	0.599	0.613

Table 4.3 Area under curve for data segments of Experiment 2 obtained using the Lasso regression model.

The predictor variables include motifs and/or domains and graphical properties of the proteins.

Experiment 2 - Data Segmentation Lasso regression - AUC values (Predictor variables include motifs, domains and graphical properties)	
Segment 1 – Motif - Topology Segment (48 proteins and 157 predictor variables)	
	Lasso ($\alpha = 1$)
In-sample predictions (LOOCV)	0.787
Segment 2 – Domain - Topology Segment (231 proteins and 157 predictor variables)	
	Lasso ($\alpha = 1$)
In-sample predictions (LOOCV)	0.597
Segment 3 – Motif and Domain - Topology Segment (35 proteins and 157 predictor variables)	
	Lasso ($\alpha = 1$)
In-sample predictions (LOOCV)	0.848
Segment 4 – Only Topology Segment (240 proteins and 157 predictor variables)	
	Lasso ($\alpha = 1$)
In-sample predictions (LOOCV)	0.669

Table 4.4 - Area under curve for random forest model from 10-fold cross-validation experiments.

Predictor variables include motifs, domains and graphical properties of the proteins

Random Forest - AUC values (Predictor variables include motifs, domains and graphical properties)					
Experiment 1 – Raw input of master dataset with imputed missing values					
	mtry = 5	mtry = 10	mtry = 12	mtry = 13	mtry = 15
10-fold cv (ntree = 500)	0.566	0.573	0.563	0.57	0.57
10-fold cv (ntree = 1000)	0.559	0.576	0.577	0.58	0.575
Experiment 2 – [Variable selection with an IV cutoff ≥ 0.056]					
	mtry = 5	mtry = 10	mtry = 12	mtry = 13	mtry = 15
10-fold cv (ntree = 500)	0.585	0.577	0.594	0.58	0.581
10-fold cv (ntree = 1000)	0.584	0.581	0.59	0.581	0.579
Experiment 3 - [Top 3-PCs on all motif/domain information without IV filtering] + [Topology/Graphical predictors]					
	mtry = 5	mtry = 10	mtry = 12	mtry = 13	mtry = 15
10-fold cv (ntree = 500)	0.564	0.57	0.554	0.569	0.569
10-fold cv (ntree = 1000)	0.567	0.569	0.571	0.569	0.569

Table 4.5 - Area under curve for GBM model from 10-fold cross-validation experiments.

Predictor variables include motifs, domains and graphical properties of the proteins

Gradient Boosting Machine (GBM) - AUC values (Predictor variables include motifs, domains and graphical properties)	
Experiment 1 – Raw input of master dataset with imputed missing values	
10-fold CV	0.591
Experiment 2 – [Variable selection with an IV cutoff ≥ 0.056]	
10-fold CV	0.6
Experiment 3 - [Top 3-PCs on all motif/domain information without IV filtering] + [Topology/Graphical predictors]	
10-fold CV	0.584

Table 4.6 – Experiment 2 – Parameter tuning for GBM

Experiment 2- Variable selection by IV ≥ 0.056 GBM parameter tuning						
	cv.folds	Interaction depth	n.tree	shrinkage	n.minobsinnode	AUC
1	10	1	5000	0.001	5	0.608
2	10	1	5000	0.001	10	0.61
3	10	1	5000	0.005	1	0.6
4	10	1	5000	0.005	5	0.6
5	10	1	5000	0.005	10	0.601
6	10	1	5000	0.01	1	0.597
7	10	1	5000	0.01	5	0.605
8	10	1	5000	0.01	10	0.607

Table 4.7 - Area under curve for GBM with data segmentation using 10-fold cross-validation.

Gradient Boosting Machine (GBM) with data segmentation- AUC values (Predictor variables include motifs, domains and graphical properties)	
GBM Segment Experiment 1 - [Variable selection of motif variables using PCA]. + [Topology/Graphical predictors]:	
10-fold CV	0.88
GBM Segment Experiment 2 - [Variable selection of domain variables using PCA]. + [Topology/Graphical predictors]:	
10-fold CV	0.549
GBM Segment Experiment 3 - [Variable selection of motif and domain variables using PCA]. + [Topology/Graphical predictors]:	
10-fold CV	0.588

Table 4.8 – Overall importance of top 5 variables in predicting proteins interacting with mutant Htt protein using GBM model.

Predictor Variable	Relative Influence
Degree	12.17
PC4	11.70
Average Shortest Path Length	10.88
PC10	9.16
PC6	8.98

Table 4.9 Genes and their encoded proteins containing motifs of importance for GBM Segment Experiment 1

Motif name	Mouse Uniprot Protein ID	Human Ortholog	Protein Function
MOTIF 153 158 Nuclear localization signal MOTIF 686 690 DXDXT motif MOTIF 697 701 LXXIL motif	Q99PI5	LPIN2	nuclear transcriptional coactivator for PPARGC1A to modulate lipid metabolism Fatty acid metabolism
MOTIF 51 59 Effector region MOTIF 51 59 Effector region	P35276 P63011 Q9CZT8	RAB3D RAB3A RAB3B	GTPase mediated signal transduction, protein (vesicular) transport Exocytosis, regulation of synaptic vesicle fusion, neurotransmitter release Protein transport (vesicular traffic of proteins)
MOTIF 55 65 HIGH region MOTIF 718 722 KMSKS region	Q8BMJ2	LARS	nucleotide binding and aminoacyl-tRNA editing activity
MOTIF 372 377 Selectivity filter MOTIF 493 495 PDZ-binding	P16388	KCNA1	ion channel activity and potassium channel activity primarily in the brain
MOTIF 152 157 Nuclear localization signal MOTIF 190 196 Nuclear localization signal	Q61937	NPM1	ribosome biogenesis, centrosome duplication, histone assembly, cell proliferation, and regulation of tumor suppressors p53/TP53

Motif name	Mouse Uniprot Protein ID	Human Ortholog	Protein Function
MOTIF 235 240 Substrate recognition	P48962	SLC25A4	Catalyzes the exchange of cytoplasmic ADP with mitochondrial ATP across the mitochondrial inner membrane.
MOTIF 235 240 Substrate recognition	P51881	SLC25A5	Role in chromosome segregation, Catalyzes the exchange of cytoplasmic ADP with mitochondrial ATP across the mitochondrial inner membrane.

Table 4.10 – Overall importance of top 10 variables in predicting proteins interacting with mutant Htt protein using Lasso Regression model.

Predictor variable	Overall importance
Betweenness Centrality	55.73068938
Closeness Centrality	3.79704057
DOMAIN 36 109 H15	0.391185768
MOTIF 564 578 Bipartite nuclear localization signal	0.262330373
MOTIF 416 421 Nuclear localization signal.	0.236135785
MOTIF 9 15 Nuclear localization signal.	0.212858651
Topological Coefficient	0.202029349
MOTIF 89 95 Required for SUMO paralog-specific binding.	0.155780768
DOMAIN 1 79 Ubiquitin-like.	0.141763696
MOTIF 261 289 Q motif	0.138626675

Table 4.11 Genes and their encoded proteins containing motifs and domains of importance found using Lasso regression for Experiment 2.

Motif/Domain name	Mouse Uniprot Protein ID	Human Ortholog	Protein Function
MOTIF 564 578 Bipartite nuclear localization signal.	Q7TMK9	HNRNPQ	RNA binding and splicing, mRNA processing, Component of the GAIT (gamma interferon-activated inhibitor of translation) complex, mediates interferon-gamma-induced translation inhibition in inflammation processes.
MOTIF 416 421 Nuclear localization signal.	P14733	LMNB1	provides a framework for the nuclear envelope, interacts with chromatin.
MOTIF 9 15 Nuclear localization signal.	O88569	HNRNPA2B1	pre-mRNA processing in the nucleus, mRNA metabolism and transport, involved in chromatin regulation and acetylation and telomere extension, protecting telomeric DNA repeat against endonuclease digestion,
MOTIF 89 95 Required for SUMO paralog-specific binding.	P57080	USP25	peptidase activity and thiol-dependent ubiquitin-specific protease activity.
MOTIF 261 289 Q motif	Q61496	DDX4	nucleic acid binding and ATP-dependent helicase activity, involved in gene silencing processes by RNA in mitotic prophase.

Motif/Domain name	Mouse Uniprot Protein ID	Human Ortholog	Protein Function
DOMAIN 36 109 H15	P15864 P43274 P43276	HIST1H1C HIST1H1E HIST1H1B	condensation of nucleosome chains, DNA, RNA and chromatin binding, DNA methylation, cellular response to stress, chromatin regulation/acetylation
DOMAIN 1 79 Ubiquitin-like	P54726 P54728	RAD23A RAD23B	nucleotide excision repair, and recognition of DNA repair and DNA damage, delivery of polyubiquitinated proteins to the proteasome,

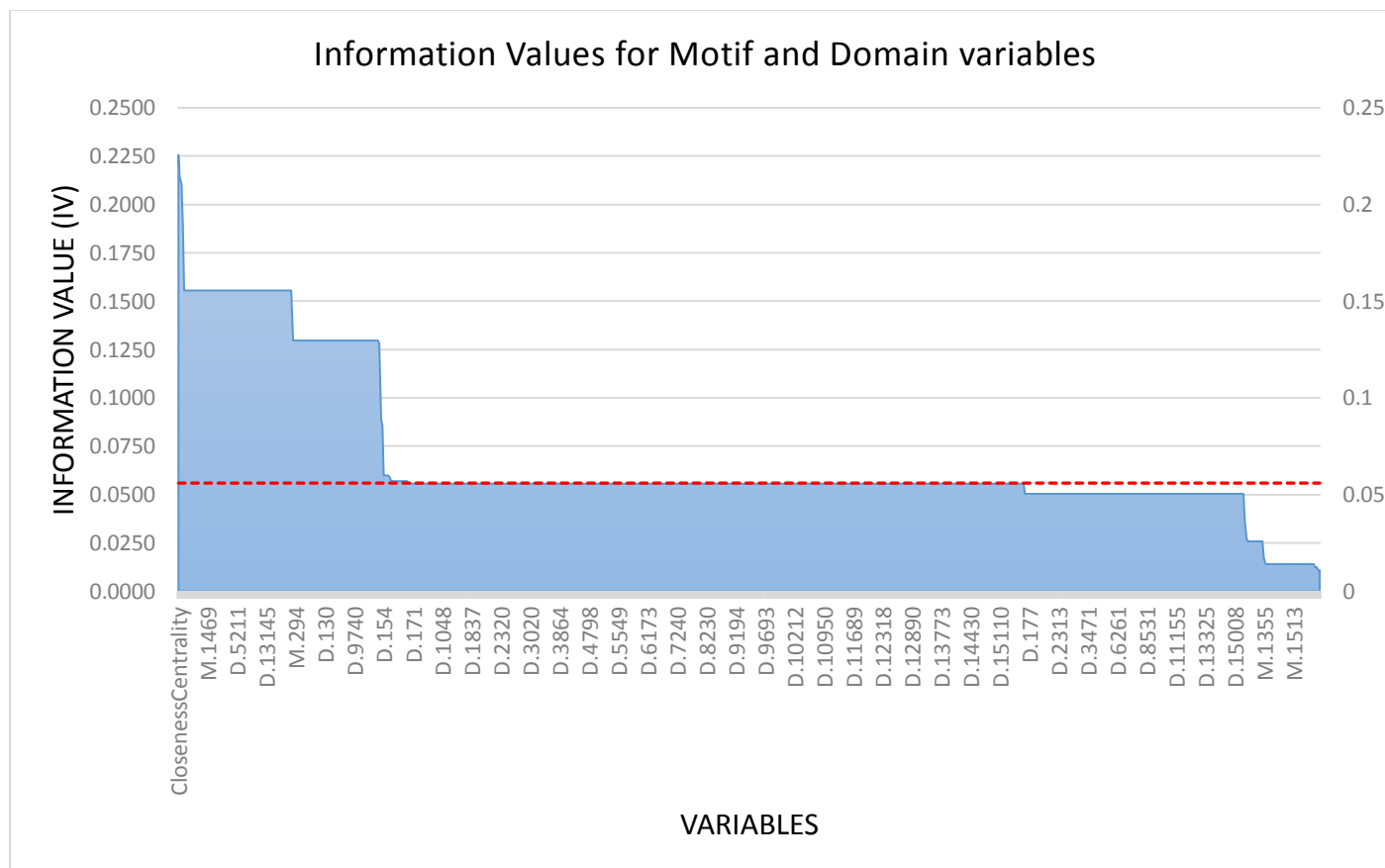


Figure 4.1- Information Value of Motif and domain variables.

The red dotted line (IV = 0.056) represents the cutoff IV selected for model building.

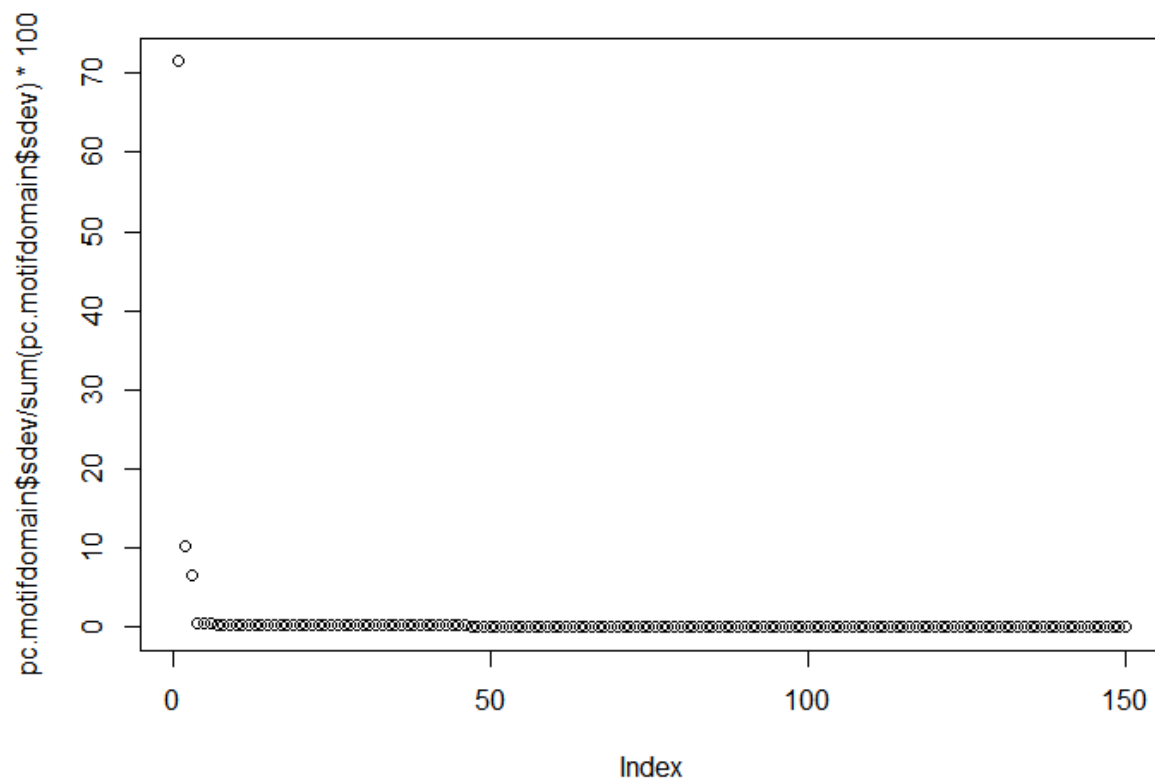


Figure 4.2 - Full scree plot of variance explained by the top 150 principal components of motif and domain variables.

