Characterization of the Huntingtin Gene Promoter and Huntingtin Transcriptional Regulation

by

Rebecca Anne Grace De Souza

B.Sc., McMaster University, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2015

Abstract

Huntington's disease (HD) is a late onset, neurological, autosomal dominant genetic disorder. Despite being associated to a defined genetic mutation within the huntingtin gene (*HTT*), little is known about its transcriptional regulation. *HTT* is expressed, at varying levels, throughout the body. At the current time, the transcriptional regulation mechanisms controlling this differential expression pattern are unknown. Previous studies have focused on the genomic region directly preceding *HTT*'s transcriptional start site. The purpose of this thesis was to utilize the current understanding of mammalian transcriptional regulation to further characterize the *HTT* promoter and to expand the search for transcriptional regulatory regions outside the promoter.

To direct this search a bioinfomatic screen was conducted, which identified 11 putative regions. Potential transcription factor binding sites (TFBSs) within these regions were identified through the use of available chIP-seq datasets. Curation of the TFBSs within the putative regions lead to selection of the 9th region, in addition to the promoter, for further study.

To test the functionality of region 9 and identified candidate transcription factors (TFs), a panel of human kidney and rat neuronal cell lines were established. These cell lines stably expressed either the *HTT* promoter or region 9 luciferase constructs. Candidate TFs were tested using siRNA mediated knockdown. Knockdown of selected candidate TFs did not modulate *HTT* promoter function.

The role of DNA methylation on transcriptional regulation of *HTT* was also explored using the Illumina 450K Methylation Array. Tissue specific DNA methylation of *HTT* using human cortex and liver tissues identified 33 differentially methylated sites. The role of the HD mutation on local and global DNA methylation was also investigated, finding no changes to local DNA and 15 differentially methylated regions globally.

In conclusion, a data driven bioinfomatic search has expanded potential regulatory regions beyond that of the promoter of the *HTT* gene. A first attempt at identifying crucial TFs involved in *HTT* regulation was not successful, however additional candidates remain to be tested. A role for DNA methylation in tissue specific regulation of *HTT* has been identified, while the HD mutation itself does not appear to affect *HTT* DNA methylation.

Preface

All of the work presented was conducted at the Center for Molecular Medicine and Therapeutics, part of the Child and Family Research Institute at the University of British Columbia, Children's and Women's Hospital campus. Methods for the collection of mice tissue was approved by the University of British Columbia's Animal Care Committee (Certificate #A14-0031). Human tissues used were collected by the 'Huntington Disease BioBank at the University of British Columbia' under the University of British Columbia, Children and Women's Hospital Ethics board (Certificate #H06-70467) and the Vancouver Coastal Health Authority Research study board (Certificate #V09-0129). Approval for the use of biobank human tissue was approved under the 'Genetic Modifiers of Huntington Disease and Huntington Disease Phenocopy Disorders' study by the University of British Columbia, Children and Women's Hospital Ethics board (Certificate #H05-70532).

Chapter 1: Sections1.2.1 - 1.2.7 have been published in De Souza, R. A. G., & Leavitt, B. R. (2014). Neurobiology of Huntington's Disease. Current Topics in Behavioral Neurosciences, (Chapter 353)(Reproduced with permission, licence no. 3553451337816).

Chapter 2: Chapter 2 contains unpublished data that is part of a manuscript in preparation, with myself as the first author. The scoring of the region was conducted by an undergraduate volunteer, Natalia Kosior, under my supervision. The scoring

paradigm was designed in the Wasserman lab and recommended by Dr. Wyeth Wasserman. The amalgamation of the separate 1 kb scored regions into the final list of candidates was conducted by myself, Natalia, and Dr. Wyeth Wasserman. Screening and identification of transcription factor binding sites, using available chIP-seq datasets, as well as modelling of DNase hypersensitivity datasets over the candidate regions was performed by Anthony Mathelier from the Wasserman Lab. Selection of Region 9 for follow up, as well as candidate transcription factors, was done by myself in consultation with Dr. Wyeth Wasserman.

Chapter 3: Chapter 3 contains unpublished data. The huntingtin promoter-luciferase plasmids were generated in the Leavitt Lab prior to my association by Dr. Kristina Becanovic. The ARE plasmid was a generous gift from the Wasserman lab. Modification of these plasmids for use in the FLP-In System (Invitrogen), as well as the creation of the Region 9 plasmid was done by myself with technical assistance from Jesslyn Beck and Pamela Wagner. Generation, propagation, characterization, and cryopreservation of the various HEK293 and ST14 stably expressing cell lines was conducted by myself, again with technical assistance from Jesslyn Beck and Pamala Wagner. The parental HEK293-FLPIn cell line was purchased fro Invitrogen while the parental ST14 cell line was a generous gift from Dr. Elena Cantaneo. The optimization of the assay used with these cell lines was done by myself.

Chapter 4: Chapter 4 contains unpublished data. The transfection, luciferase assay, quantitative real time PCR, and analysis of all the selected TFs using siRNA targets was conducted by myself with technical assistance from Austin Hill, Pamela Wagner and Kathleen De Asia. Western blots were performed by Ge Lu.

Chapter 5: Chapter 5 contains unpublished data that is part of a manuscript in preparation, with myself as the first author. The dissection of all mouse tissues was conducted by Ge Lu, followed by quantitative real time PCR conducted by myself. Selection of appropriate human samples from the UBC Huntington's Disease Bio Bank was done by myself with assistance from Dr. Sonia Franciosi. Preparation of tissues from the biobank was conducted by myself with assistance from Dr. Franciosi and Austin Hill. The genomic DNA was prepared for, and run, on the Illumina 450K DNA Methylation Array by Luchia Lam and Sarah Mah from the Kobor Lab. Analysis of the 450K array data was done, under my direction, by Sumaiya Islam from the Kobor lab. Pyrosequencing was performed by Lisa McEwen, also from the Kobor Lab. Quantitative real time PCR and subsequent analysis was conducted by myself.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A2AR	Adenosine A2A receptor
ActD	Actinomycin D
ALS	Amyotropic lateral sclerosis
ARE	antioxidant responsive element
BDNF	brain-derived neurotropic factor
C4orf44	chromosome 4 open reading frame 44
CAGE	capped analysis of gene expression
ChIP-chip	chromatin immunopreceiptation followed by chip array
ChIP-chip	chromatin immunopreciptation
ChIP-seq	chromatin immunoprecipitaiton follwed by Next-Generation sequencing
CV	coefficient of variation, experimental
CV	coefficient of variation, total
DM1	myotonic dystrophy type 1
DMRs	differentially methylated regions
GEFdb	GLUT4 enhancing factor DNA binding domain
GOI	gene of interest
HAP1	huntingtin-associated protein 1
HD	Huntinton's Disease
Hdh	mouse huntingtin gene
HEK293	human embryonic kidney cell line
HTS	high-throughput screen
HTT	huntingtin
IRF1	Interferon regulatory factor 1
JunB	jun B proto-oncogne
Luc	firefly luciferase gene
mRNA	messenger RNA
MSANTD 1	Myv/SANT DNA binding domain like 1
MSNs	medium-spiny nerons
NCoR	nuclear receptor co-repressor
Next-Gen sequencing	Next-Generation DNA sequencing
NfkB	Nuclear Factor Of Kappa Light Polypeptide Gene Enhancer In B-Cells
p53	tumor protein p53
PBF	papillomavirus binding factor
PC	principal compent
PCA	principal component analysis
PCR	polymerase chain reaction
PMI	post-mortem index
PWM	position weight matrix
qRT-PCR	quantitative realtime PCR
REST/NRSF	RE1-silencing transcription factor, also known as neuronal restrictive silencing factor
Rfx5	regulatory factor X 5
RNApolII	RNA polymerase II
Smc3	structural maintenance of chromosome 3
SNP	single polynucleotide polymorphism
ST14	rat medium spiny neuron progenitor cell line
SWAN	subset-within-array normalization
TFBSs	transcription factor binding sites
TFs	transcription factors
TSS	transcriptional start site
UTR	untranslated region
VNTR	variable number tandem repeats

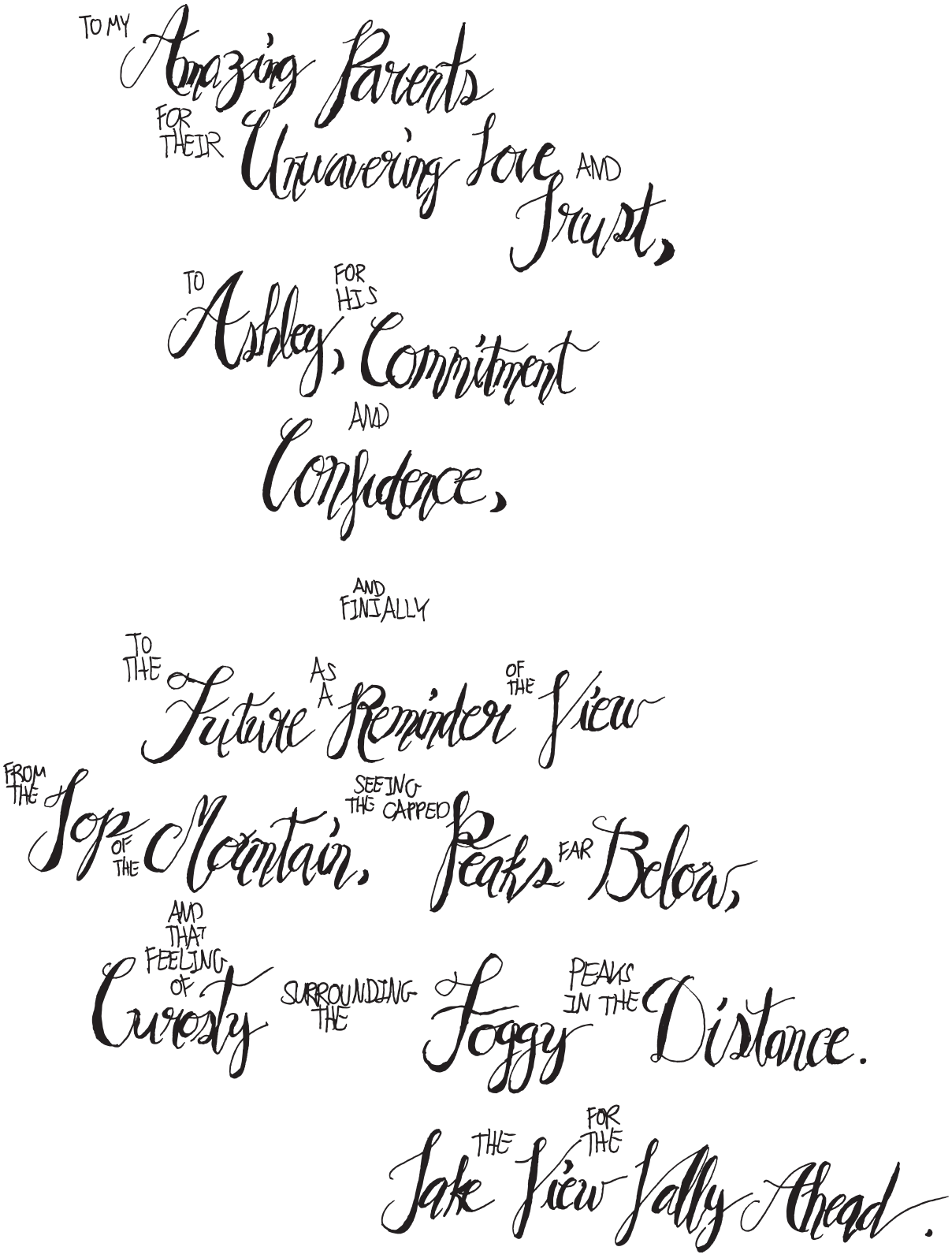
Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Blair R. Leavitt, for his guidance, council, and humour throughout the completion of this work. I would also like to thank my supervisory committee - Lynn Raymond, Wyeth Wasserman, Murry Webb, and Ryan Brinkman - for their guidance and advice.

Secondly, I thank my colleagues from the Leavitt Lab: Terri Petkau, Austin Hill, Pamela Wagner, Colum Connolly, Natalia Kosior, Ge Lu, and Kathleen De Asia for contributing their technical skills and theoretical prowess to my research. I would like to thank Kristina Becanovic, Jesslyn Beck, and Jenny Thiele for their mentorship at the beginning of my studies. I thank Angela Gurney for her help in navigating the non-science waterways and for many cups of tea. I would also like to thank my collaborators Anthony Mathelier and Sumaiya Islam, from the Wasserman and Kobor labs respectively, for their contributions to my work. I would like to thank CIHR and CHDI for funding my research.

Finally, I thank my incredible family and friends for their invaluable support and encouragement. I would especially like to thank my parents Annie and Paul De Souza and my Aunts Jovita and Bridget De Souza for, to put it simply, everything as their contributions are too numerous to name here. I thank Stephanie Dziengo for the support only a best friend can provide. I also thank Charlotte Peace and Sarah Krzyzek for numerous sanity providing craft nights. Last, but not least, I would like to thank my partner Ashley Perry for his constant positivity and his ability to dissolve my self doubt thunder clouds into supportive sunshine.

Thank you all.


TO MY Amazing Parents
FOR THEIR Unwavering Love AND Trust,
TO Ashley, FOR HIS Commitment AND Confidence,
AND FINIALLY
TO THE Future AS A Reminder OF THE View
FROM THE Top OF THE Mountain, SEEING THE CAPPED Peaks FAR Below,
AND THAT FEELING OF Curosity SURROUNDING THE Foggy PEAKS IN THE Distance.
THE Lake FOR THE View Fally Ahead.

1 Introduction

1.1 Thesis Introduction

To the families of those suffering from Huntington's Disease (HD), knowing that we as a research community have identified the causative gene and mutation is often little comfort to the daily reality of living with HD. This is not surprising given that after the discovery of the huntingtin gene (*HTT*) over 20 years ago, the prevailing notion at the time was that since we had found the gene a cure, or at the very least a treatment, would follow in short course. While we have learnt much about both the wildtype and mutant functions of the HTT protein, our hubris in assuming that our knowledge of the gene would lead swiftly to a cure has been exposed. This is not due to a lack of study indeed, the *HTT* gene is one of the most extensively studied genes with over 14752 hits in PubMed to date, but rather to our lack of knowledge of the gene and, perhaps to the optimistic assumption that HTT would have a straightforward and singular function. Unbeknownst at the time were the varied and multiple potential cellular functions of the HTT protein, not only in the brain and central nervous system, where the majority of disease pathogenesis occurs, but also during development and in peripheral tissues as well. The list of functional pathways in which HTT appears to play a role continues to grow, with new functions and pathways continually being discovered. We have also learned much about the HD mutation and the resulting impact on these cellular functions, as well as the discovery of additional toxic functions conferred on the protein by the HD

mutation. Given the large number of scientists studying both the wildtype and mutant functions of HTT, it may come as a surprise to learn that within this decade of intensive study comparatively little time has been devoted to understanding the mechanisms by which *HTT* transcription is regulated, with just over a dozen papers focused on this subject. The majority of these studies were conducted before or around the advent of genome-wide technologies, which have greatly changed our views on transcriptional regulation. As many current therapies being developed for HD center around gene therapies aimed at eliminating or reducing the mutant *HTT* protein, it is clear that having an understanding of how *HTT* is regulated and consequently expressed will be essential. In addition, and as will be discussed later in this introductory chapter, understanding how the differential expression of *HTT* across different tissues is established will also allow for the development of additional therapies to be used in conjunction with gene therapies for better regulation of HD pathogenesis.

To this end, this doctoral thesis is focused on utilizing new methods of studying mammalian transcriptional regulation to better understand the transcriptional regulatory regions of the *HTT* gene. In particular, I have broadened the somewhat narrow scope of regulatory regions studied in the past and applied the knowledge of differential expression of *HTT* across tissues to both update our understanding of *HTT* transcriptional regulation and to provide a basis for understanding its unique characteristics. In order to provide a comprehensive basis for understanding the modes and methods I have utilized it is important to introduce not only the *HTT* gene and protein but also our current understanding of

transcriptional regulation. To accomplish this I have divided this introductory chapter into two parts; the first being a introduction to HD pathogenesis and the wildtype/ mutant HTT gene/protein, with the omission of published research concerning the transcriptional regulation of *HTT*. Following this is an introduction to mammalian gene transcription including current methods of identifying transcription factor biding site motifs within the genome, definitions of regulatory regions, and epigenetic signatures of regulatory regions. Finally, a review of *HTT* transcriptional regulation in the literature is presented, highlighting our current basis of understanding *HTT* transcriptional regulation. By setting up the introductory chapter this way I hope to establish a clear reasoning for the methods of investigation I have selected and to provide a solid basis of understanding through which this research can be interpreted.

1.2 Huntington's Disease

1.2.1 Introduction

In 1872 George Huntington published a concise report entitled "On Chorea" on his study of a seemingly hereditary and discrete disorder found in several generations of an American family (Huntington 1872). It was described as a disease primarily affecting adult individuals, causing both motor and cognitive/psychiatric deficits. This clinical report so encapsulated the symptoms and hereditary nature of this particular disorder, that the disease itself became named Huntington's Chorea after George Huntington, and is known as Huntington's disease today. HD is a

progressive neurological disorder with predominantly adult-onset that occurs in all human populations (world wide prevalence of 2.71 per 100,000 individuals), but has the highest prevalence in individuals of European origin (Pringsheim et al. 2012). Due to the hereditary nature and strong penetrance of HD this disease has become the embodiment of efforts aimed at treating adult-onset neurological disorders. This introductory chapter will outline the main clinical, genetic, and neurological features of HD; review the current understanding of wildtype *HTT* gene function; and the effects of the disease-causing mutation on this gene and their potential pathogenic role in this devastating neurodegenerative disease. In addition, I will explore the relatively understudied field of *HTT* transcriptional regulation, with a particular focus on the composition and function of the *HTT* proximal promoter.

1.2.1 Clinical Features

Many of the clinical features first described by George Huntington in “On Chorea” remain part of the clinical diagnosis used today. Primary of these are the movement disorders associated with HD, with chorea being the defining feature. Chorea is described as abnormal “dance-like” movements of the limbs, trunk or face. The movement abnormalities in HD affect both voluntary and involuntary motor function, with motor disturbances often occurring in a sequential manner as disease progresses (Mahant et al. 2003). During the first stages of HD the motor symptoms primarily include involuntary movements with chorea, hypotonia and hyper-reflexia being the motor symptoms most commonly observed. As the disease progresses, voluntary movements become affected with rigidity and bradykinesia being most

prominent. Voluntary motor abnormalities become more severe later in disease progression, and contribute to the majority of functional disability described by HD patients.

Cognitive deficits also follow a similar progression, generally beginning slowly at first and worsening as the disease progresses. Early in the disease process, patients often present with a slowing of intellectual processes, changes in personality, dis-inhibition, and reduced mental flexibility (Craufurd & Snowden 2002). As the disease worsens there is a reduction in cognitive speed, flexibility, concentration, and patients progress towards “subcortical dementia” with aphasia and agnosia not commonly present. Neuropsychiatric symptoms are not uncommon in HD patients; however, they do not follow disease progression like the motor and cognitive deficits. These symptoms include depression, apathy, suicidal ideation, and anxiety (Craufurd & Snowden 2002). While the motor and cognitive symptoms described above are the major characteristics of HD and remain the focus of clinical diagnosis and treatment, other symptoms are also prevalent, and though they have historically received less attention,they still significantly impact the patient’s quality of life. These additional symptoms include sleep and circadian rhythm disorders, metabolic abnormalities (primarily weight loss despite normal caloric intake), and testicular degeneration (Van Raamsdonk et al. 2007; Craufurd & Snowden 2002).

1.2.3 Neuropathology

The most striking neuropathologic feature of HD is the relatively selective and progressive degeneration of the caudate and putamen (collectively known as the

striatum) (Vonsattel & DiFiglia 1998). This progressive shrinkage of the striatum correlates with disease progression and severity, with almost no striatal tissue remaining in very advanced stages of the disease. The neurodegeneration of the striatum is cell type specific, with the medium-spiny neurons (MSNs) being the neuronal sub-type specifically affected in HD. The other major type of neurons in the striatum, aspiny interneurons, are largely unaffected and relatively spared in HD. While other brain regions also show evidence of selective neurodegeneration in HD, none of them display degeneration to the extent seen in the striatum and most of degeneration outside of the striatum occurs later in disease progression (Hedreen et al. 1991; Spargo et al. 1993). These additional regions include the pyramidal projection neurons in layers V and VI of the cerebral cortex, the CA1 region of the hippocampus, the globus pallidus, subthalamic nucleus, substantia nigra, cerebellum and thalamus. Ubiquitinated HTT protein inclusions or aggregates are also a key neuropathologic feature of HD that were first described in mouse models of HD, and were then later observed in HD patients (Davies et al. 1997).

1.2.4 Genetics

Despite the early characterization of HD as a well-defined clinical entity and subsequent studies that clearly established HD as an autosomal dominant disorder, the discovery of the causative gene did not occur until 1993 (Knight et al. 1993). This was due in part to the novel nature (at the time) of the HD mutation, a trinucleotide repeat expansion of the nucleotides cytosine, adenine and guanine (a CAG repeat expansion). The discovery of the gene responsible for HD required large HD families

to be identified and for novel molecular techniques capable of resolving the CAG expansion to be developed. Since its discovery in 1993 we have gained a better, though by no means comprehensive, understanding of the Huntington's disease gene, renamed the huntingtin gene (or *HTT* gene) from its original designation of IT15 (standing for Interesting Transcript 15).

The CAG repeat expansion is a dynamic mutation, meaning that the length of the CAG expansion on a single allele can expand or contract as it is transmitted between generations and can vary between the cells of a single individual (Gonitel et al. 2008). In the following sections I will discuss the main features of the *HTT* gene and protein and how these features are affected in the presence of a pathogenic CAG expansion.

1.2.5 Normal Huntingtin Function

1.2.5.1 *HTT* Gene

The identification of the *HTT* gene as the mutated gene in HD in 1993 has led to 21 years of intensive study into both the wildtype and mutated function of the *HTT* gene. Despite this, there is much about the function of this protein that remains unknown. This section describes our current understanding of the wildtype function of the HTT protein. The *HTT* gene is found on chromosome 4 and is comprised of 67 exons that are translated to form a 348 kDa protein. While the *HTT* gene itself can be found in both invertebrates and vertebrates it is most highly conserved among vertebrates (80%) (Baxendale et al. 1995; Gissi et al. 2006; Tartari et al. 2008). The

protein sequence of HTT, however, bears little overall homology to any other known protein, making inferences about the function of the HTT protein through comparisons of related proteins difficult, though comparisons to known protein motifs within the HTT protein are possible.

1.2.5.2 Huntingtin Protein Sequence

Through computational analyses, searching for similarities to known protein domains, HTT has been found to contain 37 putative HEAT domains that are thought to be involved in protein-protein interactions (Andrade & Bork 1995; Takano & Gusella 2002). These 37 HEAT domains are conserved throughout vertebrates and suggest that the protein-protein interactions they dictate are also similar across vertebrates. Shortly upstream of the HEAT repeats is another region conserved only in higher vertebrates designated as the polyproline stretch (Ehrnhoefer et al. 2011). This short repeat of proline amino acids is thought to be important in the folding of the HTT protein and may function to keep the protein soluble (Bhattacharyya et al. 2006). Sequence analysis has also revealed a fully functional and active C-terminal nuclear export signal and a less active nuclear localization signal (Xia et al. 2003). The presence of the nuclear export signal and of the multiple HEAT protein-protein interaction domains suggest that HTT may be involved in transportation of molecules from the nucleus.

1.2.5.3 Post-Translational Modification of Huntingtin

Consistent with the role for HTT in molecular transport hypothesized by its HEAT domains are some of its post-translational modifications, namely palmitoylation (Young et al. 2012). Palmitoylated proteins are often involved in the assembly of vesicle trafficking control complexes and the assembly of synaptic vesicle function complexes (Huang et al. 2004). HTT is palmitoylated by huntingtin-interacting protein 14. In addition to palmitoylation, HTT is also sumoylated and ubiquitinated at N-terminal lysines K6, K9 and K15 (Kalchman et al. 1996; Steffan et al. 2004). The proteins which sumoylate or ubiquintinate these sites compete for modification, with ubiquitination being a marker for degradation of the HTT protein through the ubiquitin-ligase degradation pathway and sumoylation preventing this degradation (Ehrnhoefer et al. 2011). HTT is also phosphorylated at serines 421 and 434, and this phosphorylation influences the cellular localization, function, and cleavage of the HTT protein, discussed below (Humbert et al. 2002; Luo et al. 2005; Warby et al. 2005).

1.2.5.4 Cleavage of the Huntingtin Protein

The HTT protein contains three protease cleavage consensus sites that generate shorter fragments from the full-length HTT protein (Goldberg et al. 1996; Wellington et al. 1998). In addition to these three sites, there are an additional three caspase cleavage sites and two calpain cleavage sites N-terminal to the primary caspase cleavage sites. Both the wildtype and mutant forms of the HTT protein can

be cleaved, producing fragments of varying length, function and cellular localization. Brain region-specific cleavage of HTT protein has also been described, suggesting that different fragment lengths may have cell type specific functions, adding additional complexity to the function of this protein (Mende-Mueller et al. 2001).

1.2.5.5 Developmental Function of Huntingtin

So far I have discussed genetic and molecular features that have given us potential insights into the function of the HTT protein. I now turn to functional studies based on this information to further explore the function of this protein. First, given *HTT*'s high conservation across several phyla one might infer that loss of the *HTT* gene would have serious consequences for the organism. This has been shown to be true in mouse *HTT* "knockout" models (*Hdh* nullizygous mice) which are embryonic lethal (Nasir et al. 1995; Duyao et al. 1995; Zeitlin et al. 1995). Interestingly, embryonic lethality in these mice occurs before embryonic day 8.5. This is prior to gastrulation and formation of the nervous system, highlighting the role of HTT not only within the CNS, but in other peripheral tissues. In this case it seems that complete loss of HTT within extra-embryonic tissues results in defects in organization of these tissues causing embryonic lethality (Leavitt et al. 2001; Van Raamsdonk et al. 2005a). Other studies have shown that a reduction in embryonic HTT, below 50%, after gastrulation, while not lethal, leads to abnormalities in the formation of the epiblast (White et al. 1997). This is the structure that gives rise to the neural tube. These abnormalities in turn lead to a reduction in neurogenesis and abnormalities in the structure of the cortex and striatum. In addition to these studies,

work with chimeric mice, where *Hdh* null stem cells were introduced into a wildtype blastocyst, revealed a region-specific need for HTT in both the cortex and striatum as these regions were devoid of *Hdh* null cells (Reiner et al. 2003). Combined, these studies highlight HTT's important function early in development, both outside of the CNS and within.

1.2.5.6 Huntingtin and Cellular Survival

The studies highlighted above demonstrate HTT's important role in actively dividing cells during development, but it also appears to play a role in post-mitotic cells of the adult brain. HD is an adult-onset disorder and HTT is widely-expressed in adult tissues, especially post-mitotic neuronal cells. It is important to understand the potential differences in HTT function between these different cell types. As described previously, HTT has been suggested to play a role in cellular survival. This proposed function has been studied in both *in vitro* and *in vivo* experiments. Stable over-expression of wildtype human HTT in conditionally immortalized striatum-derived cells was able to protect against various toxic stimuli (Rigamonti et al. 2000). *In vivo* it has been found that over-expression of wildtype HTT protects against ischemic and excitotoxic injury (Yu Zhang et al. 2003; Leavitt et al. 2006). The mechanisms through which HTT conducts its pro-survival activities have been suggested to include prevention of processing of pro-caspase 9, as well as preventing the formation of the pro-apoptotic protein huntingtin-interacting protein 1 complex (Rigamonti et al. 2000; Rigamonti et al. 2001).

One of the most convincing mechanisms for HTT's pro-survival role in the CNS is via HTT's regulation of brain-derived neurotrophic factor (BDNF), a neurotrophin that plays a critical role in the survival of striatal cells. BDNF is not produced in the striatum itself, but is expressed by pyramidal cells in the cerebral cortex and anterogradely transported along cortico-striatal afferents, where it is released at axon terminals and taken up by striatal neurons (Mizuno et al. 1994; Ventimiglia et al. 1995; Altar et al. 1997). This interaction between cortical and striatal neurons allows the striatal neurons to be resistant to glutamate-mediated excitotoxic neurodegeneration (Bemelmans et al. 1999; Pineda et al. 2005). The HTT protein has also been shown to act directly on BDNF levels, with *in vivo* and *in vitro* data showing that over-expression of wildtype HTT results in an increase in BDNF levels and over-expression of mutant HTT results in a decrease in BDNF levels (Zuccato et al. 2001). Further analysis has shown that this increase is due to HTT acting directly on transcription of BDNF through HTT's interaction and sequestration of RE1-silencing transcription factor (REST; also known as neuronal restrictive silencing factor NRSF). REST/NRSF binds to a repressor element in the BDNF promoter responsible for generating the particular type of BDNF that is transported to the striatum (Zuccato et al. 2001; Zuccato et al. 2003). By sequestering REST/NRSF and not allowing it to interact with the BDNF promoter, HTT allows for increased transcription, and therefore translation, of BDNF. As the REST/NRSF transcription factor is not specific for the BDNF promoter, HTT may play a role in the regulation of other REST/NRSF genes.

1.2.5.7 Huntingtin and Vesicle Transport

As mentioned previously, HTT has been predicted to be involved in vesicular transport. This has been found to be true in the case of BDNF transport, suggesting that HTT may play a role in guiding BDNF towards striatal cells (Gauthier et al. 2004). In cultured cells, wildtype HTT increases the vesicular transport of BDNF along microtubules, but if HTT is knocked down using RNAi this transport is slowed down. HTT has also been found to interact with the p150 (Glued) subunit of dynactin, a key component of the molecular motor which moves vesicles along microtubules (Engelender et al. 1997; Imarisio et al. 2008). Through interactions with huntingtin-associated protein 1 (HAP1), Glued, and BDNF, HTT can increase the transport of BDNF along microtubules. More recently, HTT has been found to regulate the transport of the BDNF receptor TrkB in striatal neurons, further highlighting the role of HTT in BDNF function (Liot et al. 2013).

1.2.5.8 Huntingtin and the Synapse

The interaction of HTT and vesicles does not end once they have reached their intended cellular destination. Once HTT and its vesicular payload reach the cortical synapse, HTT appears to play additional roles in synaptic vesicle transmission. HTT interacts with several proteins involved in exocytosis and endocytosis at synaptic terminals, such as HIP1, HIP14, HAP1, PACSIN1, SH3Gl3, clatherin and dynamin, (reviewed in detail in Smith et al. 2005b). A key synaptic transmission molecule that HTT interacts with is PSD-95, one of a family of proteins

that binds the NMDA and Kainate receptors at the postsynaptic density (Sheng &. Kim 2002). In addition to direct interaction with proteins at the synapse, HTT appears to regulate the expression of proteins that are involved in synapse function. Complexin II and rabphilin 3A are both proteins involved in the exocytosis of vesicles, both of which display altered expression in mouse models and human cell lines of HD (Morton & Edwardson 2001; Smith et al. 2005a). Finding additional roles for HTT at the synapse reinforces the concept of HTT as a positive regulator of neuronal cell survival and function.

1.2.6 Mutant Huntingtin

1.2.6.1 The HD Mutation

As previously discussed, the causative mutation in HD is a CAG trinucleotide repeat expansion in the *HTT* gene. This CAG expansion is localized in exon 1; when the gene is translated into the HTT protein, the expansion results in an expanded polyglutamine stretch. The CAG expansion itself is a dynamic mutation, with the length of the CAG changing both between generations and within different cells in an individual. With longer expansions there is a predisposition towards lengthening of the CAG repeat versus shortening (MacDonald et al. 1999; Djoussé et al. 2004; Chattopadhyay et al. 2005). HD patients have a *HTT* allele that contains at least 39 CAG repeats, considerably longer than the average 17-20 repeats found on most WT alleles (Warby et al. 2009). Patients with longer repeats, on average, have an earlier age of onset and a more severe phenotype. This inverse correlation between

length and age of onset accounts for between 60-70% of the variance in age of onset of HD patients (Andrew et al. 1993).

The inverse correlation of CAG repeat length and age of onset, and the dynamic nature of this mutation combine to create the phenomenon of anticipation in HD families. Anticipation is defined as the phenomenon where subsequent generations of family members experience earlier ages of onset or more severe disease than their ancestors. Anticipation is also a concern for individuals carrying so called intermediate alleles, those with 27-35 CAGs. It has been found that a specific genetic haplotype is found on 95% of HD alleles, this same haplotype is over represented on intermediate alleles compared to wildtype alleles (Andrew et al. 1993). This suggests that there may be predisposing factors present in this haplotype that increase the instability of the CAG repeat. As the CAG lengths on these intermediate alleles further expand/contract as they traverse generations it is expected that they will eventually lead to fully expanded HD alleles and new proband patients and families.

1.2.6.2 Mutant Huntingtin Toxicity

Given that the CAG repeat expansion results in an expanded polyglutamine tract that affects the conformation of the HTT protein, it is important to understand how the altered polyglutamine tract affects the function of the protein (Perutz et al . 1994). As I will discuss in the following sections, the HD mutation does not appear to be a complete loss of function but is primarily a novel toxic gain of function mutation, although elements of both appear to play a role in HD pathogenesis. It's interesting

to note that while the mutation may affect certain aspects of wildtype HTT function, the mutant protein appears to retain much of the basic function of the wildtype protein. This is best displayed in *Hdh* knockout mice mentioned above. These mice are embryonic lethal; however, if mutant HTT is added back, they survive and their early development remains normal (Van Raamsdonk et al. 2005a). In addition, individuals who are homozygous for the mutant allele develop normally and have no apparent defects until later on in life (Wexler et al. 1987; Myers et al. 1989). It appears, at least early in development, that mutant HTT performs all the functions of wildtype HTT, and it is only later that the CAG expansion may interfere with wildtype function. As discussed in the previous section, post-natal removal of wildtype HTT in mature neurons results in cellular death. Similarly, mutant HTT also causes cellular death, indicating that perhaps loss of the wildtype function due to the CAG expansion is having a similar effect (Dong et al. 2011). It is likely that the combination of loss of wildtype HTT function, as well as new toxic gain of function by mutant HTT, results in the phenotype seen in HD (Cisbani & Cicchetti, 2012).

1.2.6.3 Protein Aggregation

Before furthering our discussion of the effect the CAG expansion has on the HTT protein, the issue of HTT inclusions or aggregates must be addressed. Protein aggregates that are visible on pathologic examination were first described in mouse models of HD, and were then later observed in HD patients (Davies et al. 1997). These aggregates are made up of insoluble proteins, of which HTT is the main component, though other proteins have been found in the aggregates as well. The

presence of these misfolded protein aggregates places HD within the family of neurodegenerative diseases, including Parkinson's disease, Alzheimer's disease and Amyotrophic Lateral Sclerosis (ALS), as diseases that have a proteinopathy as their putative pathologic basis. It was initially believed that these insoluble protein aggregates were directly pathogenic and contributed to the neurodegeneration seen in HD. This view has changed as it has been found that the cells containing these aggregates are not necessarily the ones that are dying. In fact, it appears that cells that do not form aggregates are more sensitive to neurotoxicity (Gauthier et al. 2004; Gutekunst et al. 1999). This suggests that the aggregates may in fact be a natural compensatory mechanism that protects neurons from degeneration. The regional expression of huntingtin and the pattern and timing of inclusion formation do not correlate with selective neurodegeneration in HD and are not thought to be primary determinants of pathology (Francelle et al. 2014; Gauthier et al. 200). In addition, not all mouse models of HD demonstrate aggregate formation concurrent with neuronal cell loss, further providing evidence that the insoluble aggregates seen in HD are not necessarily causative of disease (Arrasate et al. 2004; Van Raamsdonk et al. 2005b).

1.2.6.4 Cleavage of Mutant HTT

As discussed in the wildtype section, the HTT protein is cleaved at several sites, generating fragments of different lengths. Mutant HTT is also cleaved, although in addition to the wildtype cleavage sites it appears that mutant HTT is further cleaved and generates additional fragments. These additional fragments are

thought to be toxic species and contribute to disease pathogenesis (Ratovitski et al. 2009). Reducing the activity of caspases and calpains reduce the generation of these additional fragments and in turn delay disease progression (Gafni & Ellerby 2002; Gafni et al. 2004). In addition to altered cleavage of the HTT protein, it has been found that alternate splicing of the mutant *HTT* allele also produces a short exon1 mRNA that is later translated into a exon1 fragment (Mende-Mueller et al. 2001; Sathasivam et al. 2013).

1.2.6.5 Mutant HTT and BDNF

As discussed in the previous section, HTT plays a role in the regulation and transport of BDNF. It is therefore unsurprising to find that the mutant CAG expansion alters the relationship between HTT and BNDF. The CAG expansion prevents HTT from stimulating BDNF transcription in cortical neurons. This is due to mutant HTT's inability to sequester REST/NRSF in the cytoplasm which allows it to translocate to the nucleus and repress BDNF transcription (Zuccato et al. 2001; Zuccato et al. 2003). Mutant HTT also represses BDNF vesicular trafficking along microtubules, resulting in less BDNF being transported from the cortex to the striatum (Gauthier et al. 2004). This reduction in the amount of BDNF reaching the striatum may account for the selective degeneration of striatal neurons seen in HD. This hypothesis is supported by evidence from mice with conditional knockout of BDNF in the cortex (Baquet et al. 2004). These mice have decreased cortical and striatal volumes and differences in MSN numbers. In addition, in an exon1 fragment mouse model of HD,

known as the R6/1 model, with only one functional BDNF allele displayed an earlier onset of phenotype and enhancement of motor abnormalities compared to R6/1 mice with two functional BDNF alleles (Canals et al. 2004; Pineda et al. 2005).

1.2.6.6 Transcriptional Dysregulation

As was mentioned in the wildtype function section of this chapter, HTT regulates BDNF transcription through its interaction with the transcriptional repressor REST. Given this known interaction with a transcriptional regulator, it is unsurprising to find that in its mutated form HTT disrupts the normal expression of not only BDNF but a multitude of other genes regulated by REST, including non-coding RNAs (Zuccato et al. 2007; Johnson et al. 2008). The transcriptional dysregulation seen in HD is not localized to REST-regulated genes alone. The conformational changes conferred by the CAG repeat result in mutant HTT abnormally interacting with several other transcription factors, resulting in widespread transcriptional changes. These transcription factors include TATA-binding protein/TFIID, TAFII130 (a co-activator in CREB dependent transcription), SP1, p53, and nuclear receptor co-repressor (NCoR) (Boutell et al. 1999; Steffan et al. 2000; Shimohata et al. 2000; Dunah et al. 2002). Mutant HTT further affects transcriptional regulation by interacting with CBP and the p300/CBP-associated factor P/CAF, which are involved in chromatin remodeling through posttranslational modification of histones (Boutell et al. 1999; Steffan et al. 2000). In total, more than 81% of striatal-enriched genes are down-regulated in both mouse models and human HD caudate samples, signifying

the impact that this dysregulation has on disease pathogenesis (Desplats et al. 2006).

1.2.6.7 Mitochondrial Dysfunction

As mentioned briefly in the clinical features section of this chapter, HD patients experience metabolic abnormalities and evidence for energetic dysfunction. A main contributing factor to these symptoms is mutant HTT's effect on mitochondria. Mutant HTT has been found to increase mitochondrial fragmentation, disrupt the biogenesis and trafficking of mitochondria, and lower mitochondrial membrane potential (Panov et al. 2003; Chang et al. 2006; Milakovic et al. 2006; Wang et al. 2009). ATP production by mitochondria has also been found to be altered in the presence of mutant HTT via reduced expression of oxidative phosphorylation enzymes that mediate the production of ATP (Gu et al. 1996; Benchoua et al. 2006). In line with the transcriptional dysregulation discussed above, mutant HTT's abnormal interaction with transcription factors required for CREB dependent transcription factors results in the down regulation of PGC1-a, a well characterized regulator of mitochondrial biogenesis (Cui et al. 2006). Mutant HTT also prevents mitochondria from regulating cellular calcium homeostasis. This results in inefficient respiration, lowered levels of calcium capacity in mitochondria, and sensitivity to increases in calcium resulting from glutamate excitotoxicity (discussed below) (Milakovic et al. 2006; Fernandes et al. 2007).

1.2.6.8 Excitotoxicity in HD

Excitotoxicity refers to neuronal death induced via over-stimulation by the excitatory neurotransmitter glutamate and is proposed to be a major contributor to the selective neurodegeneration of MSNs in HD. MSNs receive excitatory glutamatergic input from the cortex and thalamus are particularly sensitive to toxicity induced from the glutamate analogues quinolinic and kainic acid (McGeer & McGeer 1976; Schwarcz et al. 1984). Glutamate activates receptors, including the NMDA receptor, which transport calcium into the neuron. Mutant HTT disrupts mitochondrial calcium homeostasis, making neurons more sensitive to changes in calcium levels, including calcium influx from NMDA receptor stimulation. NMDA receptors are composed of two subunits, one of which, NR2, has two main subtypes, A and B, that react differently to the presence of mutant HTT. NMDA receptors containing NR2B influx more calcium in the presence of mutant HTT when stimulated and this subtype can be found in tissues with higher vulnerability to mutant HTT (Li et al. 2003; Li et al. 2004). The sub-cellular localization and posttranslational modification of NMDA receptors are also known to alter the specific type of response induced by stimulation. NMDA receptors are located both in the synaptic and extra-synaptic plasma membrane, the choice of which being specified through posttranslational modification. Cleavage of the C-terminus by calpain and de-phosphorylation of NMDA receptor subunits by the phosphatase STEP result in reduced NMDA receptor localization to the synaptic membrane (Gladding et al. 2012). Both calpain and STEP show higher acivity in the presence of mutant HTT and may explain the

higher frequency of extra-synaptic NMDA receptors seen in HD mouse models (Cowan et al. 2008; Graham et al. 2009).

1.2.6.9 RNA Toxicity

The predominant view in the field of HD research is that the HD mutation confers a novel structural or functional change in the HTT protein and it is the change in the protein conformation that results in pathogenesis (Clabough 2013). This prevailing view of HD pathogenesis may be challenged by recent findings implicating a potential role of mutant RNA transcripts in mutant HTT toxicity. The majority of our understanding of RNA toxicity comes from studying the CTG expansion in myotonic dystrophy type 1 (DM1)(Fiszer & Krzyzosiak 2013). In DM1, the CTG expansion in patients ranges from 50-3,000 repeats present in the un-translated region of the dystrophia myotonica protein kinase gene, and it is the presence of this repeat in the mRNA transcript of the mutated allele that causes pathogenesis. While the threshold of repeats differs greatly between HD and DM1, with DM1 being higher, there remains a significant proportion of HD alleles with repeat sizes within the pathogenic range seen in DM1. In addition, analysis of the hairpin structures generated by CAG and CUG (remember that during transcription Ts are transcribed as Us converting a CTG repeats into a CUG repeat) repeats are similar (Jasinska et al. 2003; Sobczak et al. 2003). In both cases CAG and CUG repeats of normal length form small and unstable hairpins and larger repeats form longer and more stable ones.

By comparing RNA molecules carrying either pure CAG repeats or CAG repeats interrupted by CAA codons (both codons translate to glutamine but the CAA codon is unable to form hairpins) in Drosophlia, it was found that the CAA interrupted molecules produced less neurodegenerative features (Sobczak et al. 2003). In addition, studies looking at expanded CAG repeats in human HeLa and SK-N-MC cells showed formation of nuclear RNA foci and alternative splicing, hallmarks found in DM1 (Mykowska et al. 2011). HTT transcripts have also been found in RNA foci in human HD fibroblasts (Fiszer & Krzyzosiak 2013). Furthermore, aberrant splicing of the *HTT* gene has been attributed to the CAG expansion using a similar mechanism of RNA toxic gain of function seen in DM1 (Sathasivam et al. 2013).

RNAi mechanisms have also been implicated in HD, with studies in drosophila showing that co-expression of expanded complementary CAG and CUG repeats result in dsRNA that is cleaved by the Dicer-2 pathway, forming CAG/CUG siRNA that cause neurodegeneration (Yu et al. 2011; Lawlor et al. 2011). This finding is corroborated by studies on the antisense *HTT* transcript, which contains a CUG repeat and has been shown to regulate sense *HTT* transcription in a repeat dependent manner, a process dependent on the Dicer and RISC pathways (Chung et al. 2011). In addition to this evidence, it has been shown in human cell lines that expanded CAG repeats flanked by the *HTT* exon 1 sequence generate small RNA species through Dicer and cause a downstream silencing of genes through Ago-2 (Bañez-Coronel et al. 2012). These effects were only seen in CAG repeats and not in CAA repeats. The potential contribution of RNA toxicity to HD disease

pathogenesis is a new area of research in the field of HD, and it will take time to develop the necessary tools to study the potential role of RNA toxicity *in vivo*.

1.2.7 HD Conclusions

In this section I have endeavoured to give the reader an overview of the clinical features of HD, an appreciation of our current knowledge of the *HTT* gene and protein function, and finally a perspective on how the *HTT* gene mutation causes HD. I have highlighted the complex structure and post-translational processing of the HTT protein and outlined multiple proposed functions of the wildtype HTT protein. The potential neurotoxic effects of the CAG mutation in HD, how this may affect the HTT protein or mRNA, and how this may contribute to disease pathogenesis were reviewed.

The scientific community has known of the *HTT* gene and the CAG mutation that causes HD for over 20 years now, and it is understandable to expect that treatments and cures for HD should be progressing quickly - if not already available. Unfortunately, given the complexity and uniqueness of the HTT protein, this is not the case. There are currently no approved treatments aimed at reducing or eliminating HD pathogenesis. Current treatments for HD consist of drugs that alleviate symptoms of the disease. This shortage of treatment options is not due to lack of effort from the scientific community, but due to the unique challenges that HD presents. I hope that the reader comes away from this section with a greater appreciation of the complexities of the *HTT* gene the diverse functions of the huntingtin protein, and the unique challenges facing researchers studying the effects

of the HD mutation and developing novel treatment strategies for this devastating neurodegenerative disorder.

1.3 Fundamentals of Mammalian Transcription

1.3.1 Introduction

Before exploring our current knowledge concerning the transcriptional regulation of the *HTT* gene, it is important to first establish the current consensus surrounding mammalian transcription generally. Here I will present a short introduction of the fundamentals of mammalian gene transcription, including how our understanding of gene transcription has changed in recent years in light of genome-wide studies. By doing this before introducing the literature directly concerning transcriptional regulation of the *HTT* gene, the majority of which was published prior to the genome-wide studies to be discussed, I hope to establish a basis for understanding the paradigms that previous researchers utilized to conceptualize transcription. This is not in an effort to discredit these studies, but to highlight how our understanding has changed since their publication, thereby warranting a revision our understanding of *HTT* promoter composition and function.

As countless first year biology text books will reiterate, the “central dogma” of gene expression in mammalian cells is the process of gene transcription, from DNA to RNA, followed by the translation of the RNA template to protein. Without the transcription of a gene, sequence from the DNA in the nucleus to an RNA molecule that can be transported to the cytoplasm gene expression cannot happen. The

central protein to this process is RNA polymerase II (RNApolII), which is the main RNA polymerase used for RNA synthesis of messenger RNA (mRNA) in mammalian cells. In order for the gene transcription process to be completed successfully, RNApolII must be orientated to the correct position in the genome so as to ensure the correct transcriptional start site (TSS) is selected as well as the correct strand for the gene in question (Sandelin et al. 2007). These two steps in the transcriptional pathway are by no means trivial. The incorrect binding and initiation of RNApolII can result in a mis-transcribed gene, being either transcribed at an inappropriate time, in an inappropriate cell type, or from the wrong TSS. To ensure that the transcription process proceeds correctly, RNApolII is reliant upon a class of proteins referred to as transcription factors (TFs) that assist in the recruitment, positioning, and initiation of RNApolII (Kadonaga 2004). TFs can interact directly with DNA sequences, with each other, or with RNApolII itself to regulate its positioning and initiate elongation along the strand of DNA to be transcribed. TFs can be divided up into two main groups, those that interact directly with RNApolII, known as general transcription factors, and those that interact indirectly with RNApolII through interactions either with core TFs, a mediator protein, or other non-general TFs primarily referred to as sequence-specific TFs (Poss et al. 2013).

1.3.2 Transcription Factor Binding Site Motif Identification

TFs bind to specific DNA sequences in the genome, known as transcription factor binding sites (TFBSs). TFSBs are defined by a sequence motif, a pattern of

nucleotides the TF recognizes and utilizes to bind DNA. The most common means of identifying a TFBS is through sequencing many DNA fragments bound to the specified TF, comparing the nucleotide sequences found bound to the TF, and then using this comparison to create a position weight matrix (PWM) (Wasserman & Sandelin 2004) .Simplistically, a PWM allows for a weighted value to be generated, representing the probability that any of the 4 nucleotides will be found at each position in the motif. This PWM can then be used to screen genomic regions to identify potential TFBS. In the time before genome-wide technologies became available, the generation of a PWM was a very labor intensive process, limited by both throughput and the short length of the sequences generated (Poptsova 2014). As we have moved into a post-genomic era, where genome-wide technologies have become more cost effective, our ability to generate PWMs based upon genome-wide interrogations has increased. However, even in this post-genomic era, our understanding can be partitioned into that before and after the advent of Next-Generation sequencing (Next-Gen sequencing) technologies.

Prior to the development of Next-Generation sequencing, the primary method of PWM motif discovery utilized chromatin immunoprecipitation on a chip (ChIP-chip) technology (Poptsova 2014). ChIP-chip can be partitioned into two parts. The first step is the chromatin immunoprecipitation (ChIP), in which formaldehyde is used to crosslink any proteins bound to genomic DNA, and then the crosslinked DNA is sheared into fragments using either physical (sonication) or chemical (nuclease) treament. The crosslinked, fragmented DNA is then immunoprecipitated using an antibody specific to the TF of interest, which allows for a library of DNA sequences to

which the TF binds to be generated. The crosslinked TF is then removed and the fragmented DNA is amplified using polymerase chain reaction (PCR) and fluorescently labeled. In the second part of ChIP-chip, the fluorescently labeled DNA is hybridized onto a DNA microarray chip, on which known, tiled genomic DNA sequences (meaning that the fragments of genomic DNA contain overlapping sequences at their ends) are bound. Hybridization of the fluorescent ChIP-derived DNA to the genomic DNA on the microarray can then be detected and quantified using a fluorescence reader. As the genomic DNA attached to the microarray at each position is known, hybridized sequences can thus be identified and compiled for use in the generation of a PWM. It is clear that in ChIP-chip studies, the composition and depth of genomic DNA represented on the microarray will affect results; the more comprehensive the genome coverage, the more informative the results. In addition, the length of each individual genomic DNA sequence represented on the microarray can affect results (Carey et al. 2009). The larger the sequences represented, the more tiling of the genomic DNA becomes important, as large sequences on their own are not informative to discovery of the relativity short TFBSs. Exclusion or inclusion of tiled regions allows for a shorter hybridized sequence to be identified. Despite the limitations DNA microarray technologies impose, (*i.e.* the limited amount of genomic DNA that can be represented per array) ChIP-chip studies still provided valuable and informative results, which have been further built upon subsequent to the advent of Next-Gen sequencing technologies

With the advent and widespread acceptance of Next-Gen sequencing, the use of ChIP-chip studies to investigate TFBS has been replaced with chromatin

immunoprecipitation sequencing (ChIP-seq) studies. In ChIP-seq the initial generation of a TF binding DNA fragment library remain the same (with the small omission of fluorescent labeling), but instead of hybridization to a microarray, each fragment of genomic DNA is sequenced, creating a vast set of short DNA reads (Poptsova 2014). These short reads are then aligned to an established genome assembly; as this alignment is generated, “piles” of reads accumulate around genomic regions, signifying the presence of a TFBS in that region. In order to identify the TFBS itself, within the piles, “peaks” are identified, the peaks representing sequences common to all of the aligned sequence reads in the pile. These peaks should therefore represent the TFBS to which the TF itself was bound. By utilizing sequencing instead of microarray hybridization, biases for sequences represented on microarrays are eliminated and true representations of genome-wide binding can be generated. The greater depth of genomic information and sensitivity garnered by ChIP-seq studies also allows for the identification of TFBS subtypes, such as a TF with a preferential TFBS in one developmental stage and a differing TFBS in another. In addition, by studying the genomic regions just outside of the peaks, information about potential co-factors can also be gained. With the vast amount of data generated from ChIP-seq studies come limitations and potential pitfalls in the processing and analysis of the data. Differences in the computational pipeline through which datasets are processed can affect the resulting peaks and PWM generated from them. As the purpose of this thesis is not to provide a comparison of differential bioinfomatic data processing techniques, I will refer the reader to the following review on the subject (Poptsova 2014).

The sequence length of a TFBS is typically between 6-10 bp in length, and given the nature of DNA having only four nucleotides, by chance the occurrence of a given TFBS is quite high. It is clear, however, from both ChIP-chip and ChIP-seq studies that even though there are a vast number of potential binding sites for a given TF based on DNA sequence alone, the majority of these are not bound by a TF. For example, a 2009 study using GATA-binding factor 1 found that, on average only 1 in every 500 sequence predicted sites are actually bound by the TF (Zhang et al. 2009). Given the small number of bound TF compared to the larger number of potential sites, it is clear that sequence similarity alone is not a good predictor for a functional TFBS. Additional factors must act in unison to provide the correct set of conditions that allows that site to function as a TFBS. These factors include epigenetic factors that dictate the availability for the putative TFBS to come in contact with the TF, such as histone post-translational modifications, which can dictate nucleosome placement as well as DNA methylation (Rothbart & Strahl, 2014). The roles of histone modification and DNA methylation on the ability for TFs to access genomic DNA will be discussed in a following section. The presence and binding of other TFs nearby which recruit or reinforce binding of the TF in question to the putative TFBS may also influence the functionality of a putative TFB (Slattery et al. 2014).

1.3.3 Cis-Regulatory Regions

As noted above, TFs seldom act on their own. They are commonly found in groups of homo- or hetero-TFs all binding to a similar region in the genome and

acting as a unit to establish transcriptional control (Slattery et al. 2014). These regions of the genome are often referred to as cis-regulatory regions as they are often found on the same chromosome as the gene they regulate (Hardison & Taylor 2012). In this section, I will discuss two out of the three main groups of cis-regulatory regions, promoters and enhancers, as they form a major component of this thesis. The last group, insulators, functions as the name suggests to insulate regulatory regions from inappropriately interacting with one another as well as preventing the epigenetic marks in one region of the genome, in particular silencing marks, from spreading to active regions (Carey et al. 2009). As they were not studied in detail in this thesis, for the sake of brevity, I will say no more about them.

1.3.3.1 Promoters

Promoters can be further subdivided into two categories, the core promoter and the proximal promoter. The core promoter is comprised of TSSs as well as surrounding DNA sequence, with TFBS motifs for the general TFs to direct RNApolII to bind and initiate transcription (Juven-Gershon et al. 2008). Early studies identified several core promoter elements, such as the TATTA box and the initiation element, a combination of which were thought to be necessary for transcription to occur (reviewed in detail in Smale & Kadonaga 2003). The TATTA box in particular was thought to be an essential component of the majority of core promoters until studies revealed that only ~10-20% of promoters contain a TATTA box consensus sequence (Cooper et al. 2006; Gershenzon & Ioshikhes 2005). This paradigm shift in TATTA box prevalence may be accounted for by additional evidence regarding the

prevalence of genes with multiple TSSs. It was traditionally thought that the majority of genes have a single TSS from which transcription is always initiated; however, Next-Gen sequencing has allowed for techniques such as capped analysis of gene expression (CAGE), in which the 5' cap of mRNA is used to facilitate purification, reverse transcription, and sequencing of all the transcribed mRNA molecules in a given cell line or tissue, to be developed and have challenged this assumption. CAGE analysis has revealed that the majority of human and mouse genes are not transcribed from a single TSS at a distinct nucleotide position, but from a set of closely located TSSs (Carninci et al. 2005; Carninci et al. 2006). This distinction between promoters that utilize one versus those that utilize several TSSs further divides the core promoter class of cis-regulatory regions into sharp and broad core promoters respectively (Sandelin et al. 2007). Further analysis into the composition differences between sharp and broad promoters has highlighted the prevalence of the TATTA box with sharp core promoters and tissue specific expression, while broad core promoters are associated with CpG rich core and proximal promoter regions and ubiquitous expression. It should be noted that these characteristics for broad and sharp core promoters are not exclusive; core promoters that are exceptions to these trends do exist.

Proximal promoter sequences are found in close proximity to the core promoter, typically considered to be within the first ~300 bp of the core promoter (Carey et al. 2009). The TFBS within the proximal promoter do not contain motifs for the general TFs but for sequence specific TFs, which bind to the proximal promoter and affect the binding of general TFs and RNApolII at the core promoter. The

combination of the proximal and core promoter drive basal transcription of the gene in question and are commonly thought of together as one unit, the promoter. Given the over-representation of TFBS in the promoter, for the TFs that are utilized in proximal and core promoter at least, promoter sequences were commonly used on DNA microarrays in ChIP-chip studies (Poptsova 2014). While this did help identify TFBS utilized in promoters, it excluded much of the genome, and made the identification of regulatory regions further from the promoter difficult (Poptsova 2014). ChIP-seq, on the other hand, has the ability to identify TFBS and potential regulatory regions outside of the immediate promoter region.

1.3.3.2 Enhancers

The second type of cis-regulatory element, enhancers, act as a second level of transcriptional control, allowing more precise regulation of the promoter. Enhancers, as their name suggests, are generally thought to increase the function of their associated promoters, although there have been cases of enhancers working inversely and preventing promoter function, so called silencers (Hardison & Taylor 2012). Like promoters, enhancers are comprised of binding motifs for several TFs. Unlike promoters, proximity and orientation to the TSS that the enhancer is acting upon does not appear to be important; enhancers have been found to interact with genes over 1MB away and with promoters on a different chromosome (Hardison & Taylor 2012). As discussed above, although many potential TFBSs exist in the genome, only a small number of these are occupied by their corresponding TFs. Co-occupation of the associated TFs in an enhancer are thought to reinforce and

sustain binding (Slattery et al. 2014). Depending on the TFs associated with the enhancer region, the mechanism of enhancing transcription can differ. TFs can, through mediator proteins, encourage additional binding of RNApolII or stabilize its interaction with the promoter (Yin & Wang 2014). It is important to note that the effect of a given enhancer on its associated promoter can differ depending on the other TFs co-occupied at the same regulatory region (Diamond et al. 1990; Ezer et al. 2014). So it cannot be said that a given TF is an activator or a silencer but that the specific regulatory regions can act as enhancers or silencers. Given that proximity to a gene is not a complete indicator of interaction, it has been difficult previously to predict enhancer sequences and to identify the genes with which they interact. ChIP-chip and ChIP-seq studies have provided a way to investigate these regions by providing genome-wide information on the presence of TFBS, and information on the chromatic state of known enhancers is allowing for better prediction criteria to be set.

1.3.4 Epigenetic Regulation of Transcription

1.3.4.1 Nucleosome Positioning and Histone Modification

Enhancers, as well as promoters and insulators, bear chromatin marks that can aid in the prediction and identification of these cis-regulatory elements. Genomic DNA does not exist freely in the nucleus; it is packaged by being wound around nucleosome complexes, which are comprised of copies of 4 histone proteins, H2A, H2B, H3 and H4. These nucleosomes are then further organized to create chromatin. Areas where nucleosomes are tightly associated, creating a closed

chromatin structure, are often where non-expressed or silent genes are located, whereas the converse is true for expressed portions of the genome (Rothbart & Strahl, 2014). This organizational pattern of DNA wound around nucleosomes is often referred to as “beads on a string” as DNA between nucleosomes is not bound and free to interact with other proteins (Sajan & Hawkins 2012). In general, genomic regions occupied by nucleosomes cannot bind TFs and the components of transcription machinery (Sajan & Hawkins 2012). Identification of nucleosome-free regions is done through DNase hypersensitivity assays, in which the DNA in between nucleosomes is degraded using DNases that cannot degrade DNA bound to nucleosomes(Sajan & Hawkins 2012). The remaining DNA is then sequenced and aligned to the reference genome; portions of the genome for which DNA sequences are not found are DNase hypersensitive and unbound by nucleosomes.

The tails of the histones comprising the nucleosome can be post-translationally modified, which can further impact the placement of the nucleosome (Rothbart & Strahl, 2014). Nucleosomes can be methylated, acetylated, ubquitinylated, and/or phosphorylated (Rothbart & Strahl, 2014). The effect on nucleosome positioning is dependent not only on the type of post-translational modification, but also on which histone and which amino acid of the histone tail is modified. For example, H3K4 methylation is associated with activation of genomic regions while H3K9 methylation is associated with silencing of genomic regions (Wozniak & Strahl 2014) . By understanding the patterns in histone modification, so called histone code, we can gain information about the expression patterns of

nearby genomic regions. This can aid in the prediction of regulatory regions outside of the promoter region.

1.3.4.2 DNA Methylation

In addition to sequence variation, DNA also conveys regulatory information through nucleotide modification by the addition of a methyl group to the fifth carbon of cytosine (Jones 2012). The cytosines that are methylated are commonly found in CpG dinucleotides, although non-CpG methylation is also present in the genome (Pinney 2014). Methylated cytosines are easily deaminated to form uracil residues, which are subsequently converted to thymines during DNA repair (Cooper & Krawczak 1989). Due to this conversion, regions where cytosines have not been methylated are often CpG rich as the conversion of cytosines to thymines is less than in methylated regions. This epigenetic mark, while seemingly simple, can greatly affect nucleosome positioning, which in turn affects overall chromatin structure (Karymove et al. 2001). The additional methyl group also changes the structure of the cytosine nucleotide, which can affect TF binding (Dantas Machado et al. 2014). In general, DNA methylation is associated with a closed chromatin structure while non-methylated regions are associated with a more open one, although recent evidence suggests that this paradigm may not always hold true (Jones 2012). Gene body methylation, for example, appears to display a more complex relationship, where moderate levels of methylation are associated with increases in expression while very low and very high are associated with decreases

in expression (Jjingo et al. 2012). While some DNA methylation marks are established by parental methylation patterns, other marks are subject to change over the lifetime of an individual from both a developmental stage and an environmental viewpoint (Ci & Liu 2015; Szyf 2014). Aberrant DNA methylation has been implicated in a host of diseases, including several neurodegenerative disorders, such as Fragile X syndrome, Alzheimer's, and as will be discussed below, HD.

1.4 Transcriptional Regulation of the *HTT* Gene

1.4.1 Regional Expression of Huntingtin

Before considering how *HTT* is transcriptionally regulated, it is important to consider where and when *HTT* is spatially and temporally expressed. Despite the association of mutant *HTT* with selective neurodegeneration in specific brain regions, *HTT* is actually ubiquitously expressed, albeit at low levels, throughout the body in both humans and mice (Van Raamsdonk et al. 2007; Dixon et al. 2004; Trottier et al. 1995; Wood et al. 1996; Moreira Sousa et al. 2013; Li et al. 1993; Metzler et al. 2000). Given the large size of the transcript and protein, it would be contradictory to cellular energy conservation to think that *HTT* is transcribed and translated in the periphery but serves no important cellular function. The function of HTT outside of the CNS has not been studied extensively and we can only hypothesize, based on HTT's potential role in neuronal survival, that in peripheral cells it may have a similar primary function. It is unknown exactly what mechanism of

transcriptional regulation is responsible for the differential expression between the CNS and periphery; this aspect of HTT expression has never been investigated.

In addition to differential tissue expression, there appears to be at least one case of differential expression of a splice variant of *HTT* (Lin et al. 1993). The two splice variants identified differ only in the location of polyadenylation site while the coding region remains identical. The longer of the two splice variants appears to be primarily expressed in human brain tissue compared to lymphocytes. The functional relevance of these two transcripts remains unclear; however, in the mutant condition, an additional short splice variant encompassing exon 1 has been recently identified and may play a role in HD pathogenesis (Sathasivam et al. 2013).

1.4.2 Epigenetic Regulation of *HTT*: DNA Methylation

In addition to its widespread expression it has been suggested that *HTT* may be subject to inter-individual variability, particularly in the brain. Using inbred, genetically identical male mice, Dixon et al., 2004 found considerable amounts of transcriptional expression difference between cortex samples of different individuals. As these mice were genetically identical, this points to environmental and/or epigenetic differences that could affect *HTT* expression. In the preceding section, I discussed DNA methylation as an epigenetic mark conferred directly on genomic DNA that can have transcriptionally relevant consequences. DNA methylation changes have also be identified as a potential mechanism through which environmental factors can have a lasting role on gene expression. Changes in DNA

methylation at the *HTT* gene locus have been identified in a study of individual sperm cells from a single individual; as *HTT* is known to be highly expressed in the testicular tissues, this differential DNA methylation may play a role in the testicular degeneration seen in male HD patients (Flanagan et al. 2006). In addition, DNA methylation changes, and corresponding expression changes, are known to contribute to differences in gene expression between tissues, meaning that the differential expression of *HTT* between tissues may be attributed to differences in DNA methylation; however this has never been investigated (Wan et al. 2015).

The methylation of DNA primarily occurs on CpG dinucleotides; however non-CpG methyaltion is also found throughout the genome (Juna Lee et al. 2010; López Castel et al. 2011). This non-CpG methylation primarily occurs at CpNpG sites with N most commonly being an A nucleotide. Given that the causative mutation in HD is a trinucleotide CAG expansion, it is reasonable to hypothesize that the length of the CAG repeat may have an effect on DNA methylation of the *HTT* gene. Several studies have implicated aberrant DNA methylation as a result of the HD mutation in both mice and humans; however none comment upon changes in methylation at the *HTT* locus as the result of the HD expansion (Ng et al. 2013; Villar-Menéndez et al. 2013).

1.4.3 The *HTT* Promoter Region

To summarize previous findings concerning the *HTT* promoter I have created Figure 1.1 detailing all previously identified motifs, including TFBSs, that have been

verified beyond identification based on sequence similarity to a known TFBS. I refer the reader to this figure as they read the following section; however I must add the following notes before continuing. The studies being discussed below, unless noted, utilized promoter sequences generated from various phage and cosmid libraries and not sequences taken directly from human samples. In addition, the studies vary in utilizing the transcriptional start site or the translational start site as points of reference and these sites at times do not match up with each other or with the standard genome assembly used today from the UCSC genome browser. This is especially true for the studies conducted before or around the early 2000s, which is not surprising given that the UCSC genome browser did not go online until 2000. The discrepancies in transcriptional and translational start site most likely arise from differences in the phage and cosmid clones selected in these early studies. As many of the studies differ in terms of transcriptional start site I have used the translational start site as the positional point of reference in Figure 1.1. In this Figure I have positioned the features identified in these early studies based on their sequence and not the numeric position from the translational or transcriptional start sites mentioned in their respective studies. Finally, identified motifs that did not have a corresponding sequence identified in the UCSC genome browser assembly were not included as these missing sites are likely artifacts from the selected phage and cosmid clones.

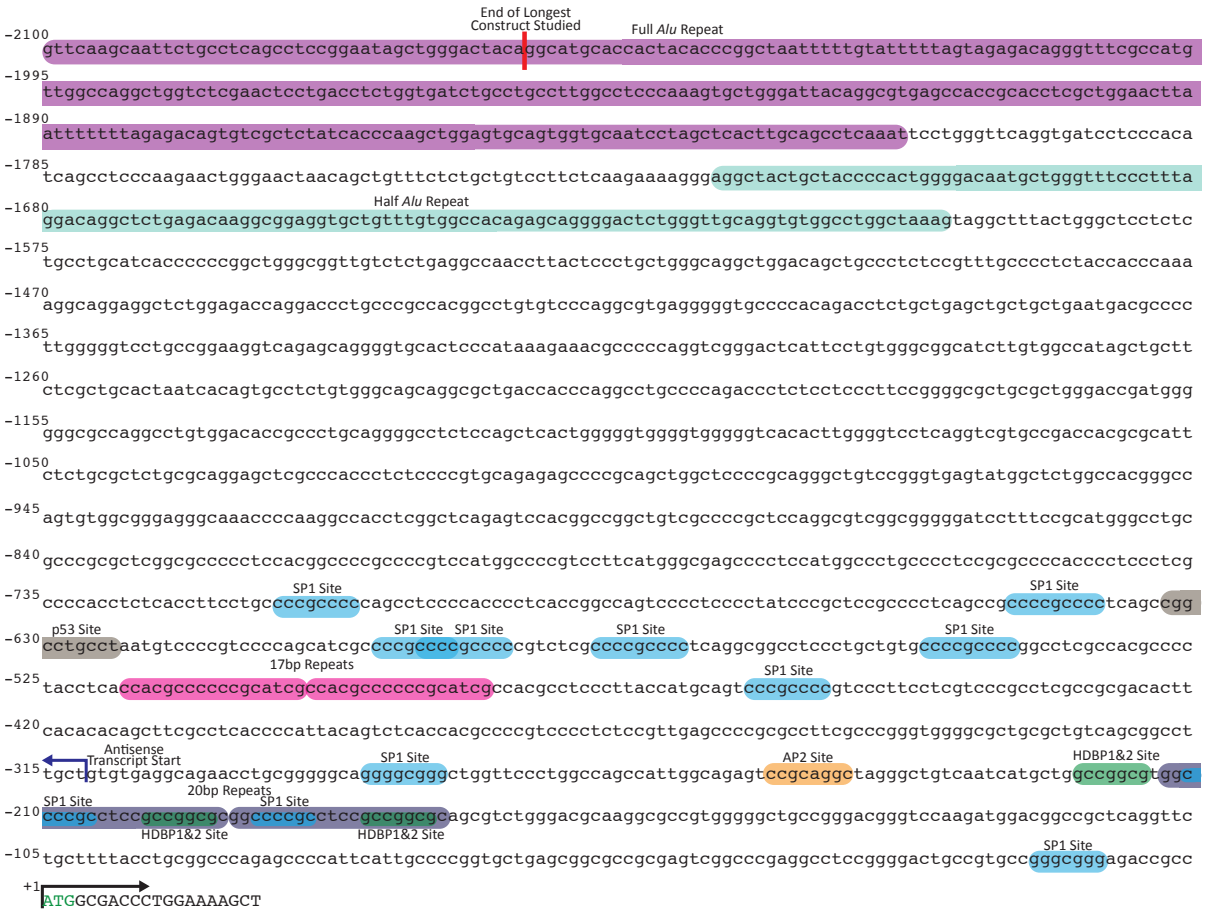


Figure 1.1 Previously Identified TFBS and HTT Promoter Features
Provided are the genomic locations of all previously identified TFBS from the papers discussed below. Shown are only TFBS for which additional supporting data was provided (*i.e* TFBS identified through sequence similarity alone were not included). +1 indicates translational start site. Sequence and figure adapted from the UCSC genome browser.

Study of the *HTT* promoter began shortly after its identification as the causative gene for HD, with the first published study appearing in 1995 (Lin et al. 1995). This first paper utilized previously established mouse and human phage and cosmid libraries as sources of *HTT* proximal promoters and the start of the gene. Using sequencing, the researchers compared the homology of the mouse *Hdh* and human *HTT* proximal promoters and found high homology, greater than 78%. Homology was particularly high ~200 bp 5' of the transcriptional start site, indicating that this region may house important regulatory elements required for *HTT* transcription. As will become more notable in subsequent discussion, the proximal promoter in both mouse and human was found to be devoid of TATTA or CAAT boxes and GC rich. Finally the researchers utilized two commercially available bioinfomatic analysis programs to probe for putative TFBSs, finding numerous putative SP1 sites and a single AP2 site. Unlike modern chIP-seq methods, in which putative sites are predicted first through identified binding of the TF to the genomic region, these bioinfomatic analysis programs used sequence similarity to pre-programmed PWM alone to identify potential TFBS. Following upon this research Coles et al., 1997 published a paper aimed at identifying potential polymorphisms in the *HTT* promoter region between different human populations. Utilizing sequences directly acquired from individuals, they identified 4 alleles based on two base pair substitutions and two variable number tandem repeats (VNTR) of either 6 or 20 bp; however, none of these 4 alleles correlated with the age of onset in HD patients. The two VNTR identified are only present in single copies in non-human primates, indicating that they arose after the divergence of the human lineage. Interestingly,

though the 20 bp VNTR was further characterized by the same group, where they utilized a luciferase assay to identify VNTR as essential to maintain minimal function of the promoter Coles et al., 1998, the 6 bp VNTR does not appear in any later publications. They also verify the function of the conserved SP1 and AP-2 site identified previously through electronic mobility shift assay. Holzmann et al., 1998 published on the rat *HTT* homologue promoter, finding similar conservation between mouse, rat and human along the same region identified in the Coles papers, as well as the lack of a TATA or CAAT box in the rat *HTT* promoter. They also identified a CRE site that is found in both rat and mouse that is not present in humans; deletion of this site reduced rat promoter function in a functional assay. In addition to the work done on the rat *HTT* promoter, Holzmann et al., 1998 also published work on the human *HTT* promoter identifying several other putative TFBS based on available bioinfomatic programs. They also generated larger reporter constructs to be used in functional assays and identified both a full and half Alu repeat ~2000 bp from the translational start site.

An antisense transcript has also been identified originating in the proximal promoter region of *HTT* gene (Chung et al. 2011). This antisense transcript originates from the translational start site and is transcribed on the opposite strand from the *HTT* mRNA from several closely placed TSS, with the majority of transcripts having a TSS at +300 relative to the transcriptional start site of *HTT*. The antisense transcript has two splice variants, one with 3 exons and one with 2. Neither are highly expressed, but the 3 exon variant is expressed in multiple tissue types and has higher expression than the 2 exon variant. This antisense transcript has been

found to reduce *HTT* mRNA expression slightly (Chung et al. 2011). The CAG repeat seems to have an effect on both antisense transcription and the ability of the antisense transcript to regulate *HTT* transcription.

1.4.4 TFs Involved in *HTT* Promoter Function

As already mentioned, many of the early studies utilized bioinfomatic programs to screen for putative TFBS based on sequence similarity alone, followed by verification through EMSA in some cases but not all. As an improvement identification by sequence similarity alone, several additional studies utilized functional assays to identify TFs and associated TFBS in the *HTT* promoter region. To date, there are 4 TFs that have been implicated in *HTT* transcriptional regulation using a functional assay, p53: Sp1, HDBP1 and HDBP2 (Ryan et al. 2006; Tanaka et al. 2004; Wang et al. 2012). In addition, work previously established by my colleages has identified NfkB as an additional regulator of *HTT* expression. p53 is a regulator of cellular survival, primarily known for its role in DNA damage repair and for its tumor suppression functions. It is involved in cellular responses to stress (Pflaum et al 2014). Its function as a regulator of *HTT* expression is consistent with *HTT*s role in cellular, particularly neuronal, survival and response to cellular stress.

SP1 is a fairly ubiquitous TF, known to bind many genes, with a predicted 12,000 binding sties in the human genome (Beishline & Azizkhan-Clifford 2015). It is involved in many cellular processes, including, cell growth/death, immune response, and development; this is once again consistent with HTTs role in cellular survival and

development (Beishline & Azizkhan-Clifford 2015). Surprisingly, despite being published in 2013, the researchers did not make reference to established online databases when designating what they established as the transcriptional start site (Wang et al. 2012). Indeed, upon further review of their constructs, it is clear that parts of the sequence they used do not match the UCSC genome assembly for the *HTT* promoter, and the TSS they designate lies upstream of the identified TSS.

Binding and the naming of HDBP1 and HDBP2 were conducted in the same study (Tanaka et al. 2004). These two proteins were identified using a yeast one-hybrid system designed specifically to screen for candidate *HTT* TFs. The two hits from this screen were designated HDBP1 and HDBP2 respectively. The sequence for HDBP1 matches that of the SLC2A4 gene, while HDBP2 matches that of ZNF395. SCL2A4 is a transmembrane protein involved with insulin-regulated glucose transport. It is unclear how this transmembrane protein would be implicated in *HTT* transcription, though the authors did note that the HDBP1 protein could also be a splice variant of GLUT4 enhancing factor DNA binding domain (GEFdb) based on amino acid homology. They did not provide evidence of the existence of this splice variant mRNA, leaving the identity of this TF still up for debate. ZNF395 was identified as identical to the papillomavirus binding factor (PBF), a known transcription factor involved in the expression of papillomaviurs genes. PBF has been implicated in cell growth (Sichtig et al. 2007).

Finally, data currently in review by my colleagues, Becanovic et al. (in review), has identified NFκB as transcriptionally regulating *HTT* expression. NFκB is well known and ubiquitously expressed TF involved in various cellular stress pathways.

Using promoter sequences from several HD patients in promoter-luciferase assays Becanovic et al. identified a single patient who's *HTT* promoter had significantly less function. Subsequent analysis revealed that this was due to a point mutation incurred during the cloning process, this mutation affected a single highly conserved nucleotide in the TFBS for NFκB in the promoter. The use of Hapmap data sets uncovered the presence of a single polynucleotide polymorphism (SNP) at another position within this same NFκB TFBS. Using three human HD cohorts with known CAG and age of onset Becanovic et. al were able to phase the NfKB SNP to either the widltype or HD allele in each individual. In two of the cohorts the SNP was found predominantly on the wildtype allele the presence of this SNP on the wildtype allele decreased the age of onset by on average 5 years compared to the predicted age of onset based on CAG size. In the third cohort the NFκB SNP was found predominantly on the HD allele, this had the inverse effect on age of onset, increasing the age of onset by on average 12 years.

1.4.5 Conclusions

As I have illustrated in the above section, while the HD community has made some strides into understanding the modes of transcriptional regulation of the *HTT* gene, the efforts made pale in comparison to the vast amount of research conducted on the protein itself. In addition, the bulk of *HTT* transcriptional research has occurred prior to advancements in sequencing technologies, which have allowed our understanding of TFs and their respective TFBS to grow. The TFBS identified in the

above studies were found using sequence similarity as their sole source based on PWM generated before genome-wide ChIP-chip and ChIP-seq based PWM were created. For the majority of these studies EMSA was used to confirm the functionality of the predicted TFBS as a binding site. While EMSA is a good starting place to establish if a TFBS can bind nuclear extracts, it is important to establish exactly what is interacting with the TFBS through a Supershift Assay, the only study to do this was one of the p53 studies. Even so both an EMSA and a Supershift Assay are both *in vitro* representations of actual binding in a cellular, and more importantly, genomic context. The focus of previous research has only focused on a very small portion of the *HTT* core and proximal promoter. We now know that long distance enhancers play a significant role in gene regulation, particularly in a differential tissue expression context. Previous research has not provided us with a mechanism through which differential tissue gene expression can occur. Finally, although there have been several studies into the genome-wide effects of the HD mutation on epigenetic modification, the majority of these studies do not report on epigenetic changes at the *HTT* locus. Given the potential for inter-individual expression variability, it is important to understand potential epigenetic effects. This is especially important for future HD genetic treatments, as variability of *HTT* expression could be influential on treatment dosage and treatment effect.

1.5 Objectives

The objectives of this thesis were 1) to conduct a bioinformatic-driven search for potential regulatory regions outside of the *HTT* promoter region, 2) to create a

stably expressing cellular screening system to test potential regulatory regions and associated TFs, and 3) to investigate the role of DNA methylation in tissue expression differences of *HTT* and the effects of the HD mutant allele on *HTT* gene methylation.

I conducted a comprehensive screen, based on available genome-wide data, of the *HTT* gene locus and identified additional putative regulatory regions using epigenetic marks and conservation. These regions were then screened for putative TFBS based on available ChIP-seq data sets. We developed several cell lines in both Human Embryonic Kidney cell (HEK293) and a rat MSN progenitor cell (ST14) that stably express various promoter - luciferase constructs. In addition to *HTT* promoter constructs, I have also established cell lines that carry one of my putative regulatory regions in front of the *HTT* promoter. These cell lines were used to develop an assay for the screening of potential regulatory TFs. Finally, I utilized human samples from the UBC HD Tissue Biobank to study DNA methylation of the *HTT* gene. Using both HD and control cortex samples, I investigated both genome-wide changes in DNA methylation and changes at the *HTT* gene locus. In addition, I used human cortex and liver samples to investigate the role of DNA methylation on differential *HTT* expression between these two tissues.

2 Bioinfomatic Assessment of the *HTT* Locus

2.1 Introduction

As elaborated upon in the introductory chapter, advances in genome-wide screening technologies have vastly altered our understanding of regulatory proteins and their effect on gene transcription; this has broadened the scope of potential regulatory regions from proximal regions close to TSSs, to regions both within the gene body itself as well as seemingly unrelated sequences hundreds, if not thousands, of kilobases away. Given that the majority of studies concerting *HTT* transcriptional regulation occurred before the expansion of genome-wide technologies it seemed appropriate to renew investigations into regulation of this gene using data generated from these new technologies. To do this I utilized selected studies within the large-scale ENCODE project to conduct a comprehensive analysis of the *HTT* locus (ENCODE Project Consortium 2004). I included datasets for histone modifications, transcription factor binding, DNase accessible open-chromatin, and sequence homology. Using these datasets I was able to identify additional regions containing potential regulatory sequences for further investigation.

To identify TFBS within my putative regulatory regions, I then utilized ChIP-seq datasets for known TFs from both ENCODE and PAZAR to scan my regions (ENCODE Project Consortium 2004; Portales-Casamar et al. 2007). This enabled me to identify TFBS based on both sequence similarity to a PWM, and also based on peaks within the ChIP-seq data set for said TF. In this way my identified TFBS are

based upon biological binding of the TF during the ChIP-seq assay. To prioritize this list of candidate TFBSs, and associated TFs, I used two approaches. First, to identify TFs within each region that may be acting synergistically to affect gene transcription I created a protein protein interaction network and then clustered the network to identify known TF interactions within my candidate TFs. This information was then used to identify clusters of TFBS in the putative regions, these TFs became my top candidates for further study. Secondly, I used the known *HTT* expression difference between CNS tissues and peripheral tissues to identify DNase hypersensitivity datasets from ENCODE from both brain derived cell lines and peripheral derived cell lines. I then compared DNase hypersensitivity sites for the brain datasets and the peripheral datasets within my regions of interest to find DNase hypersensitive sites that were unique to either dataset, and those common to both. Hypersensitive sties unique to either cell type would indicate that those sites are potentially used in tissue specific expression. Those common to both cell types would indicate regions important for expression in both the periphery and the brain.

2.2 Methods

2.2.1 Regulatory Region Heuristic Scoring Procedure

I localized my region of scoring to the first intron of the gene preceding *HTT*, *GRK4*, to the end of the gene following *HTT*, *MSANTD1*. This corresponds to region chr4:2,967,424-3,260,411, a total of 292,988 bp. As mentioned in the introductory chapter is possible that genomic regions outside of the chosen region may

contribute to *HTT* transcriptional regulation in the absence of any further information regarding interactions of more distal sequences with the *HTT* promoter limiting this initial search to the local genomic area was the best logical step. The preceding gene, *GRK4*, is part of a family of 7 G protein-coupled kinases that are involved in the physiological feed back mechanism of desensitization of G protein-coupled receptors (Watari et al. 2014). *GRK4* was first thought to be expressed only in the brain and testes, however it has subsequently been found to be expressed in many other cell types (Andresen 2010). This pattern of expression is very similar to that of the *HTT* transcript, suggesting that these two genes in this local genomic context may be regulated by similar regulatory regions. The core promoter region has been established to be within 1815 bp upstream of the transcription start site of *GRK4* (Hasenkamp et al. 2008). To avoid the *GRK4* promoter region I began my screen for *HTT* regulatory regions at the first intron of the *GRK4* gene. The gene following *HTT* is a short transcript *MSANTD1*, this transcript had been named chromosome 4 open reading frame 44 (C4orf44) until recently when it was renamed to *MSANTD1* based on protein sequence similarity to the Myb/SANT DNA-binding domain. No other information on this gene exists in the literature; I attempted to assess transcript expression of this gene by RT-qPCR in human samples utilized in chapter 5 of this thesis, but was unsuccessful (data not show). I have included this short gene in my analysis to see if any regulatory regions within this gene could be identified.

Using the tracks from UCSC genome browser, listed in table 2.1, and the accompanying criteria for scoring based upon each track, I scored the genome in 1 kb increments starting and ending at the genomic regions listed above. In addition to

noting the score for each region a list of the features within each 1 kb region contributing to the score was also kept, (file will be provided on line). This scoring was done using the available tracks and datasets as of May 2012-July 2012.

Table 2.1 UCSC Genome Browser Tracks Used and Scoring Paradigm

Track Name	Feature Indicated	Notes
Human mRNA	presence of gene	No points awarded
Human ESTs	presence of gene	No points awarded
ENCODE Regulation	DNase sensitivity, Histone Modification/Epigenetic Marks	Points added for marks for transcriptional activation but no points added or deducted for marks for transcriptional silencing, just noted.*,**
CD34 DNaseI	DNase sensitivity	Adds 1 point
CpG Islands	Presence of CpG island	Presence of island noted but no points added or subtracted
ENCODE TF Binding	TFBS	Adds 1 point
ENCODE DNA Methylation	Histone Modification/Epigenetic Marks	Points added for marks for transcriptional activation but no points added or deducted for marks for transcriptional silencing, just noted.*,**
ENCODE DNase	DNase sensitivity	Adds 1 point
ENCODE Histone	Histone Modification/Epigenetic Marks	Points added for marks for transcriptional activation but no points added or deducted for marks for transcriptional silencing, just noted.*,**
ORegAnno	TFBS	Adds 1 point
TFBS conserved	TFBS	Adds 1 point
Vista Enhancer	TFBS	Adds 1 pont
UCSF Brain Methylation	Histone Modification/Epigenetic Marks	Points added for marks for transcriptional activation but no points added or deducted for marks for transcriptional silencing, just noted.*,**
Conservation	Conservation	Set up: excluding primates, including Placental Mammal (mouse, rabbit, rat, cow, dog), including vertebrate (opossum, chicken and fugu, exclude primates) Listing Mammal Cons and Multiz Align. Presence of conservation adds 1 point
Repeat Masker	Short and Long repeats	Repeat noted but does not add or subtract a point
* Marks for activation: H3K4 methylation, H3K36 methylation, H3K4 Acetylation, H3K27 Acetylation		
** Marks for silencing: H3K9 methylation, H3K27 methylation		

2.2.2 Identification of Putative TFBS Using ChIP-seq Datasets

The identification of putative TFBS for my regions of interest was conducted in two stages. The first being the creation of PWM based upon available ChIP-seq datasets. The second, using these PWM to scan these regions for TFBS.

2.2.2.1 PWM Construction

Using all the ChIP-seq datasets obtained from ENCODE (since 2012) and PAZAR datasets listed in table 2.2 (a total of 478 ChIP-seq datasets representing 107 TFs) I constructed a PWM for each TF represented using the following pipeline:

1. 200nt around the ChIP-seq peak max positions (around the centre when peak ma positions was not available) was extracted
2. MEME was applied to the top 600 peaks (ranked using the peak signal value) (Bailey et al. 2009)
3. From each one of the top 3 most over-represented motifs found using MEME:
    a) Apply MAST (Bailey & Gribskov 1998) to the whole data set coming from 1.
    b) Construct a PFM from MAST hits
        a) Construct a PWM following (Wasserman & Sandelin 2004)
    c) Extract 500nt around the ChIP-seq peak max positions (around the centre when peak max not available) for the whole data set
    d) Predict TFBSs in the sequences coming from d) by using a 85% relative threshold

e) Compute a centrality score for quality control as defined in CentriMo (Bailey & Machanick 2012)

f) Discard the PWM if the logarithm of the centrality p-value is above -200

For each TF I kept the PWM with the best centrality score (lowest p-value) passing my quality score threshold.

2.2.2.2 TFBS Prediction

The regions identified using my regulatory region heuristic scoring procedure were scanned using the ChIP-seq datasets, and associated PWM from the PWM construction method mentioned above. I scanned each of my putative regulatory regions with my PWM using a 85% relative threshold to predict TFBS (Wasserman & Sandelin 2004). To ensure that the predicted TFBS reflected binding of TFs as identified by the ChIP-seq assays I only considered TFBS that corresponded to a ChIP-seq peak for that specific TF. For example, a predicted JunB TFBS within a peak from a ChIP-seq dataset for SMC3 would not be included. The identified TFBS within the regions of interest were used to create a custom UCSC genome browser track, this allowed me to visualize the data in the context of the regions of interest as well as other genomic features.

Table 2.2 ChIP-seq Datasets from PAZAR

Transcription Factor	ChIP-seq experiment acession number	Transcription Factor	ChIP-seq experiment acession number
CTCF	GSE20650	RAD21	GSE25021
CTCF	GSE25021	NFYA	GSE26439
CTCF	GSM325899	CEBPB	GSE31939
SP1	GSE34791	GATA1	GSE24674
GATA2	GSE29194	GATA1	GSE26501
GATA2	GSE29195	GATA1	GSE29194
REST	E-MTAB-437	GATA1	GSE29195
TAL1	GSE24674	E2F4	GSE24326
TAL1	GSE25000	ESR1	GSE24166
TAL1	GSE26014	ESR1	GSE25021
TCF7L2	GSE29194	ESR1	E-MTAB-740
TCF7L2	GSE29195	TAF1	GSE17917
HNF4A	GSE23436	SPI1	GSE25426
HNF4A	GSE25021	SPI1	GSE26014
ETS1	GSE17954	POU5F1	GSE20650
ETS1	GSE29808	POU5F1	GSM539547
GATA3	GSE29073	EGR1	GSE21665
FOXA2	GSE25836	STAT1	E-GEOD-12782
GABPA	GSE24933	ZNF263	E-GEOD-19235
GABPA	GSE29808	HNF4A	E-TABM-722
NANOG	GSE18292	ESR1	GSE 22609
NANOG	GSE20650		
NFKB1	GSE34329		
FOXA1	GSE23852		
FOXA1	GSE27823		
FOXA1	GSE28264		

2.2.3 Protein-Protein Interaction Network of Candidiate TFs Using STRING

As noted in the introductory chapter, TFs are often found to work in associating groups to affect transcriptional regulation. To identify clusters of known TF interactors within my candidate TFs, I utilized an online protein-protein interaction network database, STRING (Franceschini et al. 2012). The TF interaction network was used to create the following settings: High confidence (required confidence score 0.700), co-occurrence, co-expresion, experiments, databases. This online database allowed the creation of an interaction network based upon known and predicted interactions between my candidate TFs. I utilized the clustering feature of this database, which uses the Markov Cluster Algorithm, and an inflation setting of 3, to identify clusters of associating TFs based upon the defined interaction criteria.