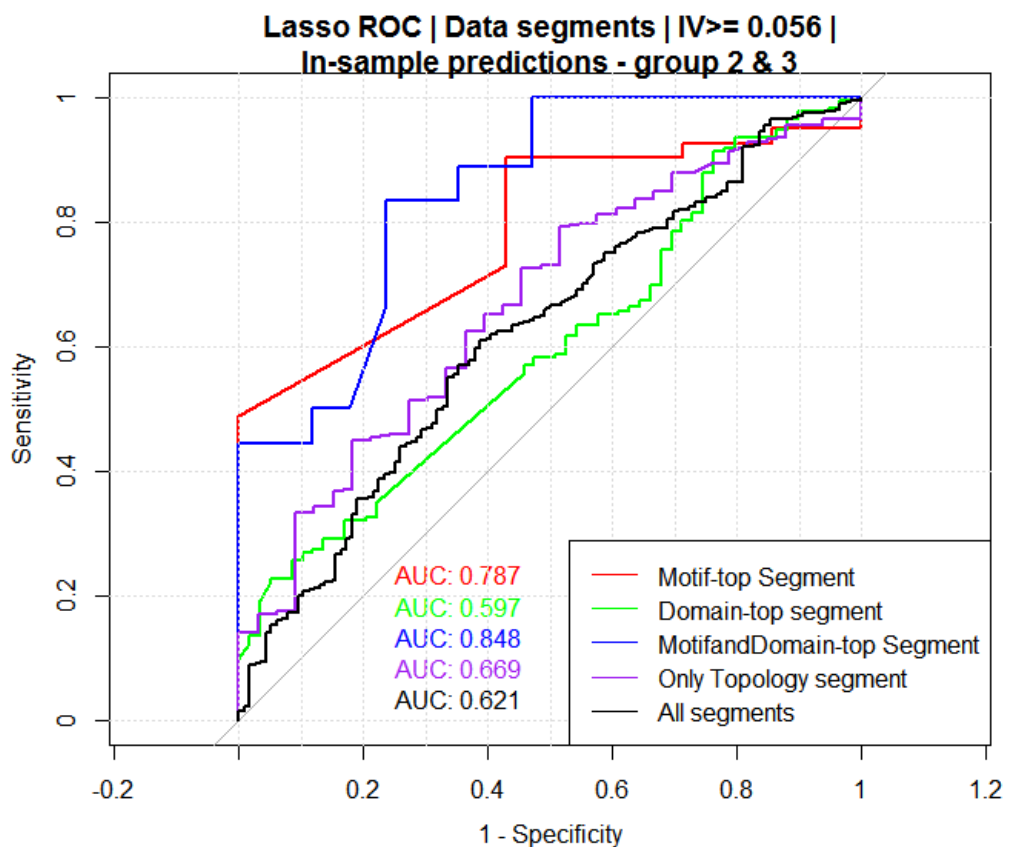


Figure 4.3- Receiver Operating Curves (ROC) of data segments for Experiment 2 using Lasso regression

Motif-topology segment (in Red), domain-topology segment (in Green) and motif and domain - topology segment (in Blue), only topology segment (purple)

Random Forest Variable Importance mutant binding(Group 2 + Group 3)

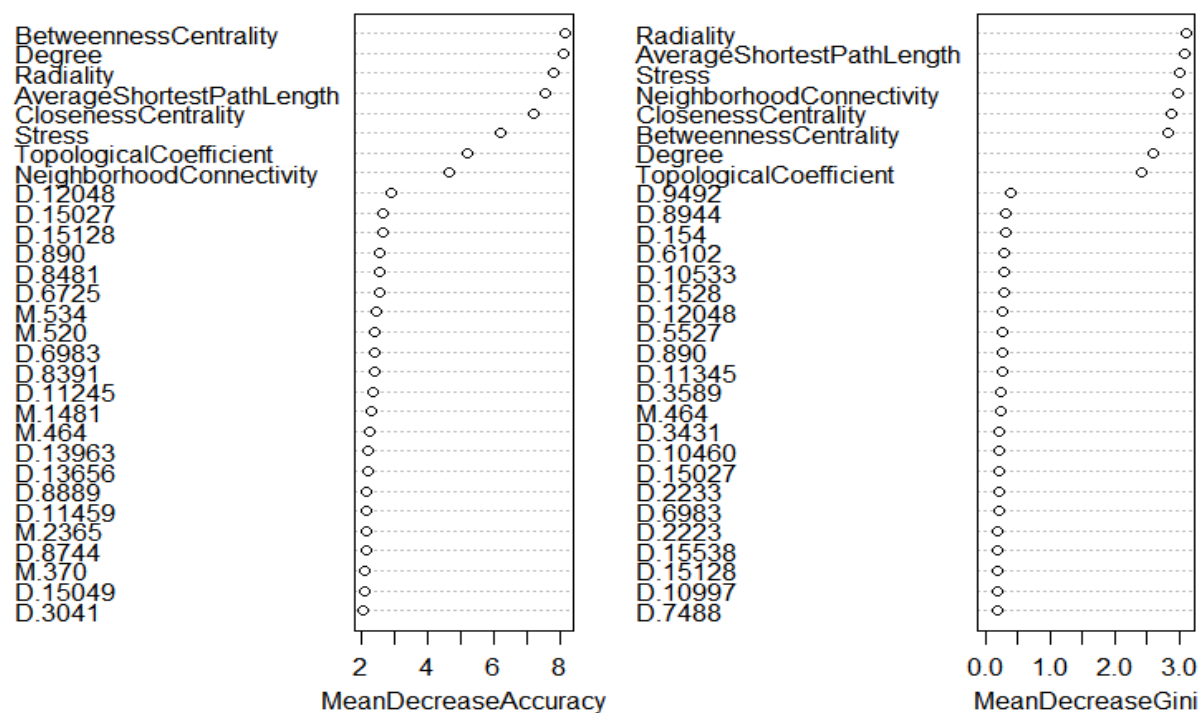


Figure 4.4 - Variable importance of proteins interacting with mutant Htt as shown by the Random Forest.

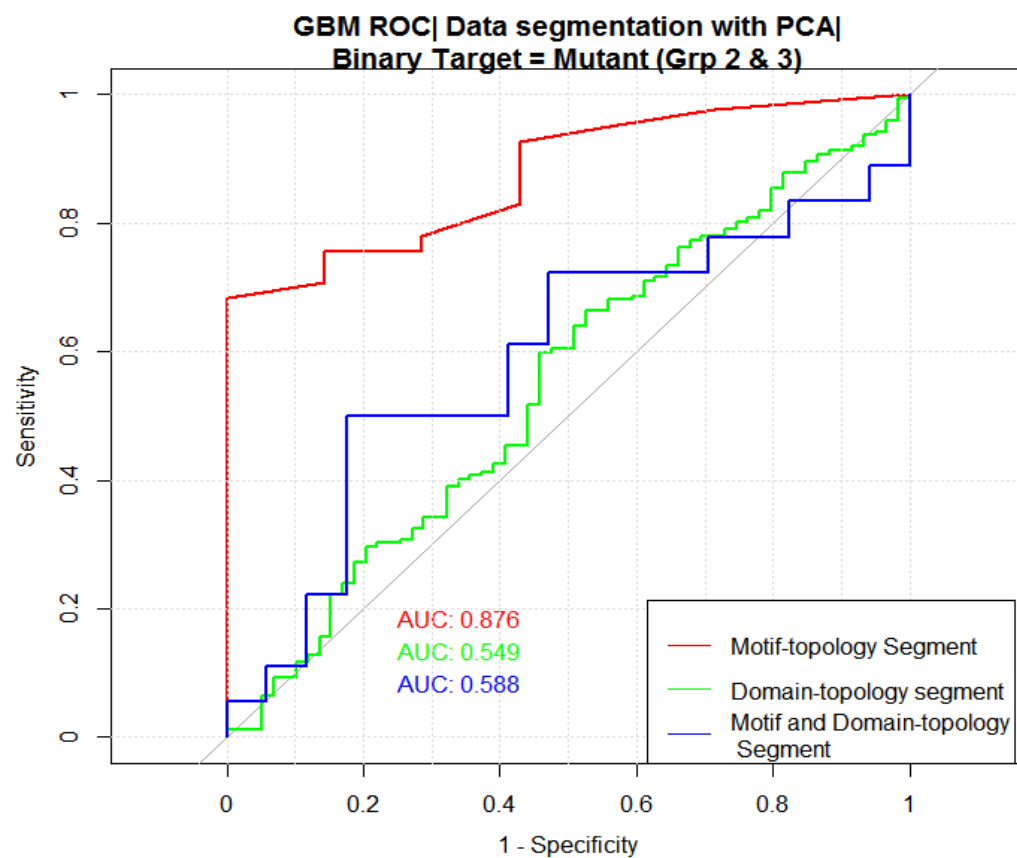


Figure 4.5 – Receiver Operating Curves (ROC) for data segments of the master dataset using GBM model.

Motif-topology segment (in Red), domain-topology segment (in Green) and motif and domain - topology segment (in Blue).

5 Summary

Although extensive studies on Huntington's disease (HD) have revealed the complex pathophysiology of this severely crippling disorder, the sequence of events through which the mutant Huntingtin (mHtt) protein executes its action still remains elusive. The complexity of the pathophysiology of HD can be attributed to the tendency of mHtt to abnormally interact with various other proteins that either do or do not interact with the wild-type Htt protein in normal conditions. The presence of Htt protein at various subcellular locations and its association with numerous other protein partners during its normal course of action also complicates the picture. The phenotype of HD is therefore an outcome of numerous processes initiated by the mHtt protein along with other proteins that act as either suppressors or enhancers of the effects of mHtt protein and PolyQ aggregates. To address this complexity, researchers have detected and analyzed proteins that physically interact with wild-type and mHtt proteins and have provided valuable information on various molecular and cellular processes affected in the mutant cells.

We hypothesized that integration of physical and genetic interactors of wild type and mHtt protein would enable us to predict unknown interactors of Htt protein using both unsupervised and supervised machine learning approaches. We built a Huntington's disease integrome (HDI) integrating human orthologs of protein

interactors of wild-type and mHtt in a mouse model of HD, with genetic modifiers of mHtt toxicity found in yeast HD models.

We used an unsupervised machine learning algorithm to partition the network into clusters and in the process discovered a novel connection linking Huntington's disease with chromosome condensation, DNA damage and apoptosis. We identified 27 candidate genes and validated three (NPLOC4, TUBGCP2 and NLRC4) of those genes in a drosophila model of HD. These findings are novel and remarkable for Huntington's disease and help establish our model implicating the role of mHtt in causing abnormal chromatin condensation, DNA damage, and neuronal cell death.

We used a separate supervised machine learning approach to create a model that built on the structural and graphical properties of protein interactors of both wild and mHtt protein. This model demonstrated that the information contained in proteins such as their motifs, domains and graphical properties have the ability to predict an interaction with Huntingtin protein, and offer a way to test and predict other interactors of wild type and mHtt protein.

Despite extensive research, researchers are still working to close gaps between the molecular processes affected in HD and their transition to clinical symptoms in HD patients. We postulated a systems biology approach utilizing machine learning techniques to reconcile the space between the HD genotype and phenotype. Indeed,

the machine learning approaches applied here put forth a system to identify molecular processes yet unknown to be involved in HD, in the hope of developing curative therapeutic options for this disabling disease.

6 Appendices

6.1 Network Properties of HDI

Given a graph G with vertices $\{v_1, v_2, \dots, v_n\}$, the adjacency matrix of G is defined to be as follows:

$$A = (a_{ij}) \text{ with } a_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \text{ is an edge in } G \\ 0 & \text{otherwise} \end{cases}$$

For an unweighted network, the adjacency $a_{ij} = 1$ if the nodes i and j are connected and 0 otherwise while for a weighted network, $0 \leq a_{ij} \leq 1$.

6.1.1 Network heterogeneity

The connectivity of a node is denoted by the number of its direct neighbors (for unweighted networks) and by the sum of the strength of its connections to other nodes (for weighted networks)(Dong and Horvath, 2007)

$$Connectivity_i = k_i = \sum_{j \neq i} a_{ij}$$

Network heterogeneity is the coefficient of variation of the connectivity.(Dong and Horvath, 2007)

$$Heterogeneity = \frac{\sqrt{variance(k)}}{mean(k)}$$

A network with high heterogeneity has a tendency to exhibit hubs in its structure.(Dong and Horvath, 2007)

6.1.2 Average number of neighbors

It denotes the average connectivity of a node in the network.(Assenov et al., 2008)

(See Network density)

6.1.3 Network density

For a given network with n nodes, the connectivity of n , is denoted by k_n , which is a set of numbers of its neighbors. The average number of neighbors of the node n , indicates its *average connectivity* (average number of neighbors) in the network. Network density is a normalized version of the average connectivity(Dong and Horvath, 2007) and is given by:

$$Network\ density = \frac{\sum_i \sum_{j \neq i} \alpha_{ij}}{n(n-1)} = \frac{S_1(k)}{n(n-1)} = \frac{mean(k)}{n-1}$$

The density of a network lies between 0 and 1; as the value leans towards 1, the density of edges in the network increases. Network density does not consider duplicated edges or self-loops.

6.1.4 Network diameter

Network diameter is the largest distance between two nodes in a network.