2.2.4 DNase Hypersensitivity Analysis

Using the available DNase I hypersensitivity datasets in ENCODE the available data was separated into two sets, those from brain derived cell lines and those from peripheral cell lines (Table 2.3). Brain specific DNase hypersensitivity sites were those that overlapped my 11 regions of interest and did not intersect with peaks from the peripheral cell line datasets. The converse is true for the peripheral DNase hypersensitivity data set. I used this information to create custom tracks in the UCSC Genome browser to visualize the DNase hypersensitive sites in context with the the 11 regions of interest and other genomic features.

Table 2.3 DNase Datasets from ENCODE

Peripheral DNase Datasets from ENCODE	Brain DNase Datasets from ENCODE
wgEncodeOpenChromDnase8988tPk.narrowPeak	wgEncodeOpenChromDnaseCerebellumocPk.narrowPeak
wgEncodeOpenChromDnaseHeartocPk.narrowPeak	wgEncodeOpenChromDnaseCerebrumfrontalocPk.narrowPeak
wgEncodeOpenChromDnaseHek293tPk.narrowPeak	wgEncodeOpenChromDnaseGlioblaPk.narrowPeak
wgEncodeOpenChromDnaseHepatocytesPk.narrowPeak	wgEncodeOpenChromDnaseMedulloPk.narrowPeak
wgEncodeOpenChromDnaseHepg2Pk.narrowPeak	wgEncodeOpenChromDnaseMedullod341Pk.narrowPeak
wgEncodeOpenChromDnaseHuh75Pk.narrowPeak	wgEncodeOpenChromDnaseSknshPk.narrowPeak
wgEncodeOpenChromDnaseHuh7Pk.narrowPeak	wgEncodeUwDnaseHacPkRep1.narrowPeak
wgEncodeOpenChromDnaseStellatePk.narrowPeak	wgEncodeUwDnaseHacPkRep2.narrowPeak
wgEncodeUwDnaseHcfPkRep1.narrowPeak	wgEncodeUwDnaseHahPkRep1.narrowPeak
wgEncodeUwDnaseHcfPkRep2.narrowPeak	wgEncodeUwDnaseHahPkRep2.narrowPeak
wgEncodeUwDnaseHcfaaPkRep1.narrowPeak	wgEncodeUwDnaseHaspPkRep1.narrowPeak
wgEncodeUwDnaseHcfaaPkRep2.narrowPeak	wgEncodeUwDnaseHaspPkRep2.narrowPeak
wgEncodeUwDnaseHcmPkRep1.narrowPeak	wgEncodeUwDnaseM059jPkRep1.narrowPeak
wgEncodeUwDnaseHcmPkRep2.narrowPeak	wgEncodeUwDnaseM059jPkRep2.narrowPeak
wgEncodeUwDnaseHepg2PkRep1.narrowPeak	wgEncodeUwDnaseNhaPkRep1.narrowPeak
wgEncodeUwDnaseHepg2PkRep2.narrowPeak	wgEncodeUwDnaseNhaPkRep2.narrowPeak
wgEncodeUwDnaseHpfPkRep1.narrowPeak	wgEncodeUwDnaseSknmcPkRep1.narrowPeak
wgEncodeUwDnaseHpfPkRep2.narrowPeak	wgEncodeUwDnaseSknmcPkRep2.narrowPeak
wgEncodeUwDnaseHrgecPkRep1.narrowPeak	wgEncodeUwDnaseSknshraPkRep1.narrowPeak
wgEncodeUwDnaseHrgecPkRep2.narrowPea	wgEncodeUwDnaseSknshraPkRep2.narrowPeak

2.3 Results

2.3.1 11 Potential Transcriptional Regulatory Regions Identified at the *HTT* locus

Using the scoring paradigm outlined in the methods section I systematically scored the genome in 1 kb intervals from the first exon of the gene preceding *HTT, GRK4*, to the end of the following gene, *MANSTD1.* Figure 2.1 provides a histogram of the scores generated for each 1 kb segment. As expected, the majority of the segments had low score, 0 or 1-2, these indicate regions with little or no markers for regulatory sequences. Based upon this histogram I selected those above a threshold of 7 for further investigation. One kb sections with this score were then further examined to collate incremental sections into larger regions, these regions were then given an overall average score based upon the score of the 1 kb sections comprising the region. One kb sections that were below my threshold of 7, but resided between sections that were over the threshold, were considered on a per case basis. For example, a section with a score of 6 flanked by two sections with scores of 8 would be included. Amalgamating the 1 kb sections that passed my threshold resulted in 11 regions of interest (Figure 2.2) with average scores for each region listed in Table 2.4. The proximal *HTT* promoter region and TSS was identified as a large region, region 6, with the highest average score. Several additional regions were identified within and shortly after the *GRK4* gene, regions 1-5. Another large high scoring region was identified in the 3' UTR of the *HTT* gene, region 11. No regions were identified in the *MASTD1* gene.

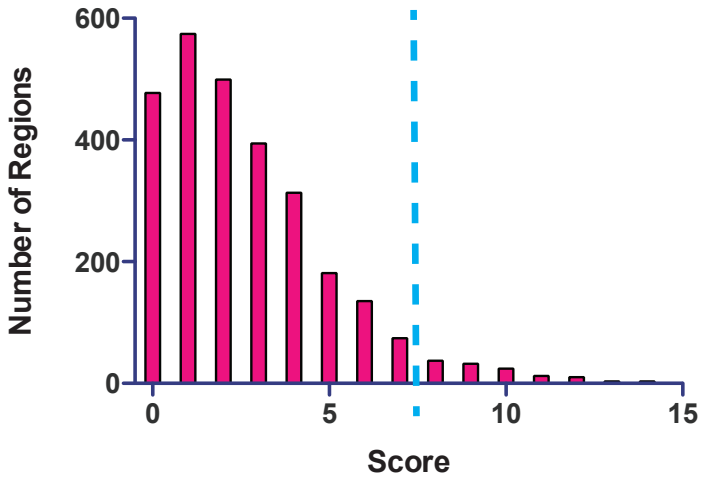


Figure 2.1 Histogram of Scores for each 1 kb region Assessed
Using the Heuristic Scoring Paradigm listed in the methods section I scored the genome between chr4:2,967,424-3,260,411. Presented are the scores all of the 1 kb regions for features associated with transcriptional regulation. A threshold of 7 features was selected to designate regions for follow up analysis.

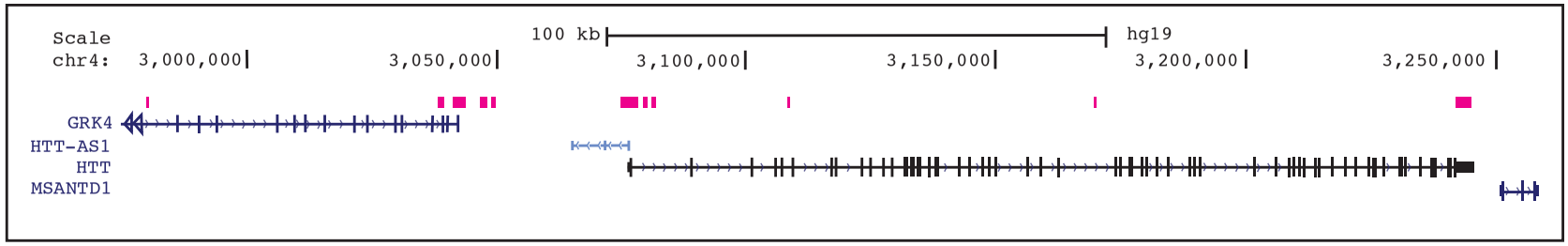


Figure 2.2 11 Putative Regions of Transcriptional Regulation

Taking the 1 kb regions whichhad a score greater than 7 features I collated the regions and combined those whichwere in sequential order. From this I created 11 putative regions of transcriptional regulation. Regions were numbered from 1-11 from left to right. Figure adapted from UCSC genome browser

Table 2.4 Putative Regulatory Region Details and Average Score

	Location	Start	End	Length(bp)	Average Score
Region 1	First intron of preceding gene (*GRK4*)	2,980,000	2,980,600	600	8
Region 2	Spans introns 13 and 14 of preceding gene (*GRK4*)	3,038,500	3,039,500	1,000	6
Region 3	Spans final intron, final exon and 3'UTR of preceding gene (*GRK4*) and inter gene space	3,041,550	3,044,000	2,450	8
Region 4	inter gene space between preceding gene (*GRK4*) and *HTT*	3,047,000	3,048,200	1,200	6
Region 5	inter gene space between preceding gene (*GRK4*) and *HTT*	3,049,000	3,050,000	1,000	8
Region 6	Spans proximal promoter region, 5'UTR, first exon and first intron of *HTT* gene	3,074,800	3,078,250	3,450	10
Region 7	First intron of *HTT* gene	3,079,300	3,080,250	950	8
Region 8	First intron of *HTT* gene	3,081,200	3,091,900	10,700	8
Region 9	Intron 5 of *HTT* gene	3,108,100	3,108,750	650	7
Region 10	Intron 28 of *HTT* gene	3,169,625	3,170,075	450	6
Region 11	3' UTR of *HTT* gene	3,242,200	3,245,200	3,000	8

2.3.2 Identification of TFBS within 11 Candidate Regions

Using the PWM and TFBS Prediction methods described in the methods section, I scanned my 11 regions of interest for TFBS using ChIP-seq datasets from ENCODE and PAZAR. In total this comprised of 479 ChIP-seq datasets representing 103 individual TFs. Of these TFs, 43 were identified in one or more of my 11 candidate regions, Table 2.4 details the TFs and the number of sites across my 11 regions that were identified. A reminder that, as mentioned in the methods section, I only considered TFBS based on the generated PWM if that TFBS occurred within an identified ChIP-seq peak for the corresponding TF to increase the quality of TFBS predictions. By doing this I limited my predicted TFBS to those where binding of the corresponding TF has been observed, eliminating TFBS that would be identified on sequence similarity alone.

Table 2.5 TFs Identified in Putative Regulatory Regions

TF	Number of Occurrences Across Putative Regulatory Regions	TF	Number of Occurrences Across Putative Regulatory Regions
Atf3	3	**Nrsf**	1
Batf	12	**Pax5**	10
Bcl11a	7	**Pbx3**	1
Bhlhe40	8	**Pou2f2**	4
Cebpb	1	**Pu1**	3
Myc	2	**Rad21**	3
E2f6	15	**Rfx5**	9
Fos	2	**Smc3**	9
Egr1	35	**Sp1**	2
Elf1	8	**Srf**	2
JunB	9	**Stat1**	2
Fosl1	8	**Stat3**	21
Gata1	1	**Tbp**	1
Hnf4	1	**Tcf12**	36
Irf1	21	**Tcf7l2**	3
Irf3	1	**Usf1**	4
JunD	5	**Usf2**	8
Maff	2	**Zbtb33**	2
Mafk	2	**Zeb1**	12
Max	5	**Znf217**	1
Mef2	1	**Znf263**	9
Nrf1	8		

2.3.3 Identifying Associating TFs Among the 44 Identified TFs

As mentioned in the introductory chapter TFs often act synergistically to promote binding and effect gene regulation. By identifying interacting TFs within my identified 44 TFs I then scanned my regions for TFBS of interacting TFs. This allowed us to prioritize my candidate TFBS, and associated TFs. Using the online protein interacting database, STRING, which contains both known and predicted protein protein interactions, I created a network of my 44 TFs (Figure 2.3). I used the Markov Clustering tool, available through String, to identify TFs within the network with high associations. Within my TFs I identified 7 clusters, each highlighted by a different colour in the figure. 16 of the TFs were identified as not having any interactions with any other of my TFs. To identify clusters of interacting TFBS within the regions I plotted the TFBS within each region and colour coded each TF based upon its association group, Figure 2.4 provides an example plot from region 6, the region surrounding the TSS of *HTT*.

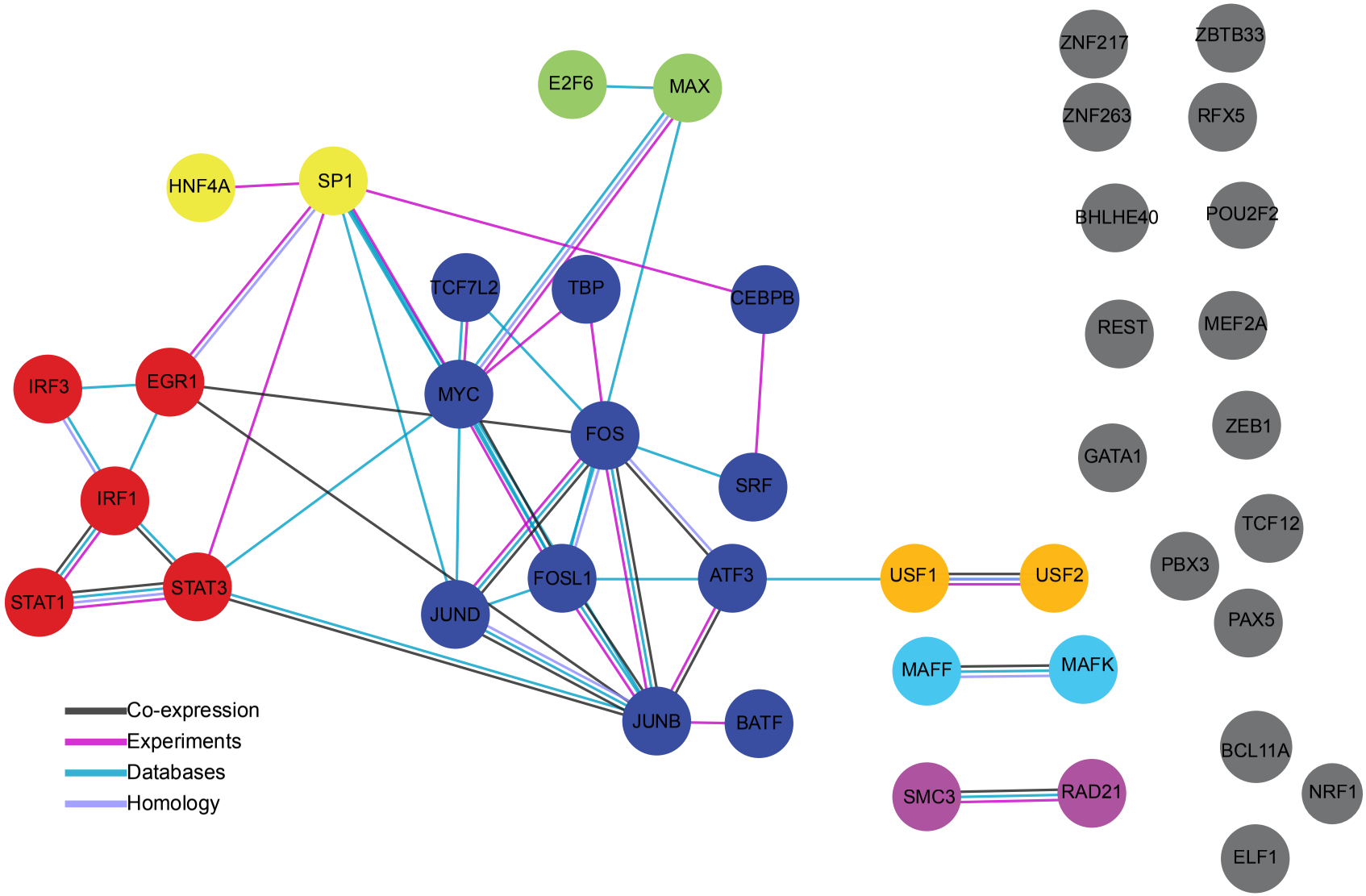

E2F6
MAX
HNF4A
SP1
TCF7L2
TBP
CEBPB
IRF3
EGR1
MYC
FOS
SRF
IRF1
STAT1
STAT3
FOSL1
ATF3
JUND
JUNB
BATF
USF1
USF2
MAFF
MAFK
SMC3
RAD21
ZNF217
ZBTB33
ZNF263
RFX5
BHLHE40
POU2F2
REST
MEF2A
GATA1
ZEB1
TCF12
PBX3
PAX5
BCL11A
NRF1
ELF1
Co-expression
Experiments
Databases
Homology

Figure 2.3 Protein-Protein Interactions Between Candidate TFs
As described in the methods section, an online database (http://string-db.org/) was used to identify known protein-protein interactions between my candidate TFs. Clusters of interacting TFs, as identified by Markov Cluster Analysis, are indicated by matching colours. Figure adapted from http://string-db.org/

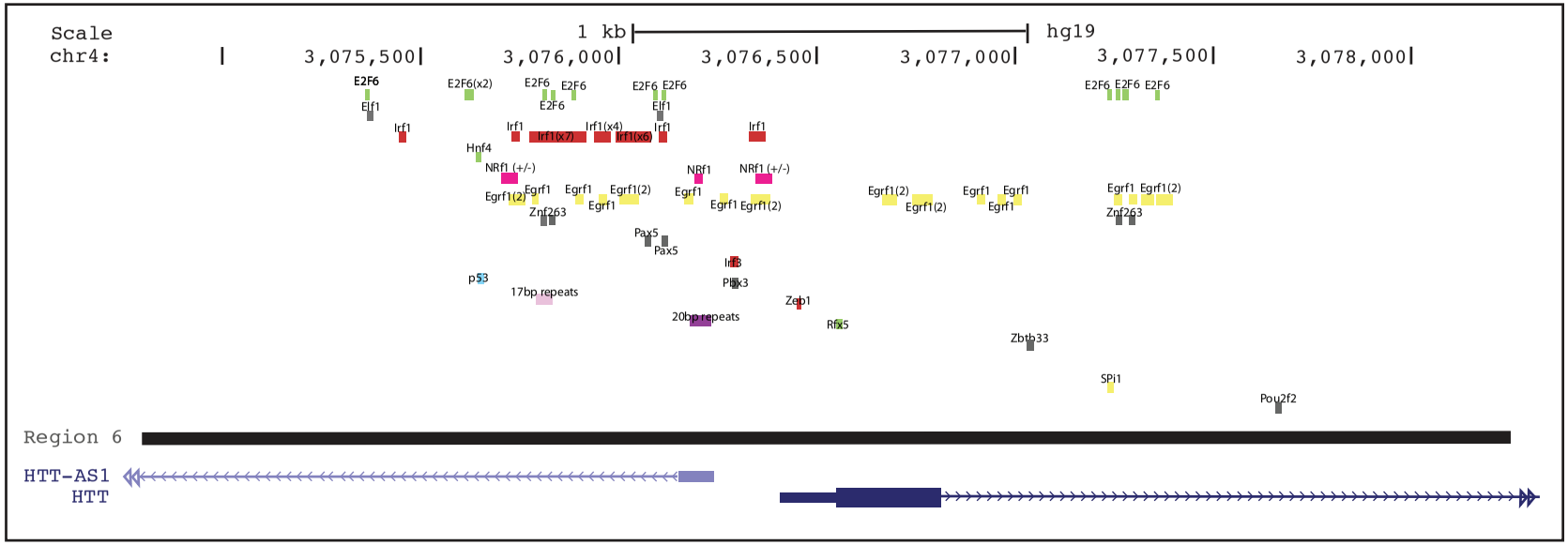


Figure 2.4 Predicted TFBS within HTT Promoter Region (Region 6)
Predicted TFBS in Region 6, the region whichcoincides withthe proximal *HTT* promoter. Numbers in brackets indicate number of sites if sites are in close proximity, +/- indicates a site on bothstrands. Also depicted are the 17 and 23 bp repeats mentioned in the introductory chapter. TFBS are colour coded according to their interacting group (figure 2.3). Figure adapted from UCSC genome browser.

2.3.4 Brain and Peripheral specific DNase Hypersensitivity

DNase hypersensitivity indicates regions of the genome that are not impaired by histones or chromatin structure and are thought to be “open” to both DNA cleavage by DNase as well as binding of TFs. As histone positioning and chromatin structure can change between cell types DNase hypersensitivity allows for identification of differentially open or closed sites based on cell type. As mentioned in the introduction, *HTT* is differentially expressed in brain versus most other peripheral tissues. Identifying which of my regions of interest, or portions of, are specifically open in either brain or peripheral cell types I can identify regions that are likely to have interactions with TFs in each cell type. This allowed me to further prioritize my list of candidate regions and TFBS to those that are likely to be open and bind TFs in either brain or peripheral tissues. Using the DNase hypersensitivity model described in the methods section I created three UCSC genome browser tracks, added to Figure 2.2 to create Figure 2.5. It is clear that some of my regions of interest are not preferentially open in either cell type, for example regions 7 and 8 are not open in either cell type using the DNase datasets I obtained.

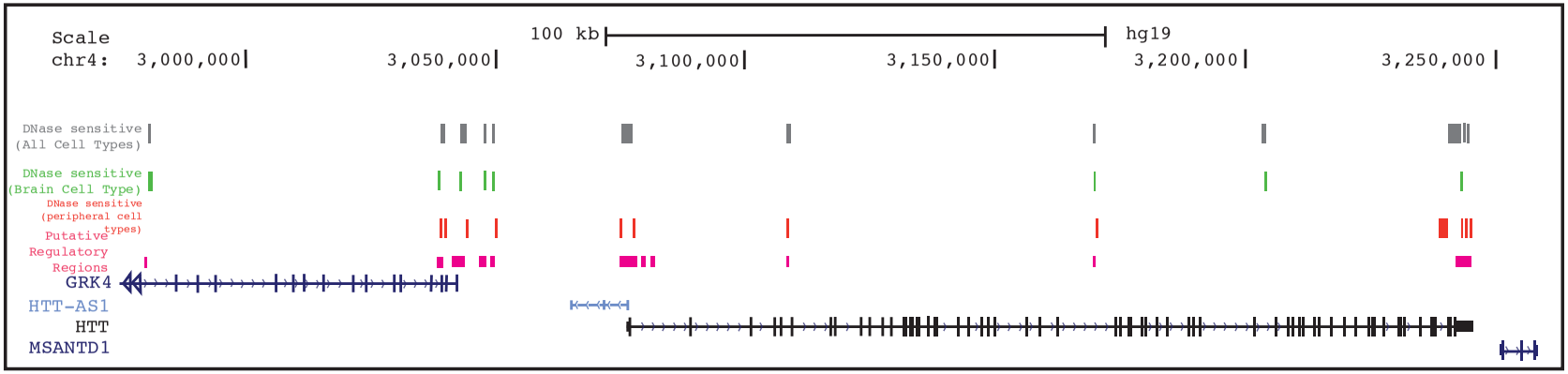


Figure 2.5 DNAse Hypersensitivity in Putative Regulatory Regions

Figure 2.2 modified to include DNase hypersensitivity sites within all cell lines tested (grey), brain cell type (green), peripheral cell type (red). Adapted from the UCSC genome browser.

2.3.5 Selection of Region 9 for Followup

Upon considering each region of interest for both clustering of interacting TFBS and DNase hypersensitivity I selected region 9 in addition to region 6, the proximal *HTT* promoter, for further research (Figure 2.6). This region contains several TFBS for transcription factors in the largest cluster of associated TFs, including JunB, Batf, Fos, and Fosl1. This region also contains a peripheral cell type specific DNase hypersensitivity site. While this site does not overlap any of the identified TFBS it is possible that this nearby open site allows for TFs to be loaded on to near by TFBS. Indeed although the prominent view concerning DNase hypersensitivity and TF binding states that the DNA must be clear of impeding proteins for TF binding to occur some TFs have been found to require the presence of nucleosomes to bind (Ballaré et al. 2013). As the primary follow up was done in a peripheral cell type, HEK293 cells, I decided to select a region for follow up that displayed a peripheral DNase hypersensitivity. This was done under the assumption that a region that is primarily open in peripheral cell types would likely have TFBS for TFs that are more active in a peripheral cell type.

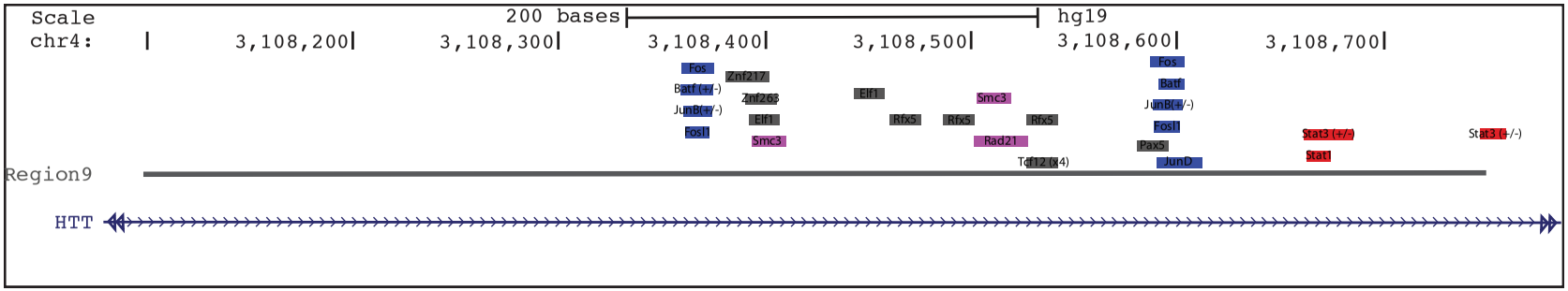


Figure 2.6 Putative TFBS within Region 9

Predicted TFBS in putative regulatory region 9. TFBS are colour coded according to their interacting group (figure 2.3). Figure adapted from UCSC genome browser.

2.5 Discussion

In my analysis of the *HTT* promoter, I have used a scoring paradigm based up on known markers of transcriptional regulation to identify regions outside of the proximal promoter that may be involved in *HTT* transcriptional regulation; this is the first study of its kind for the *HTT* gene. This systematic screening of the *HTT* gene locus has identified 11 regions which bear genomic and chromosomal marks indicative of regulatory regions. Using available datasets I have identified putative TFBS based upon peaks in ChIP-seq datasets. These TFBS represent 44 TFs. A protein-protein interaction network for these 44 TFs was created in order to investigate groups of interacting TFs within these 11 regions. Finally, I used this TFBS interaction information as well as DNase hypersensitivity differences between brain and peripheral cell types to select a genomic region of interest to conduct follow up studies.

In order to expand our knowledge concerning the transcriptional regulation of the *HTT* gene, I endeavoured to consider regions beyond the proximal promoter region previously studied. As no enhancers, or silencers, of the *HTT* gene have ever been identified, it was impossible to base my study on previously identified enhancers for this gene. As a starting point, I decided to localize this search to the local genomic context of the *HTT* gene, namely the gene up and down stream. While the gene preceding *HTT, GRK4,* does appear to have a similar expression pattern to *HTT,* making regulatory regions within and near this gene potential co-regulators for both genes, there is no reason to assume that the local genome alone is contributing

to *HTT* regulation. At this early stage of investigation, however, it would not be feasible or wise to expand the boundaries of a search without further information. Using information from genome-wide chromosome conformation datasets (4C or 5C datasets) would allow for the identification of more distal enhancers, if any do exist. In either case I would strongly suggest that consideration of cell type be used in the generation of the selected dataset. *HTT* is ubiquitously expressed, but does have tissue specific up regulation of expression in brain and testes. In addition to differentially expressed TFs between tissue types, it is likely that tissue specific enhancers are also involved in this differential expression. By using 4C or 5C datasets representing both brain and peripheral cell or tissue types, it is possible to identify interacting regions of the genome that differ between cell and tissue types. On the same topic of tissue and cell type specificity, I did not use this criteria in the selection of ChIP-seq datasets. This was done in order to include as many TFs in my screen as possible. In the future, as more ChIP-seq datasets in a more diverse set of cell and tissue types become available, it will be possible conduct a similar analysis using datasets differing by cell type for a single TF.

I have conducted an extensive search into potential TFs and TFBS within my regions of interest, but it is important to note that there are many more TFs than those represented by my selected datasets. It is entirely possible that as yet untested TFs play a large role in *HTT* gene regulation. Until these TFs have been used in a ChIP-seq study, it will be difficult to determine their role in *HTT* gene regulation.

While I have conducted an extensive search into potential TFs and TFBS within my regions of interest it is important to note that there are more TFs than those represented by my selected datasets. It is entirely possible that untested TFs play a large role in *HTT* gene regulation. Until these TFs have been used in a ChIP-seq study it will be difficult to determine their role in *HTT* gene regulation.

3 Generation of Stably Expressing Cell lines & *HTT* Expression Assay

3.1 Introduction

By systematically investigating the TFs of putative TFBS within a regulatory region, we can gain important insights into the pathways that control the expression of the gene of interest. Generating an assay that is both reproducible and quantifiable is a powerful tool towards this end as it would allow researchers to thoroughly investigate each TF in turn. The validity of this assay, however, relies heavily on the optimization steps taken through its generation to ensure the correct interpretation of the results. To investigate potential regulatory regions of *HTT* transcription I generated an *in vitro* based cellular assay using a reporter gene as a surrogate of *HTT* expression. This assay was validated according to the parameters used to optimize an assay for use in a high-throughput setting to ensure the reproducibility of the assay.

3.1.1 Reporter Gene Assay

In order to study the functionality of a putative regulatory region it is often necessary to utilize a reporter gene instead of monitoring the endogenous gene transcript. This allows the researcher to study the promoter region in isolation as well as simplifying analysis. This strategy does, of course, come with the drawback of

removing the promoter from its endogenous genomic setting. It does, however, allow for additional putative regulatory regions to be tested in individually, thereby allowing researchers to accurately identity active TFBSs in these regions. Reporter genes are often selected based upon the ability to simply and quickly quantify protein concentrations. Ideal reporter genes are those that are not found endogenously in the genome being studied, are small enough as to take up little space on recombinant plasmids, and produce a gene product that can be easily assessed and quantified. Three such reporter genes commonly used in these studies are the *E. coli lacZ* gene, the jellyfish GFP gene, and the firefly luciferase gene. All three of these genes are quantifiable using relatively simply assays as compared to direct quantification of the endogenous gene product. The selection of an appropriate reporter gene is typically influenced by the downstream applications of the assay, for the example, in my studies, I wished to assay cellular viability alongside the report gene levels so it was important to make sure that the chemistries of both assays were compatible.

As mentioned in the introductory chapter, a prior study utilizing *HTT* promoter-gene reporter constructs had been conducted by our group. The reporter gene used in that study was the firefly luciferase gene (Luc), I continued to use this reporter gene as I found it to be highly reproducible and the assay chemistry was compatible with the downstream viability assay I had selected. I also opted to retain the same *HTT* promoter region that was used in the previous study, a ~3.7 kb region upstream of the ATG start site as defined by the UCSC genome browser. This region was isolated from the HD allele of an HD patient and represents the most common

haplogroup found in the HD population (Becanovic et al., in review; Warby et al. 2009). A reminder to the reader that previous research conducted by our group found no differences in promoter function between the most common haplogroups (Becanovic et al., in review). Our constructs are unique in the field of *HTT* promoter function studies as the ATG translational start site of the luciferase gene replaces the ATG translational start site of what would be the *HTT* gene. My constructs carry the entire 5' untranslated region (UTR) of the endogenous *HTT* gene, thus allowing the constructs to be regulated more like the endogenous gene than other constructs omitting or only partially including the 5' UTR.

Once a reporter gene has been selected and the recombinant plasmids generated the next important consideration is the mode of transfection of the plasmid into the cell. Simply transfecting the plasmid into cells and quantifying the expression of the reporter gene needs to be controlled for as transfection efficiencies can vary from experiment to experiment. The variance in transcription efficiency results in differential copies of the plasmid entering and being translated by the cell which can confound results. Transfection efficiency is affected by plasmid size, making comparisons between different sized plasmids difficult. To avoid this problem many studies have used a dual transfection paradigm, utilizing a secondary reporter gene on an additional plasmid that is co-transfected into the cells and used to asses transfection efficiency and normalize the quantity of reporter gene from the primary plasmid (Allard & Kopish 2008). The copies of this secondary plasmid are not assumed to be present in equal amounts to the primary plasmid. In fact, the concentration of secondary plasmid compared to the primary one is much lower.

This is because the secondary plasmid is used solely as an indication of transfection efficiency for the secondary plasmid, the efficiency of the primary plasmid is then inferred from that of the secondary. While this technique of normalization is capable of generating estimates of efficiency between transfections of either different primary plasmids or transfections occurring on different days, the efficiency calculated will always be that of the surrogate secondary plasmid which is often different in size to that of the primary plasmid.

3.1.2 Stably Expressing Cell Lines Using 'FLP-In' System

To remove the additional source of variability occurring from transient transfection I opted to generate cell lines that stably express our constructs, thereby removing the need to transfect cells for each experiment. In order to be able to compare multiple constructs with differing regulatory regions I opted for a site directed integration method versus a random integration method which would have resulted in variable copies of each construct being integrated randomly in the genome. Figure 3.1 provides a schematic of the FLP-In System™ (Invitrogen) which I modified for my study. This system utilizes a site-specific recombinase, FLP derived from yeast, along with antibiotic resistance and *lacZ* expression to ensure single, site directed integration of the constructs of interest. This has two advantages, firstly all of the cells used will carry a single copy of the selected construct, reducing variability between experiments. Secondly, and as will be further explained below, the site directed nature of the system ensures that cell lines derived to express different constructs will always have construct integration in the same genomic location.

To briefly explain the mechanics of the FLP-InSystem™, firstly a FLP recombinase recognition site, FRT site, is randomly integrated into the genome of the cell line of interest. This is done through a plasmid that carries a single FRT site after the promoter and ATG start site of a combined *Zeocin* resistance and *lacZ* gene. Zeocin exposure selects cells in which integration has occurred and allows for clonal selection. Clones are expanded into cell lines and are doubled checked for FRT integration with *lacZ* expression with a β-glactosidase assay. These Zeocin resistant and *lacZ* expressing lines are then screened using a Southern Blot to identify those with a single FRT integration. Single integration is imperative as it allows the downstream constructs to all be integrated in the same genomic location and prevents genomic instability as multiple FRT sites within the genome could potentially recombine during insertion of the target construct.

Once a cell line with a single FRT site has been selected, this parent cell line is then used in a site-directed integration using dual transfection of two plasmids; the first carrying both an additional FRT site and the regulatory region-luciferase construct to be integrated, and the second carrying only the gene for FLP recombinase. The FRT site on the secondary plasmid is placed in front of a promoter-less and ATG-less *hygromycin* resistance gene, allowing FLP recombinase to performed a site-directed insertion of the *hygromycin* resistance gene between the genomic *zeocin-lacZ* gene and its promoter and ATG site. This pushes the *zeocin-lacZ* gene out of frame with the promoter and ATG site, and sets the previously promoterless and ATG-less hygromycin gene in frame ready to be transcribed (see Figure 3.1). Cells with successful integration are then selected

through hygromycin resistance and further tested through zeocin sensitivity and lack of lacZ expression. The FLP recombinase causes a site directed integration of between the FRT site within the genome and the FRT site on the transfected plasmid by making a double stranded break in the middle of the palindromic FRT sites and ligating each FRT site to each other (Zhu & Sadowski 1995). This form of targeted integration results in two FRT sites in the genome, because of this the FLP recombinase used has been modified to have a shorter half life to prevent the integrated construct from being excised.

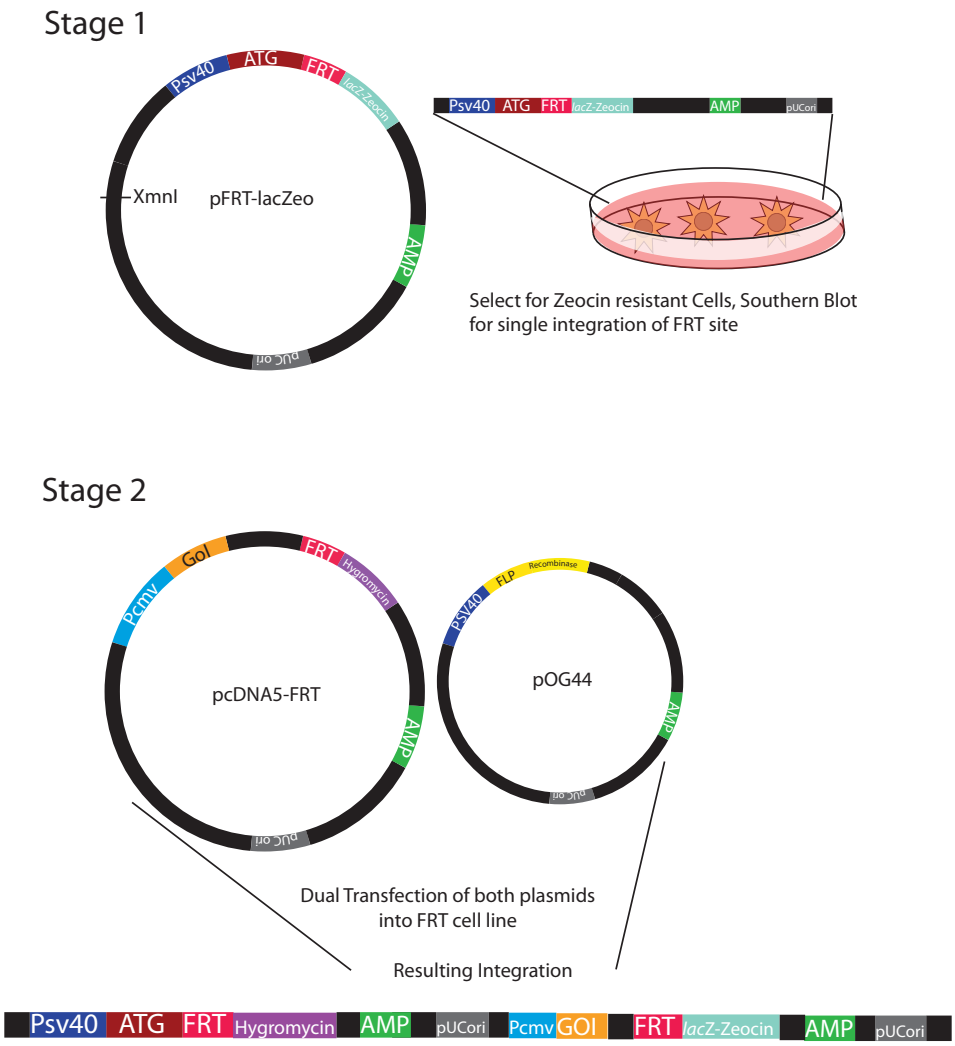


Figure 3.1 FLP-In System™
Schematic of FLP-In System™ (Invitrogen) Stage 1: Integration of FRT site, the pFRT-lacZeo plasmid is linearized using the restriction enzyme XmnI, linearized plasmid is then transfected into the cell line of interest and random integration of the construct occurs. Selection with zeocin allows for zeocin resistant colonies to be isolated, these are propagated into cell lines and tested for B-galatosidase expression and singe integration of the FRT site using Southern Blot. Stage 2: once a parental FRT cell line is selected a dual transfection of the pcDNA5-FRT plasmid (with the gene of interest (GOI)) and the pOG44 plasmid (which carries a FLP Recombinase gene) allows for site directed integration at the single FRT site in the parental cell line. This integration pushes the ATG-less lacZ-Zeocin gene out of frame with the Psv40 promoter and ATG start site and puts the ATG-less hygromycin gene in frame, allowing for selection of hygromycin resistant cells. Figure modified from Invitrogen.

3.1.3 *HTT*-Reporter Gene Constructs

The *HTT* promoter constructs previously generated in our lab represent the longest portion of the *HTT* proximal promoter studied thus far. Despite this, these constructs do not encompass the entirety of what we would now call the *HTT* promoter based on the results of my bioinfomatic assessment (chapter 2), namely what I call Region 6. Figure 3.2 gives a comparison of the location of Region 6 compared to the location and size of our *HTT* promoter construct. Most importantly my constructs end at the ATG translational start site of the *HTT* gene and omit exon 1 of *HTT* where the CAG expansion lies. It is unknown what the effect of CAG size, namely pathogenic verses non-pathogenic CAG size, could have on endogenous *HTT* gene transcription. Region 6 extends to include almost all of exon 1 so it is highly likely that additional transcriptionally relevant factors exist in this exon. I have chosen not to pursue the questions of the effect of the first exon or CAG size on promoter function at this time. It is unclear what the effect on luciferase function the inclusion of the first exon would have. In order for the luciferase assay to function the luciferase protein must catatylze the conversion of luciferin to luciferyl adenylate. The addition of the first exon may prevent the luciferase protein from performing optimally and would be interpreted as a decrease in function. Likewise, it is unclear what the effect of an expanded CAG, which would create an expanded polyglutamine tract, would have on luciferase function. It would be plausible that the larger CAG repeat would impede the function of the luciferase gene causing a

difference in luciferase quantity that is not reflective of the effect of the repeat on the transcription of the gene.

In addition to the *HTT* promoter construct, I have generated an additional construct in order to study the function of a putative regulatory region - Region 9 from my bioinfomatics assessment (chapter 2). Using the sequence for Region 9 from the UCSC Genome browser I constructed an additional FRT plasmid with the Region 9 sequence in front of my *HTT* promoter-luciferase construct. I opted for this arrangement because, in general, the orientation of enhancers does not affect their functionality. This orientation also preserves the position of the 5' UTR being in frame with the ATG start site of luciferase.

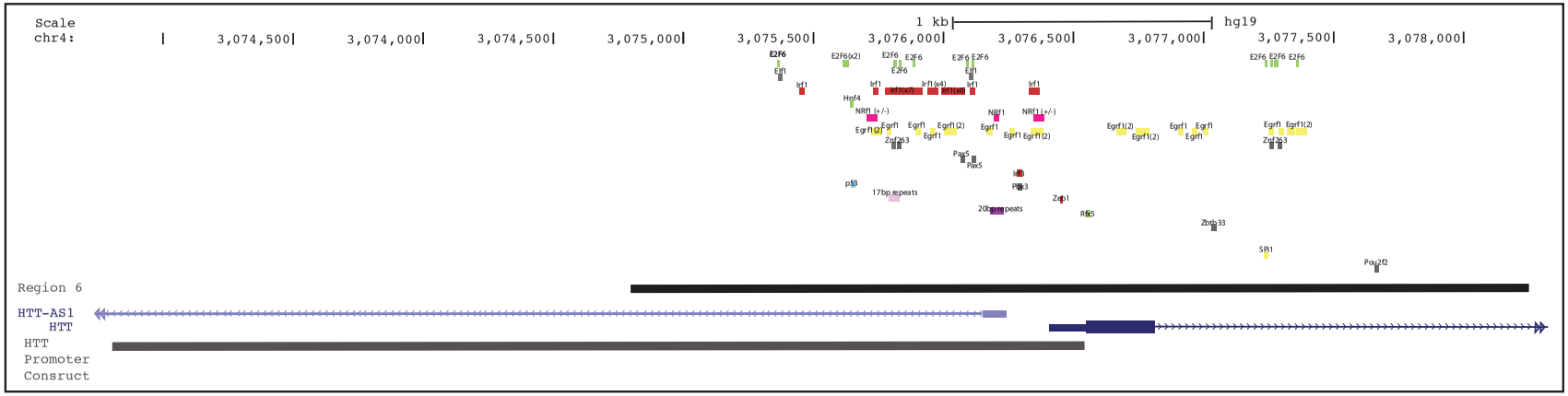


Figure 3.2 HTT Promoter Construct in Relation to Region 6 & Identified TFBS
Representation of the *HTT* promoter construct in relation to region 6 (identified in chapter 2 bioinfomatic assessment), the *HTT* gene itself, the *HTT* Antisense gene, and the identified TFBS (chapter 2). Figure is to scale. Modified from the UCSC Genome Browser.

3.1.4 Cell Lines

I selected two cell lines in which to generate my stably expressing cell lines, Human Embryonic Kidney 293 cells (HEK293) and a rat medium spiny neuronal progenitor cell line (ST14). Both cell lines have been used in our lab in the aforementioned *HTT* promote studies and have proven to be both robust and reliable in culture. Both cell lines are immortalized, with the ST14 cell line using a temperature sensitive SV40 large-T antigen to achieve immortalization allowing the cells to be differentiated allowing study in both proliferating and differentiated cell types (Hovakimyan et al. 2008; Ehrlich et al. 2001). The HEK293 cell lines allowed us to study *HTT* promoter function under the control of human TFs while the ST14 cell line allowed us to study *HTT* promoter function in a neuronal cell type.

3.2 Methods

3.2.1 Cell Culture

I purchased a previously derived HEK293 single FRT integration cell line (FLP-In™-293 cell line) from Invitrogen. The exact chromosomal location of integration is unknown but appears to be in a transcriptionally open region of the genome as evidenced by the strong lacZ expression by β-galatosidase assay. The HEK293-FRT cells were maintained on DMEM medium supplemented with 10% FBS, 2mM L-glut and 100ug/mL Zeocin. Cells were kept in a cell culture incubator maintained at 37 degrees C and with a supply of 5% CO_2. The ST14 cell lines used were a gift from the Cattenao lab. These cells were maintained on DMEM medium

supplemented with 10% FBS and 2% L-glut. Cells were kept in a cell culture incubator maintained at 33 degrees C with a supply of 5% CO_2. As the stably expressing cell lines were developed they were maintained on the appropriate antibiotic as determined by the stage of the FLP-In System™ (Invitrogen). The exception for this being when cells were transfected, as antibiotics can interfere with transfection the cells were seeded in antibiotic free medium when being used in transfection.

3.2.2 Creation of FLP-Plasmids

Table 3.1 provides details for all the plasmids utilized in this thesis. The pFRT-lacZeo, pcDNA5-FRT, and POG44 plasmids were all included in the FLP-In kit from Invitrogen. The pFRT-lacZeo construct is used to randomly integrate a FRT site in the genome of the cell line of interest as displayed in Figure 3.1. The pcDNA5-FRT plasmid carries the promoter-less and ATG-less hygromycin resistance gene with the FRT in front, this is the plasmid that is inserted in a site directed manner in the second stage of the FLP-In system. The pOG44 plasmid carries FLP recombinase. The previously generated *HTT* promoter-lucifease plasmid was created by Kristina Becanovic in our lab. It carries a portion of the *HTT* promoter 3.7 kb upstream of the translational start site. The sequence for this *HTT* promoter was isolated from an HD patient and represents haplotype A, the most common on the HD allele. This vector was created on the PGL3 vector backbone from Promega. The ARE construct was a gift from the Wasserman lab and is an antioxidant responsive element (ARE) that modifies the expression of a minimally expressing SV40 promoter also on the PGL3

vector backbone. This construct has been shown to have increased luciferase expression when cells are exposed to various forms of cellular stress (Wasserman & Fahl 1997). The promoter-less vector is the PGL3 vector backbone with the luciferase gene having no promoter to drive its expression. The Region 9 construct was generated by taking the given sequence from the UCSC genome browser for Region 9, ordering a custom made plasmid from IDT bearing the sequence with a restriction enzyme site on either end of the sequence. The region 9 sequence was restriction digested out of the custom plasmid and inserted into the full length *HTT* promoter construct which was also digested with the same restriction enzymes. Plasmids were propagated on a using an electrocompetent strain of DH5a e.coli and collected using either a mini, midi or maxi prep kit (Invitrogen). Restriction digests were carried out using restriction enzymes purchased from NEB. Agrose gel extractions of restriction digested fragments was performed using the Gel extraction kit from NucleoSpin® Gel and PCR Clean-up (Machery-Nagel). Plasmid ligations were performed using T4 DNA ligase (NEB).

Table 3.1 Plasmid Description

Plasmid	Description
***HTT* Promoter**	Proximal *HTT* promoter (as illustrated in figure 3.2) in frame withthe luciferase gene of the PGL3 plasmid (Promega)
Region 9	*HTT* promoter plasmid modified to include Region 9 sequence upstream of *HTT* promoter sequence
ARE	A generous gift from the Wasserman Group {Wasserman:1997wz}
Promoter-less	From Promega, PGL3 plasmid without any promoter sequence upstream of the luciferase gene
pFRT-lacZeo	From Invitrogen, carries a SV40 promoter infront of a ATG site-FRT site sequence followed by a fusion ATG-less LacZ-Zeocin gene
pcDNA5-FRT	From Invitrogen, carries a SV40 promoter infront of a multiple cloning site (where the user can insert their gene of interest), followed by a FRT site and a ATG-less hygromycin gene
pOG44	From Invitrogen, carries a modified FLP-recombinase gene driven by a SV40 promoter

3.2.3 Antibiotic Resistance Kill Curves

In order to perform antibiotic selection necessary for the generation of antibiotic resistant colonies each parent cell line was screened for the lowest dose of antibiotic necessary to kill all the cells after two weeks. The selection of an appropriate dose is essential as the addition of antibiotic occurs shortly after the expected integration of the FRT constructs. Too high of a dose, and cells that do have integration may not be able to produce enough of the appropriate resistance gene to survive the initial addition of antibiotic This could also favour multiple integrations over single ones as multiple integrants will produce more antibiotic resistance gene product. Too low a dose, and cells that do not have integration could survive selection and colonies would be incorrectly isolated for cell line expansion.

To perform a kill curve for FLP-In™-293 cells, cells were plated in a 6 well plate in media without zeocin so as to have a confluence of ~25% after 24 h of incubation. After 24 h the media was removed and replaced with media dosed with various concentrations of hygromycin antibiotic. Media was changed with fresh hygromycin dosed media every 3-4 days for 2 weeks. Pictures were taken for comparison at every media change using an inverted microscope. To perform a kill curve for ST14 cells the same procedure as above was conducted with the use of either zeocin or hygromycin media as pertained to the step in the FLP-In™ system being conducted.

3.2.4 Transfection, Isolation & Propagation of FRT or Stably Expressing Cell Lines

3.2.4.1 Transfection

Transfection of both HEK293 and ST14 cell lines was performed using TransIT®-LT1 reagent (Mirus). A ratio of 1 ul LT1 : 3.76 ug DNA was used in both cell types. As the amount of DNA required for transfection was greater than the amount of DNA necessary for FRT integration the balance of DNA was filled using sheered salmon sperm DNA. HEK cells were seeded in six well culture dishes for a density of ~80% at the time of transfection, in antibiotic free medium 24 h before transfection. The transfection medium was created in DMEM supplemented with 2% L-glut only the day of the transfection and added drop wise to the culture dishes. For the generation of ST14-FLP cells 150 ng of pcFRT/lacZeo lineralized using restriction enzyme XmnI (NEB) was used with 2350 ng of sheered salmon sperm DNA (NEB). For the generation of either HEK or ST14 stably expressing cell lines both the pOG44 plasmid and the construct to be stably expressed were used in a ratio of 9 pOg44 : 1 selected construct. 48 h after transfection the cells were trypsinized and passaged to a 10cm culture dish, again in media without antibiotic. The appropriate antibiotic was added at the concentration determined as described above 24 h after passaging. Cells were selected with antibiotic for 2 weeks, changing media every 3-4 days.

3.2.4.1 Isolation and Propagation of Cell Lines

After two weeks of antibiotic selection small colonies of antibiotic resistant cells developed. These colonies were isolated using glass cloning cylinders (Corning®) and Dow Corning® 976V silicone high vacuum grease. to allow individual trypsinization and passaging of each colony into a 24 well plate. Colonies were allowed to grow to fill the 24 well before being passaged to a 6 well plate, at this time the passage number was set to 1 as this was the first passage since the colonies were derived from the parental cell line. As the newly generated lines expanded they were eventually passaged to T75 flasks and a small sample was taken for β-galactosidase testing. When generating ST14-FRT cell lines only cell lines that were β-galactosidase positive at this point were retained. For the generation of ST14-Stable construct and HEK293-Stable construct lines only cell lines that were B-galactosidase negative were retained. Cell lines that successfully passed β-galactosidase testing were then cropyreserved in freezing medium, DMEM with 10% DMSO. When reviving cells from cryopervation cells were maintained for 24 h on media without antibiotic to reduce cellular stress, after which antibiotic media was added.

3.2.5 β-galactosidase Staining

β-galactosidase staining was perfumed using an existing protocol. In short, a sample of cells was plated in a 6 well plate, 24 h after plating the cells were fixed for

5 min at 4 degrees using fixative solution. The fixative was then removed and a staining solution, with X-gal, was added before the cells were incubated overnight in a 37 degree incubator. The presence of blue stained cells indicated β-galactosidase expression while unstained cells indicated inactive galactosidase.

3.2.6 Southern Blot

Genomic DNA was extracted using an existing phenol chorophorm protocol (Green & Sambrook 2012) . Southern blotting was performed using the DIG-High Prime DNA Labeling and Detection Starter Kit II (Roche). A DNA probe with a sequence complementary to the zeocin-lacZ fusion gene on the pFRT-lacZeo construct was DIG labeled using the Roche DIG Oligonucleotide 5'-End Labeling Set (Roche). gDNA was digested with the restriction enzyme HindIII as this enzyme has a cut site near the probe hybridization sequence and is known to have cut sites dispersed throughout the genome (NEB).

3.2.7 Cell Viability and Luciferase Expression Assay

Cellular viability and luciferase expression was assayed using the ONE-Glo™ + Tox Luciferase Reporter and Cell Viability Assay kit (Promega). This is a two-step assay system in which viability is first assessed using the Celltiter-Fluor reagent.Following this luciferase expression is assessed using the ONE-Glow reagent. The Celltiter-Fluor reagent is cell permeable and is cleaved by 3 proprietary proteases only expressed and active in intact, live cells. Upon cleavage the Celltiter-

Fluor reagent emits a floraphore detectible at 380–400nm$_{Ex}$/505nm$_{Em}$. The ONE-Glow reagent is a one step luciferase assay that both lyses the cells and provides the luciferin substrate necessary for the luciferase reaction. Cells were plated in white walled, clear bottom, 96 well plates from Costar. Wells in the outermost rows and columns were omitted to avoid edge effects. Three wells of the plate were designated as blank control wells to which only media without cells was added. 30 min before reading the Celltiter-Fluor reagent was added in a 1X concentration and incubated at 37^{o}C in the incubator until the plate was to be read. Follow the reading fluorescence the ONE-GLow reagent was added and incubated for 3 min at room temperature with shaking before being read. Both the fluorescent signal from the Celltiter-Fluor reagent and the luminance signal from the ONE-Glow reagent were read using the Omega Polar Star plate reader (Thermo Scientific).

3.3 Results

3.3.1 Creation of FRT Plasmids for FLP-In System

To use the *HTT* promoter construct, the ARE construct, the promoter-less construct, and the Region 9 construct in the FLP-In system it was necessary to move the various regulatory region-luciferase gene portions of the original plasmids to the pcDNA5-FRT plasmid. The intended use of the FLP-in system is to generate cell lines that constitutively express a protein of interest, as such the pcDNA5-FRT plasmid contains multiple cloning site downstream of a viral CMV promoter. In order

to use these plasmids and this system for my study, I first had to modify the pcDNA5-FRT plasmid to remove the CMV promoter and then insert the regulatory region-luciferase segments. In my first attempt to accomplish this I used a cut site in the backbone of the PGL3 vector just outside of the multiple cloning site where the regulatory elements in all of the constructs were inserted and a cut site just after the luciferase gene to excise out the regulatory region-luciferase gene portions of each original construct. I then inserted this into a pcDNA5-FRT plasmid that was also restriction digested in order to remove the CMV promoter on the pcDNA5-FRT plasmid. The new vectors are designated by the regulatory region they carried followed by -FRT. Primers were created on either side of each ligation site in the new vectors, and sequencing was performed to ensure the integrity of the ligations and to double check the sequence of the regulatory region-luciferase constructs inserted.

When these FRT vectors were tested for luciferase expression using a transient transfection no appreciable expression was found as compared to the original construct (data not shown). Further research into the original PGL3 vector backbone revealed the presence of a 'synthetic poly A pause site' which Promega claims is present on the vector backbone to prevent read through from other genes on the plasmid from interfering with the inserted regulatory region. They do not suggest that this synthetic poly A pause site is essential for luciferase reporter gene function, however a careful perusal of another study that attempted to excise the luciferase gene from the PGL3 vector suggests that these researchers also ran into similar issues which were later rectified by the inclusion of the synthetic poly A pause site (Yan et al. 2004). The PGL3 backbone vector was designed with as few

restriction enzyme cut sites outside of the multiple cloning site to make it easier for users to clone in their regulatory region of interest. This meant a restriction enzyme upstream of the synthetic poly A pause site was not present that would allow me to excise our regulatory region-luciferase constructs with the synthetic polyA pause site in one piece. So to make new FRT plasmids I excised the FRT-hygromycin sequence from the pcDNA5-FRT plasmid and inserted it after the luciferase gene on the original PGL3 plasmids. New primers were created and the ligation sites sequenced to again ensure the integrity of the plasmids. Figure 3.3 shows that the expression from these new FRT constructs is comparable to that of the original constructs as shown through luciferase expression in transient transfection.

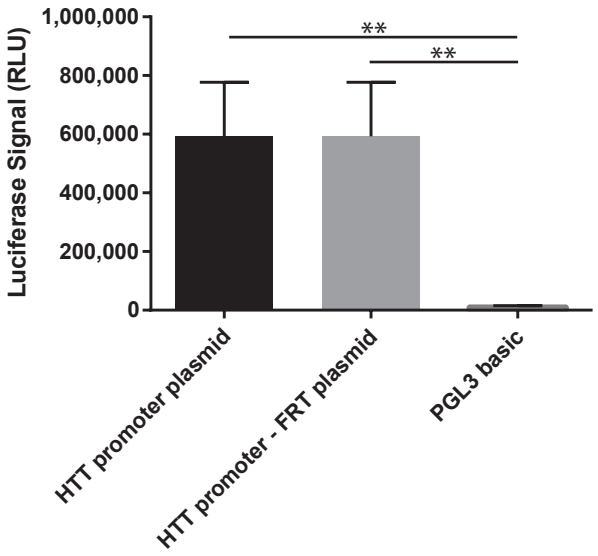


Figure 3.3 Transient Transfection of Original HTT Promoter Plasmid and HTT Promoter -FRT Plasmid

Using transient transfection of FLP-In™-293 cells the luciferase activity of both the original *HTT* promoter plasmid and the modified *HTT* promoter - FRT plasmid was compared to the PGL3 basic plasmid. One-way ANOVA **$p<0.01$

3.3.2 Generation of HEK293-FLP-In cells

Once FRT versions of the original plasmids was obtained, I was able to begin the generation of FLP-In™-293 cells for each of FRT constructs. I first conducted a kill curve assay as described in the methods section, (data not shown). A dose of 125*μ*g/mL was selected from this kill curve for the selection of hygromycin resistant cells. As specified in the methods section, cell lines were generated for each FRT construct. Since I conducted a clonal isolation from the transfection of each construct I was able to generate several cell lines for each construct. I then selected a line, at random, from each construct to move forward with for optimization and for use in subsequent experiments. Performing a secondary clonal isolation in the second step of the FLP-In™ system is a departure from the FLP-In™ literature as theoretically the only cells that should survive the hygromycin selection step should have site directed insertion of the FRT plasmid at the genomic FRT site. I found that this is not the case. In several instances, I found β-galatosidase positive colonies after hygromycin selection. While this could indicate that the FRT plasmid had randomly inserted into the genome allowing the *zeocin-lacZ* fusion gene to be expressed these cells should still be hygromycin sensitive as the *hygromycin* gene utilized lacks a promoter and a ATG start site.

3.3.3 Optimization of HEK293 Cells for Multiplexed Luciferase and Viability Assay

In order to show that the cell viability assay, Celltiter-Fluor, accurately reflects cell density and viability, I performed two validation experiments. In the first, I plated different densities of HEK293-full length *HTT* promoter cells and quantified the cells shortly after they had adhered to the plate using the Celltiter-Fluor assay (Figure 3.4A). The fluorescent signal was positively correlated with increases in cell density indicating that the assay reflects changes in cell density. In the second experiment, I seeded a consistent density of cells, 20,000 cell/well, and then treated the cells with a cytotoxic agent, digitonin, at increasing concentrations, (Figure 3.4B). In this case the florescent signal was inversely correlated with increases in digitonin indicating that as digitionin exposure increased, less viable cells remained.

The validity of these results relies on the correct optimization of my assay. To ensure the reproducibility of my assay over many repetitions I opted to treat the assay as if it was to be used in a high-throughput screen (HTS). To facilitate the optimization I conducted a Z' scoring experiment (Zhang et al. 1999). The Z' score uses a comparison between positive and negative controls to assess both the sensitivity and variability of an assay. A Z' score with a value $1 > Z' \geq 0.5$ indicates an assay is 'excellent assay' for high-throughput screening. By ensuring that my assay was optimized for HTS, I could be confident that my assay set up was robust, reproducible, and would give reliable results. Being optimized for HTS would also allow me to conduct a drug or small molecule screen in the future looking for potential drug compounds or targets that could be used to modulate *HTT* transcription. To accomplish this optimization, in two assays conducted on separate

days, I seeded three 96 well plates with 20,000 cells/well from the *HTT* promoter expressing cell line. The following day I dosed two plates with 1 ug/mL of Actinomycin D (ActD), a general transcriptional inhibitor, and the other plate with DMSO alone. After 24 h of treatment, the cells were assayed using the multiplexed Celltiter-Fluor and ONE-Glow assay (Figure 3.5). The Z' scores for these experiments were 0.709 and 0.756. While the Z' scores between days are not identical, indicating slight day to day differences, both trials had Z' within the required parameters indicating that my assay is appropriate for HTS.

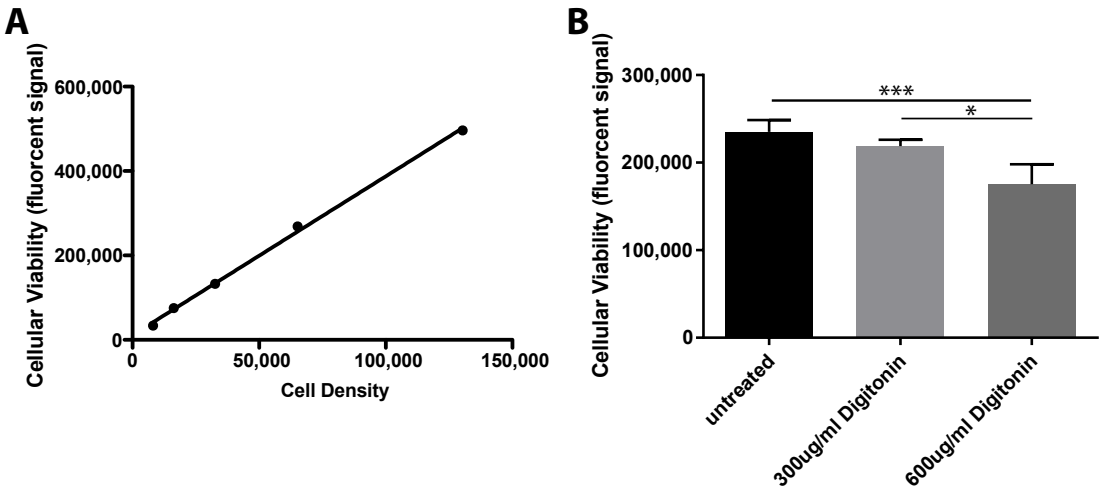


Figure 3.4 Cell Density & Cytoxicity Correlate With Cell Viability Assay
(A) Increasing densities of cells positively correlate with cell viability signal (as indicated by fluorescent signal from the assay), linear regression R= 0.998 (B) Cells were treated for 30 min with either 300 ug/mL or 600 ug/mL of Digitonin, One-way ANOVA, *p<0.05, ***p<0.001

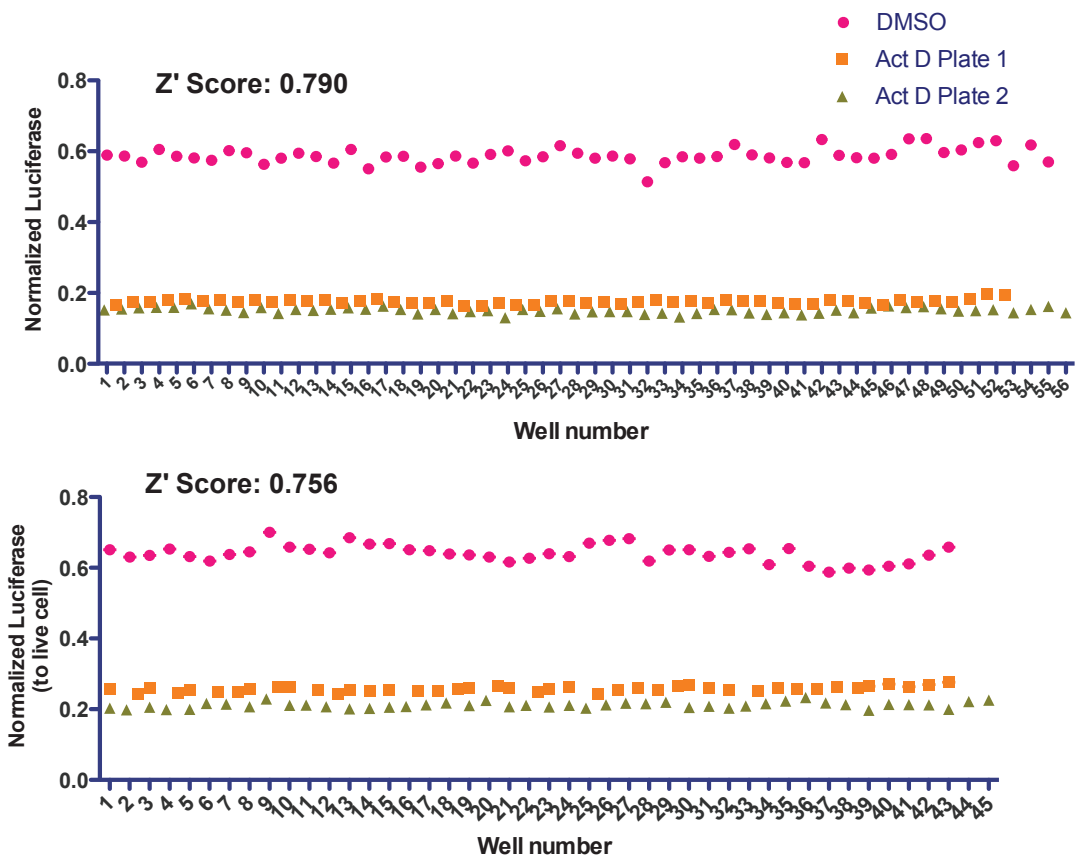


Figure 3.5 Z' Test

Z' Test performed as follows: three separate 96 well plates were seeded with 20,000 HEK293 *HTT* promoter cells, 24 h after seeding two of the plates were treated with 1 ug/mL ActD, 24 h after treatment cells were assayed using the cell viability and luciferase assay as described in the methods section. Each graph represents an independent Z' Test performed on different days. Z' score calculated as (Zhang et al. 1999).

3.3.4 Generation of ST14-FRT Cells

To generate a ST14-FRT cell line with a single integration of the FRT site I first performed a Zeocin kill curve on ST14 cells, (data not shown). Based on this kill curve a dose of 100 ug/mL was selected for zeocin election of FRT integrants. After restriction digest with the restriction enzyme XmnI several concentrations of lineraized pFRT-lacZeo were used in the random integration step of the FLP-In™ system. Transfection and isolation of colonies was conducted as written in the methods section. Cell lines that were β-galatosidase positive were kept for southern blot analysis (data not shown). In total from the 40 β-galatosidase positive colonies isolated 2 were found to have single integration of the FRT construct, of which one was randomly chosen as the parent FRT cell line to create ST14 stably expressing cell lines.

3.3.5 Generation of ST14 Stably Expressing Cell Lines

As with the generation of the FLP-In™-293 cells, a kill curve using the varying doses of hygromicin was performed, (data not shown). A dose of 125ug/mL was selected and isolation and propagation of cell lines occurred as described in the methods section. β-galatosidase staining of isolated colonies after hygromycin selection also revealed several colonies that were β-galatosidase positive while being hygromycin resistant as was seen in the HEK293-FLP-In cells. Only colonies negative for β-galatosidase staining were kept. These cells have been tested for luciferase expression and have been tested in a density assay with the Celltiter-

Fluor assay (Figure 3.6). At the present time these cell lines need to be further optimized, as was done with the HEK293-FLP-In cells, before they can be used for further experiments.

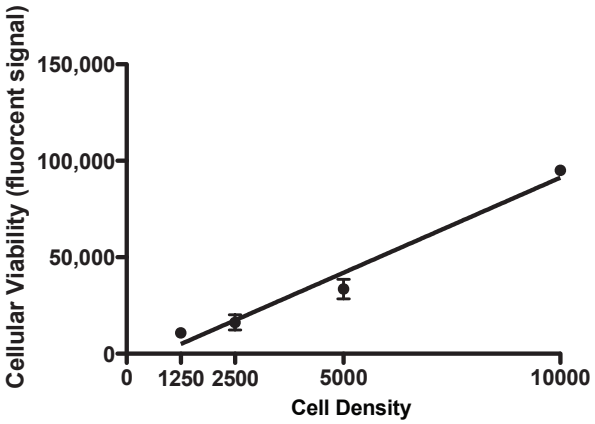


Figure 3.6 ST14 Cell Density Test
ST14 *HTT* promoter cells were seeded at increasing densities, 24 h after seeding cells were assayed using the cell viability signal (as indicated by fluorescent signal) as described in the methods section. Linear regression, R=0.95

3.4 Discussion

I have successfully generated a panel of stable cell lines expressing regulatory region-luciferase constructs and control constructs in both a human peripheral tissue cell line (HEK293) and a rat striatial neuronal cell line (ST14). I have also successfully optimized the use of the HEK293 cell lines in a multiplexed assay such that they could be used for HTS. The ST14 cell lines that have been generated require further optimization before being used in subsequent experiments, however they may be a valuable resource for future studies. ST14 cells can be differentiated from dividing progenitor neuronal cells to differentiated cells allowing for studies in both cell types to be conducted. The neuronal nature of these cells, while not human, also provides a cellular environment more akin to the endogenous neuronal setting in which *HTT* expression is highest.

Throughout the generation of these stably expressing cell lines using the FLP-In system, it has been evident that the comparatively straight-forward antibiotic selection method suggested by Invitrogen is not sufficient enough to ensure the homogeneity of subsequent cell lines. In both the FLP-In™-293 and ST14-FRT cell lines, where the *LacZeo* fusion gene should have been displaced from its original promoter and ATG start site due to the FLP recombinase driven insertion of constructs, I found hygromycin resistant, β-galatosidase positive colonies. It is conceivable, though unlikely, that the hygromycin carrying plasmids could have randomly integrated into the genome in a FLP recombinase independent fashion. These cells, however, should not have been able to survive hygromycin exposure as

the *hygromycin* gene carried on the plasmids is promoter and ATG-less meaning a functional hygromycin transcript could not have been transcribed. Conversations with Invitrogen technicians have suggested that read-through of the inserted gene of interest may cause low levels of the *LacZeo* gene to be expressed. This answer is unsatisfactory as, as noted above, FLP recombinase driven insertion separates the *LacZeo* fusion gene from a promoter and an ATG start site. This means that, even if read-through occurred, the resulting transcript would not have an ATG start site from which a functional protein could be translated. At present the protocol included with the FLP-In™ System promotes a pooling of all hygromycin resistant colonies to generate a single cell line. Based on my results, this would generate a cell line of mixed population, creating an opportunity for the random selection of cells during multiple passages which could lead to over-representation of β-galatosidase positive cells. As the biological mechanism behind the existence of these hygromycin resistant, β-galatosidase expressing cells is unclear, I would recommend that Invitrogen update its protocol to require clonal selection and β-galatosidase testing of hygromycin resistant colonies.

4 Functional Assessment of the *HTT* Promoter & Putative Regulatory Region

4.1 Introduction

4.1.1 Transcription Factors of Interest

At the end of chapter two I had identified 11 potential regions of transcriptional regulation, and highlighted putative TFBS in these regions based on available ChIP-seq datasets and noted groups of TFs with known interactions within these regions. As *HTT* is differentially expressed between CNS and peripheral tissues, I also assessed which regions were DNase hypersensitive based upon cell type. Based on this information Region 9 was selected as the top candidate region along with Region 6, the proximal promoter region, to study in peripherally derived HEK293 cells. From both of these regions I selected four TFs to test in the HEK293 stably expressing cells and the associated assay, described in chapter 3. These four candidate TFs are the jun B proto-oncogene (JunB), Structural maintenance of chromosomes 3 (Smc3), regulatory factor X 5 (Rfx5), and Interferon regulatory factor 1 (IRF1). I also selected three additional TFs, p53, SP1, and Nuclear Factor Of Kappa Light Polypeptide Gene Enhancer In B-Cells (NFkB) as they have been shown in previous studies to be regulators of *HTT* promoter expression. The locations of putative TFBSs based on the ChIP-seq data in chapter 2 in both Region 6 and Region 9 are presented in Figures 2.4 and 2.6. Results from the luciferase assay were then further validated through quantitative realtime PCR (qRT-PCR) and Western blot.

4.2 Methods

4.2.1 Transient Transfection

Transient transfection was performed in FLP-In™-293 cells, as these were the parent cells used to generate the stably expressing cells (chapter 2), they do not express luciferase. Cells were seeded at a density of 20,000 cells per well in white walled, clear bottom, 96 well plates (Costar). Cells were co-transfected 24 h after seeding using TransIT®-LT1 reagent (Mirus) and a ratio of 35 ng of luciferase construct to 5 ng of Renilla control plasmid. The balance of DNA in each transfection was filled with PGL3 basic plasmid that contains a promoter-less luciferase gene. Luciferase and Renilla levels were assayed using the Dual-Luciferase® Reporter Assay System (Promega) as per the manufacturers instructions and read on the Omega Polar Star plate reader (Thermo Scientific).

4.2.2 siRNA Transfection

siRNA constructs were purchased from OriGene for the following TFs: p53, NFkB1, Sp1, Smc3, JunB, Rfx5, and Irf1. For each siRNA 3 variants (A,B,C) were obtained, in addition control siRNAs were purchased: a universal scramble siRNA, HPRT, and a fluorescent siRNA. Sequences for each siRNA are provided in table 4.1. I used the TransIT-TKO® Transfection Reagent (Mirus) at a concentration of 0.5 ul in 96 well plates, 2.5 ul in 24 well plates, and 10 ul in 6 well plates. For transection I seeded my HEK293 stably expressing cells (generated in chapter 3) at a density of 20,000 cells per well in 96 well plates, 108,000 cells per well in 24 well plates, and

593,600 cells per well in 6 well plates. To identify the correct concentration of siRNA for use with this transfection reagent I performed a series of transfections in 24 well plates at the following concentrations (12.5 nM, 25 nM, 50 nM, 75 nM) using the fluorescent siRNA, (data not shown). 24 h after treatment I observed fluorescent intensity and found a concentration of 75 nM to have the strongest and most comprehensive transfection of my HEK293 cells.

Table 4.1 siRNA Constructs

siRNA Target	siRNA Construct A	siRNA Construct B	siRNA Construct C
p53	rCrCrArCrCrArUrCrCrArCrUrArCrArArCrUrArCrArUrGTG	rGrGrArUrUrUCrArUrCrUrCrUrUrGrUrArUrArUrGrArUGA	rGrGrArUrGrUrUrUrGrGrGrArGrArUrGrUrArArGrArArATG
NFkB1	rGrCrUrGrUrArUrArArGrUrUrArCrUrArGrArArArUrUrCCT	rGrGrGrCrUrArCrArCrCrGrArArGrCrArArUrUrGrArArGTG	rArGrUrArUrCrUrArGrCrArArUrCrArCrArArCrArCrUrGGC
SP1	rGrCrCrArGrUrArArCrUrUrArUrGrUrArCrArArGrGrArUGA	rCrCrArArGrGrArArArUrArArGrGrArCrArGrUrCrUrArGCT	rCrCrCrUrCrArArCrCrCrUrArUrUrCrArUrUrArGrCrArUTA
Smc3	rGrGrUrGrUrArArArGrUrUrCrArGrArArArUrArArGrGrUTA	rGrGrArArUrArGrArCrArGrCrArUrArArArCrArArArGrUGC	rUrCrCrArGrArCrArArUrUrArArGrGrArGrArUrGrCrUrCrAGC
JunB	rGrCrUrGrGrArArArCrArGrArCrUrCrGrArUrUrCrArUrATT	rCrGrArUrCrUrGrCrArCrArArGrArUrGrArArCrCrArCrGTG	rCrUrCrUrCrUrArCrArCrGrArCrUrArCrArArArCrUrCrCTG
Rfx5	rGrGrArUrGrArUrArCrArUrGrCrUrArAUrUrUrGrCrUrUAT	rGrCrArGrUrArArArCrCrArArCrUrArArUrArUrUrArUTG	rGrGrCrUrArArGrUrArUrGrArUrGrArArUrArUrArUrArGGT
IRF1	CCAGUGCAAUAAGGAAUUGAACUTT	rArGrUrUrUrCrUrArGrArGrUrGrArUrGrArArArUrGrCrUCT	rGrGrArArCrArUrGrCrUrUrArUrArUrArArArCrArUrArGTC
Universal Scramble	not specified	N/A	N/A
HPRT	not specified	N/A	N/A
Fluorescent siRNA	not specififed	N/A	N/A

4.2.3 Luciferase Assay

The luciferase assay was performed as detailed in chapter 3.2.7.

4.2.3 Quantitative Real-Time PCR (RT-qPCR)

Six well plates transfected with siRNA (as described above) were processed for RNA extraction using the protocol detailed in the PureLink® RNA Mini Kit (Invitrogen) with the following modifications: 1) Homogenization was achieved using a 21 gauge needle. 2) In order to increase RNA yield and purity, DNase was used to degrade any residual genomic DNA in the prep column, this was done using the the PureLink® DNase Set and the protocol for this set as described in the detailed users manual for the PureLink® RNA Mini Kit. The concentration and purity of RNA was assessed using a nanodrop spectrometer (Thermo Scientific). Reverse transcription was performed using the SuperScript® VILO™ cDNA Synthesis Kit (Invitrogen). Quantitative analysis of mRNA expression was performed using FastSYBR®green master mix according to the manufacture's instructions (Applied Biosystems). Amplification of cDNA was performed using the StepOne Plus Real-Time PCR System (Applied Biosytems). Primers used are provided in Table 4.2. Quantification of mRNA levels was calculated using the standard curve method using 10-fold serial dilutions comprised of a portion of each sample used in the study. Normalization of the quantified mRNA levels was accomplished using a normalization factor generated by the GeNorm program included in the qBase® software package

(Biogazelle). The normalization factor was generated for each sample using amplification of 3 normalization genes, ActB, PGK1, HPRT, done in separate well reactions. Primers used are provided in Table 4.2.

Table 4.2 RT-qPCR Primers

Transcript Target	Forward (5'-3')	Reverse (5'-3')
p53	AGACTGCCTTCCGGGTCACT	CAGAACGTTGTTTTCAGGAAGTAGTT
NfkB1	GCAGCACTACTTCTTGACCACC	TCTGCTCCTGAGCATTGACGTC
Sp1	ACGCTTCACACGTTCGGATGAG	ACGCTTCACACGTTCGGATGAG
Smc3	ATGCGTGGAAGTCACTGCTGGA	GGCAGAAAAGTAACCTCTCCAGG
JunB	TCATGACCCACGTCAGCAA	CAGAAGGCGTGTCCCTTGA
Rfx5	CACTGACACCTGTCTGCCAAAG	CCTTCGAGCTTTGATGTCAGGG
IRF1	GAGGAGGTGAAAGACCAGAGCA	TAGCATCTCGGCTGGACTTCGA
ActB	AGTACTCCGTGTGGATCGGC	GCTGATCCACATCTGCTGGA
HPRT	TTATGGACAGGACTGAACGTCTTG	GCACACAGAGGGCTACAATGTG
PGK1	CAAATGGAACACGGAGGATAAAG	CTTTACCTTCCAGGAGCTCCAA
HTT	TCCACCATGCAAGACTCACTTAG	TGGGATTTGACAAGATGAACGT

4.2.4 Western Blot

Six well plates were seeded and transfected with siRNA (as described above), following either 24 or 48 h of siRNA treatment cells were harvested using trypsinization and pelleted. To each pellet 65 ul of lysate buffer was added, with additional protein inhibitors and centrifuged at 15,000 rpm for 15 min. Following this lysis procedure a small portion of the supernatant was used for protein quantification using a Bradford Assay. For each sample, 90ug of protein was mixed with loading buffer and denatured by heating for 10 min at 70 Cº. The samples were then loaded into the wells of a 4-12% SDS gel and run at 200V for 1H at room temperature. The protein was transferred to a PVDF membrane a 25V overnight at 4 Cº. The membranes were blocked with 5% BAS-PBST and incubated overnight with primary antibody, 1:1000 dilution, with the following antibodies from Santa Cruz: NFkB p65 (sc-109); NFkB p50 (sc-7178) and p53 (sc-6243). GAPDH was used as a loading control (1:5000 from Abcam 9484). The secondary antibodies were Licor Goat-anti-rabbit (800) and Goat anti-mouse (680) (LiCor, Lincoln, NE). The membranes were imaged using a LiCor Odyssey scanner (LiCor, Lincoln, NE). Quantification was performed using the accompanying Odyssey 3.0 analytical software (LiCor, Lincoln, NE).

4.3 Results

4.3.1 Addition of Region 9 to the *HTT* Promoter Construct Does Not Affect Promoter Function in HEK293 Cells

Using the stably expressing HEK293 cells generated in chapter 3, I assayed the effect of the addition of region 9 on the basal function of our *HTT* promoter construct (Figure 4.1A). These initial results suggested that region 9 had a strong repressive effect on *HTT* promoter function. To further test these results, the two constructs were tested using transient transfection in FLP-In™-293 cells that were not stably expressing luciferase (Figure 4.1B). These results indicated that the addition of Region 9 had no effect on *HTT* promoter function, suggesting that the significant difference in expression seen in the stably expressing cells was due to the selection process involved in generating stable expression of the constructs as opposed to a direct effect on *HTT* promoter expression. I compared several of my HEK293 stably expressing cell lines and found differences in expression between cell lines expressing the same construct (Figure 4.1C). Given that these cell lines were derived using a site directed method in which all constructs were inserted into the same genomic region, these expression differences are unlikely to have arisen from differences in genomic environment. I therefore hypothesize that the differences in expression are due to the additional clonal selection process involved in the final stage of the FLP-In™ system. The differences in expression between cell lines derived by clonal selection would have been masked had I simply pooled all of the hygromycin resistant colonies as suggested by the FLP-In™ system. This could

have resulted in changes in expression between passages and freeze thaws due to random selection of cells during the cell passaging and cryopreservation processes. Changes in expression over time may have then been mis-attributed to the passage age of the cell line and not due to random changes in cell population.

Due to the inherent differences in expression between my stably expressing cell lines, it would be inappropriate to compare the viability normalized luciferase expression ratios between cells without further normalization. As such, I have presented the luciferase assay data in this chapter as percentages of the untreated condition for each cell line. This allows for comparison of treatment effects both between and within each cell line.

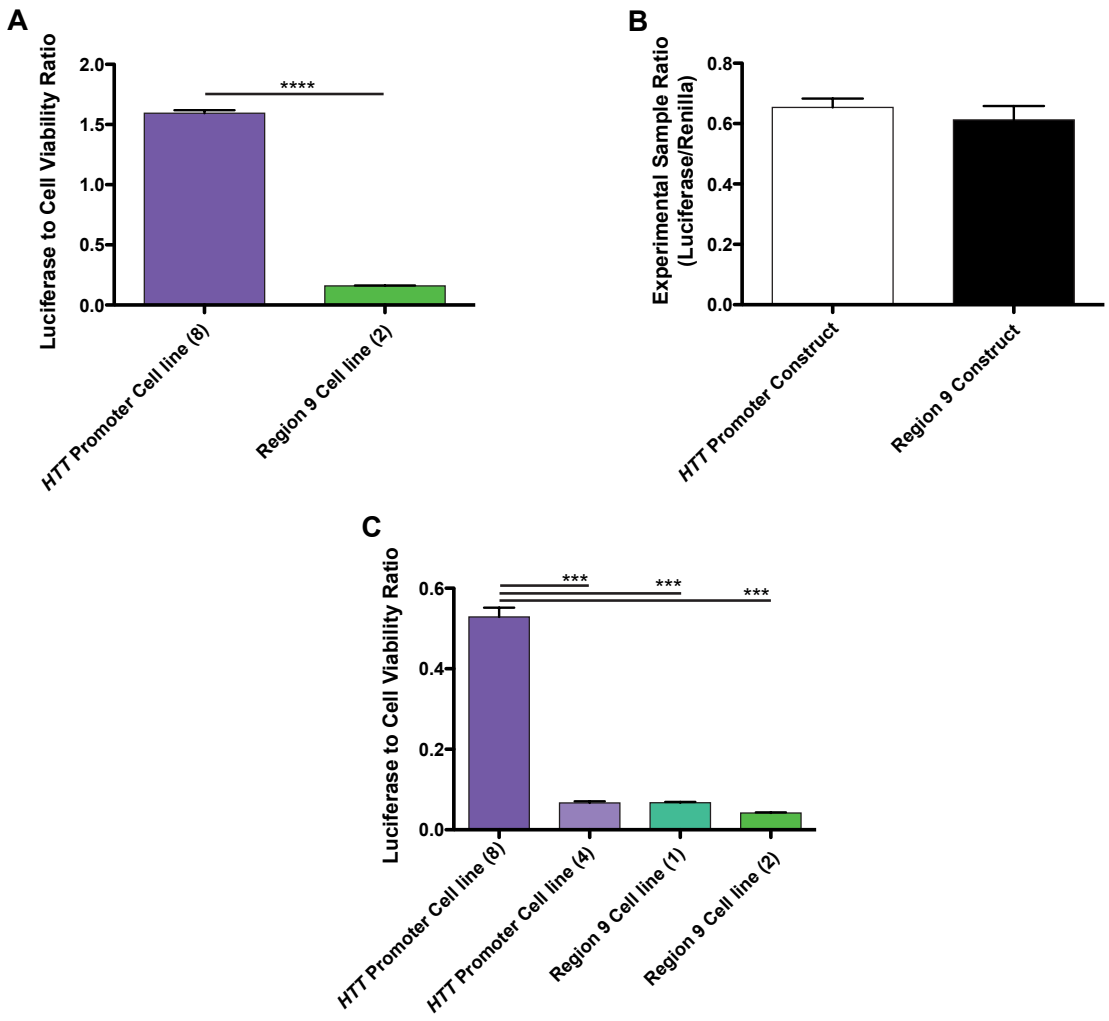


Figure 4.1 Stable and Transient Expression of HTT Promoter and Region 9 Constructs

(A) Luciferase assay of *HTT* promoter and Region 9 cell lines 24 h after seeding, n=3 per cell line, student T-test. (B) Transient transfection of *HTT* promoter and Region 9 constructs 24 h after transfection, average of two experiments n=3 per treatment per experiment, student T-Test. (C) Luciferase assay of two *HTT* promoter and two Region 9 cell lines, colony from which each line was expanded from indicated in brackets, 24 h after seeding, n=3 per cell line, One-way ANOVA with Tukey post test, ***p<0.001, ****p<0.0001

4.3.2 siRNA Knockdown of TFs Previously Identified in the Literature

As reviewed in the introduction chapter, several TFs have been previously identified in the literature as modulators of *HTT* expression, namely p53, SP1, and HDB1 & 2. These studies have focused on over-expression of these identified TFs. To to test these TFs in my system I attempted knockdown of two of these TFs, p53 and SP1. Over-expression of either TF was previously reported to increase *HTT* transcription (Ryan et al. 2006; Wang et al. 2012), I therefore hypothesized that knockdown of these TFs should result in a decrease in *HTT* promoter expression. Previous studies conducted in our lab also implicated the p50 subunit of NFkB as a regulator of *HTT* transcription. As such, I included the NFkB p50 subunit in my initial panel of TFs to test.

4.3.2.1 p53 Knockdown

Transfection protocols often recommend different collection time points based upon the target end point, namely mRNA transcript or protein. This is often attributed to the differences in half-life between mRNA and proteins of the same gene; transcripts often have shorter half-lives and are degraded quickly while proteins have longer half-lives (Vogel & Marcotte 2012). Decreases in protein levels may take longer to be observed and to have downstream effects than changes in mRNA. These are, of course, assumptions and optimal time-points for knockdown may differ between genes and knockdown methods. To establish a suitable time-point for my

siRNA knockdown experiments, I initially tested two time- points, 24 and 48 h, using the p53 siRNA constructs.

I assessed knockdown at both the transcript and protein level at each time point using RT-qPCR and Western Blot respectively (Figure 4.2). siRNA knockdown at the transcript level for p53 was observed at both 24 and 48 h. The protein level decrease in p53 were not observed until 48 h after treatment, this suggests that appreciable effect on downstream targets of p53 will not be observed until 48 h after treatment. It should be noted that in the 24 h treatment, I used single variants of each siRNA and at 48 h I used a combined pool of all three siRNAs. There is debate among scientists about the benefits and drawbacks of using single variants versus pooling siRNAs; I refer the reader to the following article for further information on the topic, (Smith 2006). In my assay I found no difference in efficacy of pooling versus using single variants using the p53 siRNAs, as well as other siRNAs targeting other TFs (data not shown). In fact I often observed slight increases in knockdown efficiency using the three variants pooled together compared to the variants alone, likely due to targeting multiple sections of the mRNA transcript in the pooled condition. As such, I have utilized pooled siRNA variants for the remainder of the experiments in this chapter.