6.1.5 Clustering coefficient

The clustering coefficient of a node is ratio N/M , where N is the total number of edges between the neighbors of n , and M is the total number of edges that can possibly exist between the neighbors of n . This is represented by the following equation(Dong and Horvath, 2007):

$$ClusterCoef_i = \frac{\sum_{l \neq i} \sum_{m \neq i, l} a_{il} a_{lm} a_{mi}}{(\sum_{l \neq i} a_{il})^2 - \sum_{l \neq i} a_{il}^2}$$

It can also be defined as (Assenov et al., 2008)

$$ClusterCoef_i = \frac{2e_i}{k_i(k_i-1)}$$

where, k_i is the number of neighbors of node i , and e_i is the total number of connected pairs between all neighbors of node i . The clustering coefficient of a node always lies between 0 and 1. The network clustering coefficient is the average of clustering coefficients of all the nodes in the network.

6.1.6 Average shortest path length

It is also known as the characteristic path length. It measures the expected distance between two connected nodes in a network.(Assenov et al., 2008)

6.2 Rscripts

To ensure reproducibility of results, Rscripts used for analysis have been documented in the form of Jupyter notebooks and uploaded on GitHub.

6.2.1 IV calculation and Lasso Regression – Experiment 2

```
#####
## STEP I - Calculate Information values of variables
## STEP II - Lasso regularization
#####
## Read files
setwd("C:/PhD folder/SVM project/AnalysisOutputs/MLProject_PhaseII")
library(dplyr)
df2 <- read.delim(file = "mergedfiles_all_withNA.txt", header = T, sep = "\t")
df2[1:5, 778:781]
rownames(df2) <- df2$Row.names
predictorX <- df2[,3:781] # identify the columns representing the variables and
convert to matrix.
dim(predictorX)
predictorX[1:5,775:779]
class(predictorX)

#### Data prep #####
#### STRATEGY - Impute missing values in the data
#### Step 1 - Impute missing values in motif-domain segments with -1
#### Step 2 - Impute missing values in topology columns with the mean of each
column.
#### Step 3 - Find Information Values of the predictor variables
#####
#### Imputation of missing values in motif, domain segments with -1
predictorX[,1:769][is.na(predictorX[,1:769])] <- -1 # replace all NA in the
categorical variables (motif and domain) with -1

# Imputation of missing values in topology columns with mean of the column
values.
library(zoo)
predictorX[,770:779] <- na.aggregate(predictorX[,770:779])
sum(is.na(predictorX)) # check ..should be zero
```

```

# Linking the uniprot IDS with their respective group labels.
predictorX$uniprot_swissprot <- row.names(predictorX) # create a column with
uniprot IDs based on the row names.
class(predictorX$uniprot_swissprot) # must be character
dataf <- read.delim(file= "all_uniprotIDs.txt", header = TRUE, sep= "\t") ## Data
frame containing Uniprot Swiss ID
dataf$uniprot_swissprot <- as.character(dataf$uniprot_swissprot) # convert factor
to character.
class(dataf$uniprot_swissprot) # must be "character"
target <- left_join (predictorX,dataf,by="uniprot_swissprot") # this function will
link the unique protein ids with their group labels !!
mytarget <- target[,c(1:780,782)]
mytargetvariable <- mytarget[,781] # the target variable

x = predictorX[1:779] #dataframe of features
y = mytargetvariable #dependent variable

# recreate binary tags
y.wtmt <- ifelse(y %in% c("group2", "group3"), 1, 0)
y.wtOnly <- ifelse(y=="group1", 1, 0)
y.mtOnly <- ifelse(y=="group2", 1, 0)
table(y.wtmt,y)
table(y.wtOnly,y)
table(y.mtOnly,y)

class(y.wtmt); class(y.wtOnly) # both (target variables) have to be numeric
class(x$M.18)
x[,c(1:769)] <- lapply(x[,c(1:769)], factor) # convert independent categorical
variables to factor
#data[cols] <- lapply(data[cols], factor)

x <- cbind(x,y.wtmt); x[1:5,775:780] # bind target variable to the predictor
variables.

#####
# Compute Information value and WOE
# NOTE: The binary target variable is set to "Binding to mutant (mutant only vs.
mutant+WT)"
#####
library (Information) # load library

```

```

library(xlsx)
IV <- create_infotables(data=x, y="y.wtmt", bins=10, parallel=FALSE) # all 554
entries
#summary of IV values of all independent variables
V_Value = data.frame(IV$Summary)
range_V_Value <- range(V_Value$IV)
#write.xlsx(V_Value, file = "IVvalues.xlsx", col.names = TRUE, row.names =
FALSE)
V_Value <- V_Value[order(- V_Value$IV),] # order the IV values in descending
order.
#plot IV values in bar plot
barplot(V_Value$IV[1:157], col = "darkgreen", xlab = "Variables", ylab = "IV
value", names.arg = names(IV$Summary$Variable), main = "Information Value
Summary")

##### Select variables with IV value >= 0.056
V_Value <- subset(V_Value, IV>= 0.056)

# Subset predictorX with the variables with IV >= 0.056
myvec <- as.vector(V_Value$Variable)
predictorX.subset <- predictorX[, myvec]; dim(predictorX.subset) # subset and
check.
# Note: the predictor subset has missing values imputed as -1 for motif domain
variables and mean of
# column values for topology variables.

#####
##### Logistic Regression on the new subset
#####
# Input Data prep
# Linking the uniprot IDS with their respective group labels.
predictorX.subset$uniprot_swissprot <- row.names(predictorX.subset) # create a
column with uniprot IDs based on the row names.
class(predictorX.subset$uniprot_swissprot) # must be character
dataf <- read.delim(file= "all_uniprotIDs.txt", header = TRUE, sep= "\t") ## Data
frame containing Uniprot Swiss ID
dataf$uniprot_swissprot <- as.character(dataf$uniprot_swissprot) # convert factor
to character.
class(dataf$uniprot_swissprot)# must be "character"
target <- left_join (predictorX.subset,dataf,by="uniprot_swissprot") # this
function will link the unique protein ids with their group labels !!

```

```

mytarget <- target[,c(1:158,160)]
mytargetvariable <- target[,160] # the target variable
x = predictorX.subset[1:157] #dataframe of features
y = mytargetvariable #dependent variable

# recreate binary tags
y.wtmt <- ifelse(y %in% c("group2", "group3"), 1, 0)
y.wtOnly <- ifelse(y=="group1", 1, 0)
y.mtOnly <- ifelse(y=="group2", 1, 0)
table(y.wtmt,y)
table(y.wtOnly,y)
table(y.mtOnly,y)

# ##### Lasso Group 2 + Group 3#####
## AUC = 0.652 for IV cutoff >= 0.056
#####
library(glmnet) # load library
set.seed(1)
lasso <- glmnet(as.matrix(x), y.wtmt, alpha = 1,family = "binomial")
plot(lasso, label = TRUE)
plot(lasso, xvar = "lambda", label = TRUE)
print(lasso)

set.seed(2)
cvfit_lasso <- cv.glmnet(as.matrix(x),y.wtmt,alpha = 1, family =
"binomial",type.measure = "deviance")
plot(cvfit_lasso)

#lambda minimum of cvfit which gives minimum mean cross-validated error.(k-
means cross validation)
cvfit_lasso$lambda.min
cvfit_lasso$lambda.1se

#coefficients for the lambda minimum (lambda.min) and most regularized
lambda (lambda.1se) values.
coef(cvfit_lasso, s = "lambda.min")
coef(cvfit_lasso, s = "lambda.1se")

## prediction using lambda min for lasso regression
mypc.comp <- as.matrix(x); class(x)

```

```

lasso_pred <- predict(lasso, newx = mypc.comp, type = "response", s =
cvfit_lasso$lambda.min)
lasso_pred
class(y.wtmt)
class(lasso_pred)

## plotting the ROC curve using the above predictions
library(pROC)
lrocobj <- roc(y.wtmt,as.numeric(lasso_pred))
plot.roc(lrocobj, print.auc = TRUE, legacy.axes = TRUE, grid=c(0.1,
0.2),print.thres=TRUE, main = "Lasso ROC - All-in-one| IV>=0.056|group 2 &
3")
# plot.roc(smooth(lrocobj))
auc(lrocobj) # area under curve for lasso ROC

# Variable importance
# use caret package
library(caret)
varimp_lasso <- varImp(lasso, lambda = cvfit_lasso$lambda.min)
# write.xlsx(varimp_lasso,file = "varimp_lasso.xlsx", sheetName =
"Varimp_lasso")

# ##Lasso - 10-fold cv - Grp 2 +Grp 3 #####
# ## AUC - 0621 for IV >= 0.056
#####
library(cvTools) #run the above line if you don't have this library
library(glmnet)

k <- 10 #the number of folds
set.seed(123)
folds <- cvFolds(NROW(x), K=k)
x$kfoldsfpred <- rep(0,nrow(x))

kfoldprediction <-
  for(i in 1:k){
    train <- x[folds$subsets[folds$which != i], -158] #Set the training set
    train_response <- y.wtmt[folds$subsets[folds$which != i]] # set the training set
    response
    validation <- x[folds$subsets[folds$which == i], -158] #Set the validation set
    lasso_newglm <- glmnet(as.matrix(train), train_response, alpha = 1,family =
"binomial") #Get your new logistic regression model (just fit on the train data)
  }

```

```

randomseed = 123 + (i-1)*10
set.seed(randomseed)
lasso_cvglm <- cv.glmnet(as.matrix(train), train_response, alpha = 1, family =
"binomial", type.measure = "deviance")
lasso_newpred <- predict(lasso_newglm, newx = as.matrix(validation), type =
"response", s = c(lasso_cvglm$lambda.min)) #Get the predictions for the
validation set (from the model just fit on the train data)
x[folds$subsets[folds$which == i], ]$kfoldpred <- lasso_newpred #Put the
hold out prediction in the data set for later use
}
as.data.frame(x$kfoldpred) # predictions for all proteins using k-fold validation !
(10-fold)

# ROC curve
class(lasso_newpred)
lrocobj1 <- roc(y.wtmt, as.numeric(x$kfoldpred))
plot.roc(lrocobj1, print.auc = TRUE, legacy.axes = TRUE, grid=c(0.1,
0.2), print.thres=TRUE, main = "Lasso ROC|All-in-one|IV>=0.056\n|k-
foldpredictions - group 2 & 3")
# plot.roc(smooth(lrocobj))
auc(lrocobj1) # area under curve for lasso ROC

```

```

#### END!! DO NOT RUN

```

```

#####

```

6.2.2 Random forest – Experiment 2

```
# Identify features and variables.
x = predictorX.subset[1:157] #dataframe of features
y = mytargetvariable #dependent variable

# ## Random Forest #####
# ## RF parameters - ntree = 500, mtry = 12
#####
library(cvTools)
k <- 10 #the number of folds
set.seed(123)
folds <- cvFolds(NROW(x), K=k)
x$kfoldspred0 <- rep(0,nrow(x))
x$kfoldspred1 <- rep(0,nrow(x))
kfoldsprediction <-
  for(i in 1:k){
    train <- x[folds$subsets[folds$which != i], c(1:157)] # training set
    train_response <- y.wtmt[folds$subsets[folds$which != i]] # training set
    response
    validation <- x[folds$subsets[folds$which == i], c(1:157)] # validation set
    rf_grp1 <- randomForest(as.factor(train_response)~., data = train, importance =
TRUE)# get the RF model (just fit on the train data)
    randomseed = 123 + (i-1)*10
    set.seed(randomseed)
    rf_newpred <- predict(rf_grp1, newdata = validation, type = "prob", norm.votes
= TRUE, predict.all = FALSE) #Get the predictions for the validation set (from
the model just fit on the train data)
    x[folds$subsets[folds$which == i], ]$kfoldspred0 <- rf_newpred[,1]
    x[folds$subsets[folds$which == i], ]$kfoldspred1 <- rf_newpred[,2]#Put the
hold out prediction in the data set for later use
  }
as.data.frame(x[,158:159]) # predictions for all proteins using 10-fold validation

# ROC curve for Random forest 10-fold cv
library(pROC)
rocobj_grp1 <- roc(y.wtmt, x$kfoldspred1)
plot.roc(rocobj_grp1, print.auc = TRUE, legacy.axes = TRUE, grid=c(0.1,
0.2),main = "Random Forest| 10-fold cv|IV>= 0.056|\nmutant binding
(Grp2+Grp3)|mtry = 12|ntree = 500")
```


6.2.3 Gradient Boosting Machine – Experiment 2

```
# Identify features and variables.
x = predictorX.subset[1:157] #dataframe of features
y = mytargetvariable #dependent variable

# ## GBM #####
# ## GBM parameters – minobsnode = 10, shrinkage =0.001, ntrees = 5000
#####
library(cvTools)
train <- data.frame(x, y.wtmt)
k <- 10 #the number of folds
set.seed(123)
folds <- cvFolds(NROW(train), K=k)
train$kfoldspred <- rep(0,nrow(train))

kfoldprediction <-
  for(i in 1:k){
    training <- train[folds$subsets[folds$which != i], -159] #Set the training set
    training_response <- y.wtmt[folds$subsets[folds$which != i]] # set the training
    set response
    validation <- train[folds$subsets[folds$which == i], -159] #Set the validation
    set
    fit.gbm <- gbm(y.wtmt~., data=training, distribution = "bernoulli",
                  n.trees = 5000,
                  interaction.depth = 1,
                  n.minobsinnode = 10,
                  shrinkage = 0.001,
                  cv.folds = 10
    )
    #gbm.perf(fit.gbm)
    randomseed = 123 + (i-1)*10
    set.seed(randomseed)
    gbm_newpred <- predict(fit.gbm, newdata=validation,
gbm.perf(fit.gbm,plot.it=F),type="response") #Get the predicitions for the
validation set (from the model just fit on the train data)
    train[folds$subsets[folds$which == i], ]$kfoldspred <- gbm_newpred #Put the
hold out prediction in the data set for later use
  }
as.data.frame(train$kfoldspred) # predictions for all proteins using 10-fold
validation
```

```

# ROC curve for GBM 10-fold cv
library(pROC)
roc(train$y.wtmt, train$kfoldsfpred, plot= TRUE,
     legacy.axes = TRUE, grid=c(0.1, 0.2), print.auc = T,
     main = "GBM | All-in-one | Binary Target=Mutant (Grp 2+Grp 3) \n| 10-fold
cv | n.minobs = 10, shrinkage = 0.001, ntrees = 5000")

### create a graph with all ROC curves for Exp 2 - Lasso, Random Forest, GBM
roc(train$y.wtmt, train$kfoldsfpred, plot= TRUE,
     legacy.axes = TRUE, grid=c(0.1, 0.2), print.auc = T,
     main = "Experiment 2 | ROC curves\n Lasso, Random Forest and GBM |Binary
Target=Mutant (Grp 2+Grp 3)")
plot.roc(lrocobj1, print.auc = TRUE, add = TRUE, col = "red", legacy.axes =
TRUE, grid=c(0.1, 0.2),print.auc.y = 0.8,print.auc.x = 0.2)
plot.roc(rocobj_grp1, print.auc = TRUE,add = TRUE, col = "green", legacy.axes
= TRUE, grid=c(0.1, 0.2), print.auc.y = 0.6,print.auc.x = 0.4)
legend("bottomright", col = c("black", "red", "green"), legend = c("GBM",
"Lasso", "Random Forest"), lty = 1)