Using the luciferase assay and RT-qPCR of the luciferase gene, I assessed the effect of p53 knockdown on the function of both the *HTT* promoter and Region 9 constructs at both the protein and mRNA level. The luciferase gene is known to have a half-life in mammalian cells of ~3-4 h, suggesting that changes in transcript level will be quickly mirrored in luciferase protein levels (Thompson et al. 1991). At 24 h I

found a small, yet significant, increase in *HTT* promoter function in both the *HTT* promoter cell line and the region 9 cell line (Figure 4.3A). These changes were likely due to transfection effects on the luciferase protein itself, as the reader may recall p53 protein levels at this time-point were not significantly altered (Figure 4.2C). This hypothesis is supported by the lack of change at the transcript level for the luciferase gene in the *HTT* promoter cell line (Figure 4.3C). There was a significant change in luciferase transcript in the Region 9 cell line at 24 h , however again this change was not significant compared to the scramble treatment suggesting that it was again due to transfection effects (Figure 4.3B). At 48 h the luciferase assay showed a significant decrease in *HTT* promoter function in the *HTT* promoter cell line compared to untreated (Figure 4.3D). This treatment difference, however, was not significantly different to the scramble control indicating that this decrease was due to siRNA transfection. This was also seen at the luciferase transcript level, although paradoxically the effect is in the opposite direction, with p53 and scramble siRNAs inducing an increase in transcript (Figure 4.3E). There was no significant difference between the p53 siRNA and Scramble siRNA suggesting that, although the effect was in the opposite direction, it was due to siRNA transfection. At 48 h in the Region 9 cell line there was a significant increase in *HTT* promoter function, but again this difference was not significantly different as compared to the scramble control (Figure 4.3E). This trend was mirrored in the luciferase transcript level, as well as in endogenous *HTT* expression (Figure 4.3F).

I also examined the effects of p53 knockdown on the ARE Cell line (Figures 4.3A & D). This cell line (described in Chapter 3) carries an enhancer element which

can be regulated by cellular stress. This allowed interrogation of the effect of the siRNA transfection process and knockdown of the selected TFs on cellular stress itself. I found no change at the luciferase protein level at 24 h (Figure 4.2A), and a small but significant decrease in protein level at 48 h with p53 siRNA treatment as compared to both untreated cells and scramble siRNA treated (Figure 4.2D). As p53 plays a well-known role in the apoptotic cell death pathway, decreasing p53 levels may result in this small decrease in inferred cellular stress at 48 h. Interestingly, unlike the *HTT* promoter and Region 9 cell lines, scramble treatment did not have an effect at either time point on ARE cell line luciferase protein levels. This suggest that the pathways stimulated by scramble treatment do not affect the ARE enhancer. The differential effect of the scramble treatment between my three cell lines will be further explored later in this chapter.

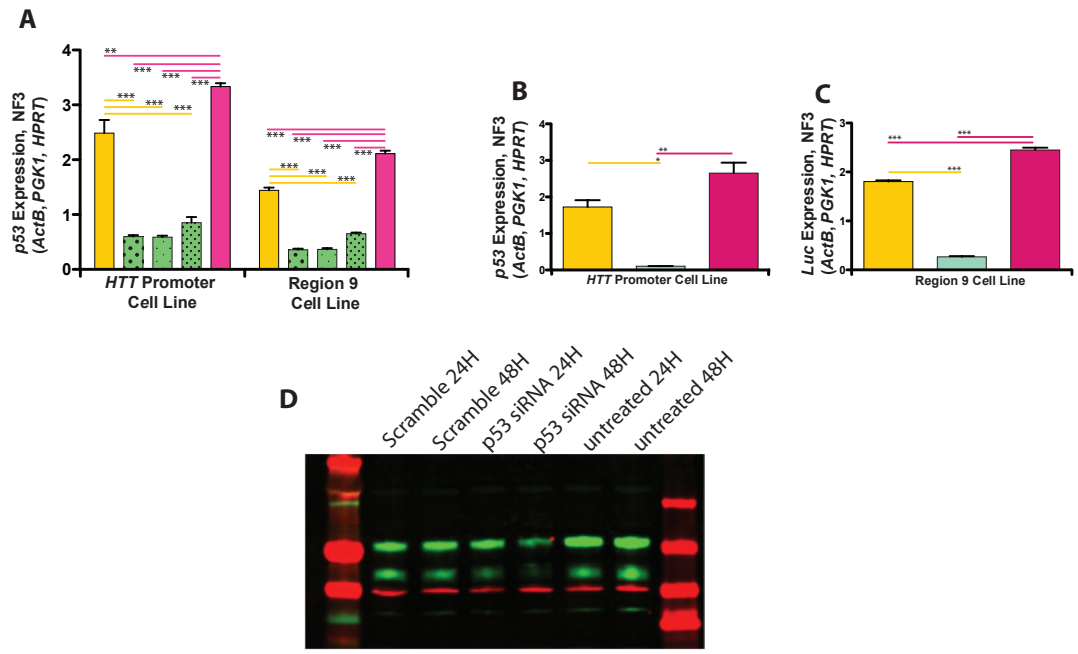


Figure 4.2 p53 Expression 24 & 48 h after p53 siRNA Transfection
A, B &C) RT-qPCR p53 expression at 24 h , (A) and 48 h, (B&C), respectively, after p53 siRNA Transfection in both the *HTT* promoter and Region 9 cell lines. One-way ANOVA with Tukey post test, *p<0.5, **p<0.01, ***p<0.001 (D) Western Blot of p53 expression at 24 and 48 h of p53 siRNA transfection in *HTT* promoter cell line.

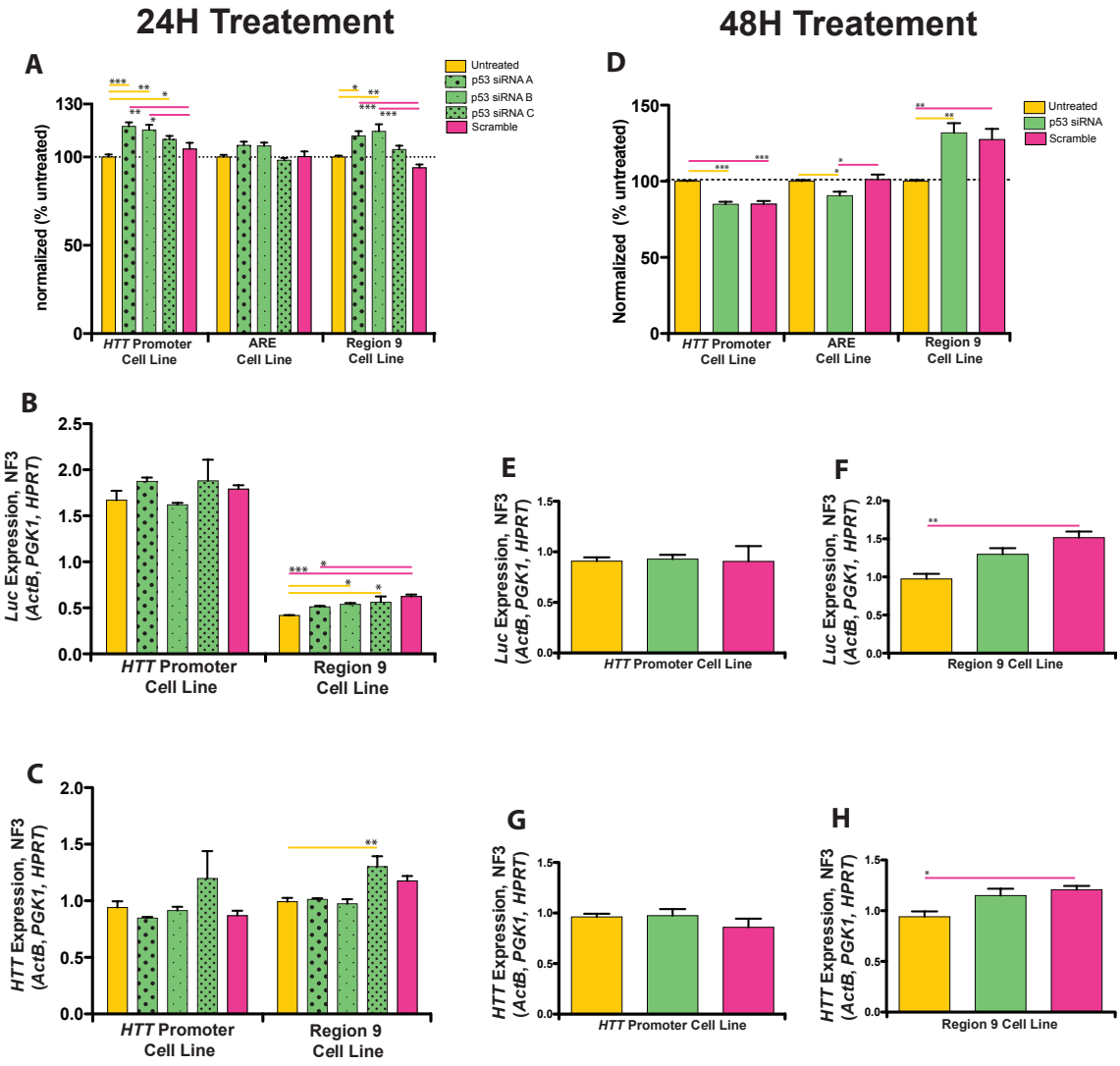


Figure 4.3 Luciferase and HTT Expression 24 & 48 h After p53 siRNA Transfection

(A, B & C) 24 h p53 siRNA treatment (D, E, F, G & H) 48 h siRNA treatment (A &D) Luciferase Assay in all three cell lines, shown as % of untreated for each cell line, average of 3 experiments with n=3 per condition in each experiment. (B, E &F) RT-qPCR Luciferase expression for *HTT* promoter and Region 9 cell lines, n=3 per condition. (D, G & H) RT-qPCR Endogenous *HTT* expression for *HTT* promoter and Region 9 cell lines, 48 h p53 siRNA treatment, n=3 per condition. One-way ANOVA with Tukey post test for each cell line, * = p<0.0.5, ** = p<0.01, *** = p<0.001

4.3.2.2 NFkB Knockdown

I again elected to test both the 24 and 48 h time points for both NFkB transcript and protein levels following NFkB siRNA treatment in order to further establish the experimental paradigm (Figure 4.4). As with p53 siRNA treatment, I saw significant knockdown of the NFkB transcript at both time points but only at 48 h was a decrease in NFkB protein level detected as compared to both scramble and untreated. I also again saw a small but significant increase in the luciferase assay in both *HTT* promoter and Region 9 cell lines at 24 h (Figure 4.5A), similar to that seen in the p53 siRNA treated cells at 24 h. This effect was, again, not replicated in the luciferase transcript or endogenous *HTT* expression at 24 h (Figures 4.5B &C). At 48 h there was no change in the luciferase assay, as compared to scramble treated, in the *HTT* promoter, ARE, and Region 9 cell lines (Figure 4.5D). This was replicated in both luciferase transcript (Figures 4.5E &F) and endogenous *HTT* transcript (Figures 4.5H &G) in both the *HTT* promoter and Region 9 cell lines.

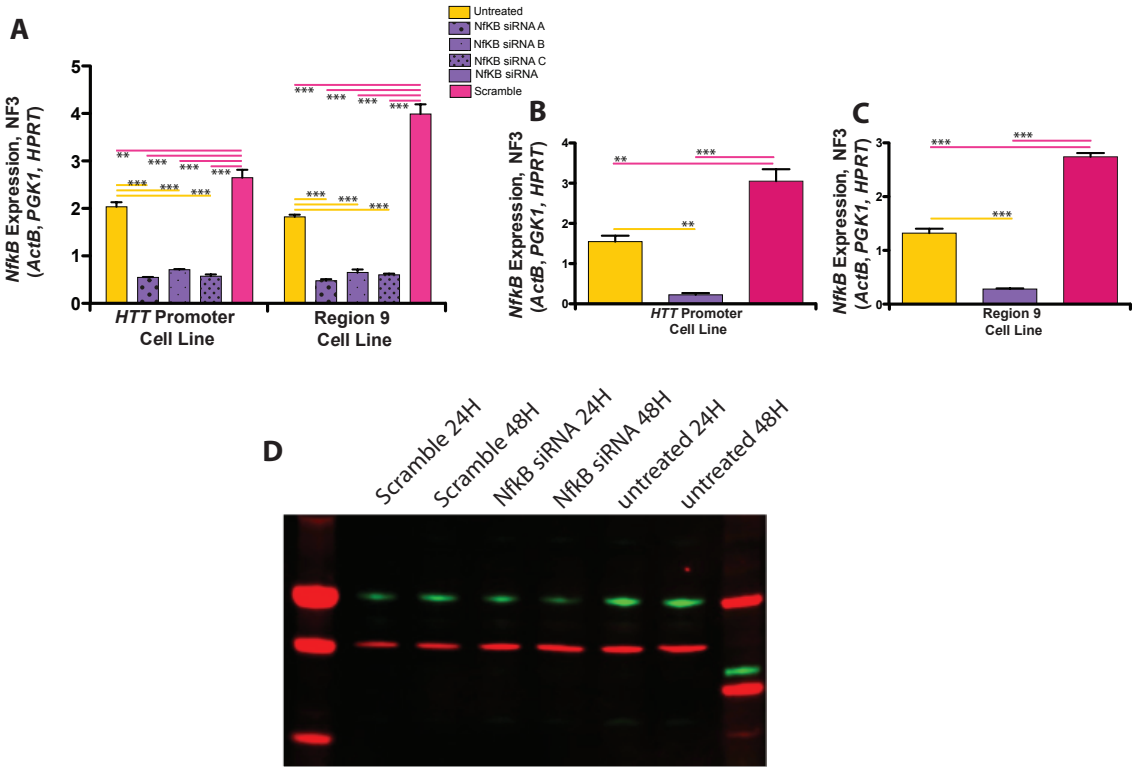


Figure 4.4 NFkB Expression 24 & 48 h After NFkB siRNA Transfection
A, B &C) RT-qPCR NfkB expression at 24 h , (A) and 48 h, (B&C), respectively, after NfkB siRNA Transfection in both the *HTT* promoter and Region 9 cell lines. One-way ANOVA with Tukey post test, *p<0.5, **p<0.01, ***p<0.001 (D) Western Blot of p50 subunit of NfkB expression at 24 and 48 h of NFkB siRNA transfection in *HTT* promoter cell line.

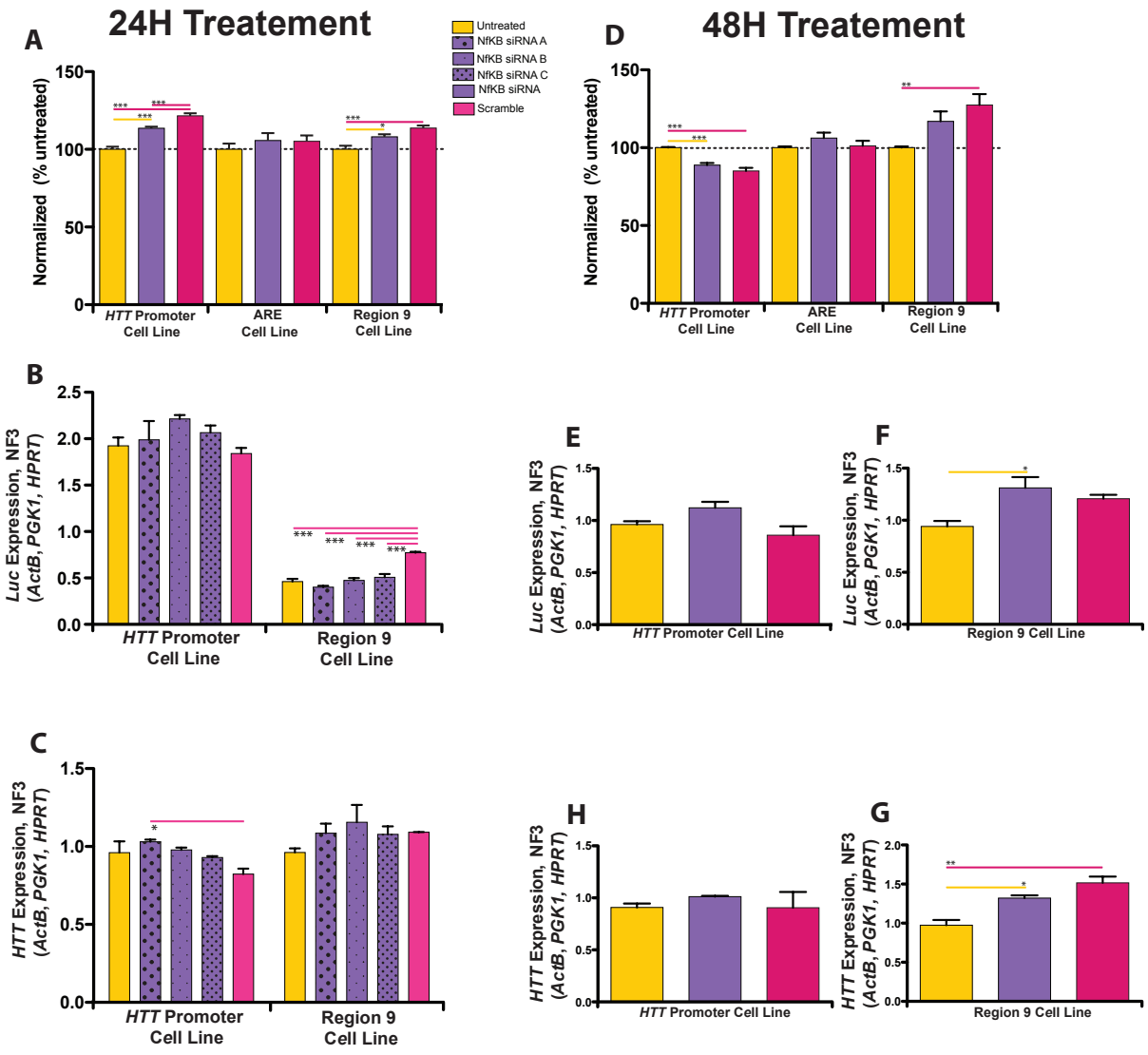


Figure 4.5 Luciferase and HTT Expression 24 & 48 h After NFkB siRNA Transfection

(A, B & C) 24 h NfkB siRNA treatment (D, E, F, G & H) 48 h siRNA treatment (A &D) Luciferase Assay in all three cell lines, shown as % of untreated for each cell line, average of 3 experiments with n=3 per condition in each experiment. (B, E &F) RT-qPCR Luciferase expression for *HTT* promoter and Region 9 cell lines, n=3 per condition. (D, G & H) RT-qPCR Endogenous *HTT* expression for *HTT* promoter and Region 9 cell lines, 48 h NFkB siRNA treatment, n=3 per condition. One-way ANOVA with Tukey post test for each cell line, * = p<0.0.5, ** = p<0.01, *** = p<0.001

4.3.2.3 SP1 Knockdown

At 48 h of treatment with Sp1 siRNA, I saw significant knockdown of Sp1 transcript in both the *HTT* promoter and Region 9 cell lines (Figure 4.6A). In the HTT promoter cell line SP1 knockdown resulted in a significant decrease in the luciferase assay as compared to untreated cells, but this was not significantly different from scramble treated cells (Figure 4.6B). This again indicates that this difference is due to siRNA treatment. In addition no effect in luciferase transcript was seen (Figure 4.6C). In the Region 9 cell line, an increase in the luciferase assay in SP1 siRNA treated cells compared to untreated was observed, but this was not significantly different from scramble treated cells (Figure 4.6B). Luciferase transcript levels indicate that SP1 knockdown decreased luciferase expression compared to the scramble treatment (Figure 4.6C), suggesting that perhaps a slightly longer time-point might reflect this in the luciferase assay. This would also suggest that the addition of region 9 is necessary for SP1 to have its regulatory effect. Interestingly, the endogenous *HTT* transcript was unaffected compared to scramble treated cells in both cell lines, although there is a trend towards an increase in transcript (Figure 4.6D). This effect may become more pronounced with additional experiments and longer time points. SP1 knockdown did not affect ARE cell line luciferase protein (Figure 4.6A), indicating that SP1 knockdown did not affect cellular stress.

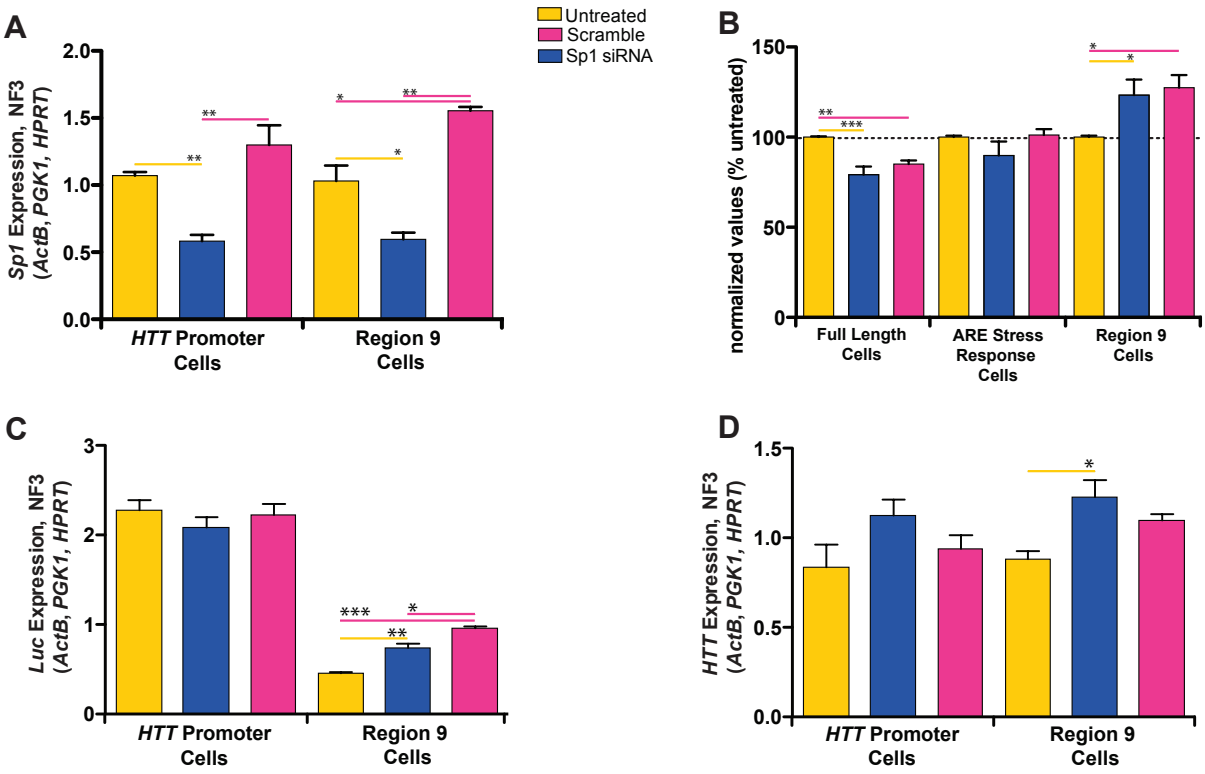


Figure 4.6 Sp1 siRNA Knockdown at 48 h, Luciferase and HTT Expression
(A) RT-qPCR Sp1 expression following 48 h of SP1 siRNA transfection in the *HTT* promoter and R9 cell line, n=3 per condition. (B) Luciferase Assay, 48 h of Sp1 siRNA transfection in all three cell lines, shown as % of untreated for each cell line, average of 3 experiments with n=3 per condition in each experiment. (C) RT-qPCR Luciferase expression for *HTT* promoter and Region 9 cell lines, 48 h Sp1 siRNA treatment, n=3 per condition. (D) RT-qPCR Endogenous *HTT* expression for *HTT* promote and Region 9 cell lines, 48 h Sp1 siRNA treatment, n=3 per condition. One-way ANOVA with Tukey post test for each cell line, * = p<0.0.5, ** = p<0.01, *** = p<0.001

4.3.3 siRNA Knockdown of Candidate TFs from Bioinfomatic Assessment of the *HTT* Gene Locus

In chapter 2 I generated a candidate list of TFs that, based on my identification of potential regulatory regions and ChIP-seq dataset searches, I believe may be involved in transcriptional regulation. I then selected of these 4 transcription factors to test in the stably expressing cells at the protein, luciferase assay, transcript, and qPCR level.

4.3.3.1 siRNA Knockdown of SMC3

SMC3 is more commonly known for its role in the formation of the cohesin complex, and is important for maintaining sister chromatid cohesion during the arrangement of the sister chromatids on the metaphase plate during mitosis. Once the cell begins to progress through the metaphase-to-anaphase transition, the cohesin complex is removed from the sister chromatids. Recently it has been found that SMC3 binding to DNA does not disappears after mitosis is complete, instead the cohesin complex re-binds DNA shortly after telophase is complete and remains throughout interphase of the cell cycle (Dorsett 2009). This additional binding has been found to contribute to gene silencing in yeast, aid in CTCF transcriptional insulation in mice and humans, and to positively regulate c-myc transcription (Wendt & Peters 2009; Gartenberg 2009; Rhodes et al. 2010).

SMC3 siRNA decreased SMC3 transcript significantly at 48 h in both cell lines as compared to untreated and scramble treated cells (Figure 4.7A). As SMC3 binding was only observed by ChIP-seq in Region 9, I did not anticipate any

changes due to SMC3 knockdown in the *HTT* promoter cell line. Contrary to this, SMC3 siRNA treatment had a small but significant decrease in the luciferase assay in the *HTT* promoter cell line as compared to both untreated and scramble siRNA treatments. This decrease was also present at the luciferase transcript level (Figure 4.7B). SMC3 siRNA treatment in the Region 9 cell line did not have any effect on the luciferase assay, and while luciferase transcript was significantly decreased compared to scramble in this cell line it was unchanged as compared to untreated cells (Figure 4.7B &C). It is possible that there are SMC3 binding sites present in the *HTT* promoter region that were not detected in the ChIP-seq dataset or were not actively bound in the cell line used to generate the ChiP-seq dataset. SMC treatment did not effect endogenous *HTT* gene transcription in the HEK293 cell lines I created (Figure 4.7D).

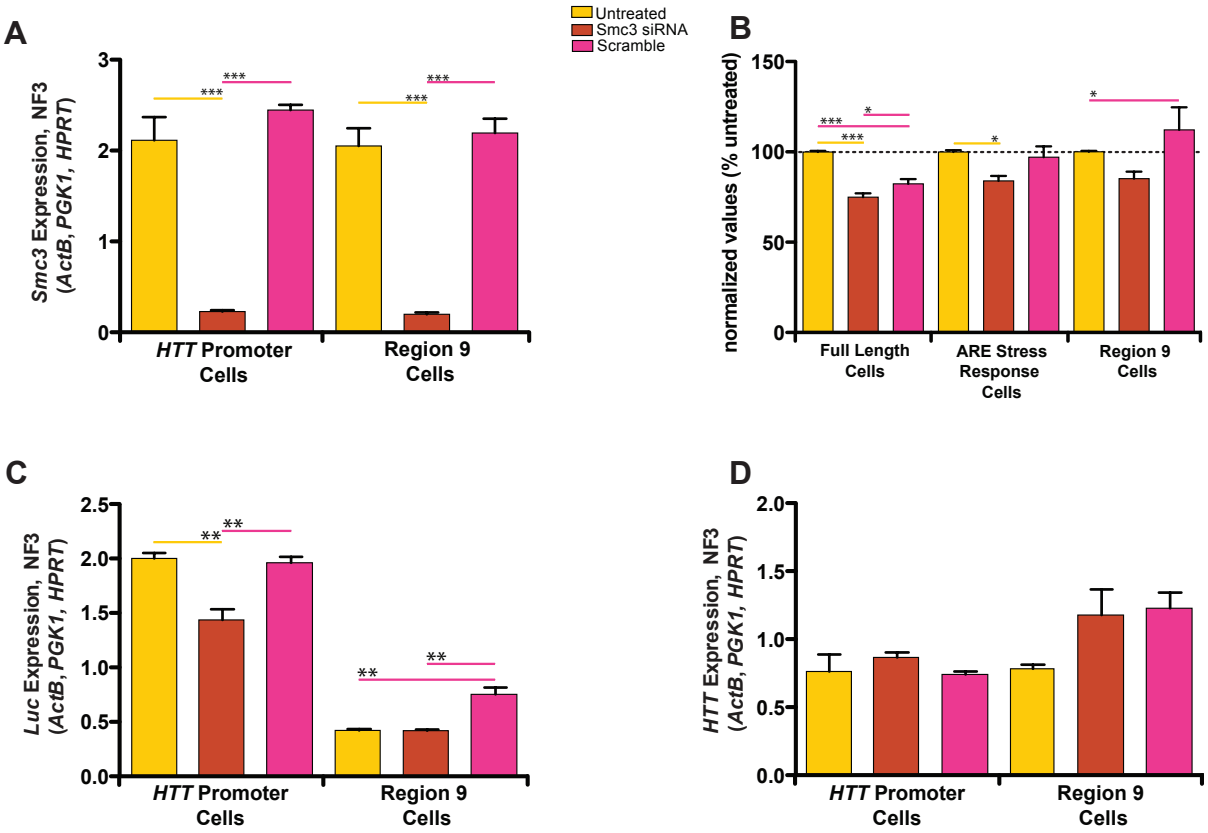


Figure 4.7 Smc3 siRNA Knockdown at 48 h, Luciferase and HTT Expression
A) RT-qPCR Smc3 expression following 48 h of Smc3 siRNA transfection in the *HTT* promoter and R9 cell line, n=3 per condition. (B) Luciferase Assay, 48 h of Smc3 siRNA transfection in all three cell lines, shown as % of untreated for each cell line, average of 3 experiments with n=3 per condition in each experiment. (C) RT-qPCR Luciferase expression for *HTT* promoter and Region 9 cell lines, 48 h Smc3 siRNA treatment, n=3 per condition. (D) RT-qPCR Endogenous *HTT* expression for *HTT* promote and Region 9 cell lines, 48 h Smc3 siRNA treatment, n=3 per condition. One-way ANOVA with Tukey post test for each cell line, *p<0.0.5, **p<0.01, ***p<0.001

4.3.3.2 siRNA Knockdown of JunB and Rfx5

JunB was identified in region 9 as part of a cluster of closely related TFs from the Jun and Fos family of transcription factors (Figure 2.3) from Chapter 2. TFs from both of these families can act in homo and heterodimers to form AP-1 proteins that are known to act in both activating and repressive capacities in a wide range of cellular processes including cell proliferation, differentiation, cell survival and apoptosis (Vesely et al. 2009). As elaborated upon in the introduction chapter *HTT* is known to be involved in cell viability and apoptosis making investigations into TFs part of the AP-1 family of proteins of interest.

Rfx5 is is the DNA binding subunit of the heterotrimer RFX, a TF complex known to be involved in the regulation of class II major histocompatability complex genes which are important for immune system function (Chakraborty et al. 2010). Literature searchers did not indicate genes in any other cell pathways as being regulated by Rfx5. Regulation of *HTT* by Rfx5 would implicate that this immune system specific TF has additional regulatory roles

At 48 h of treatment I found that treatment with scramble siRNA significantly increased both JunB and Rfx5 transcript levels, indicating that either siRNA transfection itself or the scramble siRNA affects both JunB and Rfx5 expression (Figure 4.8). To test whether the siRNA transfection or the scramble siRNA is causing transcript of either TF to increase I re-ran my p53 siRNA treated samples using JunB and Rfx5 primers in RT-qPCR (Figure 4.9). The fold change in both JunB and Rfx5 transcript levels in the p53 siRNA and scramble siRNa from the p53siRNA experiment appears to be the same as in either JunB siRNA or Rfx5 siRNA

experiments. This indicates that both JunB and Rfx5 transcript levels are affected by the siRNA transfection itself. In either case, the presence of the corresponding siRNA knocked down this increased expression to untreated levels. This is reflected in the luciferase assay and luciferase transcript expression as treatment with either siRNA resulted in similar effects (Figure 4.10A&B). These effects also reflect changes seen with scramble siRNA treatment alone. No effect on endogenous *HTT* transcript was observed in the *HTT* promoter cell line for either TF (Figure 4.10C&D). There was a small, but significant, increase in endogenous *HTT* transcript in the Region 9 cell line, seen in both TFs, (Figure 4.10E&F). These increases were similar to those seen with scramble alone, the lack of statistical significance of the scramble treatment as compared to untreated is most likely due to the increase in variability seen in this cell line.

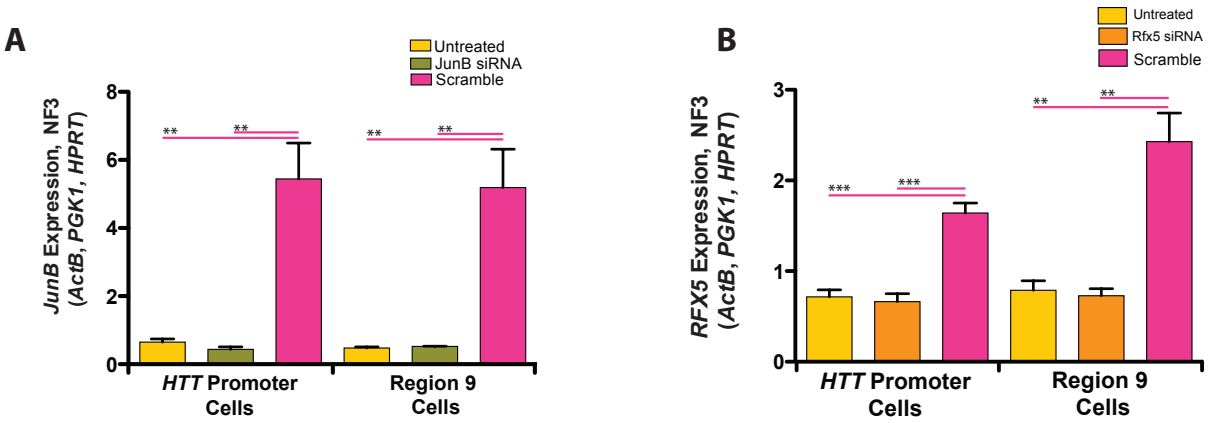


Figure 4.8 JunB and Rfx5 Expression Following Targeted siRNA treatment
(A) RT-qPCR JunB expression following 48 h of JunB siRNA transfection in the *HTT* promoter and Region 9 cell lines. (B) RT-qPCR Rfx5 expression following 48 h of Rfx5 siRNA in the *HTT* promoter and Region 9 cell lines. n=3 for each condition. One-way ANOVA with Tukey post test for each cell line, *p<0.0.5, **p<0.01, ***p<0.001

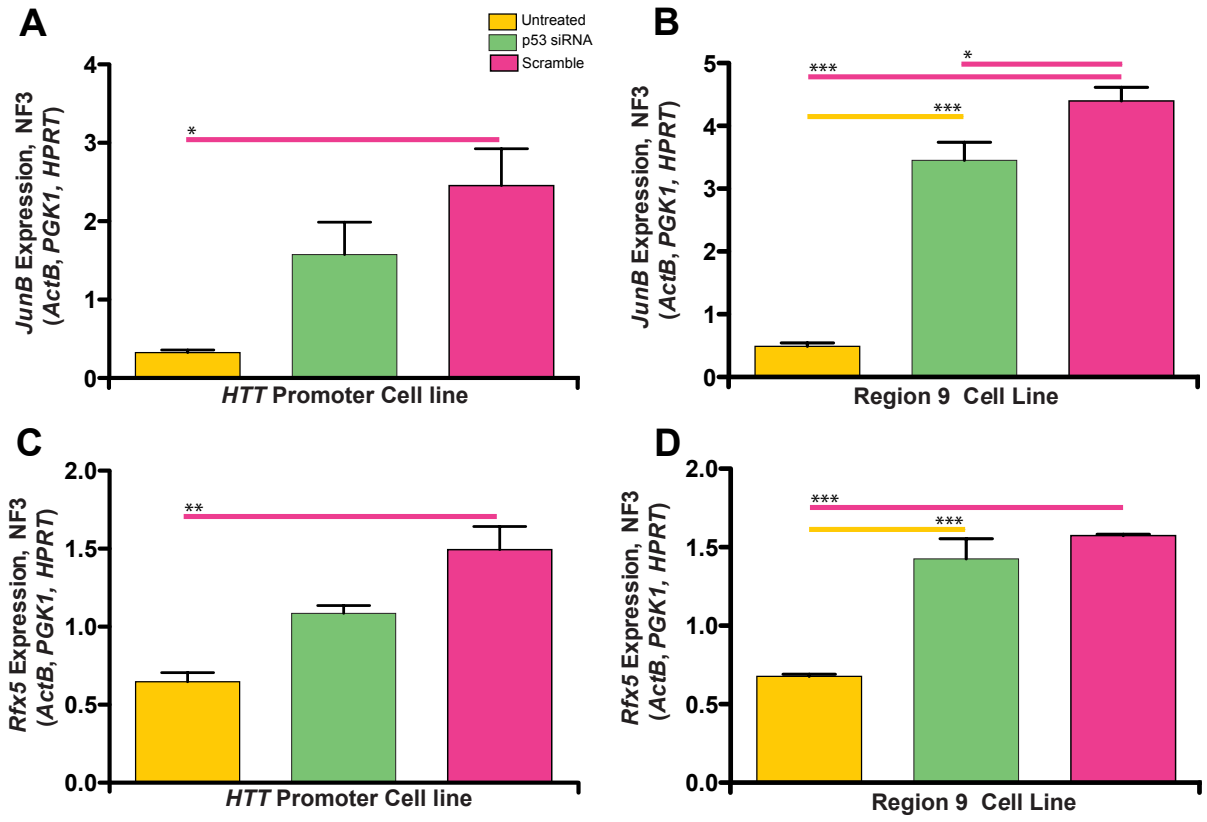


Figure 4.9 JunB and Rfx5 Expression in p53 siRNA Treated Cells
(A) RT-qPCR JunB expression following 48 h of p53 siRNA transfection in the *HTT* promoter and R9 cell lines. (B) RT-qPCR Rfx5 expression following 48 h of p53 siRNA in the *HTT* promoter and Region 9 cell lines. n=3 for each condition. One-way ANOVA with tukey post test for each cell line, * = p<0.0.5, ** = p<0.01, *** = p<0.001

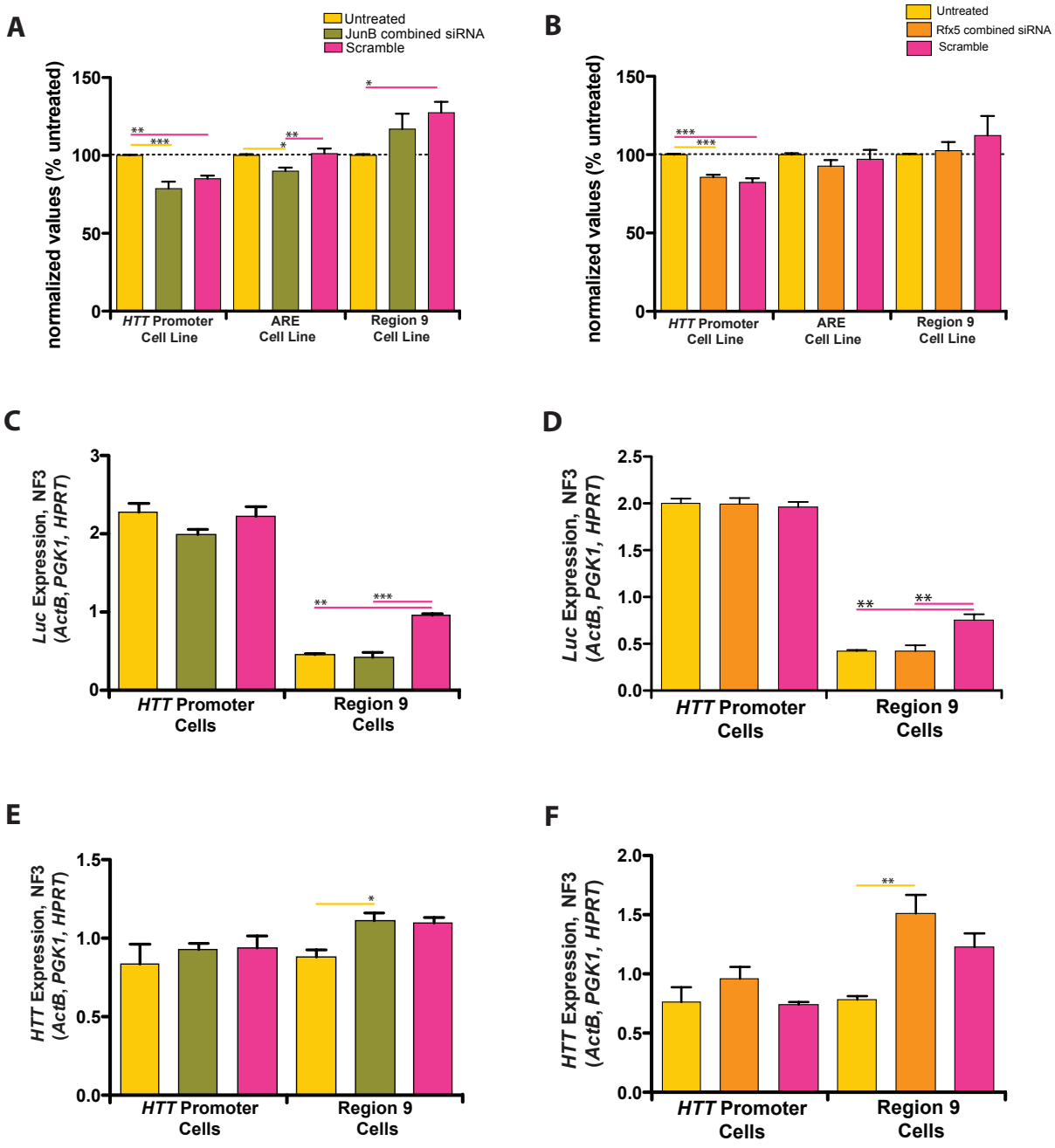


Figure 4.10 Luciferase Assay, Luciferase Transcript Expression and HTT Expression Following JunB and Rfx5 siRNA Treatment

(A & B)Luciferase Assay following 48 h of JunB (A) or Rfx5 (B) siRNA treatment in *HTT* promtoer, ARE and Region 9 Cell lines, shown as % of the untreated condition, average of 3 experiments with n=3 per condition in each experiment. (C & D) RT-qPCR Luciferase transcript expression in *HTT* promoter and region 9 cell lines, 48 h JunB (C) and Rfx5 (D). (E & F) RT-qPCR Endogenous *HTT* transcript expression in *HTT* promoter and region 9 cell lines, 48 h JunB (E) and Rfx5 (F). n=3 for each condition. One-way ANOVA with Tukey post test for each cell line, *p<0.0.5, **p<0.01, ***p<0.001

4.3.3.3 siRNA Knockdown of IRF1

IRF1 is a well-known transcription factor that was first characterized as a transcriptional activator of the cytokine Interferon beta (Kröger et al. 2002). It is known to regulate target genes in many cellular pathways including cell cycle arrest, apoptosis, differentiation and immune responses (Tamura et al. 2008; Honda & Taniguchi 2006;; Romeo et al. 2002). In my bioinfomatic screen I found several IRF1 binding sites within the *HTT* promoter region.

IRF1 siRNA treatment for 48 h resulted in a decrease in luciferase expression as compared to untreated cells but an increase as compared to the scramble treated cells in the luciferase assay (Figure 4.11A). The region 9 cell line displayed a significant decrease in luciferase as compared to both the untreated and scramble treated cells in the luciferase assay (Figure 4.11A). While there were no IRF1 binding sites identified in region 9, binding sites for interactors STAT3 and STAT1, were found in region 9. No effect on the ARE cell line was detected with IRF1 siRNA treatment (Figure 4.11A). At 48 h of IRF1 siRNA treatment I observed an apparent increase in IRF1 transcript as compared to untreated cells (Figure 4.11B). The interferon pathway has been known to be activated by siRNA transfection (Sledz et al. 2003). This may be in part due to the length of the siRNA used, research has shown that a longer 27mer length, such as I used, does not induce the interferon pathway (Kim et al. 2005). However Kim et al. 2005 utilized much lower concentrations of siRNA, 25 nM, whereas I used 75 nM. This difference may explain why I see an increase in IRF1 transcript following IRF1 siRNA treatment as interferon pathway stimulation by siRNA is known to be dosage dependent (Sledz et

al. 2003). As with JunB and Rfx5 I examined IRF1 expression in p53 siRNA treated samples to determine if the increase in IRF1 expression seen after IRF1 siRNA treatment was siRNA dependent (Figure 4.11E). IRF1 expression was elevated in both p53 siRNA and scramble treated *HTT* promoter cells indicating that siRNA treatment. Unlike JunB and Rfx5, where I also saw an increase in TF transcript with scramble treatment, the increase in IRF1 was not knocked down by the IRF1 siRNA to untreated levels. This suggest that either the IRF1 siRNAs themselves are ineffectual at knocking down transcript or that they do not target the IRF1 transcript at all. Mis-targeting of the IRF1 siRNAs to an unrelated gene would explain the decrease in luciferase assay in the Region 9 cell line (Figure 4.11A). siRNA/DICER mediated knockdown of non-targeted genes by the sense strand of the double stranded siRNA has been observed (Nolte et al. 2013). As I currently do not have data for IRF1 transcript levels in the Region 9 cell line it is unclear if the decrease in the luciferase assay is due to an apparent increase in IRF1 transcript, and potential off-target effects, or if IRF1 is in fact knocked down in the Region 9 cell line.

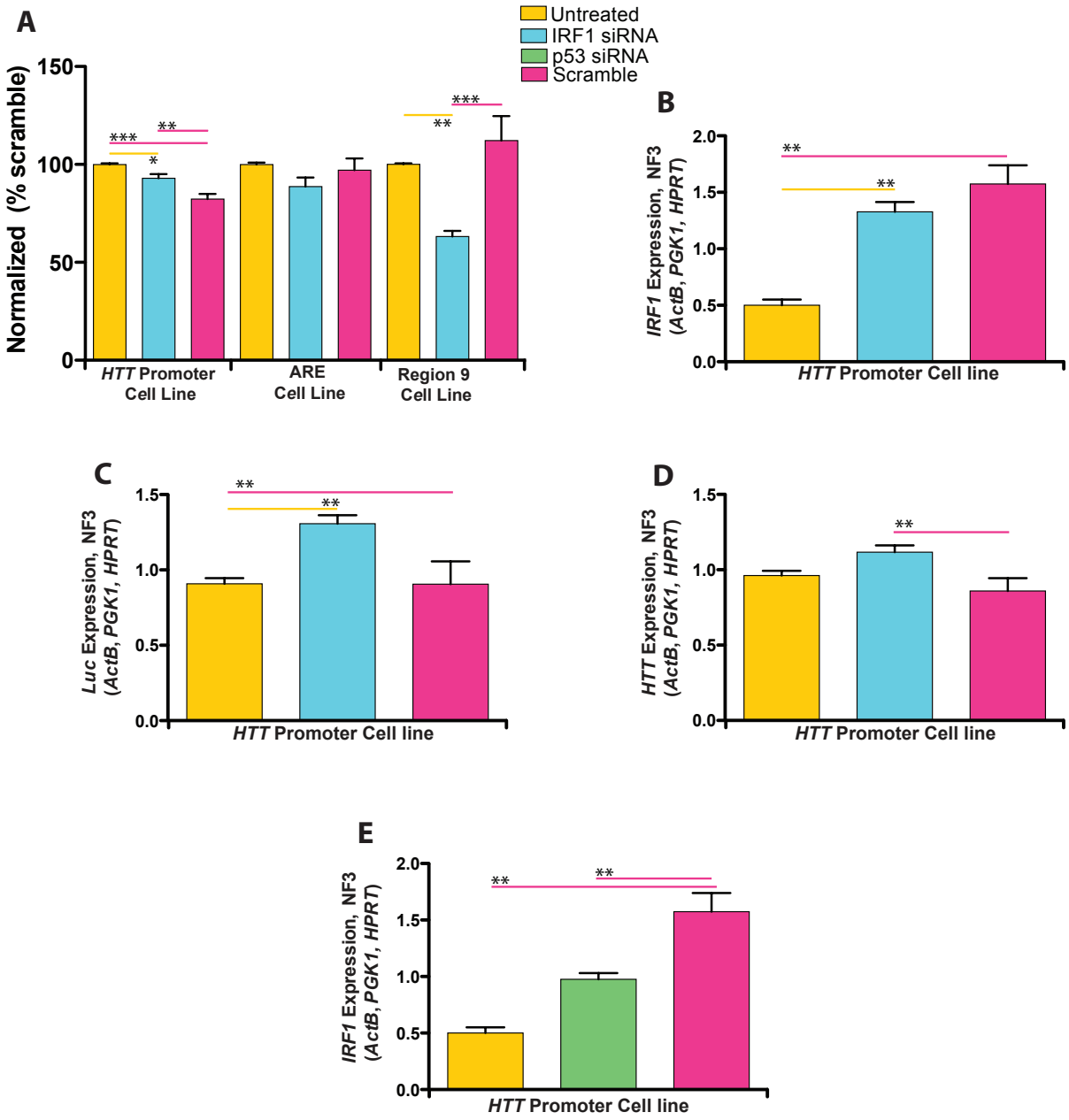


Figure 4.11 Luciferase Assay, 48 h IRF1 siRNA Treatment
(A)Luciferase Assay following 48 h of IFR1 siRNA treatment, shown as % of the untreated condition, average of 3 experiments with n=3 per condition in each experiment. (B, C &D) RT-qPCR IRF1, luciferase, and *HTT* transcript expression respectively in the *HTT* promoter cell line, 48 h of IRF1 siRNA treatment. (E) RT-qPCR IRF1 transcript expression in 48 h p53 siRNA treated *HTT* promoter cells. One-way ANOVA with Tukey post test for each cell line, *p<0.0.5, **p<0.01, ***p<0.001

4.3.4 Differential Effect of siRNA Treatment on *HTT* promoter and Region 9 Constructs in HEK293 Stably Expression Cell Lines

At the 48 h time point a consistent trend was observed in the luciferase assay data. The *HTT* promoter cell line and the Region 9 cell line each had a consistent change in the scramble treated cells as compared to untreated cells. The effect in the *HTT* promoter cell line was a reduction in the luciferase assay, whereas in the Region 9 cells scramble treatment caused an up regulation. Taking the untreated and scramble treated wells across all experiments (n=18 for each treatment) for each cell line revealed how consistent this trend was (Figure 4.12 A). It is evident that TF binding to Region 9 causes the *HTT* promoter to act differentially when siRNA transfection occurs. To see if this trend was also apparent in Luciferase transcript and endogenous *HTT* expression I combined all of my untreated and scramble treated data (n=8-9) at 48 h (Figure 4.12B). The increase in the Region 9 cell line in the luciferase assay is reflected in luciferase transcript. The decrease in the *HTT* promoter cell line was not replicated in luciferase transcript. Endogenous *HTT* expression was also not affected in the *HTT* promoter cell line, however a similar increase to that seen in the luciferase assay and transcript was seen in the Region 9 cell line (Figure 4.12C). This indicates that the region 9 cell line itself has a differential response to siRNA treatment.

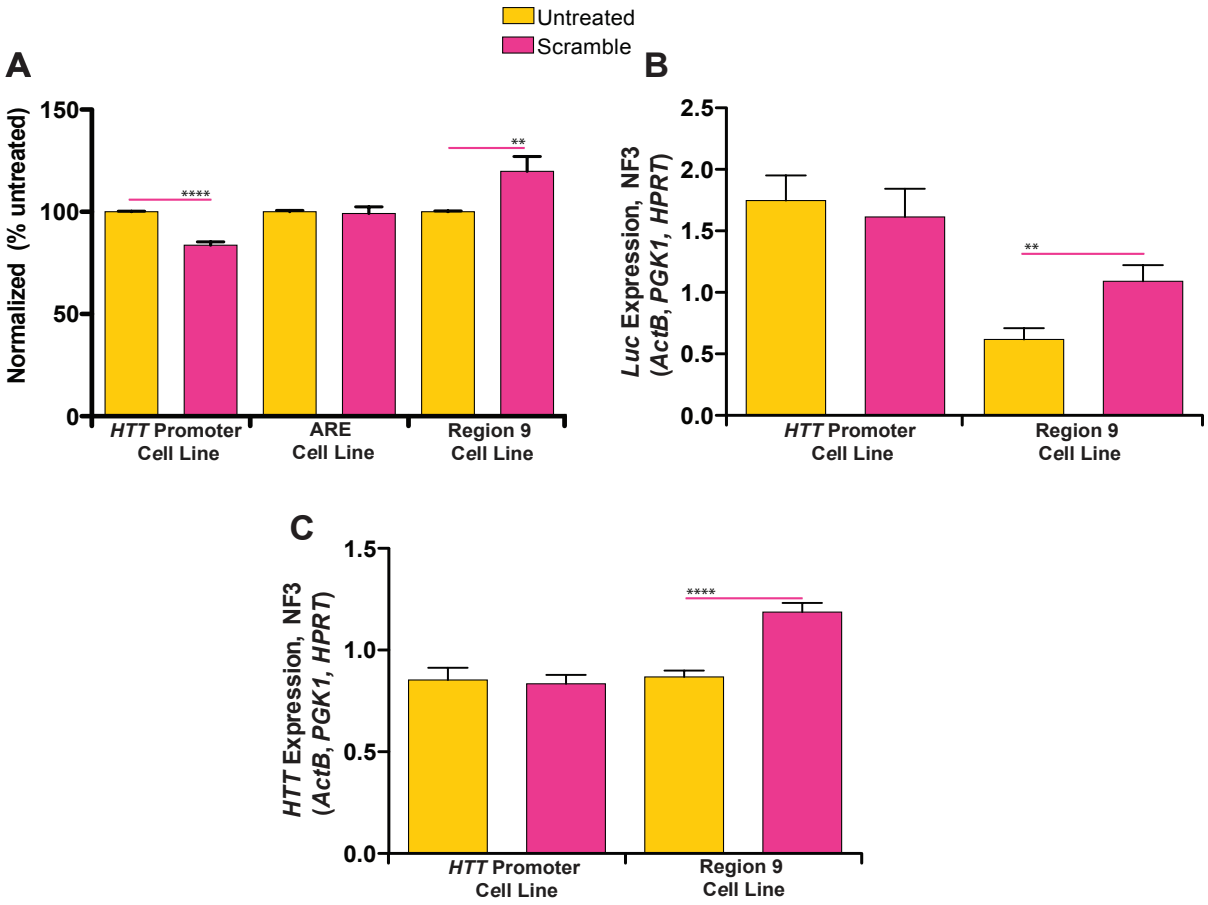


Figure 4.12 HTT Promoter Construct and Region 9 Construct Are Differentially Effected by siRNA Transfection

(A) Untreated and scramble treated wells from all 48 h luciferase assay experiments presented for the *HTT* promoter, ARE and Region 9 Cell lines, shown as % of untreated, n=18 per treatment. (B) RT-qPCR Luciferase expression in all 48 h untreated and scramble treated experiments, n= 8-9 per treatment. (C) RT-qPCR Endogenous *HTT* expression in all 48 h untreated and scramble treated experiments, n=18 per treatment. Student's t-test per cell line, **p<0.01, ****p<0.0001

4.4 Discussion

Using my HEK293 stable expressing cell lines I tested several previously identified TFs, as well as TFs identified in my bioinfomatic assessment, by using targeted siRNAs against these TFs. I assessed the level of knockdown using transcript levels and, in two cases, protein levels. Assessment of protein levels allowed us to establish a time point, 48 h, at which TF protein levels would likely mirror the transcript expression levels. I looked at the effect on *HTT* promoter function through both my established luciferase assay as well as through direct transcript measures of luciferase transcript. I also assessed the effect of siRNA knockdown on the endogenous HEK293 *HTT* gene transcript.

In my experimental paradigm it was evident that the *HTT* promoter and Region 9 constructs, stably expressed in their respective cell lines, are differentially affected by siRNA treatment. The *HTT* promoter cell line displays a consistent baseline decrease in scramble treated cells in the luciferase assay, while the Region 9 cell line consistently displays an increase. The increase in the luciferase assay in the Region 9 cell line is mirrored in luciferase transcript levels in this cell line. The differential effect of siRNA treatment on the two constructs does confound analysis of the data, making verification of gene specific knockdown especially important. This is clear in the case of JunB, Rfx5, and IRF1 where siRNA treatment itself resulted in a significant increase of all three TFs in cells treated with an siRNA targeting a different TF, p53. In addition, it appears that the Region 9 cell line itself may have a differential response to siRNA transfection related stressors, the

endogenous *HTT* gene transcript is elevated in scramble treated cells while untreated Region 9 cells have a similar endogenous *HTT* level to untreated and scramble treated *HTT* promoter cells. The source of this differential response remains unknown, it is possible that the stringent selection that these cell underwent during their derivation into these cell lines is responsible. Further inquiry into this differential response is required, for example by examining the response to siRNA transfection stress in the additional *HTT* promoter and Region 9 cell lines. Global transcriptional differences between the cell lines used here may also indicate baseline differences that may explain their differential response.

Of the TFs that were not elevated by siRNA treatment itself, namely NFkB, p53, Sp1, and Smc3, only Smc3 showed an effect on *HTT* promoter function in both the luciferase assay and the luciferase transcript level. Surprisingly this was only seen in the *HTT* promoter cell line despite the identified Smc3 site being in Region 9. It is possible that additional Smc3 sites do reside in the promoter region but were not identified in the ChIP-seq studies I used due differences in cell type or tissue used in those studies. ChIP-seq itself is inherently a 'snap shot' of the proteins bound to genomic DNA at the time the cells were formaldehyde treated (Poptsova 2014). As Smc3 is well known to be involved in the cohesion of sister chromatids before metaphase in mitosis it is feasible that, depending on the phase of the cell cycle the majority of the cells in the sample are in, SMC3 binding may differ depending on the time point of cell collection (Dorsett 2009). Follow up investigation into the role of SMC3 in *HTT* promoter function is warranted based upon my siRNA results. Interestingly, endogenous HEK293 *HTT* transcript levels were unaffected by SMC3

siRNA treatment. While the *HTT* promoter fragment utilized to generate the *HTT* promoter-luciferase constructs that I used were taken from an HD allele with the most common *HTT* haplogroup, it is unknown at the current time what the haplogroup of the endogenous HEK293 *HTT* is. It is possible that a difference in haplogroup is responsible for the differential response to SMC3 siRNA knockdown between the *HTT* promoter construct and endogenous *HTT* expression.

In this siRNA knockdown paradigm, knockdown of both of the published TFs, p53 and SP1, did not result in significant changes in luciferase assay, luciferase transcript, or endogenous *HTT* expression. Subsequent to the findings concerning p53's role in *HTT* promoter function, attempts to repeat the results found through over expression of p53 were conducted by others in the Leavitt lab, using both our own longer constructs as well as the shorter *HTT* promoter constructs used in the Ryan et al., 2006 and Feng et al., 2006 papers. Using the same concentrations of p53 over expression plasmid as the Ryan group, we were unable to replicate the results found in the Ryan 2006 paper, (data not shown). The lack of effect of knockdown of p53 on both *HTT* promoter function in my stably expressing cell lines as well as endogenous HEK293 *HTT* expression is consistent with my lab's previous data. Based upon my results I conclude that p53 is not a regulator of *HTT* transcription in this cell type.

Sp1 was identified as a transcriptional regulator of the *HTT* promoter in a 2012 report and was also identified as a putative TFBS in the region containing the *HTT* promoter in my bioinfomatic assay (Wang et al. 2012). The Wang et. al 2012 report found several putative SP1 TFBSs both up and down stream of a TSS that

they identified and is different than the one identified in the UCSC genome browser. In contrast, I only found one SP1 TFBS downstream of the first exon of *HTT*. This difference in the number of identified SP1 putative TFBSs is likely due to the difference in identification of TFBS, namely I based my TFBS on biding in ChIP-seq assays while t Wang et. al 2012 used a purely bioinfomatic method based upon a PWM for SP1 and sequence similarity alone. It is possible that the additional SP1 sites identified in Wang et. al 2012 are functional and bind SP1, but due to cell type differences used in the ChIP-seq assay they were not identified in my screen. While the use of an over-expression paradigm, also in HEK293 cells, to test the role of SP1 in *HTT* promoter function does suggest that SP1 positively regulates the *HTT* promoter, this may be an effect of the over-expression of SP1 and may not accurately reflect SP1's function under normal conditions. In my siRNA-mediated knockdown paradigm I found that knockdown of SP1 did not effect *HTT* promoter function using either the luciferase assay or activity of the endogenous HEK293 *HTT* promoter. This indicates that under normal cellular conditions, i.e. endogenous HEK293 SP1 expression levels, the contribution of SP1 to *HTT* promoter function is minimal. While Wang et. al 2012 did identify a SP1 putative TFBS in their minimal promoter region they also identified additional TFBS for other TFs (as I did in my bioinfomatic screen) in this region. These TFBS, and associated TFs, may play a larger role in *HTT* promoter function than supposed by Wang et. al 2012. Site directed mutagenesis of the of the SP1 sites identified by t Wang et. al 2012 coupled with over expression of SP1 could provide greater clarity as to the role of SP1 in *HTT* promoter function. As both Wang et. al 2012 and my research was conducted in

HEK293 cell lines, which are derived from human kidney cells, it would be of interest to identify if SP1 expression levels in HEK293 cells is comparable to that found in neuronal cell lines. If the effect of SP1 on *HTT* promoter function requires higher levels of SP1 than found in endogenously in HEK293 cells, and SP1 is found to be increased in neuronal cell types, then it is possible that SP1 may play a role in tissue specific regulation of *HTT* expression.

5 DNA Methylation at the *HTT* Locus and in HD

5.1 Introduction

Given that CAG repeat length only accounts for 60-70% of the variability in age of onset in HD patients it is clear that additional factors (potentially epigenetic factors) play a role in HD progression and pathogenesis (Andrew et al. 1993). As our understanding of epigenetic regulation mechanisms becomes increasingly developed, questions concerning the role of these mechanisms in HD pathogenesis have arisen (Junghee Lee et al. 2013). To date, little attention has been given to epigenetic changes at the *HTT* locus itself. Individual differences in *HTT* expression could explain the remaining 30% of variability in age of onset as manipulation of mutant *HTT* expression has been found to affect HD progression in mice (Graham et al. 2006). There is evidence of inter-individual *HTT* expression variation in neuronal tissues in a study of endogenous *Hdh* expression in inbred mice, finding a considerable amount of *Hdh* brain expression variation between mice (Dixon et al. 2004). As these mice were all genetically identical, it is reasonable to suggest epigenetic variance between these individuals as a major contributing factor to expression variance. A prominent source of epigenetic variation comes in the form of DNA methylation, where cytosines, primarily in CpG dinucleotides, are modified by the addition of a methyl group. The *HTT* promoter is known to be CpG rich suggesting DNA methylation as a potential source of transcriptional regulation between individuals. Furthering this speculation are findings suggesting the

prevalence of non-CpG methylation in CpNpG trinucleotides where N is primarily an A (Juna Lee et al. 2010; Lister et al. 2009). Given the close proximity of the CAG repeat to the *HTT* promoter it is feasible that any additional methylation of the repeat due to an increase in repeat length could affect the *HTT* promoter and potentially affect transcriptional regulation.

Epigenetic differences between tissue types has also been found to regulate differential expression of genes (Wan et al. 2015). It has been noted that while the HD mutation is primarily thought to affect neuronal cells, specifically MSNs in the caudate and putamen, *HTT* itself is ubiquitously expressed with highest expression in both the CNS and testicular tissues in both protein and RNA studies (Van Raamsdonk et al. 2007; Dixon et al. 2004; Li et al. 1993). This pattern of expression has never been investigated beyond the identification of its existence and we know nothing concerning how this pattern of expression is established. As the expression of *HTT* is greatly reduced in non-neuronal or testicular tissues, understanding the mechanism of transcriptional regulation between tissue types will allow for a greater understanding of the role of *HTT* in these tissues as well as open avenues for potential therapeutic intention.

Two previous studies have assessed DNA methylation changes in the context of the HD mutation. The first utilized a genome-wide approach to assess DNA methylation changes in an immortalized mouse striatal cell line, expressing either the mutant or wildtype human protein (Ng et al. 2013). The second utilized both human HD putamen samples and samples from the R6/2 mouse model of HD to assess the effect of the HD mutation on the DNA methylation of a single gene, the

Adenosine A2A receptor (Villar-Menéndez et al. 2013). Both of these studies have laid the groundwork on which to further investigate DNA methylation changes in HD. Caution, however, is warranted when considering these studies, as both studies utilized transgenic mouse models expressing either a fragment of the mutated HD first exon, the Villar-Menendez study, or immortalized mouse striatal cells expressing a full length mutated *HTT* construct, the Ng study. While both of these mouse models do recapitulate some aspects HD pathogenesis, it is unclear to what extent DNA methylation changes in the R6/2 model are a result of over-expression of the toxic first exon fragment, and in the immortalized cells to what extent immortalization has changed genome-wide DNA methylation. The Villar-Mendez study also utilized human HD striatum samples, given the advanced stage of HD in these patients (the majority of samples had a Vonsattel stage rating of 3-4, Vonsattel stage rating is based upon striatal atrophy (Vonsattel et al. 1985) it is likely that neurodegeneration in the striatum at these stages would be extensive. This raises questions as to the proportion of neuronal versus glial cells remaining in the tissues studied.

To better understand the role of DNA methylation both in HD and in the regulation of the *HTT* locus, I designed a study examining *HTT* expression differences within and between tissues of individuals and methylation changes between individuals both globally and at the *HTT* locus. Contradictory to the Dixon 2004 study, I found a singular lack of inter-individual *Hdh* and *HTT* expression difference in inbred mice and discovered an error in the selection of a normalization that resulted in a misinterpretation of the findings published in Dixon 2004. Consistent with published literature I find considerable difference in *HTT* expression

between tissue types, specifically between cortex and liver samples. I addressed two separate yet related questions with regards to DNA methylation in HD using the Illumina 450K DNA methylation bead array in both a genome-wide and candidate gene (*HTT* locus specific) approach. Firstly, using age and sex matched samples I sought to explore the global effect on DNA methylation in HD and control cortex samples. In this same cohort of samples I also queried the local effect on DNA methylation at the *HTT* locus. Sub-setting out individuals from this first cohort I selected those for which I had matching liver tissue in order to address the second question of tissue specific DNA methylation at the *HTT* locus. With regards to my cortex only cohort, at the *HTT* locus there was no difference in DNA methylation between control and HD cortex samples. Globally I found 15 putative differentially methylated regions (DMRs) between HD and control cortex samples. With the second cohort of tissue matched samples, I report the first evidence of DNA methylation changes at the *HTT* locus between cortex and liver tissue types.

5.2 Methods

5.2.1 Real-Time Quantitative PCR (qRT-PCR)

Mouse tissues were isolated from FVB, littermate, male mice at 3 months of age according to guidelines established in the following UBC ethics certificate: A14-0031. Tissues were then flash frozen following dissection and stored at -80 until processed. Human samples were taken from the UBC HD Bio bank under the ethics reference: 'Huntington Disease BioBank at the University of British Columbia' UBC C

& W Research Ethics Board certificate H06-70467 and Vancouver Coastal Health Authority Research Study #V09-0129. These samples were frozen at the time of their collection and stored at -120 for long-term storage and then at -80 prior to collection. A table detailing the human samples used, age, sex, CAG size on both alleles, post-mortem index (PMI) is given in Table 5.1. Age of Onset relative to CAG size was determined using the methods provided in Langbehn et al 2004. Both the mouse and human samples were then processed for RNA extraction using the protocol detailed in the the PureLink® RNA Mini Kit (Invitrogen) with the following modifications: 1) Tissue homogenization was achieved using a Fastprep Homogenizer (Thermo Scientific). 2) In order to increase RNA yield and purity DNase was used to degrade any residual genomic DNA in the prep column, this was done using the the PureLink® DNase Set (Invitrogen) and the protocol for this set as described in the detailed users manual for the PureLink® RNA Mini Kit (Invitrogen). The concentration and purity of RNA was assessed using a nanodrop spectrometer (Thermo Scientific). Reverse transcription was performed using the e SuperScript® VILO™ cDNA Synthesis Kit (Invitrogen). Quantitative analysis of mRNA expression was performed using FastSYBR®green master mix according to the manufacture's instructions (Applied Biosystems). Amplification of cDNA was performed using the StepOne Plus Real-Time PCR System (Applied Biossytems). Primers used are provided in Table 5.2. Quantification of mRNA levels was calculated using the standard curve method using 10-fold serial dilutions comprised of a portion of each sample used in the study. Normalization of the quantified mRNA levels was

accomplished using a normalization factor generated by the GeNorm program provided in the qBase® software package (Biogazelle). The normalization factor was generated for each sample using amplification of a series of control genes in separate wells.

Table 5.1 Human Samples

Sample ID	CAG Size (HD first if applicable)	Sex	Age	Age of Onset	Age of Onset Relative to CAG size	Symptomatic HD Years	PMI (H)	Used in 450K DNA Methylation Array	Used in *HTT* Expression RT-1PCR	Used in Pyrosequcing
COB 05	19/20	M	75	NA	NA	NA	10.8	Y		
COB 20/30	17/20	M	74	NA	NA	NA	6.25	Y		
COB 22/52	18/23	F	77	NA	NA	NA	12	Y		
COB 51	17/23	M	54	NA	NA	NA	12.5	Y		
COB 59	17/19	M	21	NA	NA	NA	8.5	Y		
COB 125	15/17	M	74	NA	NA	NA	2	Y		
HDB 119	41/17	M	74	62	Late	12	3	Y	Y	Y
HDB 159	42/21	M	69	54	Mean	15		Y		
HDB 162	42/15	F	69	46	Early	23	7	Y		
HDB 165	42/31	F	71	37	Early	34	15	Y		
HDB 166	43/17	F	72	61	Late	11	8	Y	Y	Y
HDB 167	50/23	M	52	35	Late	17	15	Y	Y	Y
HDB 176	62/19	M	29	23	Mean	6	3.5	Y	Y	Y
Mean	**Control= HD =**	**M=9 F=4**	**62.4**	**45.4**		**16.9**	**8.6**			
HDB 129	48	M	52	42	Mean	10	48		Y	
HDB 156	51	M	26	pre-symptomatic	pre-symptomatic	pre-symptomatic	48		Y	
HDB 175	53	M	39	24	Mean	15	3.5		Y	
HDB 178	44	M	68	47	Mean	21	14		Y	
Mean	**49**	**M=4**	**46.3**	**37.7**		**15.3**	**28.4**			

Table 5.2 RT-qPCR Primers

Human Primers	Forward	Reverse
HTT	TCCACCATGCAAGACTCACTTAG	TGGGATTTGACAAGATGAACGT
ActB	AGTACTCCGTGTGGATCGGC	GCTGATCCACATCTGCTGGA
RGAG4	GGACAGCGCCCAACATTG	CTGGCTACCCTTTAGGCAACA
Ddah1	TTTAAGGACTATGCAGTCTCCACAGT	AGCCATGCTGCAGAAACTCTTC
HPRT	TTATGGACAGGACTGAACGTCTTG	GCACACAGAGGGCTACAATGTG
PGK1	CAAATGGAACACGGAGGATAAAG	CTTTACCTTCCAGGAGCTCCAA
Mouse Primers	**Forward**	**Reverse**
Hdh	CATCCTGGAAGCCATTGCA	TTTGTATATCTGAGTCTACTTCCTCCTTTC
ActB	CCAGCCTTCCTTCTTGGGTAT	TGTGTTGGCATAGAGGTCTTTACG
PGK1	CCCCAAGTGGAGGGAAGTACA	TGCCCAGCCGATAGACATC
HPRT	CGTCGTGATTAGCGATGATGA	TCCAAATCCTCGGCATAATGA
18S	AGAAACGGCTACCACATCCAA	GGGTCGGGAGTGGGTAATTT
Csnk2a2	CCACATAGACCTAGATCCACACTTCA	AGGTGCCTGTTCTCACTATGGATAA

5.2.2 Illumina 450K Methylation Bead Array

DNA was isolated using the Quiagen DNeasy Blood & Tissue kit as per manufacture's instructions. In short, tissue samples of equivalent size were taken from the UBC HD biobank and homogenized while frozen using a handheld homoginizer before being processed using the kit protocol. Sample yield and purity was assed using Nanodrop N-1000 (Thermo Scientific). 750 ng of DNA was bisulfite converted using the Zymo Research EZ DNA Methylation Kit (Zymo Reserach). After bisulfite conversion, 160 ng of DNA was applied to the Illumina 450K Methylation Array, as per manufacture's protocols (Illumina).

5.2.3 Pre-processing and Normalization of 450K Methylation Array Data

Pre-processing of the raw data generated from the 450K Methylation Array was conducted and resulted in the elimination of 48,542 probes, leaving 437,035 probes for further analysis. Table 5.3 provides a detailed breakdown of the probes eliminated in pre-processing. Normalization of the remaining probes was done in R (Pinheiro et al. 2014) through colour correction background adjustment using control probes contained on the 450K array.

Table 5.3 Probes Removed During Pre-Processing

Reason for Probe removal	Probes Removed
p-values > 0.05, missing beta values, less than 3 beads	6,144
probes on sex chromosomes	11,648
SNP probes	65
polymorphic probes	20,150
probes with non-specific binding to X or Y chromosomes (Price et al. 2013)	10,535
Total	48,542
Remaning Probes	*437,035*

5.2.4 Sub-setting of 450K Methylation Array Dataset by Research Question and Further Normalization

My first question examined HD and control cortex tissues for both global and *HTT* locus specific DNA methylation changes. To address this question, a “cortex only dataset” was generated containing only HD and control cortex tissue samples. My second question addressed methylation changes between high and low *HTT* expressing tissues, cortex and liver. To address this question, a “matched dataset” was generated which contains only cortex and liver samples matched between individuals. Once both of my datasets were generated each was individual subjected to quantile normalization using the lumi R package followed by subset-within-array normalization (SWAN) (Du et al. 2008; Maksimovic et al. 2012) Figure 5.1A depicts a pipeline diagram of my pre-processing and normalization strategies (including those utilized below) for greater clarity. Figure 5.1B depicts the distribution of my cortex samples across my two 450K chips. As the samples used in the cortex only dataset were not run on the same chip ComBat was used to remove chip-to-chip effects (Johnson et al. 2007). The samples used in the matched dataset were all run on the same chip, as such removal of chip-to-chip effects was not necessary.

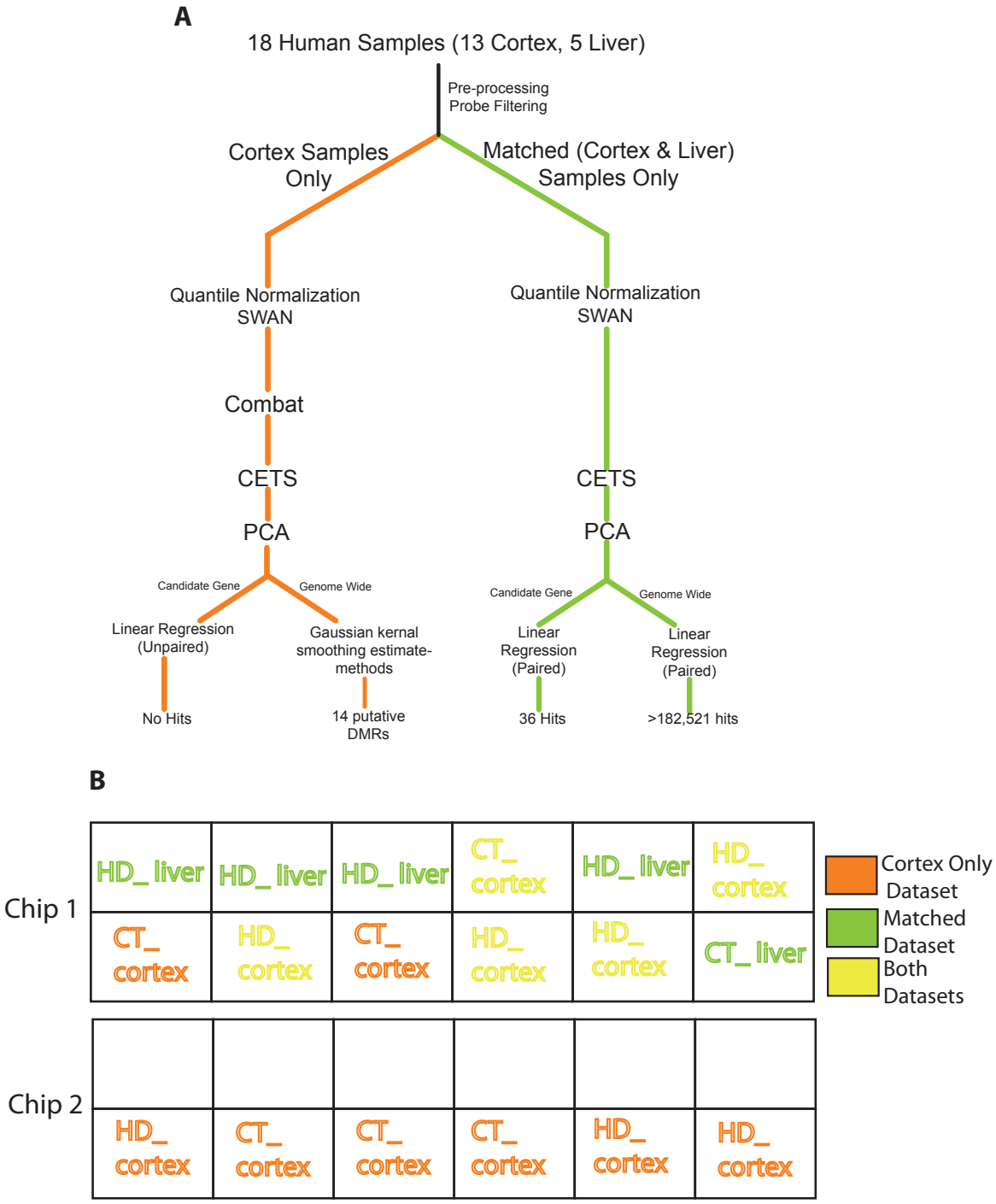


Figure 5.1 Schematic Pipeline of 450K Methylation Array Data Normalization & Analysis

(A) In order to ensure the correct normalization of the data based upon separate research questions the dataset was duplicated after pre processing and probe filtering. The orange portion of the flow chart depicts the normalization steps taken in the cortex only dataset while the green portion depicts the normalization of the matched (cortex-liver) dataset. (B) This schematic depicts the arrangement of the samples across two 450K methylation bead array chips.

5.2.5 Principal Component Analysis and Neuron/Glia Cell-Type Correction

Principal Component Analysis (PCA) decomposes the measured methylation patterns into a set of linearly independent principal component (PC) patterns that are ranked according to how much variance in the data they represent. The top ranked PCs can often be correlated with known traits in the cohort, such as tissue type, cellular composition, or disease state (Jones et al. 2013). To investigate the effect of cell type, particularly in the cortex samples included in both datasets, I utilized the CETS algorithm which estimates neuron versus glia proportions in brain tissue samples using methylation profiles of 10,000 brain cell-type specific 450K probes (Guintivano et al. 2013). In the cortex only dataset in particular, subsequent PCA of the dataset following CETS analysis revealed that neuron/glial proportions was significantly associated with the top ranking PC (accounting for 24.1% of the variance in the dataset). To account for changes in DNA methylation due to differences in inter-individual brain cell type composition, a multiple linear regression model was built with different cell components (ie. percentage of neurons and glia, respectively) for each of the 450K probes for the cortex samples in both the cortex only and matched datasets (Lam et al. 2012). The residuals for each regression model were applied to the mean value of each data series to obtain the “corrected” methylation data. PCA was subsequently used to check that the correlation of the cell-type components was minimal, in the cortex only dataset subsequent to CETS correction cell-type composition correlated with PC12, accounting for a negligible 0.64% of the variance in the dataset.

5.2.6 Differential Methylation Analysis

All statistical analysis on normalized and corrected data was performed using R statistical software (version 3.0). Probes with DNA methylation levels significantly different between HD cases and controls were identified first using the R lima package's moderated unpaired t-tests with empirical Baysian variance method (Smyth 2004). In the matched dataset, differentially methylated probes between cortex versus liver samples were identified using paired analysis with moderated t-statistics estimated by empirical Bayesian modeling in limma. Benjamini-Hochberg correction to control the false discovery rate at 0.05. All statistical analysis was performed on transformed M-values (Du et al. 2010).

5.2.8 Pyrosequencing

Genomic DNA was isolated from the samples used in the 450K array using the same method. The DNA was bisulfite convered using the Zymo Research EZ-DNA Methylation™ kit. PyroMark Assay Design 2.0 (Qiagen) software was used to design the bisulfite pyrosequencing assays, primers listed in Table 5.4. HotstarTaq DNA polymerase kit (Qiagen) was used to amplify the target region using the biotinylated primer set. PCR conditions were as follows: 5 minutes at 95°C, 45 cycles of 95°C for 30s, 58°C for 30s, and 72°C for 30s, and a 5 minute 72°C extension step. The amplicon was then electroplated on an agrose gel for confirmation of both the presence and quality of the product. Streptavidin-coated Sepharose beads were used to bind the biotinylated strand of the PCR product,

these were then washed and denatured to yield single stranded DNA. Sequencing primers were introduced to allow for pyrosquencing (Pyromark™ Q96 MD pyrosequencer, Qiagen). Pyro Q-CpG software (Qiagen) was used to generate quantitative methylation levels of the targeted CpG dinucleotides of interest.

Table 5.4 Pyrosequencing Primers

Assay 1	
Target Sequence Forward Primer	TGGATGTTTTGATGAAGTTAGTTGTTATGT
Target Sequence Reverse Primer	CCCAACTTAACCAACTCCACTT (Biotin)
Sequencing Primer	GTTATGTTGGAGAGGT
Assay 2	
Target Sequence Forward Primer	TTGATGGGGAGGTTAATTGT
Target Sequence Reverse Primer	ACTTCCTAACTCCTACTATACACT (Biotin)
Sequencing Primer	GAAATAGGAAAAGAGAGATTATTAA
Assay 3	
Target Sequence Reverse Primer	TATAGGTGTAGGGTTTAGTAGTGAGTAGAT (Biotin)
Target Sequence Reverse Primer	CTAACATTTCCCTATCCCCTTCC
Sequencing Primer	CCCTACTTTAAAATTCCTC

5.3 Results

5.3.1 *Hdh* is Stably Expressed in Brain Tissues Between Individuals

In the 2004 Dixon paper the researchers endeavoured to separate the variability in endogenous mouse *Hdh* expression from experimental variability, namely in efficiencies in cDNA synthesis. To do this they established a testing paradigm wherein the variability contributed by experimental variation, referred to as coefficient of variation (CV_{exp}), could be subtracted from the total variation between individuals, referred to as total coefficient of variation (CV_{total}). Dixon et al., 2004 reported a CV_{exp} of 14% and a CV_{total} of 30% indicating that over half of the variation within the study was a result of inter-individual variability. In order to establish a similar degree of variability between our own mice strains, I first repeated the Dixon et al., 2004 experiment utilizing whole forebrain samples from two sets of 4 wild type FVB male littermate mice. As described in the methods section, I utilized a gene normalization method to normalize my results (Figure 5.2A). I found a CV_{exp} of 5.4% and a CV_{total} of 6.3%, indicating that inter-indivdual variability comprised only 0.9% of the total variability seen, a considerably different result from the Dixon paper. Further consideration into the methods used in the Dixon paper revealed their use of a single normalization gene, 18S, compare to my use of a normalization factor based on several reference genes. To address this I re-normalized the data to 18S alone (Figure 5.2B) finding a CV_{exp} of 16.8% and a CV_{total} of 27.8%. This result is more in line with the results reported in the Dixon 2004 paper suggesting that the variability reported by the researchers was due to variation in their normalization gene not in

Hdh expression. Furthermore, when I normalized 18S gene expression using my established normalization factor (Figure 5.2C) I found a CV_{exp} for the 18S gene of 11.03% and a CV_{total} of 18.24%. Indicating that 7.21% of the variability is due to inter-individual differences, much greater than the 0.9% found with the *Hdh* gene. This indicates 18S is differentially expressed between individuals and thus a poor reference gene for the analysis of *HTT* expression data.

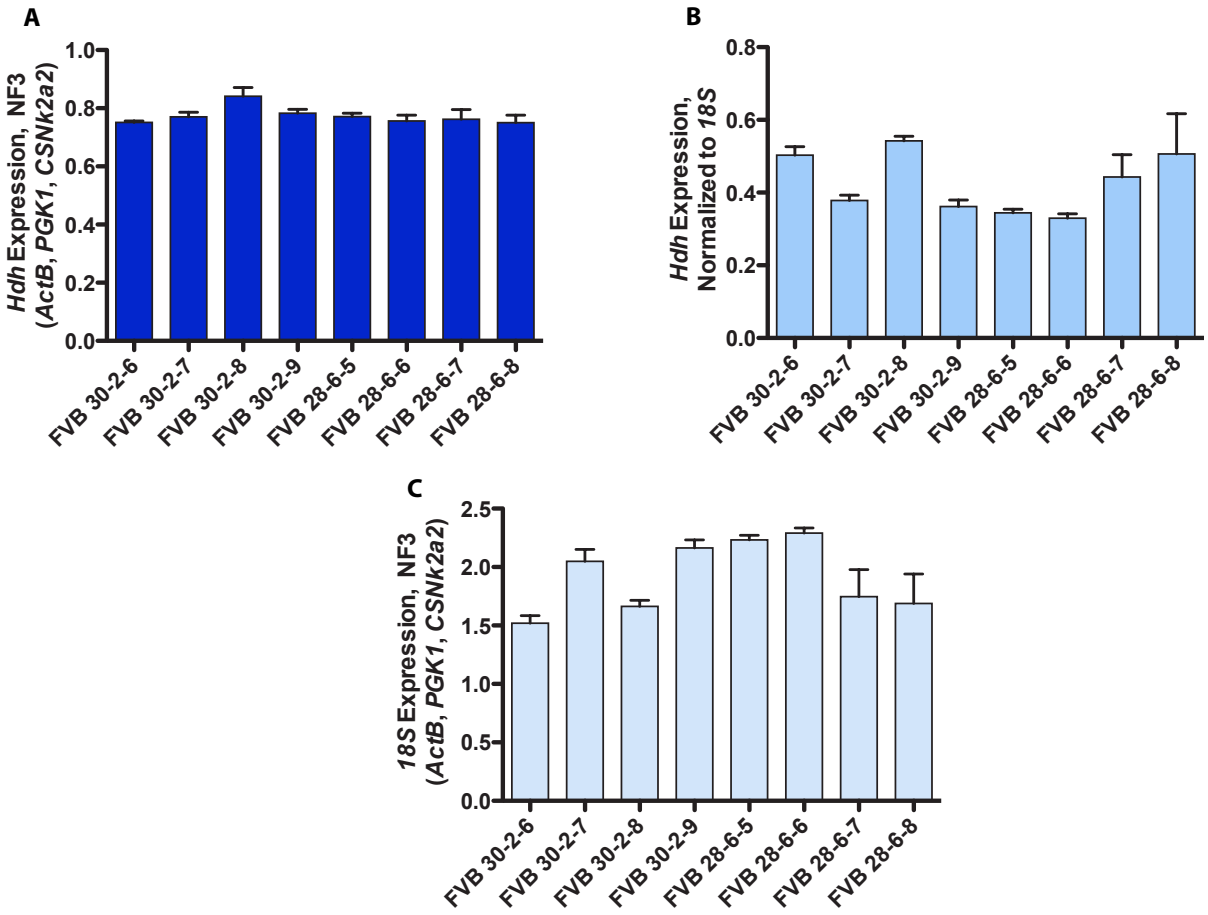


Figure 5.2 Hdh Expression in WT FVB Forebrain
(A) *Hdh* expression in WT, FVB forebrain samples, qRT-PCR, normalized to a normalization factor using three normalization genes *ActB, PGK1, CSNK2a2* as described in the methods section 5.2.1. (B) The same samples as (A) normalized to 18S alone. (C) 18S expression in the same samples as (A) normalized to the normalization factor of 3 used in (A) .

5.3.2 *Hdh* is Expressed at Higher Levels in Brain & Testes Compared to Liver

The Dixon et al. 2004 paper, as well as several other publications, (Van Raamsdonk et al. 2007; Li et al. 1993) have established a pattern of differential expression of both *Hdh* and *HTT* in protein and mRNA levels. To confirm these results in my mouse and human samples, I first examined the mRNA expression difference between cortex, liver, and testes in 11 wildtype, FVB, male mice (two sets of 4 litter mates each and one set of 3 littermates) (Figure 5.3). As expected, I found *HTT* expression to be highest in the cortex samples, followed by the testes, and lowest expression in the liver.

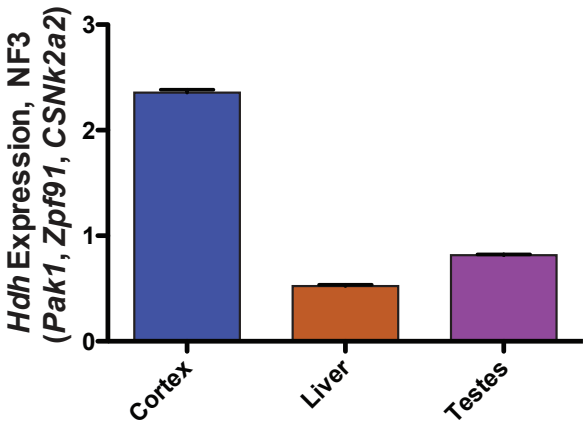


Figure 5.3 Hdh Expression in FVB Cortex, Liver and Testes
Hdh expression in cortex, liver, and testes tissues from 11 male WT FVB mice (two sets of 4 litter mates and one set of 3 litter mates) qRT-PCR, Normalized to a normalization factor using three normalization genes *Pak1, Zpf91, CSNK2a2* as described in the methods section 5.2.1.

5.3.2 Human Samples Used for DNA Methylation Analysis

My study addressed two separate yet related questions with regards to DNA methylation changes in HD, because of this it was important to parse my 450K Methylation Array dataset to ensure the correct normalization methods were used to address each question. Figure 5.4 provides a Venn diagram depicting the human HD and control, cortex and liver samples used and their distribution in my parsed datasets. Table 5.4 provides further details of the individuals used in each dataset.