```

6.2.4 Gradient Boosting Machine – with Data segmentation

```
## GBM for Motif-Topology segment
#####
# Identify features and variables
x = predictorX[1:60] #dataframe of features
y = mytargetvariable #dependent variable

# Creating binary tags
# group 2 and group 3 proteins are tagged with 1 and group 1 proteins with 0
y.wtmt <- ifelse(y %in% c("group2", "group3"), 1, 0)
y.wtOnly <- ifelse(y=="group1", 1, 0)
table(y.wtmt,y)
table(y.wtOnly,y)

train5 <- data.frame(y.wtmt, x)

# .. PCA analysis to reduce motif variables
pc.motif <- prcomp(train5[,12:61])
summary(pc.motif)
plot(pc.motif, type='l') # .. scree plot
# check R's PCA list object
names(pc.motif)
# variable loadings: projections of the original variable onto the PC-space
head(pc.motif$rotation)
head(pc.motif$x)
# variance explained by top PCs
plot(pc.motif$sdev/sum(pc.motif$sdev)*100) # .. full scree plot
# keep only top 12 PCs
pcs <- pc.motif$x
pcs12 <- pcs[,1:12]
#####
Repeated 10-fold cross validation - Motif- Topology Segment
# AUC 0.88 for nsim = 1
# AUC - 0.86 for nsim = 100
#####
library(gbm)
library(cvTools)
library(ROCR)
library(pROC)
library(xlsx)
```

```

# create empty roc plot to plot roc curves
plot.roc(0:1, 0:1, type = "n", legacy.axes = TRUE, main = "GBM | Binary
Target=MT binding with PCA | repeated 10-Fold CV")

k <- 10 #the number of folds
x = cbind(predictorX[1:10], pcs12)
y <- y.wtmt
set.seed(1234)

folds <- cvFolds(NROW(x), K=k)
nsim <- 1 # number of repetitions
myauc <- rep(0, nsim)
mypreds <- data.frame(matrix(0, nrow(x), ncol = 100)) # create a dataframe to
store results of all 100 nsim repetitions
row.names(mypreds) <- row.names(x) # row names for the dataframe
names(mypreds) <- paste("K", (1:100), sep = "") # column names
j <- 1
x$kfolds1pred <- rep(0, nrow(x)) # append a column to original dataframe to
temporarily store results of each k-fold
ptm <- proc.time()

repeatcv <- function(){
  while (j <= nsim){
    for(i in 1:k){
      train <- x[folds$subsets[folds$which != i], -23] #Set the training set
      train_response <- y[folds$subsets[folds$which != i]] # set the training set
      response
      validation <- x[folds$subsets[folds$which == i], -23] #Set the validation set
      randomseed = 7842 + (i-1)*10 + j
      set.seed(randomseed)
      new_gbm.fit <- gbm(train_response~., data=train, distribution = "bernoulli",
                        n.trees = 5000,
                        interaction.depth = 1,
                        n.minobsinnode = 1,
                        shrinkage = 0.001,
                        cv.folds = 10)
      new_gbmpred <- predict(new_gbm.fit, newdata=validation,
                            gbm.perf(new_gbm.fit, plot.it=F),
                            type="response")
      x[folds$subsets[folds$which == i],]$kfolds1pred <- new_gbmpred
    }
  }
}

```

```

    mypreds[,j] <- x$foldlpred
    rocobj2 <- roc(y.wtmt, as.numeric(mypreds[,j]))
    myauc[j] <- rocobj2$auc # assign auc value to the jth item of your numeric
vector 'myauc'
    plot.roc(rocobj2,add = TRUE)
    j <- j+1
  }
  predictions <- as.data.frame(mypreds[,1],row.names = row.names(mypreds))
  write.xlsx(predictions, file = "predictions_gbmMotifTop.xlsx", col.names =
TRUE, row.names = TRUE)
  returnlist = list(predictions,myauc,mean(myauc), sd(myauc))
  returnlist
  roc(train5$y.wtmt, mypreds[,1], plot= TRUE,legacy.axes = TRUE, grid=c(0.1,
0.2), print.auc = T, main = "GBM|Motif-Top with PCA| Binary Target =
Mutant(Grp2+Grp3)")
}
repeatcv()
proc.time() – ptm

# create ROC chart for motif-topology segment
gbm_motifTop <- read.xlsx(file = "predictions_gbmMotifTop.xlsx", sheetIndex =
1, sheetName = "Sheet1")
roc(train5$y.wtmt, gbm_motifTop$mypreds...1., plot = TRUE, col = "red",
legacy.axes = TRUE, grid=c(0.1, 0.2),print.auc = TRUE, print.auc.y =
0.2,print.auc.x = 0.75, main = "GBM ROC| Data segmentation with PCA|nBinary
Target = Mutant (Grp 2 & 3)")

```

7 Bibliography

A. Theofilatos, K., M. Dimitrakopoulos, C., K. Tsakalidis, A., D. Likothanassis, S., T. Papadimitriou, S., and P. Mavroudi, S. (2011). Computational Approaches for the Prediction of Protein-Protein Interactions: A Survey. *Curr. Bioinforma.* 6, 398–414.

Aldecoa, R., and Marín, I. (2010). Jerarca: Efficient Analysis of Complex Networks Using Hierarchical Clustering. *PLoS ONE* 5, e11585.

Andersson, U., and Scarpulla, R.C. (2001). PGC-1-Related Coactivator, a Novel, Serum-Inducible Coactivator of Nuclear Respiratory Factor 1-Dependent Transcription in Mammalian Cells. *Mol. Cell. Biol.* 21, 3738–3749.

Arnau, V., Mars, S., and Marín, I. (2005). Iterative Cluster Analysis of Protein Interaction Data. *Bioinformatics* 21, 364–378.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics* 24, 282–284.

Aylward, E.H., Sparks, B.F., Field, K.M., Yallapragada, V., Shpritz, B.D., Rosenblatt, A., Brandt, J., Gourley, L.M., Liang, K., Zhou, H., et al. (2004). Onset and rate of striatal atrophy in preclinical Huntington disease. *Neurology* 63, 66–72.

Bader, G.D., and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4, 2.

Bae, B.-I., Xu, H., Igarashi, S., Fujimuro, M., Agrawal, N., Taya, Y., Hayward, S.D., Moran, T.H., Montell, C., Ross, C.A., et al. (2005). p53 Mediates Cellular Dysfunction and Behavioral Abnormalities in Huntington's Disease. *Neuron* 47, 29–41.

Baldo, B., Weiss, A., Parker, C.N., Bibel, M., Paganetti, P., and Kaupmann, K. (2012). A Screen for Enhancers of Clearance Identifies Huntingtin as a Heat Shock Protein 90 (Hsp90) Client Protein. *J. Biol. Chem.* 287, 1406–1414.

Basu, M., Bhattacharyya, N.P., and Mohanty, P.K. (2013). Comparison of Modules of Wild Type and Mutant Huntingtin and TP53 Protein Interaction Networks: Implications in Biological Processes and Functions. *PLoS ONE* 8, 1–11.

Beauchamp, M.A. (1965). An improved index of centrality. *Behav. Sci.* 10, 161–163.

Beconi, M.G., Yates, D., Lyons, K., Matthews, K., Clifton, S., Mead, T., Prime, M., Winkler, D., O'Connell, C., Walter, D., et al. (2012). Metabolism and Pharmacokinetics of JM6 in Mice: JM6 Is Not a Prodrug for Ro-61-8048. *Drug Metab. Dispos.* 40, 2297–2306.

Benchoua, A., Trioulier, Y., Zala, D., Gaillard, M.-C., Lefort, N., Dufour, N., Saudou, F., Elalouf, J.-M., Hirsch, E., Hantraye, P., et al. (2006). Involvement of Mitochondrial Complex II Defects in Neuronal Death Produced by N-Terminus Fragment of Mutated Huntingtin. *Mol. Biol. Cell* 17, 1652–1663.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300.

Benn, C.L., Landles, C., Li, H., Strand, A.D., Woodman, B., Sathasivam, K., Li, S.-H., Ghazi-Noori, S., Hockly, E., Faruque, S.M.N.N., et al. (2005). Contribution of nuclear and extranuclear polyQ to neurological phenotypes in mouse models of Huntington's disease. *Hum. Mol. Genet.* 14, 3065–3078.

Benn, C.L., Sun, T., Sadri-Vakili, G., McFarland, K.N., DiRocco, D.P., Yohrling, G.J., Clark, T.W., Bouzou, B., and Cha, J.-H.J. (2008). Huntingtin Modulates Transcription, Occupies Gene Promoters In Vivo, and Binds Directly to DNA in a Polyglutamine-Dependent Manner. *J. Neurosci.* 28, 10720–10733.

Bezprozvanny, I., and Hayden, M.R. (2004). Deranged neuronal calcium signaling and Huntington disease. *Biochem. Biophys. Res. Commun.* 322, 1310–1317.

Bizat, N., Hermel, J.-M., Boyer, F., Jacquard, C., Créminon, C., Ouary, S., Escartin, C., Hantraye, P., Krajewski, S., and Brouillet, E. (2003). Calpain Is a Major Cell Death Effector in Selective Striatal Degeneration Induced In Vivo by 3-Nitropropionate: Implications for Huntington's Disease. *J. Neurosci.* 23, 5020–5030.

- Blank, M., Lerenthal, Y., Mittelman, L., and Shiloh, Y. (2006). Condensin I recruitment and uneven chromatin condensation precede mitotic cell death in response to DNA damage. *J. Cell Biol.* 174, 195–206.
- Borrelli, E., Nestler, E.J., Allis, C.D., and Sassone-Corsi, P. (2008). Decoding the Epigenetic Language of Neuronal Plasticity. *Neuron* 60, 961–974.
- Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. (2004). GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 20, 3710–3715.
- Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32.
- Cattaneo, E., Rigamonti, D., Goffredo, D., Zuccato, C., Squitieri, F., and Sipione, S. (2001). Loss of normal huntingtin function: new developments in Huntington's disease research. *Trends Neurosci.* 24, 182–188.
- Cattaneo, E., Zuccato, C., and Tartari, M. (2005). Normal huntingtin function: an alternative approach to Huntington's disease. *Nat. Rev. Neurosci.* 6, 919–930.
- Chai, Y., Koppenhafer, S.L., Bonini, N.M., and Paulson, H.L. (1999). Analysis of the Role of Heat Shock Protein (Hsp) Molecular Chaperones in Polyglutamine Disease. *J. Neurosci.* 19, 10338–10347.
- Chakraborty, J., Rajamma, U., and Mohanakumar, K.P. (2014). A mitochondrial basis for Huntington's disease: therapeutic prospects. *Mol. Cell. Biochem.* 389, 277–291.
- Chandrasekaran, S., and Bonchev, D. (2016). Network analysis of human post-mortem microarrays reveals novel genes, microRNAs, and mechanistic scenarios of potential importance in fighting huntington's disease. *Comput. Struct. Biotechnol. J.* 14, 117–130.
- Chang, D.T.W., Rintoul, G.L., Pandipati, S., and Reynolds, I.J. (2006). Mutant huntingtin aggregates impair mitochondrial movement and trafficking in cortical neurons. *Neurobiol. Dis.* 22, 388–400.
- Chatterjee, R.S., Lui, G., Cha-um, M., and Patra, B.N. (2013). Overexpression of Ribosomal Genes Suppress Poly-Q (Glutamine)Induced Toxicity of Human Huntington's Disease Protein in Yeast. *J. Life Sci. Technol.* 228–232.

- Chen, X.-W., and Liu, M. (2005). Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics* 21, 4394–4400.
- Chen, S., Lu, F.F., Seeman, P., and Liu, F. (2012). Quantitative Proteomic Analysis of Human Substantia Nigra in Alzheimer’s Disease, Huntington’s Disease and Multiple Sclerosis. *Neurochem. Res.* 37, 2805–2813.
- Choo, Y.S., Johnson, G.V.W., MacDonald, M., Detloff, P.J., and Lesort, M. (2004). Mutant huntingtin directly increases susceptibility of mitochondria to the calcium-induced permeability transition and cytochrome c release. *Hum. Mol. Genet.* 13, 1407–1420.
- Chopra, V., Fox, J.H., Lieberman, G., Dorsey, K., Matson, W., Waldmeier, P., Housman, D.E., Kazantsev, A., Young, A.B., and Hersch, S. (2007). A small-molecule therapeutic lead for Huntington’s disease: Preclinical pharmacology and efficacy of C2-8 in the R6/2 transgenic mouse. *Proc. Natl. Acad. Sci.* 104, 16685–16689.
- Cisbani, G., and Cicchetti, F. (2012). An in vitro perspective on the molecular mechanisms underlying mutant huntingtin protein toxicity. *Cell Death Dis.* 3, e382.
- Costa, V., Giacomello, M., Hudec, R., Lopreiato, R., Ermak, G., Lim, D., Malorni, W., Davies, K.J.A., Carafoli, E., and Scorrano, L. (2010). Mitochondrial fission and cristae disruption increase the response of cell models of Huntington’s disease to apoptotic stimuli. *EMBO Mol. Med.* 2, 490–503.
- Cowan, C.M., and Raymond, L.A. (2006). Selective Neuronal Degeneration in Huntington’s Disease. In *Current Topics in Developmental Biology*, G.P. Schatten, ed. (Academic Press), pp. 25–71.
- Craufurd, D., and Snowden, J. (2002). Neuropsychological and neuropsychiatric aspects of Huntington’s disease. *Oxf. Monogr. Med. Genet.* 45, 62–94.
- Cui, L., Jeong, H., Borovecki, F., Parkhurst, C.N., Tanese, N., and Krainc, D. (2006). Transcriptional Repression of PGC-1 α by Mutant Huntingtin Leads to Mitochondrial Dysfunction and Neurodegeneration. *Cell* 127, 59–69.
- Culver, B.P., Savas, J.N., Park, S.K., Choi, J.H., Zheng, S., Zeitlin, S.O., Yates, J.R., and Tanese, N. (2012). Proteomic Analysis of Wild-type and Mutant Huntingtin-associated Proteins in Mouse Brains Identifies Unique Interactions and Involvement in Protein Synthesis. *J. Biol. Chem.* 287, 21599–21614.

- Damiano, M., Galvan, L., Déglon, N., and Brouillet, E. (2010). Mitochondria in Huntington's disease. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* 1802, 52–61.
- Davies, S.W., Turmaine, M., Cozens, B.A., DiFiglia, M., Sharp, A.H., Ross, C.A., Scherzinger, E., Wanker, E.E., Mangiarini, L., and Bates, G.P. (1997). Formation of Neuronal Intranuclear Inclusions Underlies the Neurological Dysfunction in Mice Transgenic for the HD Mutation. *Cell* 90, 537–548.
- DiFiglia, M., Sapp, E., Chase, K., Schwarz, C., Meloni, A., Young, C., Martin, E., Vonsattel, J.-P., Carraway, R., Reeves, S.A., et al. (1995). Huntingtin is a cytoplasmic protein associated with vesicles in human and rat brain neurons. *Neuron* 14, 1075–1081.
- Djousse, L., Knowlton, B., Hayden, M., Almqvist, E. w., Brinkman, R., Ross, C., Margolis, R., Rosenblatt, A., Durr, A., Dode, C., et al. (2003). Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease. *Am. J. Med. Genet. A.* 119A, 279–282.
- Dobrynin, G., Popp, O., Romer, T., Bremer, S., Schmitz, M.H.A., Gerlich, D.W., and Meyer, H. (2011). Cdc48/p97–Ufd1–Npl4 antagonizes Aurora B during chromosome segregation in HeLa cells. *J Cell Sci* 124, 1571–1580.
- Dong, J., and Horvath, S. (2007). Understanding network concepts in modules. *BMC Syst. Biol.* 1, 24.
- Dougherty, S.E., Reeves, J.L., Lesort, M., Detloff, P.J., and Cowell, R.M. (2013). Purkinje cell dysfunction and loss in a knock-in mouse model of Huntington Disease. *Exp. Neurol.* 240, 96–102.
- Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M., et al. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat. Genet.* 4, 387–392.
- Duyao, M.P., Auerbach, A.B., Ryan, A., Persichetti, F., Barnes, G.T., McNeil, S.M., Ge, P., Vonsattel, J.P., Gusella, J.F., Joyner, A.L., et al. (1995). Inactivation of the mouse Huntington's disease gene homolog Hdh. *Science* 269, 407–410.
- Ehrnhoefer, D.E., Butland, S.L., Pouladi, M.A., and Hayden, M.R. (2009). Mouse models of Huntington disease: variations on a theme. *Dis. Model. Mech.* 2, 123–129.

- Enokido, Y., Tamura, T., Ito, H., Arumughan, A., Komuro, A., Shiwaku, H., Sone, M., Foulle, R., Sawada, H., Ishiguro, H., et al. (2010). Mutant huntingtin impairs Ku70-mediated DNA repair. *J. Cell Biol.* *189*, 425–443.
- Faber, P.W., Barnes, G.T., Srinidhi, J., Chen, J., Gusella, J.F., and MacDonald, M.E. (1998). Huntingtin Interacts with a Family of WW Domain Proteins. *Hum. Mol. Genet.* *7*, 1463–1474.
- Faber, P.W., Alter, J.R., MacDonald, M.E., and Hart, A.C. (1999). Polyglutamine-mediated dysfunction and apoptotic death of a *Caenorhabditis elegans* sensory neuron. *Proc. Natl. Acad. Sci.* *96*, 179–184.
- Faber, P.W., Voisine, C., King, D.C., Bates, E.A., and Hart, A.C. (2002). Glutamine/proline-rich PQE-1 proteins protect *Caenorhabditis elegans* neurons from huntingtin polyglutamine neurotoxicity. *Proc. Natl. Acad. Sci.* *99*, 17131–17136.
- Fan, M.M.Y., and Raymond, L.A. (2007). N-Methyl-d-aspartate (NMDA) receptor function and excitotoxicity in Huntington's disease. *Prog. Neurobiol.* *81*, 272–293.
- Fan, H.-C., Ho, L.-I., Chi, C.-S., Chen, S.-J., Peng, G.-S., Chan, T.-M., Lin, S.-Z., and Harn, H.-J. (2014). Polyglutamine (PolyQ) Diseases: Genetics to Treatments. *Cell Transplant.* *23*, 441–458.
- Farrer, L.A., Cupples, L.A., Kiely, D.K., Conneally, P.M., and Myers, R.H. (1992). Inverse relationship between age at onset of Huntington disease and paternal age suggests involvement of genetic imprinting. *Am. J. Hum. Genet.* *50*, 528–535.
- Fatoba, S.T., and Okorokov, A.L. (2011). Human SIRT1 associates with mitotic chromatin and contributes to chromosomal condensation. *Cell Cycle* *10*, 2317–2322.
- Fienberg, A.A., Hiroi, N., Mermelstein, P.G., Song, W.-J., Snyder, G.L., Nishi, A., Cheramy, A., O'Callaghan, J.P., Miller, D.B., Cole, D.G., et al. (1998). DARPP-32: Regulator of the Efficacy of Dopaminergic Neurotransmission. *Science* *281*, 838–842.
- Freeman, L.C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry* *40*, 35–41.
- Friedel, C.C., Krumsiek, J., and Zimmer, R. (2009). Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast. *J. Comput. Biol.* *16*, 971–987.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29, 1189–1232.

Fujita, K., Nakamura, Y., Oka, T., Ito, H., Tamura, T., Tagawa, K., Sasabe, T., Katsuta, A., Motoki, K., Shiwa, H., et al. (2013). A functional deficiency of TERA/VCP/p97 contributes to impaired DNA repair in multiple polyglutamine diseases. *Nat. Commun.* 4, 1816.

Gafni, J., and Ellerby, L.M. (2002). Calpain Activation in Huntington's Disease. *J. Neurosci.* 22, 4842–4849.

Gafni, J., Hermel, E., Young, J.E., Wellington, C.L., Hayden, M.R., and Ellerby, L.M. (2004). Inhibition of Calpain Cleavage of Huntingtin Reduces Toxicity ACCUMULATION OF CALPAIN/CASPASE FRAGMENTS IN THE NUCLEUS. *J. Biol. Chem.* 279, 20211–20220.

Gauthier, L.R., Charrin, B.C., Borrell-Pagès, M., Dompierre, J.P., Rangone, H., Cordelières, F.P., De Mey, J., MacDonald, M.E., Leßmann, V., Humbert, S., et al. (2004). Huntingtin Controls Neurotrophic Support and Survival of Neurons by Enhancing BDNF Vesicular Transport along Microtubules. *Cell* 118, 127–138.

Gelperin, D.M. (2005). Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. *Genes Dev.* 19, 2816–2826.

Gemayel, R., Vences, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annu. Rev. Genet.* 44, 445–477.

Gemayel, R., Chavali, S., Pougach, K., Legendre, M., Zhu, B., Boeynaems, S., van der Zande, E., Gevaert, K., Rousseau, F., Schymkowitz, J., et al. (2015). Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity. *Mol. Cell* 59, 615–627.

Giorgini, F., Guidetti, P., Nguyen, Q., Bennett, S.C., and Muchowski, P.J. (2005). A genomic screen in yeast implicates kynurenine 3-monooxygenase as a therapeutic target for Huntington disease. *Nat. Genet.* 37, 526–531.

Girstmair, H., Saffert, P., Rode, S., Czech, A., Holland, G., Bannert, N., and Ignatova, Z. (2013). Depletion of Cognate Charged Transfer RNA Causes Translational Frameshifting within the Expanded CAG Stretch in Huntingtin. *Cell Rep.* 3, 148–159.

Godin, J.D., Colombo, K., Molina-Calavita, M., Keryer, G., Zala, D., Charrin, B.C., Dietrich, P., Volvert, M.-L., Guillemot, F., Dragatsis, I., et al. (2010). Huntingtin Is Required for Mitotic Spindle Orientation and Mammalian Neurogenesis. *Neuron* 67, 392–406.

Goehler, H., Lalowski, M., Stelzl, U., Waelter, S., Stroedicke, M., Worm, U., Droege, A., Lindenberg, K.S., Knoblich, M., and Haenig, C. (2004). A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol. Cell* 15, 853–865.

Goffredo, D., Rigamonti, D., Tartari, M., Micheli, A.D., Verderio, C., Matteoli, M., Zuccato, C., and Cattaneo, E. (2002). Calcium-dependent Cleavage of Endogenous Wild-type Huntingtin in Primary Cortical Neurons. *J. Biol. Chem.* 277, 39594–39598.

Graham, R.K., Deng, Y., Slow, E.J., Haigh, B., Bissada, N., Lu, G., Pearson, J., Shehadeh, J., Bertram, L., Murphy, Z., et al. (2006). Cleavage at the caspase-6 site is required for neuronal dysfunction and degeneration due to mutant huntingtin. *Cell* 125, 1179–1191.

Gray, M., Shirasaki, D.I., Cepeda, C., André, V.M., Wilburn, B., Lu, X.-H., Tao, J., Yamazaki, I., Li, S.-H., Sun, Y.E., et al. (2008). Full-Length Human Mutant Huntingtin with a Stable Polyglutamine Repeat Can Elicit Progressive and Selective Neuropathogenesis in BACHD Mice. *J. Neurosci.* 28, 6182–6195.

Greengard, P., Allen, P.B., and Nairn, A.C. (1999). Beyond the Dopamine Receptor. *Neuron* 23, 435–447.

Grisson, A., Mantovani, F., Comel, A., Agostoni, E., Gustincich, S., Persichetti, F., and Sal, G.D. (2011). Ser46 phosphorylation and prolyl-isomerase Pin1-mediated isomerization of p53 are key events in p53-dependent apoptosis induced by mutant huntingtin. *Proc. Natl. Acad. Sci.* 108, 17979–17984.

Gutkunst, C.-A., Li, S.-H., Yi, H., Mulroy, J.S., Kuemmerle, S., Jones, R., Rye, D., Ferrante, R.J., Hersch, S.M., and Li, X.-J. (1999). Nuclear and Neuropil Aggregates in Huntington's Disease: Relationship to Neuropathology. *J. Neurosci.* 19, 2522–2534.

Guzhova, I.V., Lazarev, V.F., Kaznacheeva, A.V., Ippolitova, M.V., Muronetz, V.I., Kinev, A.V., and Margulis, B.A. (2011). Novel mechanism of Hsp70 chaperone-mediated prevention of polyglutamine aggregates in a cellular model of huntington disease. *Hum. Mol. Genet.* 20, 3953–3963.

Hackam, A.S., Yassa, A.S., Singaraja, R., Metzler, M., Gutekunst, C.-A., Gan, L., Warby, S., Wellington, C.L., Vaillancourt, J., Chen, N., et al. (2000). Huntingtin Interacting Protein 1 Induces Apoptosis via a Novel Caspase-dependent Death Effector Domain. *J. Biol. Chem.* 275, 41299–41308.

Haeusler, A.R., Donnelly, C.J., Periz, G., Simko, E.A.J., Shaw, P.G., Kim, M.-S., Maragakis, N.J., Troncoso, J.C., Pandey, A., Sattler, R., et al. (2014). C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* 507, 195–200.

Harris, C.A., Miranda, A.F., Tanguay, J.J., Boegman, R.J., Beninger, R.J., and Jhamandas, K. (1998). Modulation of striatal quinolinate neurotoxicity by elevation of endogenous brain kynurenic acid. *Br. J. Pharmacol.* 124, 391–399.

Hausladen, A., and Fridovich, I. (1994). Superoxide and peroxynitrite inactivate aconitases, but nitric oxide does not. *J. Biol. Chem.* 269, 29405–29408.

Heng, M.Y., Tallaksen-Greene, S.J., Detloff, P.J., and Albin, R.L. (2007). Longitudinal Evaluation of the Hdh (CAG)150 Knock-In Murine Model of Huntington's Disease. *J. Neurosci.* 27, 8989–8998.

Hickey, M.A., Kosmalska, A., Enayati, J., Cohen, R., Zeitlin, S., Levine, M.S., and Chesselet, M.-F. (2008). Extensive early motor and non-motor behavioral deficits are followed by striatal neuronal loss in knock-in Huntington's disease mice. *Neuroscience* 157, 280–295.

Higashiyama, H., Hirose, F., Yamaguchi M, Inoue, Y.H., Fujikake, N., Matsukage A, and Kakizuka A (2002). Identification of ter94, *Drosophila* VCP, as a modulator of polyglutamine-induced neurodegeneration. *Publ. Online* 21 Febr. 2002 Doi101038sjcdd4400955 9.

Holbert, S., Denghien, I., Kiechle, T., Rosenblatt, A., Wellington, C., Hayden, M.R., Margolis, R.L., Ross, C.A., Dausset, J., Ferrante, R.J., et al. (2001). The Gln-Ala repeat transcriptional activator CA150 interacts with huntingtin: neuropathologic and genetic evidence for a role in Huntington's disease pathogenesis. *Proc. Natl. Acad. Sci.* 98, 1811–1816.

Hosp, F., Vossfeldt, H., Heinig, M., Vasiljevic, D., Arumughan, A., Wyler, E., Landthaler, M., Hubner, N., Wanker, E.E., Lannfelt, L., et al. (2015). Quantitative Interaction Proteomics of Neurodegenerative Disease Proteins. *Cell Rep.* 11, 1134–1146.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 417–441.

Huntington Study Group COHORT Investigators, and Dorsey, E.R. (2012). Characterization of a large group of individuals with huntington disease and their relatives enrolled in the COHORT study. *PloS One* 7, e29522.

Hwang, Y.J., Han, D., Kim, K.Y., Min, S.-J., Kowall, N.W., Yang, L., Lee, J., Kim, Y., and Ryu, H. (2014). ESET methylates UBF at K232/254 and regulates nucleolar heterochromatin plasticity and rDNA transcription. *Nucleic Acids Res.* 42, 1628–1643.

Imai, S., Armstrong, C.M., Kaeberlein, M., and Guarente, L. (2000). Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. *Nature* 403, 795–800.