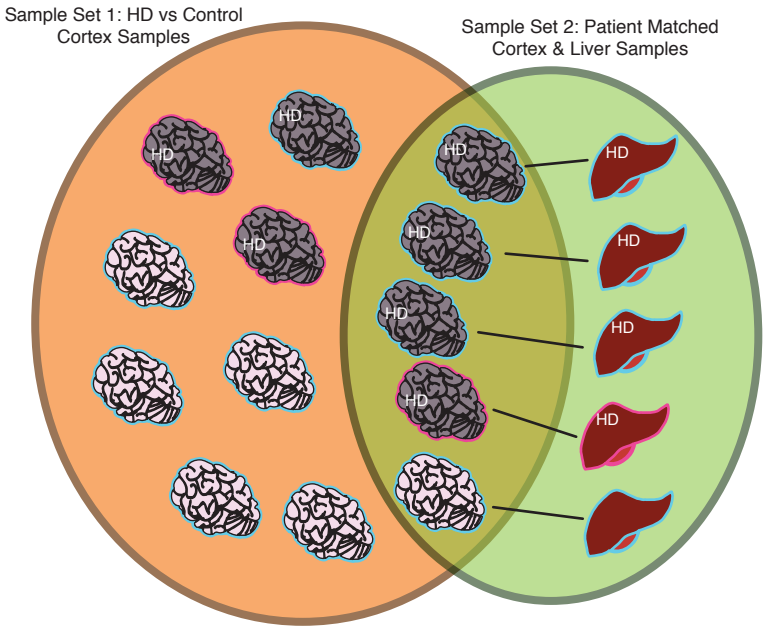


Figure 5.4 Ven Diagram of Distribution of Human Samples in Dataset Generation
Ven Diagram depicting the distribution of the human samples used in the creation of the cortex only and matched datasets.

Table 5.5 Human Sample Information By Dataset

Cortex Samples	N = 13
Sex (M/F)	10/3
Disease status (HD/Control)	7/6
HTT CAG length (avg. repeats)(HD/control)	46/17.33
WT CAG length (avg. repeats)	20.62
Age at Death (avg. yrs)	62.23*
Age at Onset (avg. yrs)	44.71

Matched Samples	N = 5 Liver, N = 5 Cortex
Sex (M/F)	4/1
Disease status (HD/Control)	4/1
HTT CAG length (avg. repeats) (HD/control)	48.75/17
WT CAG length (avg. repeats)	19.4
Age at Death (avg. yrs)	55.6*
Age at Onset (avg. yrs)	40.25

5.3.3 Global Methylation Changes Between HD and Control Cortex

As described in the Methods section, following pre-processing, SWAN, ComBat, and colour correction, I utilized a cell type correction algorithm, CETS, which allowed me to estimate the neuron versus glial proportions present in the individual cortex samples of my cortex only dataset (Guintivano et al. 2013). This algorithm utilizes 10,000 probes found on the 450K array that are known to have differential methylation between neuronal and glial cell type populations. Figure 5.5 details the neuronal proportions in control versus HD tissues, while there is not a significant difference in neuronal proportions between the two groups, there is a trend towards a decrease in neuronal proportions in the HD cortices compared to control. Conversely, there is a trend towards an increase in gliosis in the HD samples as compared to control (data not shown). This is not unexpected as neuronal loss and gliosis are know to occur outside of the striatum in late stages of HD, including in the cortex. Given that HD progression from the initial diagnosis of symptoms to patient death is ~20 years (Rinaldi et al. 2012) and the average length of time between age of onset and age of death in my samples in this dataset is 15.3 years, it is likely that the majority of the individuals included in my study were in late stages of HD at the time of their death. This would explain the differences in neuronal versus glial cell type proportions in my cortex samples. As detailed in the methods section I generated a linear regression based model to correct for individual differences in cell type proportions within my samples. The probes that were utilized by the CETS

algorithm to estimate the neuronal versus glial proportions were removed from further analysis.

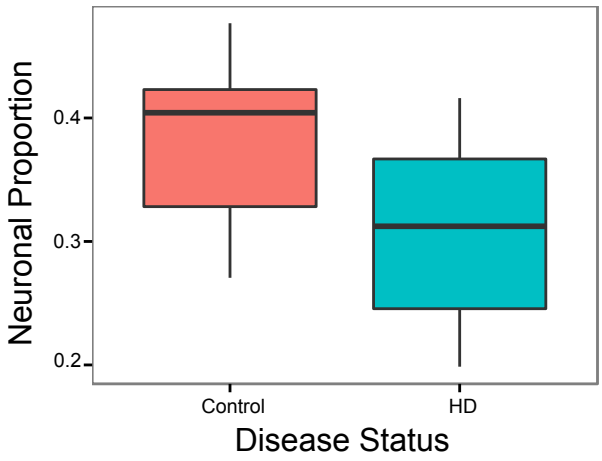


Figure 5.5 Neuronal Proportions in Human Cortex Samples
The CETS algorithm, which utilizes 10,000 probes on the 450k methylation array known to be differentially methylated between neuronal and glial cell types was used to estimate the proportions of these cell types in the cortex samples used. Neuronal proportions were not significantly different by student's t-test.

5.3.4 DNA Methylation at *HTT* Locus is Unaffected by the HD Mutation

Using the same genomic coordinates as described in chapter 2, I assessed DNA methylation changes at the 86 CpG sites and a single non-CpG site within the *HTT* locus region represented by probes in the Illumina 450K Methylation Array. Using the statistical tests described in the methods section I assessed the methylation differences between HD and control cortex for each probe. When tested individually none of the 87 probes in this genomic region are differentially methylated between HD and control cortex samples (Figure 5.6). In addition, unsupervised hierarchical clustering with average linkage did not establish groupings between the HD and cortex samples.

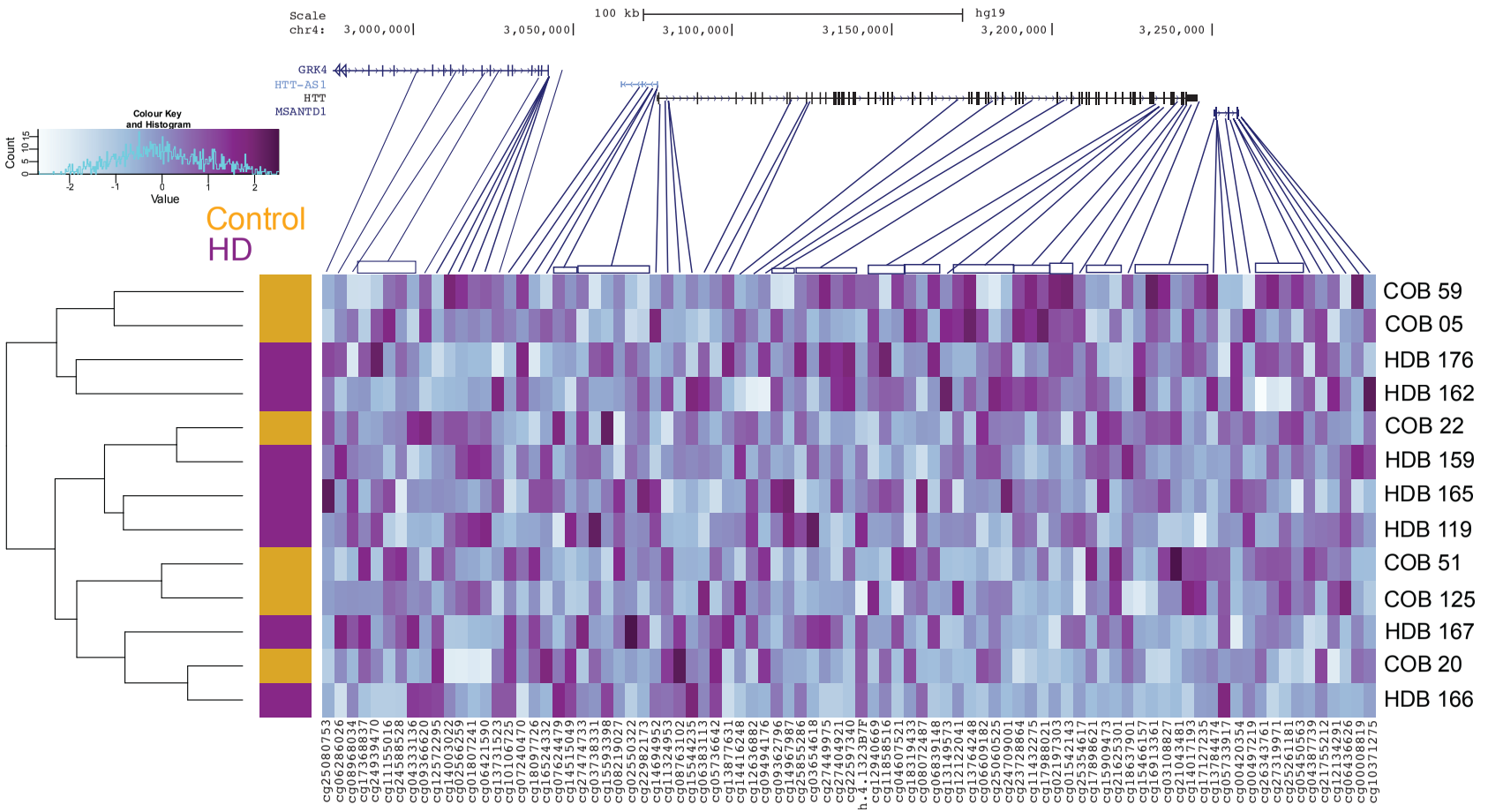

Scale
chr4:
3,000,000
3,050,000
3,100,000
3,150,000
3,200,000
3,250,000
100 kb
hg19
GRK4
HTT-AS1
HTT
MSANTD1
Colour Key and Histogram
Count
Value
Control
HD
COB 59
COB 05
HDB 176
HDB 162
COB 22
HDB 159
HDB 165
HDB 119
COB 51
COB 125
HDB 167
COB 20
HDB 166

Figure 5.6 DNA Methylation is Not Significantly Altered at the HTT Gene Locus Between HD and Control Cortex

Heatmap representing the 87 probes represented on the 450K methylation array in the genomic region screened in chapter 2. Features of the Genomic region are represented in the upper portion of the figure. None of the represented probes are differentially methylated between control and HD cortex. Unsupervised hierarchical clustering did not segregate samples based on disease status.

5.3.5 15 Potential Differentially Methylated Regions (DMRs) Identified Between HD and Control Cortex

Using the HD and control cortex dataset, a DMR searching algorithm, DMRcate, was used to identify potential DMRs across the genome (Peters et al. 2015). At a FDR of 0.01 and a threshold of max beta fold changes > 0.01, 14 potential DMRs were found (Figure 5.7) and are summarized in Table 5.5. At present this is only a preliminary analysis of global methylation changes between HD and control cortex, and further follow-up analysis is necessary.

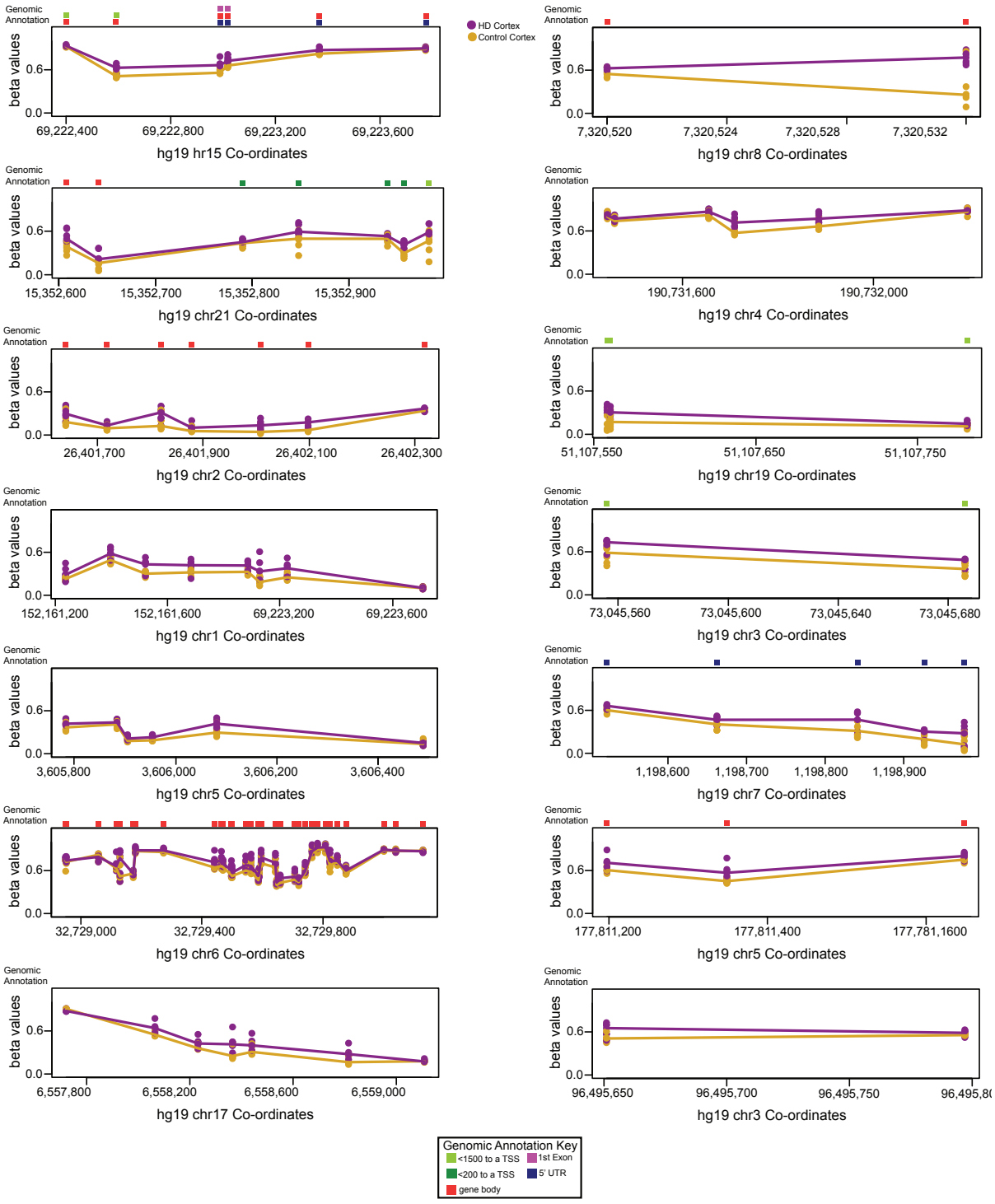


Figure 5.7 14 Identified DMRs Between HD and Control Cortex

Shown are the 14 identified DMRs between HD and Control Cortex as identified by the DMRcate algorithm (Peters et al. 215). Genomic context for each DMR is provided by the annotation key and accompanying boxes above each graph.

Table 5.6 DMRs Detailed Information

DMR Number	Genes Associated	hg19 Co-Ordinates	Number of Probes	Mean p Value across region
1	MIR548 h4, SPESP1, NOX5	chr15:69,222,400-69,223,774	6	0.000392022
2	C21orf81	chr21:15,352,608-15,352,983	7	0.00000000297
3	FAM59B	chr2:26,401,640-26,402,318	7	0.00000148
4		chr1:152,161,237-152,162,507	8	0.000718683
5		chr5:3,605,783-3,606,489	6	0.00165268
6	HLA-DQB2	chr6:32,728,951-32,730,130	33	0.000308365
7		chr17:6,557,720-6,559,109	7	0.001612758
8	SPAG11B	chr8:7,320,520-7,320,532	2	0.000262143
9		chr20:25,834,845-25,834,845	1	0.00111286
10		chr4:190,731,443-190,732,196	6	0.003839515
11	SNAR-F	chr19:51,107,558-51,107,781	3	0.004880532
12	PPP4R2	chr3:73,045,556-73,045,686	2	0.006150947
13	ZFAND2A	chr7:1,198,522-1,198,977	5	0.006515939
14	COL23A1	chr5:177,811,198-177,811,648	3	0.008242616

5.3.5 The *HTT* locus is Differentially methylated at 32 CpGs between Cortex and Liver Samples

To assess *HTT* locus differences between cortex and liver I used my matched dataset (I again refer the reader to Figure 5.4 which establishes the samples used in this dataset). As with the cortex only dataset, a CETS correction was performed to correct for cell proportion differences in the cortex samples (data not shown). These cell proportion differences were corrected using the methods described above. Using the same 87 probes as above, the methylation differences between cortex and liver tissues were assessed for each probe (Figure 5.8). 32 probes were identified as being significantly differentially methylated between cortex and liver tissues, highlighted in red in Figure 5.8. Unsupervised hierarchical clustering with average linkage also clustered the samples based on tissue type for this region, indicating that these regions were differentially methylated between cortex and liver tissues. Figure 5.9 provides Delta beta differences (cortex-liver) for the probes in this region. Interestingly, the majority of the differentially methylated probes are found within the gene body, primarily towards the 3' end of the *HTT* gene. This suggests that regions within the 3' end may be involved in differential expression between liver and cortex tissues. These probes also have a similar methylation pattern as probes in the preceding *MANSTD1,* the gene following *HTT*. As mentioned in Chapter 2, considerably little is known about the *MANSTD1* gene, until recently it was designated as C4orf44. It is possible that this gene is also differentially expressed in human liver and cortex samples. Of the probes within the proximal promoter region

only two are differentially methylated. This suggests that the proximal promoter itself may not be regulated by direct differential methylation.

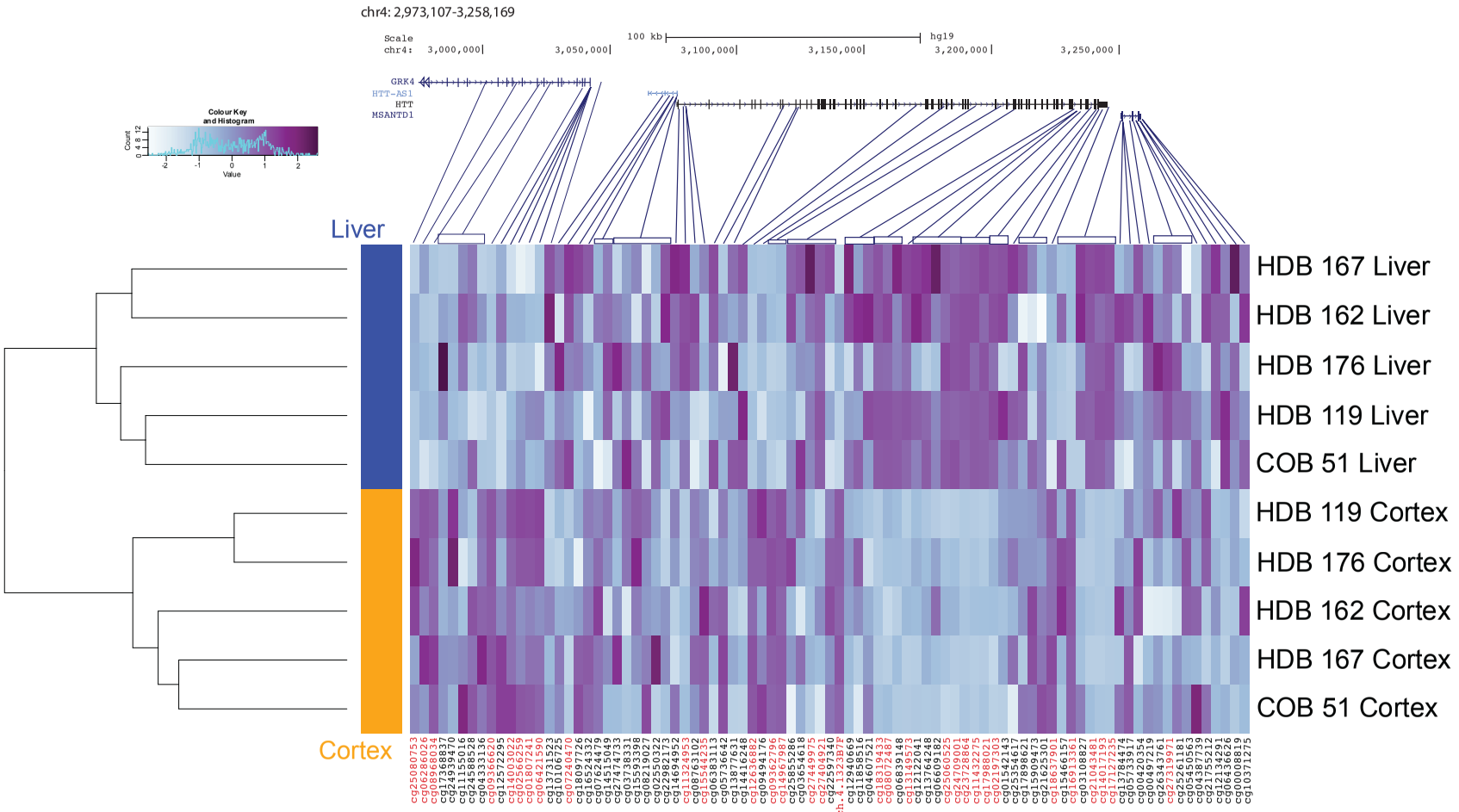
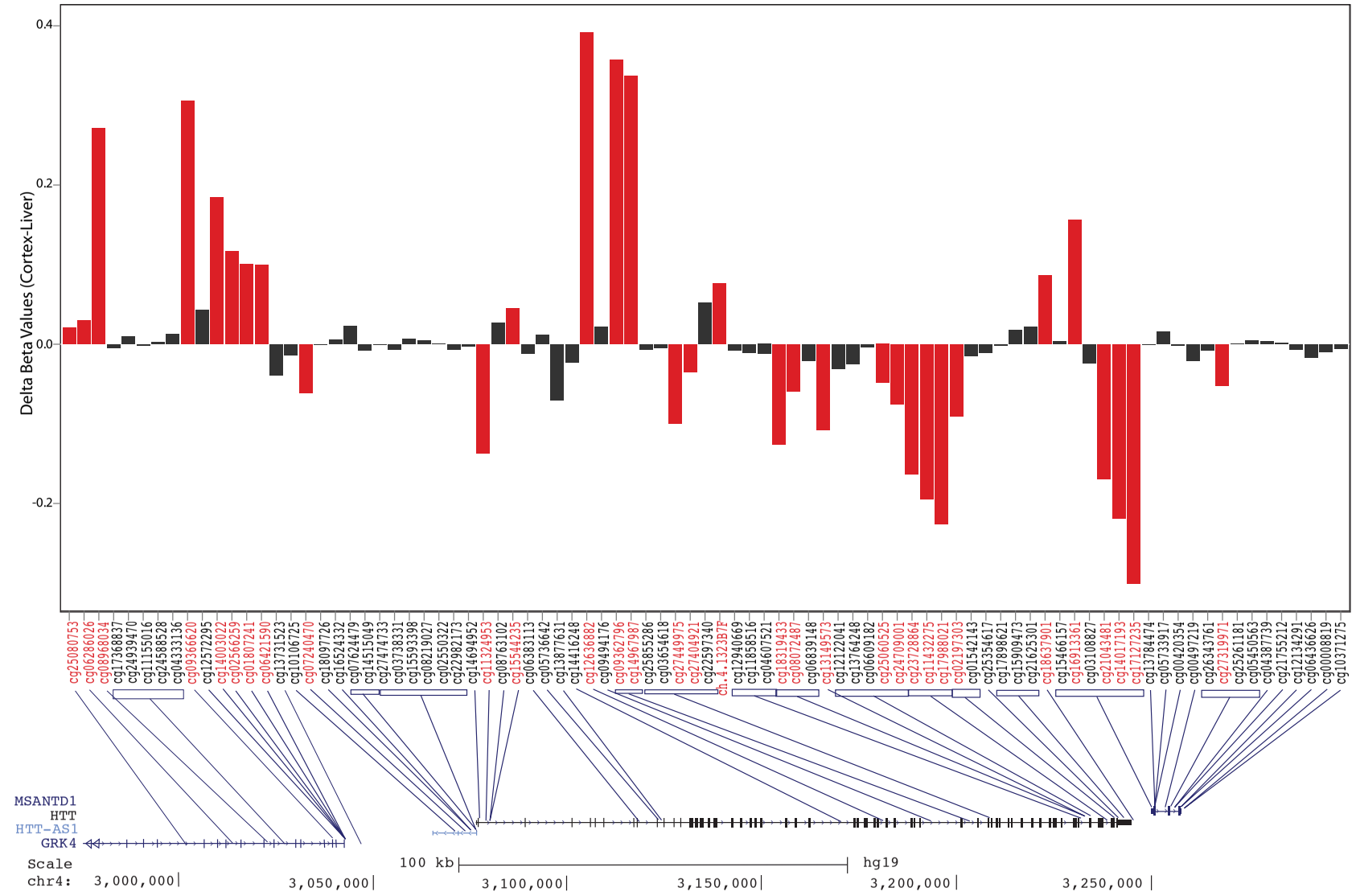

chr4: 2,973,107-3,258,169
Scale
chr4:
100 kb
hg19
3,000,000
3,050,000
3,100,000
3,150,000
3,200,000
3,250,000
GRK4
HTT-AS1
HTT
MSANTD1
Colour Key and Histogram
Count
Value
Liver
Cortex
HDB 167 Liver
HDB 162 Liver
HDB 176 Liver
HDB 119 Liver
COB 51 Liver
HDB 119 Cortex
HDB 176 Cortex
HDB 162 Cortex
HDB 167 Cortex
COB 51 Cortex

Figure 5.8 36 Probes Identified in the HTT Gene Locus Region as Differentially Methylated Between Liver and Cortex Samples

Heatmap representing the 87 probes represented on the 450K methylation array in the genomic region screened in chapter 2. Features of the Genomic region are represented in the upper portion of the figure. 32 of the probes are differentially methylated between matched cortex and liver samples, probes are highlighted in red at the bottom of the figure. Unsupervised hierarchical clustering successfully segregated samples based on tissue type, suggesting that this genomic region is indeed a DMR between cortex and liver.



chr4: 2,973,107-3,258,169

Figure 5.9 Delta Beta Values (Cortex- Liver)
Delta Beta values (cortex- liver) for each probe in the *HTT* gene region as represented in detail in figure 5.8. The same genomic representation schematic as 5.8 is shown at the bottom of the figure to orientate the reader. Probes that are significantly differently methylated between cortex and liver are again shown shown in Red.

5.3.4 Validation of 450K Results using Pyrosequencing

To validate the 450K cortex and liver dataset, additional samples were processed from each individual used in the matched dataset for use in pyrosequencing assays. I selected three probes from the 450K analysis that were differentially methylated between cortex and liver, cg15544235, cg0720470, and cg11324953 (Figure 5.7). In the 450K array I used the CETS method to normalize each sample based upon neuronal versus glial proportions. As the DNA used in the pyrosequencing assay was from new genomic extractions from additional samples, although from the same individuals, I was unable to perform normalization for cell type proportions. Despite this, the pyrosequencing assays showed a strong correlation with the values generated for the same CpG in the 450K array. This indicates that the 450K methylation array data directly reflects genomic methylation levels.

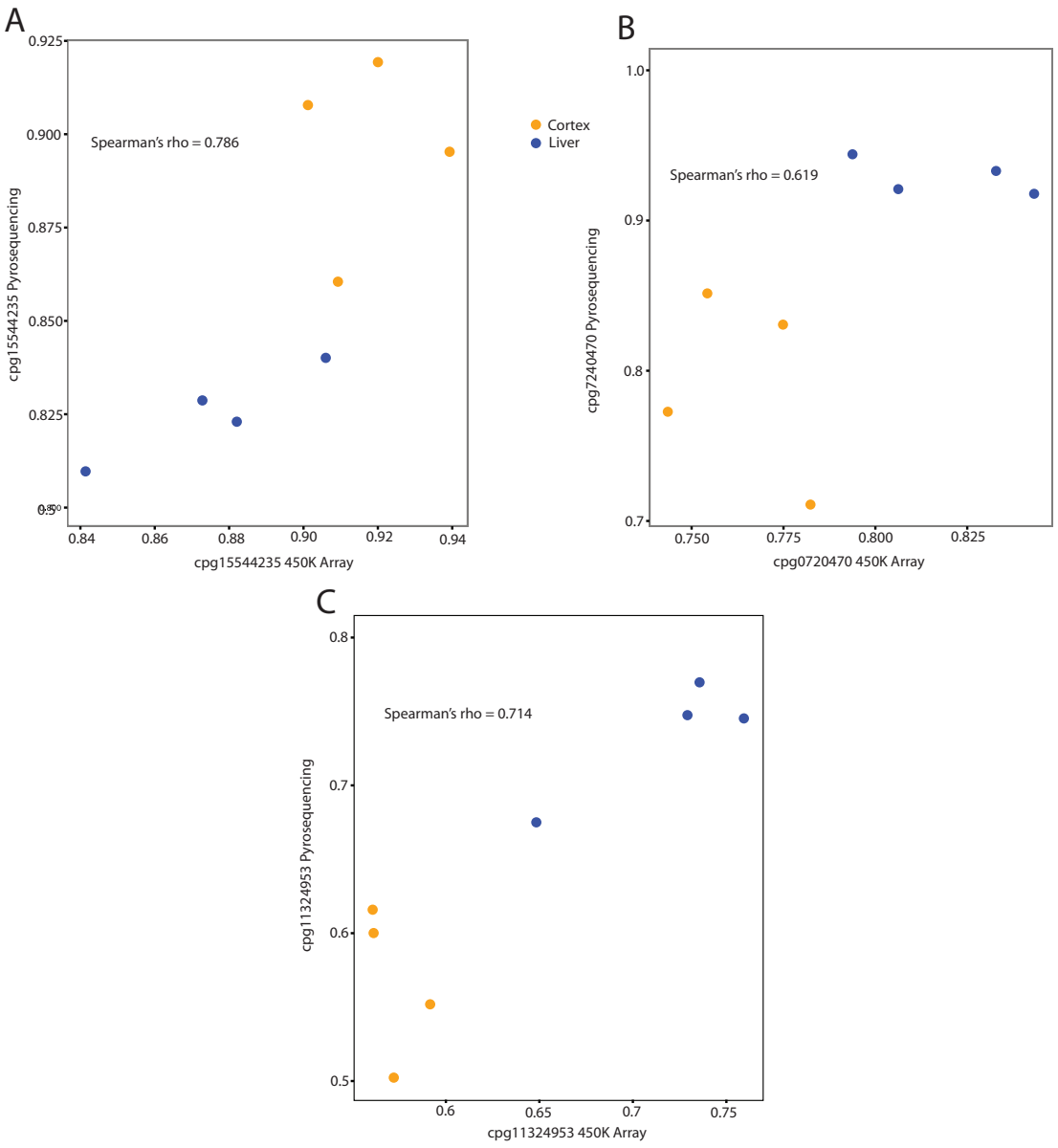


Figure 5.10 Pryosequencing Validation of Selected 450k Probes Between Cortex and Liver Samples

Using additional gDNA samples from the same individuals used in the 450k methylation array pyrosequencing was performed on 3 of the differentially methylated probes.. Plotted are the methylation values from the 450k array on the X-axis and the methylation values from pyrosequencing on the Y-axis. Spearman's rho values for each graph are provided in the figure.

5.3.5 Expression of *HTT* in Human Cortex and Liver Samples

To establish if the differentially methylation of the probes within the *HTT* locus correlate with differences in *HTT* expression, I first determined *HTT* expression within my samples using RT-qPCR as described in the methods section. Unlike the mouse *Hdh* expression analysis, I anticipated a considerable amount of variation in *HTT* expression between the individuals in this study for two reasons. Firstly, unlike the inbred FVB, littermate mice, I anticipated genetic and environmental differences between individuals would contribute to differences in *HTT* expression. Secondly, due to the nature of human sample donation, it can be difficult to obtain samples consistently from the same region of an organ between individuals. In my mouse study all the dissections were performed by the same technician, with PMI of less than 20 minutes, and I could be certain that the samples obtained were representative of the same geographic location of an organ between individuals. While all of my human samples were from either frontal cortex, liver, or testes, I could not be certain of the exact location within each organ the samples were generated from. Dissection of the human tissue was also done by different individuals, with varying PMI. I anticipated that all of these factors would contribute to an increase in *HTT* expression variability between individuals. To test this I obtained 3 RNA samples per tissue per individual in order to performed a similar test of variability between individuals within a tissue type as was done by the Dixon 2004 group and with my own mice, above. Figure 5.11 shows a comparison between all individuals between all tissues as well as a comparison between individuals within a tissue. In my human samples I found a CV_{exp} - 12.4%, CV_{total} - 16.4% for cortex,

CV_{exp} - 8.5%, CV_{total} - 29.3% for liver, and CV_{exp} - 14%, CV_{total} - 29.8% for testes. As anticipated, I found more variation in my human samples than in the mouse samples. As in the FVB mice, I found comparatively little variation in *HTT* expression between individuals within the cortex, only 4% of the total variability was due to inter-individual differences. In both liver and testes tissues, I found considerably more variation, with experimental variation contributing to almost a third of the total variation in the liver and just over half of the variation in the testes. I hypothesize that the increase in variation within the testes tissue was primarily due to the considerable heterogeneity of the testes tissue itself. Interestingly, in my human samples I found that the expression of *HTT* within the testes was at least equal to, if not greater than, the expression in matched cortex tissue. I again suspect that the heterogeneity within the testes tissue may contribute, in part, to this effect. With the mouse testes samples, due to the size, I processed the entire testicle for each mouse, thereby generating results that represent the average *Hdh* expression across the testicle.

In addition to the 4 individuals for whom I had matching liver and testes tissues used in the 450K DNA methylation analysis, I was able to obtain and additional 4 HD patients for whom matching cortex, liver, and testes tissues was available. These individuals are listed at the bottom of table 5.1. Two of the individuals had PMIs of 48 h, much higher than the PMIs of the other individuals. Longer PMIs can result in an increase in RNA degradation, however I did not find a significant difference between the RNA extractions from these two individuals compared to the rest of the samples. One of the individuals with prolonged PMIs

was also a pre-symptomatic patient. I conducted the same *HTT* expression analysis as performed for the individuals above (Figure 5.12). Immediately apparent is the extremely high expression of *HTT* in the liver of HDB129, the level of expression in this individual was more than double that of their cortex or testes. The cortex and testes tissues from HDB129 were not significantly different from the other individuals, indicating that the variability seen in the liver was not inherent in the individual itself. Further research into the health of this individual revealed the presence of liver sclerosis secondary to a Hepatitis infection that the individual had incurred earlier in their life. Chronic inflammation of the liver may have resulted in the elevated *HTT* expression in this individual. Further study utilizing additional individuals with chronic liver inflammation, both with and without HD, would indicate if the observed increase in *HTT* expression is a consequence of chronic liver inflammation alone or a result of the presence of the HD mutation and chronic liver inflammation. While the presence of the HD mutation has been associated with increases in neuronal inflammation, an association to inflammation in peripheral tissues has not previously been observed (Hsiao & Chern 2010).

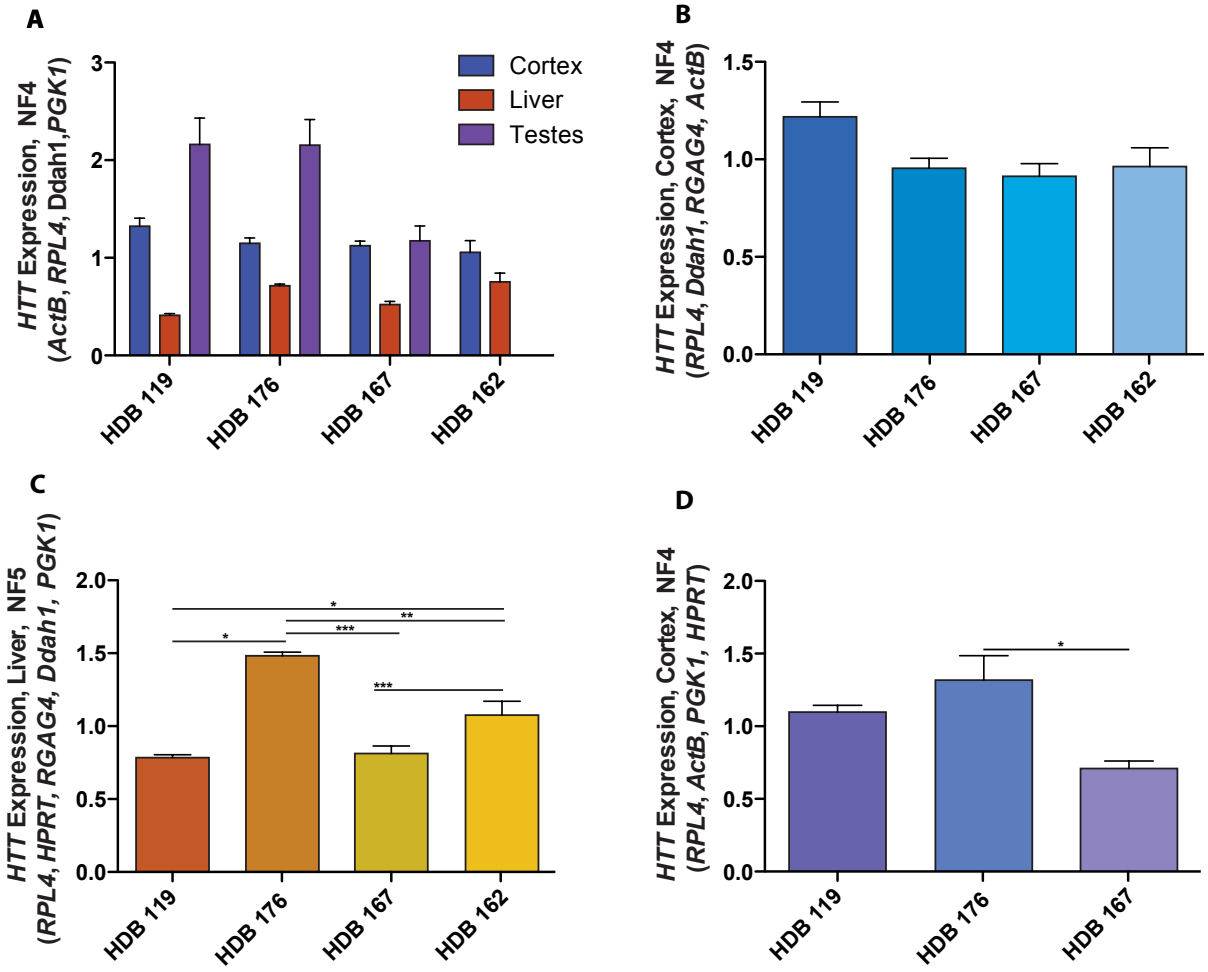


Figure 5.11 HTT Expression in Human Samples Used in 450K DNA Methylation Array

RT-qPCR analysis of *HTT* transcript levels in individuals utilized in the 450k methylation array, normalized by a normalization factor generated by at least 4 normalization genes as described in 5.2., normalization genes used are indicated on the Y axis of each graph. (A) Individual *HTT* expression in cortex, liver, and testes compared across individuals. Comparisons of *HTT* expression separated by tissue type (B) cortex, (C) liver, (D) Testes. Analysis in tissue specific graphs (B, C, &D) One-way ANOVA with Tukey post test, *p<0.5, **p<0.01, ***p<0.001.

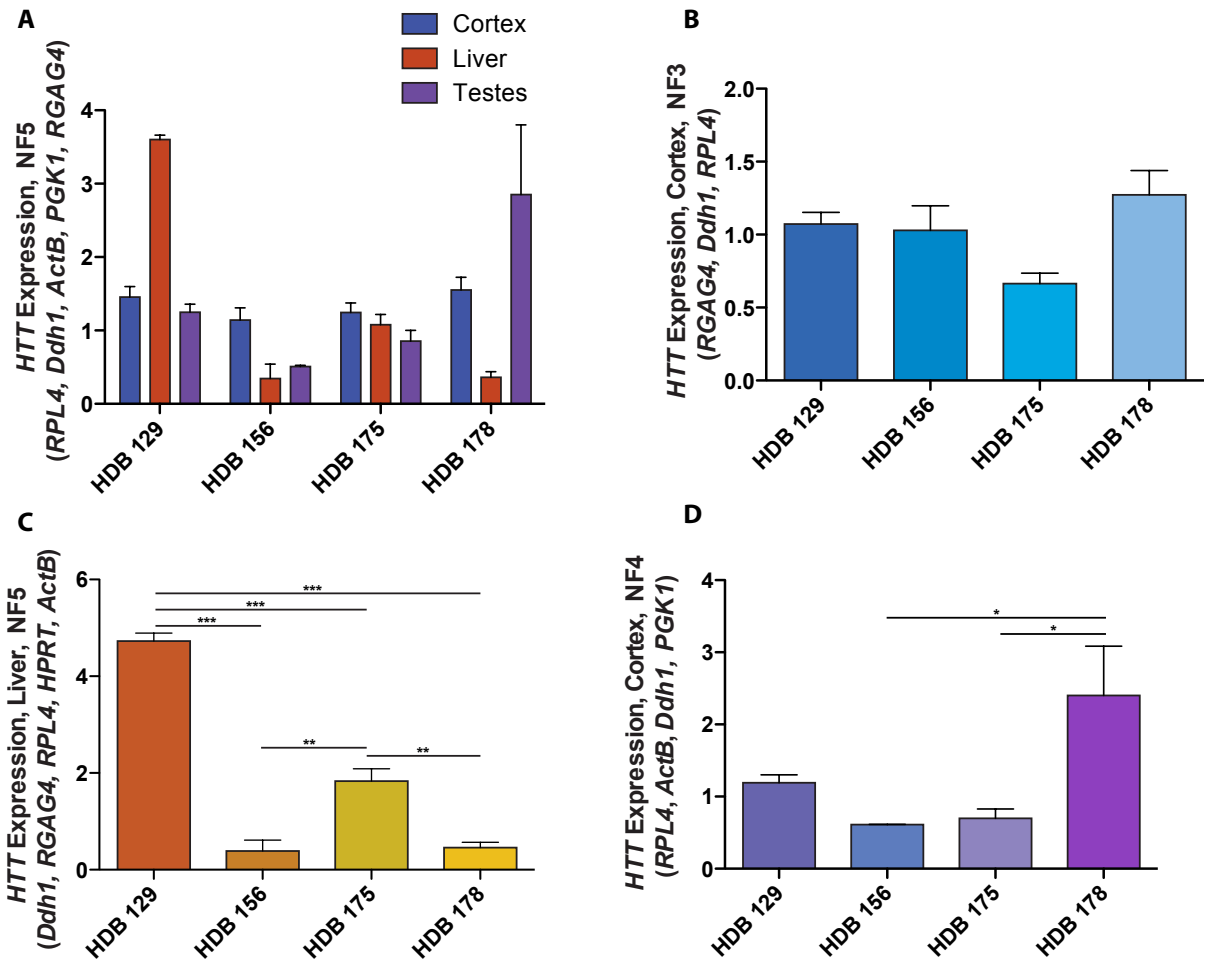


Figure 5.12 Additional Human Samples, HTT Expression
RT-qPCR analysis of *HTT* transcript levels in additional human samples not utilized in the 450K analysis, normalized by a normalization factor generated by at least 3 normalization genes as described in 5.2., normalization genes used are indicated on the Y axis of each graph. (A) Individual *HTT* expression in cortex, liver, and testes compared across individuals. Comparisons of *HTT* expression separated by tissue type (B) cortex, (C) liver, (D) Testes. Analysis in tissue specific graphs (B, C, &D) One-way ANOVA with Tukey post test, *p<0.5, **p<0.01, ***p<0.001.

5.4 Discussion

In this study, I present the first analysis of global methylation changes in human HD brain samples. I found that the CAG expansion on HD alleles does not impact local DNA methylation of the *HTT* gene. Fifteen putative DMRs were identified between HD and control cortex across the genome. Using matched cortex and liver samples I identified 32 CpG sites within the local *HTT* gene region that were differentially methylated between these tissues. I established that, contrary to previous reports, mouse *Hdh* levels do not vary between inbred individuals in cortex, liver and testes tissues. In line with the mouse data, I found that human cortex expression of *HTT* is relatively stable between individuals, although an increase in variability between individuals in both liver and testes tissues was observed.

I initially sought to replicate the results found in the Dixon 2004 paper in order to explore changes in endogenous mouse *Hdh* DNA methylation as a way to inform my human DNA methylation studies. I found, however, that the inter-individual differences seen by the Dixon group was likely due to their choice of normalization gene, 18S. When I utilized 18S as the sole normalization gene in my samples I found CV_{exp} and CV_{total} values comparable to those seen by the Dixon group. This inter-individual variability disappears when the data was normalized to a normalization factor based on the expression of 3 housekeeping genes instead of just 18S. The choice of appropriate reference gene(s) is vital to the correct analysis of RT-qPCR data (Vandesompele et al. 2002). There is a common misconception that once a housekeeping gene has been identified as being stably expressed, it can

be safely used in various tissues and research paradigms. This is not the case, many commonly used housekeeping genes can differ in expression and variability between tissues (Vandesompele et al. 2002). Caution must therefor be exercised when using housekeeping genes shown to be stably expressed in other tissues or treatment paradigms. To mitigate this risk in my study, I tested several housekeeping genes in each of my experiments and utilized a normalization program, geNorm, to assess the relative stability of each housekeeping gene before selecting the most stable to generate a normalization factor on a per-sample basis (Hellemans et al. 2007). Using multiple housekeeping genes to generate a normalization factor reduced the effect of different tissues and treatments compared to a single housekeeping used on its own. This strategy was especially important when I normalized data across multiple tissue types, as it is known that housekeeping genes, while stable within a tissue, can vary widely between tissues.