Imamura, T., Fujita, K., Tagawa, K., Ikura, T., Chen, X., Homma, H., Tamura, T., Mao, Y., Taniguchi, J.B., Motoki, K., et al. (2016). Identification of hepta-histidine as a candidate drug for Huntington's disease by in silico-in vitro- in vivo-integrated screens of chemical libraries. *Sci. Rep.* 6, 33861.

Jaeger, C., Glaab, E., Michelucci, A., Binz, T.M., Koeglsberger, S., Garcia, P., Trezzi, J.-P., Ghelfi, J., Balling, R., and Buttini, M. (2015). The Mouse Brain Metabolome. *Am. J. Pathol.* 185, 1699–1712.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. (2003). A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science* 302, 449.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature* 407, 651–654.

Jeong, H., Mason, S.P., Barabási, A.-L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41–42.

Jeong, H., Cohen, D.E., Cui, L., Supinski, A., Savas, J.N., Mazzulli, J.R., Iii, J.R.Y., Bordone, L., Guarente, L., and Krainc, D. (2012). Sirt1 mediates neuroprotection from mutant huntingtin by activation of the TORC1 and CREB transcriptional pathway. *Nat. Med.* 18, 159–165.

Jiang, M., Wang, J., Fu, J., Du, L., Jeong, H., West, T., Xiang, L., Peng, Q., Hou, Z., Cai, H., et al. (2012). Neuroprotective role of Sirt1 in mammalian models of

Huntington's disease through activation of multiple Sirt1 targets. *Nat. Med.* *18*, 153–158.

Jimenez-Sanchez, M., Lam, W., Hannus, M., Sönnichsen, B., Imarisio, S., Fleming, A., Tarditi, A., Menzies, F., Ed Dami, T., Xu, C., et al. (2015). siRNA screen identifies QPCT as a druggable target for Huntington's disease. *Nat. Chem. Biol.* *11*, 347–354.

Johri, A., and Beal, M.F. (2010). Hunting-ton for New Proteases: MMPs as the New Target? *Neuron* *67*, 171–173.

Johri, A., Chandra, A., and Flint Beal, M. (2013). PGC-1 α , mitochondrial dysfunction, and Huntington's disease. *Free Radic. Biol. Med.* *62*, 37–46.

Jordan, G.E., and Piel, W.H. (2008). PhyloWidget: web-based visualizations for the tree of life. *Bioinformatics* *24*, 1641–1642.

Kalchman, M.A., Koide, H.B., McCutcheon, K., Graham, R.K., Nichol, K., Nishiyama, K., Kazemi-Esfarjani, P., Lynn, F.C., Wellington, C., Metzler, M., et al. (1997). HIP1, a human homologue of *S. cerevisiae* Sla2p, interacts with membrane-associated huntingtin in the brain. *Nat. Genet.* *16*, 44–53.

Kaltenbach, L.S., Romero, E., Becklin, R.R., Chettier, R., Bell, R., Phansalkar, A., Strand, A., Torcassi, C., Savage, J., and Hurlburt, A. (2007). Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PLoS Genet.* *3*, e82.

Kelley, R., and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* *23*, 561–566.

Kelly, D.P., and Scarpulla, R.C. (2004). Transcriptional regulatory circuits controlling mitochondrial biogenesis and function. *Genes Dev.* *18*, 357–368.

Kim, Y.J., Yi, Y., Sapp, E., Wang, Y., Cuiffo, B., Kegel, K.B., Qin, Z.-H., Aronin, N., and DiFiglia, M. (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. *Proc. Natl. Acad. Sci.* *98*, 12784–12789.

Kitamura, A., Sasaki, Y., Abe, T., Kano, H., and Yasutomo, K. (2014). An inherited mutation in NLRC4 causes autoinflammation in human and mice. *J. Exp. Med.* *211*, 2385–2396.

Kops, G.J.P.L., Dansen, T.B., Polderman, P.E., Saarloos, I., Wirtz, K.W.A., Coffey, P.J., Huang, T.-T., Bos, J.L., Medema, R.H., and Burgering, B.M.T. (2002). Forkhead transcription factor FOXO3a protects quiescent cells from oxidative stress. *Nature* 419, 316–321.

Kreiner, G., Bierhoff, H., Armentano, M., Rodriguez-Parkitna, J., Sowodniok, K., Naranjo, J.R., Bonfanti, L., Liss, B., Schütz, G., Grummt, I., et al. (2013). A neuroprotective phase precedes striatal degeneration upon nucleolar stress. *Cell Death Differ.* 20, 1455–1464.

Kukidome, D., Nishikawa, T., Sonoda, K., Imoto, K., Fujisawa, K., Yano, M., Motoshima, H., Taguchi, T., Matsumura, T., and Araki, E. (2006). Activation of AMP-Activated Protein Kinase Reduces Hyperglycemia-Induced Mitochondrial Reactive Oxygen Species Production and Promotes Mitochondrial Biogenesis in Human Umbilical Vein Endothelial Cells. *Diabetes* 55, 120–127.

Künig, G., Leenders, K.L., Sanchez-Pernaute, R., Antonini, A., Vontobel, P., Verhagen, A., and Günther, I. (2000). Benzodiazepine receptor binding in Huntington's disease: [11C]Flumazenil uptake measured using positron emission tomography. *Ann. Neurol.* 47, 644–648.

Laccone, F., Engel, U., Holinski-Feder, E., Weigell-Weber, M., Marczynek, K., Nolte, D., Morris-Rosendahl, D.J., Zühlke, C., Fuchs, K., Weirich-Schwaiger, H., et al. (1999). DNA analysis of Huntington's disease Five years of experience in Germany, Austria, and Switzerland. *Neurology* 53, 801–801.

Lawrence, A.D., Weeks, R.A., Brooks, D.J., Andrews, T.C., Watkins, L.H., Harding, A.E., Robbins, T.W., and Sahakian, B.J. (1998). The relationship between striatal dopamine receptor binding and cognitive performance in Huntington's disease. *Brain* 121, 1343–1355.

Lee, S.-T., and Kim, M. (2006). Aging and neurodegeneration: Molecular mechanisms of neuronal loss in Huntington's disease. *Mech. Ageing Dev.* 127, 432–435.

Lee, J., Hwang, Y.J., Boo, J.H., Han, D., Kwon, O.K., Todorova, K., Kowall, N.W., Kim, Y., and Ryu, H. (2011). Dysregulation of upstream binding factor-1 acetylation at K352 is linked to impaired ribosomal DNA transcription in Huntington's disease. *Cell Death Differ.* 18, 1726–1735.

Lee, J.-J., Park, J.K., Jeong, J., Jeon, H., Yoon, J.-B., Kim, E.E., and Lee, K.-J. (2013). Complex of Fas-associated Factor 1 (FAF1) with Valosin-containing

Protein (VCP)-Npl4-Ufd1 and Polyubiquitinated Proteins Promotes Endoplasmic Reticulum-associated Degradation (ERAD). *J. Biol. Chem.* 288, 6998–7011.

Legendre, M., Pochet, N., Pak, T., and Verstrepen, K.J. (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.* 17, 1787–1796.

Lerner, R.P., Trejo Martinez, L. del C.G., Zhu, C., Chesselet, M.-F., and Hickey, M.A. (2012). Striatal atrophy and dendritic alterations in a knock-in mouse model of Huntington's disease. *Brain Res. Bull.* 87, 571–578.

Li, H., Li, S.-H., Johnston, H., Shelbourne, P.F., and Li, X.-J. (2000). Amino-terminal fragments of mutant huntingtin show selective accumulation in striatal neurons and synaptic toxicity. *Nat. Genet.* 25, 385–389.

Li, X.-J., Li, S.-H., Sharp, A.H., Nucifora, F.C., Schilling, G., Lanahan, A., Worley, P., Snyder, S.H., and Ross, C.A. (1995). A huntingtin-associated protein enriched in brain with implications for pathology. *Nature* 378, 398–402.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News* 2, 18–22.

Lin, C.-H., Tallaksen-Greene, S., Chien, W.-M., Cearley, J.A., Jackson, W.S., Crouse, A.B., Ren, S., Li, X.-J., Albin, R.L., and Detloff, P.J. (2001). Neurological abnormalities in a knock-in mouse model of Huntington's disease. *Hum. Mol. Genet.* 10, 137–144.

Lokhande, S., Patra, B.N., and Ray, A. (2016). A link between chromatin condensation mechanisms and Huntington's disease: connecting the dots. *Mol. BioSyst.* 12, 3515–3529.

Lopes, T.J.S., Schaefer, M., Shoemaker, J., Matsuoka, Y., Fontaine, J.-F., Neumann, G., Andrade-Navarro, M.A., Kawaoka, Y., and Kitano, H. (2011). Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics* 27, 2414–2421.

Lucas, E.K., Dougherty, S.E., McMeekin, L.J., Trinh, A.T., Reid, C.S., and Cowell, R.M. (2012). Developmental Alterations in Motor Coordination and Medium Spiny Neuron Markers in Mice Lacking PGC-1 α . *PLOS ONE* 7, e42878.

Luthi-Carter, R., Strand, A., Peters, N.L., Solano, S.M., Hollingsworth, Z.R., Menon, A.S., Frey, A.S., Spektor, B.S., Penney, E.B., Schilling, G., et al. (2000).

Decreased expression of striatal signaling genes in a mouse model of Huntington's disease. *Hum. Mol. Genet.* 9, 1259–1271.

MacDonald, M.E. (2003). Huntingtin: Alive and Well and Working in Middle Management. *Sci STKE* 2003, pe48-pe48.

Maehama, T., Kawahara, K., Nishio, M., Suzuki, A., and Hanada, K. (2014). Nucleolar Stress Induces Ubiquitination-independent Proteasomal Degradation of PICT1 Protein. *J. Biol. Chem.* 289, 20802–20812.

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database* 2011.

Mangiarini, L., Sathasivam, K., Seller, M., Cozens, B., Harper, A., Hetherington, C., Lawton, M., Trottier, Y., Lehrach, H., Davies, S.W., et al. (1996). Exon 1 of the HD Gene with an Expanded CAG Repeat Is Sufficient to Cause a Progressive Neurological Phenotype in Transgenic Mice. *Cell* 87, 493–506.

Marcora, E., Gowan, K., and Lee, J.E. (2003). Stimulation of NeuroD activity by huntingtin and huntingtin-associated proteins HAP1 and MLK2. *Proc. Natl. Acad. Sci.* 100, 9578–9583.

Margolis, R.L., Holmes, S.E., Rosenblatt, A., Gourley, L., O'Hearn, E., Ross, C.A., Seltzer, W.K., Walker, R.H., Ashizawa, T., Rasmussen, A., et al. (2004). Huntington's disease-like 2 (HDL2) in North America and Japan. *Ann. Neurol.* 56, 670–674.

Marsh, J.L., Pallos, J., and Thompson, L.M. (2003). Fly models of Huntington's disease. *Hum. Mol. Genet.* 12, R187–R193.

Martindale, D., Hackam, A., Wieczorek, A., Ellerby, L., Wellington, C., McCutcheon, K., Singaraja, R., Kazemi-Esfarjani, P., Devon, R., Kim, S.U., et al. (1998). Length of huntingtin and its polyglutamine tract influences localization and frequency of intracellular aggregates. *Nat. Genet.* 18, 150–154.

Maslov, S., and Sneppen, K. (2002). Specificity and Stability in Topology of Protein Networks. *Science* 296, 910–913.

Mason, R.P., and Giorgini, F. (2011). Modeling Huntington disease in yeast: Perspectives and future directions. *Prion* 5, 269–276.

Mason, R.P., Casu, M., Butler, N., Breda, C., Campesan, S., Clapp, J., Green, E.W., Dhulkhed, D., Kyriacou, C.P., and Giorgini, F. (2013). Glutathione peroxidase

activity is neuroprotective in models of Huntington's disease. *Nat. Genet.* 45, 1249–1254.

McFarland, K.N., Das, S., Sun, T.T., Leyfer, D., Xia, E., Sangrey, G.R., Kuhn, A., Luthi-Carter, R., Clark, T.W., Sadri-Vakili, G., et al. (2012). Genome-Wide Histone Acetylation Is Altered in a Transgenic Mouse Model of Huntington's Disease. *PLOS ONE* 7, e41423.

McNeil, S.M., Novelletto, A., Srinidhi, J., Barnes, G., Kornbluth, I., Altherr, M.R., Wasmuth, J.J., Gusella, J.F., MacDonald, M.E., and Myers, R.H. (1997). Reduced penetrance of the Huntington's disease mutation. *Hum. Mol. Genet.* 6, 775–779.

Meadows, J.P., Guzman-Karlsson, M.C., Phillips, S., Holleman, C., Posey, J.L., Day, J.J., Hablitz, J.J., and Sweatt, J.D. (2015). DNA methylation regulates neuronal glutamatergic synaptic scaling. *Sci Signal* 8, ra61-ra61.

Menalled, L.B., Sison, J.D., Wu, Y., Olivieri, M., Li, X.-J., Li, H., Zeitlin, S., and Chesselet, M.-F. (2002). Early Motor Dysfunction and Striosomal Distribution of Huntingtin Microaggregates in Huntington's Disease Knock-In Mice. *J. Neurosci.* 22, 8266–8276.

Menalled, L.B., Sison, J.D., Dragatsis, I., Zeitlin, S., and Chesselet, M.-F. (2003). Time course of early motor and neuropathological anomalies in a knock-in mouse model of Huntington's disease with 140 CAG repeats. *J. Comp. Neurol.* 465, 11–26.

Michishita, E., Park, J.Y., Burneski, J.M., Barrett, J.C., and Horikawa, I. (2005). Evolutionarily Conserved and Nonconserved Cellular Localizations and Functions of Human SIRT Proteins. *Mol. Biol. Cell* 16, 4623–4635.

Miller, J.P., Holcomb, J., Al-Ramahi, I., de Haro, M., Gafni, J., Zhang, N., Kim, E., Sanhueza, M., Torcassi, C., Kwak, S., et al. (2010). Matrix Metalloproteinases Are Modifiers of Huntingtin Proteolysis and Toxicity in Huntington's Disease. *Neuron* 67, 199–212.

Mojsilovic-Petrovic, J., Nedelsky, N., Boccitto, M., Mano, I., Georgiades, S.N., Zhou, W., Liu, Y., Neve, R.L., Taylor, J.P., Driscoll, M., et al. (2009). FOXO3a Is Broadly Neuroprotective In Vitro and In Vivo against Insults Implicated in Motor Neuron Diseases. *J. Neurosci.* 29, 8236–8247.

- Murphy, S.M., Urbani, L., and Stearns, T. (1998). The Mammalian γ -Tubulin Complex Contains Homologues of the Yeast Spindle Pole Body Components Spc97p and Spc98p. *J. Cell Biol.* *141*, 663–674.
- Nahhas, F., Garbern, J., Feely, S., and Feldman, G.L. (2009). An intergenerational contraction of a fully penetrant Huntington disease allele to a reduced penetrance allele: interpretation of results and significance for risk assessment and genetic counseling. *Am. J. Med. Genet. A.* *149A*, 732–736.
- Nasir, J., Floresco, S.B., O’Kusky, J.R., Diewert, V.M., Richman, J.M., Zeisler, J., Borowski, A., Marth, J.D., Phillips, A.G., and Hayden, M.R. (1995). Targeted disruption of the Huntington’s disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. *Cell* *81*, 811–823.
- Natekin, A., and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Front. Neurorobotics* *7*.
- Nemoto, S., Fergusson, M.M., and Finkel, T. (2005). SIRT1 Functionally Interacts with the Metabolic Regulator and Transcriptional Coactivator PGC-1 α . *J. Biol. Chem.* *280*, 16456–16460.
- Nepusz, T. (2016). ClusterONE algorithm - Validating results.
- Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* *9*, 471–472.
- O’Brien, K.P., Remm, M., and Sonnhammer, E.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* *33*, D476–D480.
- Oda, E., Ohki, R., Murasawa, H., Nemoto, J., Shibue, T., Yamashita, T., Tokino, T., Taniguchi, T., and Tanaka, T. (2000a). Noxa, a BH3-Only Member of the Bcl-2 Family and Candidate Mediator of p53-Induced Apoptosis. *Science* *288*, 1053–1058.
- Oda, K., Arakawa, H., Tanaka, T., Matsuda, K., Tanikawa, C., Mori, T., Nishimori, H., Tamai, K., Tokino, T., Nakamura, Y., et al. (2000b). p53AIP1, a Potential Mediator of p53-Dependent Apoptosis, and Its Regulation by Ser-46-Phosphorylated p53. *Cell* *102*, 849–862.
- Orr, A.L., Li, S., Wang, C.-E., Li, H., Wang, J., Rong, J., Xu, X., Mastroberardino, P.G., Greenamyre, J.T., and Li, X.-J. (2008). N-Terminal Mutant Huntingtin

Associates with Mitochondria and Impairs Mitochondrial Trafficking. *J. Neurosci.* 28, 2783–2792.

Panegyres, P.K., and Goh, J.G.S. (2011). The neurology and natural history of patients with indeterminate CAG repeat length mutations of the Huntington disease gene. *J. Neurol. Sci.* 301, 14–20.

Panov, A.V., Gutekunst, C.-A., Leavitt, B.R., Hayden, M.R., Burke, J.R., Strittmatter, W.J., and Greenamyre, J.T. (2002). Early mitochondrial calcium defects in Huntington's disease are a direct effect of polyglutamines. *Nat. Neurosci.* 5, 731–736.

Panov, A.V., Lund, S., and Greenamyre, J.T. (2005). Ca²⁺-induced permeability transition in human lymphoblastoid cell mitochondria from normal and Huntington's disease individuals. *Mol. Cell. Biochem.* 269, 143–152.

Parker, J.A., Connolly, J.B., Wellington, C., Hayden, M., Dausset, J., and Neri, C. (2001). Expanded polyglutamines in *Caenorhabditis elegans* cause axonal abnormalities and severe dysfunction of PLM mechanosensory neurons without cell death. *Proc. Natl. Acad. Sci.* 98, 13318–13323.

Parker, J.A., Holbert, S., Lambert, E., Abderrahmane, S., and Néri, C. (2004). Genetic and pharmacological suppression of polyglutamine-dependent neuronal dysfunction in *Caenorhabditis elegans*. *J. Mol. Neurosci.* 23, 61–67.

Passani, L.A. (2000). Huntingtin's WW domain partners in Huntington's disease post-mortem brain fulfill genetic criteria for direct involvement in Huntington's disease pathogenesis. *Hum. Mol. Genet.* 9, 2175–2182.

Patra, B., Kon, Y., Yadav, G., Sevold, A.W., Frumkin, J.P., Vallabhajosyula, R.R., Hintze, A., Østman, B., Schossau, J., Bhan, A., et al. (2013). A genome wide dosage suppressor network reveals genetic robustness and a novel mechanism for Huntington's disease. *ArXiv13112554 Q-Bio*.

Patra, B., Kon, Y., Yadav, G., Sevold, A.W., Frumkin, J.P., Vallabhajosyula, R.R., Hintze, A., Østman, B., Schossau, J., Bhan, A., et al. (2017). A genome wide dosage suppressor network reveals genomic robustness. *Nucleic Acids Res.* 45, 255–270.

Paulsen, J.S., Zimbelman, J.L., Hinton, S.C., Langbehn, D.R., Leveroni, C.L., Benjamin, M.L., Reynolds, N.C., and Rao, S.M. (2004). fMRI Biomarker of Early Neuronal Dysfunction in Presymptomatic Huntington's Disease. *Am. J. Neuroradiol.* 25, 1715–1721.

Paulsen, J.S., Langbehn, D.R., Stout, J.C., Aylward, E., Ross, C.A., Nance, M., Guttman, M., Johnson, S., MacDonald, M., Beglinger, L.J., et al. (2008). Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *J. Neurol. Neurosurg. Psychiatry* 79, 874–880.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 2, 559–572.

Peng, K., Li, Y., Long, L., Li, D., Jia, Q., Wang, Y., Shen, Q., Tang, Y., Wen, L., Kung, H., et al. (2010). Knockdown of FoxO3a induces increased neuronal apoptosis during embryonic development in zebrafish. *Neurosci. Lett.* 484, 98–103.

Petrascu-Parwez, E., Nguyen, H.-P., Löbbecke-Schumacher, M., Habbes, H.-W., Wiczorek, S., Riess, O., Andres, K.-H., Dermietzel, R., and Von Hörsten, S. (2007). Cellular and subcellular localization of Huntington aggregates in the brain of a rat transgenic for Huntington disease. *J. Comp. Neurol.* 501, 716–730.

Pouladi, M.A., Morton, A.J., and Hayden, M.R. (2013). Choosing an animal model for the study of Huntington's disease. *Nat. Rev. Neurosci.* 14, 708–721.

Pringsheim, T., Wiltshire, K., Day, L., Dykeman, J., Steeves, T., and Jette, N. (2012). The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. *Mov. Disord.* 27, 1083–1091.

Project*, T.U.S.-V.C.R., and Wexler, N.S. (2004). Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc. Natl. Acad. Sci. U. S. A.* 101, 3498–3503.

Quarrell, O.W.J., Rigby, A.S., Barron, L., Crow, Y., Dalton, A., Dennis, N., Fryer, A.E., Heydon, F., Kinning, E., Lashwood, A., et al. (2007). Reduced penetrance alleles for Huntington's disease: a multi-centre direct observational study. *J. Med. Genet.* 44, e68.

Raamsdonk, J.M.V.J. (2005). Experimental models of Huntington's disease. *Drug Discov. TodayDisease Models* 2, 291; 291-297; 297.

Radulescu, A.E., and Cleveland, D.W. (2010). NuMA after 30 years: the matrix revisited. *Trends Cell Biol.* 20, 214–222.

Ranen, N.G., Stine, O.C., Abbott, M.H., Sherr, M., Codori, A.-M., Franz, M.L., Chao, N.I., Chung, A.S., Pleasant, N., Callahan, C., et al. (1995). Anticipation and Instability of IT-15 (CAG)N Repeats in Parent-Offspring Pairs with Huntington Disease. *Am. J. Hum. Genet.* 57, 593–602.

Rao, R.V., and Bredesen, D.E. (2004). Misfolded proteins, endoplasmic reticulum stress and neurodegeneration. *Curr. Opin. Cell Biol.* 16, 653–662.

Ratovitski, T., Chighladze, E., Arbez, N., Boronina, T., Herbrich, S., Cole, R.N., and Ross, C.A. (2012). Huntingtin protein interactions altered by polyglutamine expansion as determined by quantitative proteomic analysis. *Cell Cycle* 11, 2006–2021.

Ratovitski, T., Chaerkady, R., Kammers, K., Stewart, J.C., Zavala, A., Pletnikova, O., Troncoso, J.C., Rudnicki, D.D., Margolis, R.L., Cole, R.N., et al. (2016). Quantitative Proteomic Analysis Reveals Similarities between Huntington's Disease (HD) and Huntington's Disease-Like 2 (HDL2) Human Brains. *J. Proteome Res.*

Reik, W. (1988). Genomic imprinting: a possible mechanism for the parental origin effect in Huntington's chorea. *J. Med. Genet.* 25, 805–808.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* 44, W83–W89.

Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052.

Ridley, R.M., Frith, C.D., Farrer, L.A., and Conneally, P.M. (1991). Patterns of inheritance of the symptoms of Huntington's disease suggestive of an effect of genomic imprinting. *J. Med. Genet.* 28, 224–231.

Riechers, S.-P., Butland, S., Deng, Y., Skotte, N., Ehrnhoefer, D.E., Russ, J., Laine, J., Laroche, M., Pouladi, M.A., Wanker, E.E., et al. (2016). Interactome network analysis identifies multiple caspase-6 interactors involved in the pathogenesis of HD. *Hum. Mol. Genet.* 25, 1600–1618.

Riva, L., Koeva, M., Yildirim, F., Pirhaji, L., Dinesh, D., Mazor, T., Duennwald, M.L., and Fraenkel, E. (2012). Polyglutamine Expanded Huntingtin Dramatically Alters the Genome-Wide Binding of HSF1. *J. Huntingt. Dis.* 1, 33–45.

Robinow, S., and White, K. (1988). The locus elav of *Drosophila melanogaster* is expressed in neurons at all developmental stages. *Dev. Biol.* 126, 294–303.

Romberg, N., Al Moussawi, K., Nelson-Williams, C., Stiegler, A.L., Loring, E., Choi, M., Overton, J., Meffre, E., Khokha, M.K., Huttner, A.J., et al. (2014).

Mutation of NLRC4 causes a syndrome of enterocolitis and autoinflammation. *Nat. Genet.* 46, 1135–1139.

Rosas, H.D., Koroshetz, W.J., Chen, Y.I., Skeuse, C., Vangel, M., Cudkowicz, M.E., Caplan, K., Marek, K., Seidman, L.J., Makris, N., et al. (2003). Evidence for more widespread cerebral pathology in early HD An MRI-based morphometric analysis. *Neurology* 60, 1615–1620.

Rosenblatt, A., Kumar, B.V., Mo, A., Welsh, C.S., Margolis, R.L., and Ross, C.A. (2012). Age, CAG repeat length, and clinical progression in Huntington's disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* 27, 272–276.

Santamaría, A., Ríos, C., Solís-Hernández, F., Ordaz-Moreno, J., González-Reynoso, L., Altagracia, M., and Kravzov, J. (1996). Systemic dl-Kynurenine and probenecid pretreatment attenuates quinolinic acid-induced neurotoxicity in rats. *Neuropharmacology* 35, 23–28.

Sapp, E., Penney, J., Young, A., Aronin, N., Vonsattel, J.-P., and DiFiglia, M. (1999). Axonal Transport of N-terminal Huntingtin Suggests Early Pathology of Corticostriatal Projections in Huntington Disease. *J. Neuropathol. Exp. Neurol.* 58, 165–173.

Savas, J.N., Makusky, A., Ottosen, S., Baillat, D., Then, F., Krainc, D., Shiekhatter, R., Markey, S.P., and Tanese, N. (2008). Huntington's disease protein contributes to RNA-mediated gene silencing through association with Argonaute and P bodies. *Proc. Natl. Acad. Sci.* 105, 10820–10825.

Sawa, A., Wiegand, G.W., Cooper, J., Margolis, R.L., Sharp, A.H., Lawler, J.F., Greenamyre, J.T., Snyder, S.H., and Ross, C.A. (1999). Increased apoptosis of Huntington disease lymphoblasts associated with repeat length-dependent mitochondrial depolarization. *Nat. Med.* 5, 1194–1198.

Schaefer, M.H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E.E., and Andrade-Navarro, M.A. (2012). HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores. *PLoS ONE* 7, e31826.

Schaffar, G., Breuer, P., Boteva, R., Behrends, C., Tzvetkov, N., Strippel, N., Sakahira, H., Siegers, K., Hayer-Hartl, M., and Hartl, F.U. (2004). Cellular Toxicity of Polyglutamine Expansion Proteins: Mechanism of Transcription Factor Deactivation. *Mol. Cell* 15, 95–105.

Schaper, E., Gascuel, O., and Anisimova, M. (2014). Deep Conservation of Human Protein Tandem Repeats within the Eukaryotes. *Mol. Biol. Evol.* 31, 1132–1148.

Schilling, G., Becher, M.W., Sharp, A.H., Jinnah, H.A., Duan, K., Kotzuc, J.A., Slunt, H.H., Ratovitski, T., Cooper, J.K., Jenkins, N.A., et al. (1999). Intranuclear Inclusions and Neuritic Aggregates in Transgenic Mice Expressing a Mutant N-Terminal Fragment of Huntingtin. *Hum. Mol. Genet.* 8, 397–407.

Schilling, G., Savonenko, A.V., Klevytska, A., Morton, J.L., Tucker, S.M., Poirier, M., Gale, A., Chan, N., Gonzales, V., Slunt, H.H., et al. (2004). Nuclear-targeting of mutant huntingtin fragments produces Huntington's disease-like phenotypes in transgenic mice. *Hum. Mol. Genet.* 13, 1599–1610.

Schneider, S.A., and Bird, T. (2016). Huntington's Disease, Huntington's Disease Look-Alikes, and Benign Hereditary Chorea: What's New? *Mov. Disord. Clin. Pract.* 3, 342–354.

Schönberger, S.J., Jezdic, D., Faull, R.L.M., and Cooper, G.J.S. (2013). Proteomic Analysis of the Human Brain in Huntington's Disease Indicates Pathogenesis by Molecular Processes Linked to other Neurodegenerative Diseases and to Type-2 Diabetes. *J. Huntingt. Dis.* 2, 89–99.

Sequeiros, J., Ramos, E.M., Cerqueira, J., Costa, M.C., Sousa, A., Pinto-Basto, J., and Alonso, I. (2010). Large normal and reduced penetrance alleles in Huntington disease: instability in families and frequency at the laboratory, at the clinic and in the population. *Clin. Genet.* 78, 381–387.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504.

Shirasaki, D.I., Greiner, E.R., Al-Ramahi, I., Gray, M., Boontheung, P., Geschwind, D.H., Botas, J., Coppola, G., Horvath, S., and Loo, J.A. (2012). Network organization of the huntingtin proteomic interactome in mammalian brain. *Neuron* 75, 41–57.

Silva, M.C., Fox, S., Beam, M., Thakkar, H., Amaral, M.D., and Morimoto, R.I. (2011). A Genetic Screening Strategy Identifies Novel Regulators of the Proteostasis Network. *PLoS Genet.* 7, e1002438.

Singh-Taylor, A., Korosi, A., Molet, J., Gunn, B.G., and Baram, T.Z. (2015). Synaptic rewiring of stress-sensitive neurons by early-life experience: A mechanism for resilience? *Neurobiol. Stress* 1, 109–115.

Solans, A., Zambrano, A., Rodríguez, M., and Barrientos, A. (2006). Cytotoxicity of a mutant huntingtin fragment in yeast involves early alterations in mitochondrial OXPHOS complexes II and III. *Hum. Mol. Genet.* 15, 3063–3081.

Sorolla, M.A., Reverter-Branchat, G., Tamarit, J., Ferrer, I., Ros, J., and Cabiscol, E. (2008). Proteomic and oxidative stress analysis in human brain samples of Huntington disease. *Free Radic. Biol. Med.* 45, 667–678.

Soto, C. (2003). Unfolding the role of protein misfolding in neurodegenerative diseases. *Nat. Rev. Neurosci.* 4, 49–60.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., et al. (2005). A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell* 122, 957–968.

St-Pierre, J., Lin, J., Krauss, S., Tarr, P.T., Yang, R., Newgard, C.B., and Spiegelman, B.M. (2003). Bioenergetic Analysis of Peroxisome Proliferator-activated Receptor γ Coactivators 1 α and 1 β (PGC-1 α and PGC-1 β) in Muscle Cells. *J. Biol. Chem.* 278, 26597–26603.

Strand, A.D., Baquet, Z.C., Aragaki, A.K., Holmans, P., Yang, L., Cleren, C., Beal, M.F., Jones, L., Kooperberg, C., Olson, J.M., et al. (2007). Expression Profiling of Huntington's Disease Models Suggests That Brain-Derived Neurotrophic Factor Depletion Plays a Major Role in Striatal Degeneration. *J. Neurosci.* 27, 11758–11768.

Stroedicke, M., Bounab, Y., Strempel, N., Klockmeier, K., Yigit, S., Friedrich, R.P., Chaurasia, G., Li, S., Hesse, F., Riechers, S.-P., et al. (2015). Systematic interaction network filtering identifies CRMP1 as a novel suppressor of huntingtin misfolding and neurotoxicity. *Genome Res.* 25, 701–713.

- T, N., K, I., S, Y., N, A., M, A., and Y, Y. (1985). [An autopsy case of dentatorubropallidoluysian atrophy (DRPLA) clinically diagnosed as Huntington's chorea]. *No To Shinkei* 37, 767–774.
- Tabrizi, S.J., Cleeter, M.W., Xuereb, J., Taanman, J.W., Cooper, J.M., and Schapira, A.H. (1999). Biochemical abnormalities and excitotoxicity in Huntington's disease brain. *Ann. Neurol.* 45, 25–32.
- Tabrizi, S.J., Workman, J., Hart, P.E., Mangiarini, L., Mahal, A., Bates, G., Cooper, J.M., and Schapira, A.H. (2000). Mitochondrial dysfunction and free radical damage in the Huntington R6/2 transgenic mouse. *Ann. Neurol.* 47, 80–86.
- Tagawa, K., Marubuchi, S., Qi, M.-L., Enokido, Y., Tamura, T., Inagaki, R., Murata, M., Kanazawa, I., Wanker, E.E., and Okazawa, H. (2007). The Induction Levels of Heat Shock Protein 70 Differentiate the Vulnerabilities to Mutant Huntingtin among Neuronal Subtypes. *J. Neurosci.* 27, 868–880.
- Talwar, P., Silla, Y., Grover, S., Gupta, M., Agarwal, R., Kushwaha, S., and Kukreti, R. (2014). Genomic convergence and network analysis approach to identify candidate genes in Alzheimer's disease. *BMC Genomics* 15, 199.
- Tassicker, R., Marshall, P., Liebeck, T., Keville, M., Singaram, B., and Richards, F. (2006). Predictive and pre-natal testing for Huntington Disease in Australia: results and challenges encountered during a 10-year period (1994-2003). *Clin. Genet.* 70, 480–489.
- Tauchi, H., Kobayashi, J., Morishima, K., van Gent, D.C., Shiraishi, T., Verkaik, N.S., vanHeems, D., Ito, E., Nakamura, A., Sonoda, E., et al. (2002). Nbs1 is essential for DNA repair by homologous recombination in higher vertebrate cells. *Nature* 420, 93–98.
- Thalappilly, S., Sadasivam, S., Radha, V., and Swarup, G. (2006). Involvement of caspase 1 and its activator Ipaf upstream of mitochondrial events in apoptosis. *FEBS J.* 273, 2766–2778.
- The Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 267–288.
- Toshiyuki, M., and Reed, J.C. (1995). Tumor suppressor p53 is a direct transcriptional activator of the human bax gene. *Cell* 80, 293–299.

- Tourette, C., Li, B., Bell, R., O'Hare, S., Kaltenbach, L.S., Mooney, S.D., and Hughes, R.E. (2014). A Large Scale Huntingtin Protein Interaction Network Implicates Rho GTPase Signaling Pathways in Huntington Disease. *J. Biol. Chem.* 289, 6709–6726.
- Toyoshima, Y., Yamada, M., Onodera, O., Shimohata, M., Inenaga, C., Fujita, N., Morita, M., Tsuji, S., and Takahashi, H. (2004). SCA17 homozygote showing Huntington's disease-like phenotype. *Ann. Neurol.* 55, 281–286.
- Trettel, F., Rigamonti, D., Hilditch-Maguire, P., Wheeler, V.C., Sharp, A.H., Persichetti, F., Cattaneo, E., and MacDonald, M.E. (2000). Dominant phenotypes produced by the HD mutation in STHdhQ111 striatal cells. *Hum. Mol. Genet.* 9, 2799–2809.
- Trushina, E., Dyer, R.B., Badger, J.D., Ure, D., Eide, L., Tran, D.D., Vrieze, B.T., Legendre-Guillemain, V., McPherson, P.S., Mandavilli, B.S., et al. (2004). Mutant Huntingtin Impairs Axonal Trafficking in Mammalian Neurons In Vivo and In Vitro. *Mol. Cell. Biol.* 24, 8195–8209.
- Tsoi, H., Lau, T.C.-K., Tsang, S.-Y., Lau, K.-F., and Chan, H.Y.E. (2012). CAG expansion induces nucleolar stress in polyglutamine diseases. *Proc. Natl. Acad. Sci. U. S. A.* 109, 13428–13433.
- Valle, I., Álvarez-Barrientos, A., Arza, E., Lamas, S., and Monsalve, M. (2005). PGC-1 α regulates the mitochondrial antioxidant defense system in vascular endothelial cells. *Cardiovasc. Res.* 66, 562–573.
- Valor, L.M. (2014). Transcription, Epigenetics and Ameliorative Strategies in Huntington's Disease: a Genome-Wide Perspective. *Mol. Neurobiol.* 1–18.
- Valor, L.M., and Guiretti, D. (2014). What's wrong with epigenetics in Huntington's disease? *Neuropharmacology* 80, 103–114.
- Valor, L.M., Guiretti, D., Lopez-Atalaya, J.P., and Barco, A. (2013). Genomic Landscape of Transcriptional and Epigenetic Dysregulation in Early Onset Polyglutamine Disease. *J. Neurosci.* 33, 10471–10482.
- Velier, J., Kim, M., Schwarz, C., Kim, T.W., Sapp, E., Chase, K., Aronin, N., and DiFiglia, M. (1998). Wild-Type and Mutant Huntingtins Function in Vesicle Trafficking in the Secretory and Endocytic Pathways. *Exp. Neurol.* 152, 34–40.
- Wacker, J.L., Huang, S.-Y., Steele, A.D., Aron, R., Lotz, G.P., Nguyen, Q., Giorgini, F., Roberson, E.D., Lindquist, S., Masliah, E., et al. (2009). Loss of Hsp70

Exacerbates Pathogenesis But Not Levels of Fibrillar Aggregates in a Mouse Model of Huntington's Disease. *J. Neurosci.* 29, 9104–9114.

Walker, F.F.O. (2007). Huntington's disease. *Lancet Br. Ed.* 369, 218; 218-228; 228.

Wang, N., Lu, X.-H., Sandoval, S.V., and Yang, X.W. (2013). An Independent Study of the Preclinical Efficacy of C2-8 in the R6/2 Transgenic Mouse Model of Huntington's Disease. *J. Huntingt. Dis.* 2, 443–451.

Wang, R.-H., Sengupta, K., Li, C., Kim, H.-S., Cao, L., Xiao, C., Kim, S., Xu, X., Zheng, Y., Chilton, B., et al. (2008). Impaired DNA Damage Response, Genome Instability, and Tumorigenesis in SIRT1 Mutant Mice. *Cancer Cell* 14, 312–323.

Wanker, E.E., Rovira, C., Scherzinger, E., Hasenbank, R., Walter, S., Tait, D., Colicelli, J., and Lehrach, H. (1997). HIP-I: A huntingtin interacting protein isolated by the yeast two-hybrid system. *Hum. Mol. Genet.* 6, 487–495.

Watts, D.J., and Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442.

Watts, R.L., and Koller, W.C. (1997). *Movement disorders: neurologic principles and practice* (New York: McGrawHill, Health Professions Division).

Wedderburn, R.W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika* 61, 439–447.

Weiner, W.J., and Lang, A.E. (1989). *Movement disorders: a comprehensive survey* (Mount Kisco, N.Y.: Futura Pub. Co.).

Wellington, C.L., Ellerby, L.M., Hackam, A.S., Margolis, R.L., Trifiro, M.A., Singaraja, R., McCutcheon, K., Salvesen, G.S., Propp, S.S., Bromm, M., et al. (1998). Caspase Cleavage of Gene Products Associated with Triplet Expansion Disorders Generates Truncated Fragments Containing the Polyglutamine Tract. *J. Biol. Chem.* 273, 9158–9167.

Wheeler, V.C., Gutekunst, C.-A., Vrbanc, V., Lebel, L.-A., Schilling, G., Hersch, S., Friedlander, R.M., Gusella, J.F., Vonsattel, J.-P., Borchelt, D.R., et al. (2002). Early phenotypes that presage late-onset neurodegenerative disease allow testing of modifiers in Hdh CAG knock-in mice. *Hum. Mol. Genet.* 11, 633–640.

White, J.K., Auerbach, W., Duyao, M.P., Vonsattel, J.-P., Gusella, J.F., Joyner, A.L., and MacDonald, M.E. (1997). Huntingtin is required for neurogenesis and is

not impaired by the Huntington's disease CAG expansion. *Nat. Genet.* *17*, 404–410.

Willingham, S. (2003). Yeast Genes That Enhance the Toxicity of a Mutant Huntingtin Fragment or -Synuclein. *Science* *302*, 1769–1772.

Wu, Z., Puigserver, P., Andersson, U., Zhang, C., Adelmant, G., Mootha, V., Troy, A., Cinti, S., Lowell, B., Scarpulla, R.C., et al. (1999). Mechanisms Controlling Mitochondrial Biogenesis and Respiration through the Thermogenic Coactivator PGC-1. *Cell* *98*, 115–124.

Wytenbach, A., Swartz, J., Kita, H., Thykjaer, T., Carmichael, J., Bradley, J., Brown, R., Maxwell, M., Schapira, A., Orntoft, T.F., et al. (2001). Polyglutamine expansions cause decreased CRE-mediated transcription and early gene expression changes prior to cell death in an inducible cell model of Huntington's disease. *Hum. Mol. Genet.* *10*, 1829–1845.

Yamamoto, A., Cremona, M.L., and Rothman, J.E. (2006). Autophagy-mediated clearance of huntingtin aggregates triggered by the insulin-signaling pathway. *J. Cell Biol.* *172*, 719–731.

Young, A.B., Greenamyre, J.T., Hollingsworth, Z., Albin, R., D'Amato, C., Shoulson, I., and Penney, J.B. (1988). NMDA receptor losses in putamen from patients with Huntington's disease. *Science* *241*, 981–983.

Yu, J., Zhang, L., Hwang, P.M., Kinzler, K.W., and Vogelstein, B. (2001). PUMA Induces the Rapid Apoptosis of Colorectal Cancer Cells. *Mol. Cell* *7*, 673–682.

Yuan, Z., Zhang, X., Sengupta, N., Lane, W.S., and Seto, E. (2007). SIRT1 Regulates the Function of the Nijmegen Breakage Syndrome Protein. *Mol. Cell* *27*, 149–162.

Zala, D., Benchoua, A., Brouillet, E., Perrin, V., Gaillard, M.-C., Zurn, A.D., Aebischer, P., and Déglon, N. (2005). Progressive and selective striatal degeneration in primary neuronal cultures using lentiviral vector coding for a mutant huntingtin fragment. *Neurobiol. Dis.* *20*, 785–798.

Zeitlin, S., Liu, J.-P., Chapman, D.L., Papaioannou, V.E., and Efstratiadis, A. (1995). Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. *Nat. Genet.* *11*, 155–163.

Zhang, S., Binari, R., Zhou, R., and Perrimon, N. (2010). A Genomewide RNA Interference Screen for Modifiers of Aggregates Formation by Mutant Huntingtin in *Drosophila*. *Genetics* *184*, 1165–1179.

Zhang, X., Smith, D.L., Meriin, A.B., Engemann, S., Russel, D.E., Roark, M., Washington, S.L., Maxwell, M.M., Marsh, J.L., Thompson, L.M., et al. (2005). A potent small molecule inhibits polyglutamine aggregation in Huntington's disease neurons and suppresses neurodegeneration in vivo. *Proc. Natl. Acad. Sci. U. S. A.* *102*, 892–897.

Zwilling, D., Huang, S.-Y., Sathyaikumar, K.V., Notarangelo, F.M., Guidetti, P., Wu, H.-Q., Lee, J., Truong, J., Andrews-Zwilling, Y., Hsieh, E.W., et al. (2011). Kynurenine 3-Monooxygenase Inhibition in Blood Ameliorates Neurodegeneration. *Cell* *145*, 863–874.