In comparison to my mouse samples, the human samples had considerably more variation within tissue types. I hypothesize that this result was due to increased heterogeneity of the human samples, both genetically and environmentally, compared to the inbred, littermate mice. Even so, I observed comparatively little variation between cortex samples of different individuals, and found greater variation in both the liver and testes samples. Due to the nature of human tissue bio-banking, I suspect that this variability (even within an individual) in testes samples was due to heterogeneity in cell composition in my samples both within and between individuals. In the mice I was able to assay the entire testicle, creating a sample which represented the average *Hdh* expression across the entire organ. As specific cell

types within the testis have been found to be preferentially affected by the HD mutation it is reasonable to suspect that *HTT* levels may vary between cell types within the testes (Van Raamsdonk et al. 2007). The liver, as a vital organ involved in many processes including removal of toxins from blood, infection resistance, and conversion of and storage of fats, is subject to many environmental factors. This increased exposure to environmental changes sheds some light on the inter-individual variability I observed in my samples. Of especial interest is the abnormally high *HTT* expression levels found in the liver of individual HDB129. This individual had, at one point in his lifetime, a hepatitis infection, and liver cirrhosis as well as decreased liver size was noted by the reporting pathologist. Differential gene expression has been noted in livers with cirrhosis resulting from a hepatitis infection (Iizuka et al. 2003). Given the suggested roles for *HTT* in response to cellular stress and cell viability in neuronal cell types, presented in the introductory chapter, it seems plausible that *HTT* also plays a similar role in response to the cellular stress of infection seen here in the liver. It is also possible that this increase in *HTT* transcript due to hepatitis infection was a result of the HD mutation itself. Again, as mentioned in the introductory chapter, the presence of mutant *HTT* can affect global cell responses. It is therefor possible that the liver of this HD individual, already primed due to the stress of the mutant *HTT* protein, elicits this increase response. Testing of additional livers from non-HD patients with cirrhosis subsequent to a Hepatitis infection would indicate if the increase in *HTT* transcript seen here is due to the presence of the HD allele.

Contrary to my initial hypothesis, the presence of the CAG repeat did not affect local DNA methylation between HD and control cortex. While none of the probes present on the 450K DNA methylation array directly assay DNA methylation of the CAG expansion, I anticipated that due to the difference in CAG size between HD and WT alleles increases in DNA methylation due to the additional CAGs would be reflected in the expansion of DNA methylation to nearby probed CpGs. Nonetheless, until direct assessment of DNA methylation of the CAG tract is performed, we will not know for certain if differential DNA methylation of the CAG tract occurs. Further inquiry into this question may have to wait for additional improvements in DNA methylation assays. At present, pyrosequencing of the CAG tract would most likely produce inconsistent results due to the large number of repeats. In addition, as most HD individuals are heterozygous for the HD mutation, any future work into CAG tract specific DNA methylation differences would necessitate phasing the DNA methylation levels to either the HD or WT allele.

Globally I found 14 putative DMRs between HD and control cortex. Given that the methodology used to identify these DMRs are still under peer review it is important to treat these identified DMRs as preliminary. Comparison of the DMRs identified here in my human samples to the differentially methylated sites identified in mouse using reduced bisulfite sequencing will have to wait until these 15 DMRs are further validated (Ng et al. 2013). Interestingly, despite the Adenosine A2A receptor (A2AR) being identified in human striatum samples as being differentially methylated between HD and control samples this gene was not identified in my dataset (Villar-Menéndez et al. 2013). As MSNs of the striatum experience selective

neurodegeneration in HD, assessing differential methylation of these cells is important. However, due to the selective neurodegeneration it is vital to assess cell type composition of each sample. Increased gliosis, particularly in the striatum, is known to occur in HD, this would have a significant impact on cell type composition and potentially on the identification of DMRs. It is likely that the majority of cells assayed in the Villar-Menendez study were either of the interneuron cell type (which is relativity spared in HD) or glial type cells. In my study, while I did not use striatum samples, I removed the effects of cell type composition in the cortex samples through the CETS algorithm. Even in the cortex, which is not primarily affected by the HD mutation, I found a trend towards a decrease in neuronal cell type and an increase in glial cell type in my HD samples. The affect of cell type composition in future studies of DNA methylation in HD is an important consideration.

Between matched liver and cortex samples, I identified 32 differentially methylated sites in the local *HTT* gene region. Interestingly of the probes contained withinRegion 6, which contains the proximal promoter and TSS of *HTT*, only one is differentially methylated. This differentially methylated site was not contained within the two CpG islands present in the *HTT* promoter region. The majority of differentially methylated sites exist within the gene body of *HTT*, with a large proportion of them found towards the 3' end of the gene. This suggests that while DNA methylation may play a role in the tissue specific expression levels of *HTT*, this role is not mediated through classical differential methylation of the promoter. The role of gene body DNA methylation is still poorly understood, a recent metastudy of

gene body methylation highlights the complexity of this epigenetic mark (Jjingo et al. 2012).

In contrast to the relatively straightforward negative correlation of high DNA methylation of promoter regions being correlated with decreases in expression, the researchers propose a more complicated model of gene body methylation. In this model low levels of gene body DNA methylation are associated with both very high and very low levels of transcription. The mechanism behind this seemingly paradoxical duality of low levels of gene body DNA methylation is the occupancy of DNA by dense nucleosome packaging, in the case of low expression, and its occupancy with multiple RNA polymerase II proteins when the gene is highly transcribed. The occupancy of the DNA by either nucleosomes or RNA polymerase II prevents DNA methylation by DNA methyltransferases. In the case of mid-level gene expression, i.e. the gene is not completely silenced nor is it highly expressed, the intermittent exposure of the DNA during the occasional removal of nucleases and passing of RNA polymerase II allows the DNA to be methylated by DNA methyltransferases. The majority of differentially methylated sites in the gene body of *HTT* display a high level of DNA methylation in liver as compared to cortex. As levels of *HTT* expression in the cortex are much higher than that in the liver, which one could consider to have mid level expression, it appears that *HTT* gene body DNA methylation follows this model of gene body DNA methylation. To further understand this complex relationship correlations between differential expression of *HTT* in these two tissues further correlations between to the levels of gene body DNA

methylation found in the 450K array and *HTT* expression must be done. These additional analyses remain ongoing at the present time.

6 Discussion

6.1 Study Objectives

At the start of this thesis I established three objectives; 1) to conduct a bioinfomatic-driven search of potential regulatory regions outside of the *HTT* promoter region, 2) to create a stably expressing cell-based screening system to test potential regulatory regions and associated TFs, and 3) to investigate the role of DNA methylation in tissue expression differences of *HTT* and HD mutant allele effects on *HTT* gene methylation. The overarching purpose of these objectives was to further our understanding of the regulation of the *HTT* gene given recent advancements in our understanding of transcript regulation. Ultimately I hope to apply the knowledge gained from this study to assist the implementation of translational research aimed at curing or alleviating the symptoms of HD. Here I further discuss the findings of our study, as well as future directions for this work.

6.2 Additional Transcriptional Regulation Regions of the *HTT* Gene

Using publicly available datasets, I conducted a comprehensive screen of the *HTT* gene locus for additional regulatory regions. It is important to bear in mind that this analysis was limited by the available datasets. In particular, of the over 1,300 known TFs, I was able to assess ChIP-seq binding of only a small subset of these TFs in my regions of interest (Vaquerizas et al. 2009). This means that while the

identified TFs may indeed play a role in *HTT* transcriptional regulation, this list not exclusive and additional TFs, that may play a larger role, may exist. As it would be impossible and impractical to test the role of each of the known TFs in *HTT* transcriptional regulation, I will need to rely on additional ChIP-seq data to identify additional TFs. Given the differential tissue expression of *HTT*, it is likely that differential expression and binding of TFs between tissue types is responsible (at least in part) for establishing these distinct patterns of tissue expression. In my current ChIP-seq screen of the regulatory regions of interest I did not take into account the tissue type used in each ChIP-seq assay. This was due to the limited number of datasets available and was allowed for the greatest number of TFs to be screened. As additional TFs and additional tissue types for each TF are identified using ChIP-seq, more selective tissue and cell specific analysis of TF binding will be possible.

In my bioinfomatic assessment I limited the search to the genomic region stretching from the genes immediately preceding and following the *HTT* gene. In the absence of additional chromosomal conformation data, these arbitrary boundaries were as appropriate as any. Further analysis of 4C and 5C datasets will allow for a more precise and biologically driven selection of transcriptional regulation boundaries. Given the current limitations of these assays, it is likely that additional association domains encompassing a large numbers of genes will be identified in these studies, adding additional complexity to follow-up studies. Through the use of smaller, target specific 3C assays to confirm the interaction of putative regulatory regions with the *HTT* promoter it will be possible to confirm the function of not only

the identified regions here, but also other regions identified by 4 and 5C assays. Given the tissue type specific expression differences observed for *HTT*, selection of 4 and 5C datasets of both neuronal cell type and peripheral cell type will also allow for the identification of tissue specific regulatory regions.

6.3 Generation of Stably Expressing Cell Lines

At the onset of this project the proposition of generating multiple cell lines, each with a site-specific integration of different *HTT* promoter constructs appeared to have a clear advantage over the use of cell lines with random integration of *HTT* promoter constructs. As elaborated throughout this thesis, the theoretical advantages of the FLP-In system for the generation of these stably expressing cell lines was not entirely fulfilled by the practical application of the system. In particular, questions remain as to the relevance and implications of the existence of hygromycin resistant cell lines that are B-galactosidase positive (after the second stage of the FLP-In system), and zeocin resistant cell lines that are B-galactosidase negative (after the first stage of the FLP-In system). Further exploration into to the mechanisms that contributed to creation of these anomalous cell lines is necessary. I would also suggest that in future updates of the FLP-In system, the additional clonal selection step that I performed, should be standard. Particularly if the occurrence of anomalous cell lines such as I have described cannot be prevented using this system.

Baseline luciferase expression differences were observed between different clonal cell lines for both the *HTT* promoter-luciferase construct expression as well as

for the responses to siRNA-induced stress. The extreme “bottleneck effect” incurred by clonal selection may, in part, explain some of these baseline differences. Clonal selection necessitates the propagation of a single cell into a genetically identical population of cells, which is then expanded to created a cell line. Due to this derivation from a single cell into an entire cell line, small difference between cells in the parent population can be exacerbated as they are propagated throughout the entire cell population of each new cell line. This results in cell lines that differ not only at the site directed insertion site, if different constructs are used, but also potentially at multiple sites throughout the genome. Further investigation into the differences, and similarities, between my different cell lines is a necessary step in order to have confidence in these assays as well as future experimental paradigms based on this system.

My cell-based studies have mainly focused on a peripheral cell type, HEK293 cells. Although I did establish stably expressing cell lines in a rat neuronal cell line these lines need further validation before their use. While these cell lines will better represent cell conditions in neuronal cells, as compared to the HEK293 cells, they are not human. As such, differences in TF binding site affinities between rat and human TFs may confound results. I did attempt to create a set of human neuronal stably expressing cell lines, however due to the nature of these cells clonal isolation was not possible. Other methods of stable integration or a return to transient transfection may be necessary to allow for future work in human neuronal cell lines.

6.4 Effect of siRNA Knockdown of Selected TFs

Given the baseline differences in my stably expressing cell lines, the analysis and interpretation of my siRNA knockdown data required a cautious approach. Due to the differential responses to siRNA treatment exhibited by the different cell lines, it was important to examine these results through the lens of each cell line in isolation and each treatment as compared to the scramble treated cells. Keeping these points in mind, I found that SMC3 knockdown had a negative impact on *HTT* promoter function. As elaborated in the discussion of my siRNA chapter, the lack of effect on endogenous HEK293 *HTT* expression may be due to differences in haplotype as well as context specific regulation, i.e interaction of regulatory regions specific to the genomic location of the *HTT* promoter. Follow-up of this result through the identification of additional SMC3 sites within the *HTT* promoter region, as well as sites for SMC3 interactors, will allow confirmation of the functionality of SMC3 in *HTT* transcriptional regulation through additional assays.

I also observed a significant decrease in *HTT* promoter function in the Region 9 cells treated with IRF1 siRNA. At the current time it is unclear whether the region 9 cell line also displays an increase in IRF1 expression due to siRNA treatment. If this is the case it is possible that off-target effects elicited by the IRF1 siRNA specifically are causative of this decrease in *HTT* promoter function. Further exploration into this cell line, and possibly Region 9, specific down-regulation is required to further investigate this finding.

6.5 DNA Methylation in HD and at the *HTT* Gene Locus

I conducted the first genome-wide study, in human tissue, of DNA methylation in HD and control cortex. While this is study novel in and of itself, I also looked at differences at the *HTT* gene locus specifically in both my HD and control cortices as well as in liver versus cortex samples. This is, surprisingly, rarely done in HD genome-wide studies. For example, the genome wide methylation study conduced in a mouse model of HD referenced above, (Ng et al. 2013), offered no comment on the methylation status of mouse *Hdh* at all. This may be due to a resistance of investigators to be seen as "cherry picking" of their data, when potential changes at the *HTT* gene locus did not appear in their top ranked candidate hits. In my opinion, this conservative approach does a disservice to the HD research community as a whole and a clearly defined candidate gene approach is appropriate in this case. HD is caused from a single well-defined genetic mutation; which cannot be said of many other neurodegenerative diseases, including those that are often considered to be within the same family of late-onset, neurodegenerative diseases, such as Parkinson's or Alzheimer's disease. By failing to comment directly on any effects at the locus of the causative *HTT* gene, these genome-wide HD studies failed to improve our understanding of the both the function and regulation of the *HTT* gene. This has forced other researchers, such as myself, to mine these genome-wide datasets, potentially using methods that differ from those used in the original analysis. A simple statement by groups conducting genome-wide studies in HD, as

to the state of the *HTT* gene locus, would prevent the necessity of others re-analyzing these large datasets.

My analysis of genome-wide changes in HD cortex identified 14 putative DMRs. As discussed in the DNA methylation chapter, these results are very preliminary. The methodology utilized to identify these DMRs has yet to be published. At the *HTT* gene locus local DNA methylation was not affected by the presence of the HD CAG expansion. While this was not what I had predicted at the onset of the study, this turned out to be advantageous for my cortex versus liver analysis. Due to the limited availability of matched cortex-liver samples from control individuals in our bio-bank, the majority of the samples used in the matched tissue dataset were from HD individuals. As I was confident from the HD versus control cortex dataset that the CAG expansion did not have an effect on local DNA methylation of *HTT* locus, I could assume that any changes seen in the cortex versus liver dataset were due to tissue specific methylation changes. I identified 32 differentially-methylated sites between cortex and liver in the local *HTT* gene genomic region. Follow-up of correlations between tissue specific expression and these methylation changes is currently ongoing in the lab.

Follow-up analysis of additional HD individuals to confirm the DNA methylation changes seen in my original cohort uncovered an interesting finding with regards to liver expression in a single individual, HDB129. Subsequent research into the general health of this subject revealed that a hepatitis infection had resulted in severe cirrhosis of the liver. This acute inflammatory infection may have contributed to the abnormally high expression of *HTT*, greater in fact than I observed in any of

my other cortex or testes samples, in the liver of this individual. Further follow-up of this interesting finding is required to ascertain the role of the HD mutation itself, and chronic inflammation, in the induction of this observed *HTT* over-expression. As I suspect that inflammation of the liver is causative to the observed increase in expression, additional analysis into the expression of *HTT* in brain-specific chronic inflammation would be of great interest.

6.6 Concluding Statements

I have successfully accomplished the objectives set out at the beginning of my thesis. This work has expanded the analysis of the previously limited regions of transcriptional control of the *HTT* gene to include regions beyond the proximal promoter. I have assessed several previously published, and several novel putative TFs for their role in *HTT* transcriptional regulation. Finally, I have explored the role of epigenetic control, specifically DNA methylation, at the *HTT* gene locus. The methodologies utilized by this study, as well as the results generate, create a basis upon which future studies into transcriptional regulation of the *HTT* gene can be based.

References

Allard, S.T.M. & Kopish, K., 2008. Luciferase reporter assays: Powerful, adaptable tools for cell biology research. *Cell Notes: Promega*, 21, pp. 23–26.

Altar, C.A. et al., 1997. Anterograde transport of brain-derived neurotrophic factor and its role in the brain. *Nature*, 389(6653), pp.856–860.

Andrade, M.A. & Bork, P., 1995. HEAT repeats in the Huntington's disease protein. *Nature Genetics*, 11(2), pp.115–116.

Andresen, B.T., 2010. Characterization of G protein-coupled receptor kinase 4 and measuring its constitutive activity *in vivo*. *Methods in enzymology*, 484, pp.631–651.

Andrew, S.E. et al., 1993. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nature Genetics*, 4(4), pp.398–403.

Arrasate, M. et al., 2004. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature*, 431(7010), pp.805–810.

Bailey, T.L. & Gribskov, M., 1998. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics (Oxford, England)*, 14(1), pp.48–54.

Bailey, T.L. & Machanick, P., 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic acids research*, 40(17), pp.e128–e128.

Bailey, T.L. et al., 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(Web Server issue), pp.W202–8.

Ballaré, C. et al., 2013. More help than hindrance: nucleosomes aid transcriptional regulation. *Nucleus (Austin, Tex.)*, 4(3), pp.189–194.

Bañez-Coronel, M. et al., 2012. A pathogenic mechanism in Huntington's disease involves small CAG-repeated RNAs with neurotoxic activity. C. E. Pearson, ed. *PLoS genetics*, 8(2), p.e1002481.

Baquet, Z.C. et al., 2004. Early striatal dendrite deficits followed by neuron loss with advanced age in the absence of anterograde cortical brain-derived neurotrophic factor. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 24(17), pp.4250–4258.

Baxendale, S. et al., 1995. Comparative sequence analysis of the human and pufferfish Huntington's disease genes. *Nature Genetics*, 10(1), pp.67–76.

Beishline, K., & Azizkhan-Clifford, J., 2015. Sp1 and the 'hallmarks of cancer'. *The FEBS Journal*, 282(2), pp. 224–258.

Bemelmans, A.P. et al., 1999. Brain-derived neurotrophic factor-mediated protection of striatal neurons in an excitotoxic rat model of Huntington's disease, as demonstrated by adenoviral gene transfer. *Human gene therapy*, 10(18), pp.2987–2997.

Benchoua, A. et al., 2006. Involvement of mitochondrial complex II defects in neuronal death produced by N-terminus fragment of mutated huntingtin. *Molecular biology of the cell*, 17(4), pp.1652–1663.

Bhattacharyya, A et al., 2006. Oligoproline effects on polyglutamine conformation and aggregation. *Journal of Molecular Biology*, 355(3), pp. 524–535.

Boutell, J.M. et al., 1999. Aberrant interactions of transcriptional repressor proteins with the Huntington's disease gene product, *huntingtin*., 8(9), pp.1647–1655.

Canals, J.M. et al., 2004. Brain-derived neurotrophic factor regulates the onset and severity of motor dysfunction associated with enkephalinergic neuronal degeneration in Huntington's disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 24(35), pp. 7727–7739.

Carey, M.F. et al.,, 2009. *Transcriptional regulation in eukaryotes*, Cold Spring Harbor Laboratory Pr.

Carninci, P. et al., 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6), pp.626–635.

Carninci, P. et al., 2005. The transcriptional landscape of the mammalian genome. *Science (New York, NY)*, 309(5740), pp.1559–1563.

Chakraborty, M. et al., 2010. DNA binding domain of RFX5: interactions with X-box DNA and RFXANK. *Biochimica et biophysica acta*, 1804(10), pp.2016–2024.

Chang, D.T.W. et al., 2006. Mutant huntingtin aggregates impair mitochondrial movement and trafficking in cortical neurons., 22(2), pp.388–400.

Chattopadhyay, B. et al., 2005. Modulation of age at onset of Huntington disease patients by variations in TP53 and human caspase activated DNase (hCAD) genes. *Neuroscience letters*, 374(2), pp.81–86.

Chung, D.W. et al., 2011. A natural antisense transcript at the Huntington's disease repeat locus regulates HTT expression., 20(17), pp.3467–3477.

Ci, W., & Liu, J., 2015. Programming and Inheritance of Parental DNA Methylomes in Vertebrates. *Physiology* (, 30(1), pp. 63–68.

Cisbani, G. & Cicchetti, F., 2012. An in vitro perspective on the molecular mechanisms underlying mutant huntingtin protein toxicity. *Cell Death & Disease*, 3(8), e382.

Clabough, E. B. D., 2013. Huntington's disease: the past, present, and future search for disease modifiers. *The Yale Journal of Biology and Medicine*, 86(2), pp. 217–233.

Coles, R., Caswell, R. & Rubinsztein, D.C., 1998. Functional analysis of the Huntington's disease (HD) gene promoter. 7(5), pp.791–800.

Coles, R., Leggo, J. & Rubinsztein, D.C., 1997. Analysis of the 5“ upstream sequence of the Huntington”s disease (HD) gene shows six new rare alleles which are unrelated to the age at onset of HD. *Journal of medical genetics*, 34(5), pp.371–374.

Cooper, D. N., & Krawczak, M., 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Human Genetics*, 83(2), pp. 181–188.

Cooper, S.J. et al., 2006. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome research*, 16(1), pp.1–10.

Cowan, C.M. et al., 2008. Polyglutamine-modulated striatal calpain activity in YAC transgenic huntington disease mouse model: impact on NMDA receptor function and toxicity. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(48), pp. 12725–12735.

Cui, L. et al., 2006. Transcriptional repression of PGC-1alpha by mutant huntingtin leads to mitochondrial dysfunction and neurodegeneration. *Cell*, 127(1), pp.59–69.

Dantas Machado, A. C. et al., 2014. Evolving insights on how cytosine methylation affects protein-DNA binding. *Briefings in Functional Genomics,* 14(1), pp. 61-73

Davies, S.W. et al., 1997. Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. *Cell*, 90(3), pp.537–548.

Desplats, P.A. et al., 2006. Selective deficits in the expression of striatal-enriched mRNAs in Huntington's disease. *Journal of neurochemistry*, 96(3), pp.743–757.

Diamond, M. I. et al., 1990. Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science*, 249(4974), 1266–1272.

Dixon, K.T. et al., 2004. Mouse Huntington's disease homolog mRNA levels: variation and allele effects. *Gene expression*, 11(5-6), pp.221–231.

Djoussé, L. et al., 2004. Evidence for a modifier of onset age in Huntington disease linked to the HD gene in 4p16. *Neurogenetics*, 5(2), pp.109–114.

Dong, G. et al., 2011. Modeling Pathogenesis of Huntington's Disease with Inducible Neuroprogenitor Cells. *Cellular and Molecular Neurobiology,* 31(5), pp.737-747.

Dorsett, D., 2009. Cohesin, gene expression and development: lessons from Drosophila. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 17(2), pp.185–200.

Du, P., et al., 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1), p.587.

Du, P., Kibbe, W.A. & Lin, S.M., 2008. lumi: a pipeline for processing Illumina microarray. *Bioinformatics (Oxford, England)*, 24(13), pp.1547–1548.

Dunah, A.W. et al., 2002. Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. *Science (New York, NY)*, 296(5576), pp.2238–2243.

Duyao, M.P. et al., 1995. Inactivation of the mouse Huntington's disease gene homolog *Hdh*. *Science (New York, NY)*, 269(5222), pp.407–410.

Ehrlich, M.E. et al., 2001. ST14A cells have properties of a medium-size spiny neuron. *Experimental Neurology*, 167(2), pp.215–226.

Ehrnhoefer, D.E., Sutton, L. & Hayden, M.R., 2011. Small changes, big impact: posttranslational modifications and function of huntingtin in Huntington disease. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 17(5), pp.475–492.

ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, NY)*, 306(5696), pp.636–640.

Engelender, S. et al. 1997. Huntingtin-associated protein 1 (HAP1) interacts with the p150Glued subunit of dynactin. *Human Molecular Genetics*, 6(13), pp.2205–2212.

Ezer, D. et al., 2014. Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. *Computational and Structural Biotechnology Journal*, 10(17), pp. 63–69.

Feng, Z. et al., 2006. p53 tumor suppressor protein regulates the levels of huntingtin gene expression. *Oncogene*, 25(1), pp.1–7.

Fernandes, H.B. et al., 2007. Mitochondrial sensitivity and altered calcium handling underlie enhanced NMDA-induced apoptosis in YAC128 model of Huntington's disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(50), pp.13614–13623.

Fiszer, A. & Krzyzosiak, W.J., 2013. RNA toxicity in polyglutamine disorders: concepts, models, and progress of research. *Journal of Molecular Medicine*, 91(6), pp.683–691.

Flanagan, J.M. et al., 2006. Intra- and interindividual epigenetic variation in human germ cells. *American journal of human genetics*, 79(1), pp.67–84.

Francelle, L. et al., 2014. Possible involvement of self-defense mechanisms in the preferential vulnerability of the striatum in Huntington's disease. *Frontiers in Cellular Nueroscience,* 8, article 295

Franceschini, A. et al., 2012. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1), pp.D808–D815.

Gafni, J. & Ellerby, L.M., 2002. Calpain activation in Huntington's disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 22(12), pp.4842–4849.

Gafni, J. et al., 2004. Inhibition of calpain cleavage of huntingtin reduces toxicity: accumulation of calpain/caspase fragments in the nucleus. *The Journal of biological chemistry*, 279(19), pp. 20211–20220.

Gartenberg, M., 2009. Heterochromatin and the cohesion of sister chromatids. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 17(2), pp.229–238.

Gauthier, L.R. et al., 2004. Huntingtin controls neurotrophic support and survival of neurons by enhancing BDNF vesicular transport along microtubules. *Cell*, 118(1), pp.127–138.

Gershenzon, N.I. & Ioshikhes, I.P., 2005. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics (Oxford, England)*, 21(8), pp.1295–1300.

Gissi, C. et al., 2006. Huntingtin gene evolution in Chordata and its peculiar features in the ascidian Ciona genus. *BMC genomics*, 7(1), p.288.

Gladding, C.M. et al., 2012. Calpain and STriatal-Enriched protein tyrosine phosphatase (STEP) activation contribute to extrasynaptic NMDA receptor localization in a Huntington's disease mouse model., 21(17), pp.3739–3752.

Goldberg, Y.P. et al., 1996. Cleavage of huntingtin by apopain, a proapoptotic cysteine protease, is modulated by the polyglutamine tract. *Nature Genetics*, 13(4), pp.442–449.

Gonitel, R. et al., 2008. DNA instability in postmitotic neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9), pp.3467–3472.

Graham, R.K. et al., 2009. Differential susceptibility to excitotoxic stress in YAC128 mouse models of Huntington disease between initiation and progression of disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(7), pp.2193–2204.

Graham, R.K. et al., 2006. Levels of mutant huntingtin influence the phenotypic severity of Huntington disease in YAC128 mouse models., 21(2), pp.444–455.

Green, M.R & Sambrook, J., 2012. Molecualr Cloning: A Laboratory Manual (Fourth Edition)., CSHL Press.

Gu, M. et al., 1996. Mitochondrial defect in Huntington's disease caudate nucleus. *Annals of neurology*, 39(3), pp.385–389.

Guintivano, J., et al., 2013. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics : official journal of the DNA Methylation Society*, 8(3), pp.290–302.

Gutekunst, C.-A., et al., 1999. Nuclear and Neuropil Aggregates in Huntington's Disease: Relationship to Neuropathology. T*he Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, 19(7), pp. 2522–2534.

Hsiao, H.-Y., & Chern, Y., 2010. Targeting glial cells to elucidate the pathogenesis of Huntington's disease. *Molecular Neurobiology*, 41(2-3), 248–255.

Hardison, R.C. & Taylor, J., 2012. Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, 13(7), pp.469–483.

Hasenkamp, S. et al., 2008. Characterization and functional analyses of the human G protein-coupled receptor kinase 4 gene promoter. *Hypertension*, 52(4), pp.737–746.

Hedreen, J.C. et al., 1991. Neuronal loss in layers V and VI of cerebral cortex in Huntington's disease. *Neuroscience letters*, 133(2), pp.257–261.

Hellemans, J., et al., 2007. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biology*, 8(2), R19.

Holzmann, C. et al., 1998. Isolation and characterization of the rat huntingtin promoter. *The Biochemical journal*, 336 (Pt 1), pp.227–234.

Honda, K. & Taniguchi, T., 2006. IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. *Nature reviews. Immunology*, 6(9), pp.644–658.

Hovakimyan, M. et al., 2008. In vitro characterization of embryionic ST14A-cells. *The International journal of neuroscience*, 118(11), pp.1489–1501.

Huang, K. et al., 2004. Huntingtin-interacting protein HIP14 is a palmitoyl transferase involved in palmitoylation and trafficking of multiple neuronal proteins. *Neuron*, 44(6), pp.977–986.

Humbert, S. et al., 2002. The IGF-1/Akt pathway is neuroprotective in Huntington's disease and involves Huntingtin phosphorylation by Akt. *Developmental cell*, 2(6), pp.831–837.

Huntington, G., 1872. On Chorea. *The Medical and Surgical Reporter: A Weekly Journal*, 26(15), pp. 317-321.

Iizuka, N. et al., 2003. Differential gene expression in distinct virologic types of hepatocellular carcinoma: association with liver cirrhosis. *Oncogene*, 22(19), pp.3007–3014.

Imarisio, S. et al., 2008. Huntington's disease: from pathology and genetics to potential therapies. *The Biochemical journal*, 412(2), pp.191–209.

Jasinska, A. et al., 2003. Structures of trinucleotide repeats in human transcripts and their functional implications. *Nucleic acids research*, 31(19), pp.5463–5468.

Jjingo, D. et al., 2012. On the presence and role of human gene-body DNA methylation. *Oncotarget*, 3(4), pp.462–474.

Johnson, R. et al., 2008. A microRNA-based gene dysregulation pathway in Huntington's disease., 29(3), pp.438–445.

Johnson, W.E., et al., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, 8(1), pp.118–127.

Jones, M.J. et al., 2013. Distinct DNA methylation patterns of cognitive impairment and trisomy 21 in Down syndrome. *BMC medical genomics*, 6(1), p.58.

Jones, P.A., 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7), pp. 484–492.

Juven-Gershon, T. et al., 2008. The RNA polymerase II core promoter - the gateway to transcription. *Current opinion in cell biology*, 20(3), pp.253–259.

Kadonaga, J.T., 2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, 116(2), pp.247–257.

Kalchman, M.A. et al., 1996. Huntingtin is ubiquitinated and interacts with a specific ubiquitin-conjugating enzyme. *The Journal of biological chemistry*, 271(32), pp.19385–19394.

Karymov, M. A. et al., 2001. DNA methylation-dependent chromatin fiber compaction *in vivo* and *in vitro*: requirement for linker histone. *The FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, 15(14), pp. 2631–2641.

Kim, D.-H. et al., 2005. Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy. *Nature biotechnology*, 23(2), pp.222–226.

Knight, S.J.L. et al., 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington“s disease chromosomes. The Huntington”s Disease Collaborative Research Group. *Cell*, 72(6), pp.971–983.

Kröger, A. et al., 2002. Activities of IRF-1. *Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research*, 22(1), pp.5–14.

Lam, L.L. et al., 2012. Factors underlying variable DNA methylation in a human community cohort. *Proceedings of the National Academy of Sciences of the United States of America*, 109, pp. 17253–17260.

Langbehn, D. R. et al., 2004. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length, *Clinical Genetics,* 65(4), pp. 267–277.

Lawlor, K.T. et al., 2011. Double-stranded RNA is pathogenic in Drosophila models of expanded repeat neurodegenerative diseases., 20(19), pp.3757–3768.

Leavitt, B.R. et al., 2006. Wild-type huntingtin protects neurons from excitotoxicity. *Journal of neurochemistry*, 96(4), pp.1121–1129.

Leavitt, B.R. et al., 2001. Wild-type huntingtin reduces the cellular toxicity of mutant huntingtin *in vivo*. *American journal of human genetics*, 68(2), pp.313–324.

Lee, Juna et al., 2010. Presence of 5-methylcytosine in CpNpG trinucleotides in the human genome. *Genomics*, 96(2), pp.67–72.

Lee, Junghee et al., 2013. Epigenetic mechanisms of neurodegeneration in Huntington's disease. *Neurotherapeutics*, 10(4), pp.664–676.

Li, L. et al., 2004. Enhanced striatal NR2B-containing N-methyl-D-aspartate receptor-mediated synaptic currents in a mouse model of Huntington disease. *Journal of neurophysiology*, 92(5), pp.2738–2746.

Li, L. et al., 2003. Role of NR2B-type NMDA receptors in selective neurodegeneration in Huntington disease. *Neurobiology of aging*, 24(8), pp.1113–1121.

Li, S.H. et al., 1993. Huntington's disease gene (IT15) is widely expressed in human and rat tissues. *Neuron*, 11(5), pp.985–993.

Lin, B. et al., 1993. Differential 3' polyadenylation of the Huntington disease gene results in two mRNA species with variable tissue expression., 2(10), pp.1541–1545.

Lin, B. et al., 1995. Structural analysis of the 5' region of mouse and human Huntington disease genes reveals conservation of putative promoter region and di- and trinucleotide polymorphisms. *Genomics*, 25(3), pp.707–715.

Liot, G. et al., 2013. Mutant Huntingtin alters retrograde transport of TrkB receptors in striatal dendrites. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(15), pp.6298–6309.

Lister, R. et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), pp. 315–322.

López Castel, A. et al., 2011. Identification of restriction endonucleases sensitive to 5-cytosine methylation at non-CpG sites, including expanded (CAG)n/(CTG)n repeats. *Epigenetics : official journal of the DNA Methylation Society*, 6(4), pp.416–420.

Luo, S. et al., 2005. Cdk5 phosphorylation of huntingtin reduces its cleavage by caspases: implications for mutant huntingtin toxicity. *The Journal of cell biology*, 169(4), pp.647–656.

MacDonald, M.E. et al., 1999. Evidence for the GluR6 gene associated with younger onset age of Huntington's disease. *Neurology*, 53(6), pp.1330–1332.

Mahant, N. et al., 2003. Huntington's disease: clinical correlates of disability and progression. *Neurology*, 61(8), pp.1085–1092.

Maksimovic, et al., 2012. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome biology*, 13(6), p.R44.

McGeer, E.G. & McGeer, P.L., 1976. Duplication of biochemical changes of Huntington's chorea by intrastriatal injections of glutamic and kainic acids. *Nature*, 263(5577), pp.517–519.

Mende-Mueller, L.M. et al., 2001. Tissue-specific proteolysis of Huntingtin (htt) in human brain: evidence of enhanced levels of N- and C-terminal htt fragments in Huntington's disease striatum. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 21(6), pp.1830–1837.

Metzler, M. et al., 2000. Huntingtin is required for normal hematopoiesis. *Human molecular genetics*, 9(3), pp.387–394.

Milakovic, T., et al., 2006. Mutant huntingtin expression induces mitochondrial calcium handling defects in clonal striatal cells: functional consequences. *The Journal of biological chemistry*, 281(46), pp.34785–34795.

Mizuno, K. et al., 1994. Brain-derived neurotrophic factor promotes differentiation of striatal GABAergic neurons. *Developmental biology*, 165(1), pp.243–256.

Moreira Sousa, C. et al., 2013. The Huntington disease protein accelerates breast tumour development and metastasis through ErbB2/HER2 signalling. *EMBO molecular medicine*, 5(2), pp.309–325.

Morton, A.J. & Edwardson, J.M., 2001. Progressive depletion of complexin II in a transgenic mouse model of Huntington's disease. *Journal of neurochemistry*, 76(1), pp.166–172.

Myers, R.H. et al., 1989. Homozygote for Huntington disease. *American journal of human genetics*, 45(4), pp.615–618.

Mykowska, A. et al., 2011. CAG repeats mimic CUG repeats in the misregulation of alternative splicing. *Nucleic acids research*, 39(20), pp.8938–8951.

Nasir, J. et al., 1995. Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. *Cell*, 81(5), pp.811–823.

Ng, C.W. et al., 2013. Extensive changes in DNA methylation are associated with expression of mutant huntingtin. *Proceedings of the National Academy of Sciences of the United States of America*, 110(6), pp.2354–2359.

Nolte, A. et al., 2013. Modification of small interfering RNAs to prevent off-target effects by the sense strand. *New biotechnology*, 30(2), pp.159–165.

Panov, A.V. et al., 2003. In vitro effects of polyglutamine tracts on Ca2+-dependent depolarization of rat and human mitochondria: relevance to Huntington's disease. *Archives of biochemistry and biophysics*, 410(1), pp.1–6.

Perutz, M. F. et al., 1994. Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 91(12), pp. 5355–5358.

Peters, T. J. et al. (2015). De novo identification of differentially methylated regions in the human genome. Epigenetics & Chromatin, 8(6).

Pflaum, J. et al., 2014. p53 Family and Cellular Stress Responses in Cancer. *Frontiers in Oncology*, 4, 285.

Pineda, J.R. et al., 2005. Brain-derived neurotrophic factor modulates dopaminergic deficits in a transgenic mouse model of Huntington's disease. *Journal of neurochemistry*, 93(5), pp.1057–1068.

Pinheiro, J. et al., 2014. *R Core Team (2013) nlme: Linear and nonlinear mixed effects models R package version 3.1-117*, URL: http://cran. r-project. org/web/ ….

Pinney, S. E., 2014. Mammalian Non-CpG Methylation: Stem Cells and Beyond. *Biology*, 3(4), pp. 739–751.

Poptsova, M.S., 2014. *Genome Analysis*, Horizon Scientific Press.

Portales-Casamar, E. et al., 2007. PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome biology*, 8(10), p.R207.

Poss, Z.C. et al., 2013. The Mediator complex and transcription regulation. *Critical reviews in biochemistry and molecular biology*, 48(6), pp.575–608.

Price, M. E. et al. (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & Chromatin*, 6(1).

Pringsheim, T. et al., 2012. The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. *Movement Disorders*, 27(9), pp.1083–1091.

Rothbart, S. B., & Strahl, B. D., 2014. Interpreting the language of histone and DNA modifications. *Biochimica Et Biophysica Acta*, 1839(8), pp. 627–643.

Ratovitski, T. et al., 2009. Mutant Huntingtin N-terminal Fragments of Specific Size Mediate Aggregation and Toxicity in Neuronal Cells. *The Journal of Biological Chemistry*, 284(16), pp . 10855–10867.

Rinaldi, C. et al., 2012. Predictors of survival in a Huntington's disease population from southern Italy. *The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques*, 39(1), pp. 48–51.

Reiner, A. et al., 2003. Wild-type huntingtin plays a role in brain development and neuronal survival. *Molecular neurobiology*, 28(3), pp.259–276.

Rhodes, J.M. et al., 2010. Positive regulation of c-Myc by cohesin is direct, and evolutionarily conserved. *Developmental biology*, 344(2), pp.637–649.

Rigamonti, D. et al., 2001. Huntingtin's neuroprotective activity occurs via inhibition of procaspase-9 processing. *The Journal of biological chemistry*, 276(18), pp.14545–14548.

Rigamonti, D. et al., 2000. Wild-type huntingtin protects from apoptosis upstream of caspase-3. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 20(10), pp. 3705–3713.

Romeo, G. et al., 2002. IRF-1 as a negative regulator of cell proliferation. *Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research*, 22(1), pp.39–47.

Ryan, A.B., Zeitlin, S.O. & Scrable, H., 2006. Genetic interaction between expanded murine *Hdh* alleles and p53 reveal deleterious effects of p53 on Huntington's disease pathogenesis., 24(2), pp.419–427.

Sajan, S. A., & Hawkins, R. D., 2012. Methods for identifying higher-order chromatin structure. *Annual Review of Genomics and Human Genetics*, 13, pp. 59–82.

Sandelin, A. et al., 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics*, 8(6), pp.424–436.

Sathasivam, K. et al., 2013. Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. *Proceedings of the National Academy of Sciences of the United States of America*, 110(6), pp.2366–2370.

Schwarcz, R. et al., 1984. Excitotoxic models for neurodegenerative disorders. *Life sciences*, 35(1), pp.19–32.

Sheng, M. & Kim, M.J., 2002. Postsynaptic signaling and plasticity mechanisms. *Science (New York, NY)*, 298(5594), pp.776–780.

Shimohata, T. et al., 2000. Expanded polyglutamine stretches interact with TAFII130, interfering with CREB-dependent transcription. *Nature Genetics*, 26(1), pp.29–36.

Sichtig, N., Silling, S. & Steger, G., 2007. Papillomavirus binding factor (PBF)-mediated inhibition of cell growth is regulated by 14-3-3β. *Archives of biochemistry and biophysics*, 464(1), pp.90–99.

Slattery, M. et al., 2014. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, 39(9), pp. 381–399.

Sledz, C.A. et al., 2003. Activation of the interferon system by short-interfering RNAs. *Nature cell biology*, 5(9), pp.834–839.

Smale, S. T., & Kadonaga, J. T., 2003. The RNA polymerase II core promoter. *Annual Review of Biochemistry*, 72(1), pp. 449–479.

Smith, C., 2006. Sharpening the tools of RNA interference. *Nature methods*.

Smith, R. et al., 2005. Depletion of rabphilin 3A in a transgenic mouse model (R6/1) of Huntington's disease, a possible culprit in synaptic dysfunction., 20(3), pp.673–684.

Smith, R. et al., 2005. Synaptic dysfunction in Huntington's disease: a new perspective. *Cellular and Molecular Life Sciences : CMLS*, 62(17), pp. 1901–1912.

Smyth, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), pp.Article3–25.

Sobczak, K. et al., 2003. RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic acids research*, 31(19), pp.5469–5482.

Spargo, E. et al., 1993. Neuronal loss in the hippocampus in Huntington's disease: a comparison with HIV infection. *Journal of neurology, neurosurgery, and psychiatry*, 56(5), pp.487–491.

Steffan, J.S. et al., 2004. SUMO modification of Huntingtin and Huntington's disease pathology. *Science (New York, NY)*, 304(5667), pp.100–104.

Steffan, J.S. et al., 2000. The Huntington's disease protein interacts with p53 and CREB-binding protein and represses transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12), pp.6763–6768.

Szyf, M., 2014. Nongenetic inheritance and transgenerational epigenetics. *Trends in Molecular Medicine,* 21(2) pp. 134-144.

Takano, H. & Gusella, J.F., 2002. The predominantly HEAT-like motif structure of huntingtin and its association and coincident nuclear entry with dorsal, an NF-kB/Rel/dorsal family transcription factor. *BMC neuroscience*, 3, p.15.

Tamura, T. et al., 2008. The IRF family transcription factors in immunity and oncogenesis. *Annual review of immunology*, 26(1), pp.535–584.

Tanaka, K. et al., 2004. Novel nuclear shuttle proteins, HDBP1 and HDBP2, bind to neuronal cell-specific cis-regulatory element in the promoter for the human Huntington's disease gene. *The Journal of biological chemistry*, 279(8), pp.7275–7286.

Tartari, M. et al., 2008. Phylogenetic comparison of huntingtin homologues reveals the appearance of a primitive polyQ in sea urchin. *Molecular biology and evolution*, 25(2), pp.330–338.

Thompson, J.F. et al., 1991. Modulation of firefly luciferase stability and impact on studies of gene regulation. *Gene*, 103(2), pp.171–177.

Trottier, Y. et al., 1995. Cellular localization of the Huntington's disease protein and discrimination of the normal and mutated form. *Nature Genetics*, 10(1), pp.104–110.

Van Raamsdonk, J.M. et al., 2007. Testicular degeneration in Huntington disease., 26(3), pp.512–520.

Van Raamsdonk, J.M. et al., 2005a. Loss of wild-type huntingtin influences motor dysfunction and survival in the YAC128 mouse model of Huntington disease., 14(10), pp.1379–1392.

Van Raamsdonk, J.M. et al., 2005b. Cognitive dysfunction precedes neuropathology and motor abnormalities in the YAC128 mouse model of Huntington's disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(16), pp.4169–4180.

Vandesompele, J. et al., 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology*, 3(7), p.RESEARCH0034.

Vaquerizas, J.M. et al., 2009. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4), pp.252–263.

Ventimiglia, R. et al., 1995. The neurotrophins BDNF, NT-3 and NT-4/5 promote survival and morphological and biochemical differentiation of striatal neurons in vitro. *The European journal of neuroscience*, 7(2), pp.213–222.

Vesely, P.W. et al., 2009. Translational regulation mechanisms of AP-1 proteins. *Mutation research*, 682(1), pp.7–12.

Villar-Menéndez, I. et al., 2013. Increased 5-methylcytosine and decreased 5-hydroxymethylcytosine levels are associated with reduced striatal A2AR levels in Huntington's disease. *Neuromolecular medicine*, 15(2), pp.295–309.

Vogel, C., & Marcotte, E. M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4), pp. 227–232.

Vonsattel, J.P. et al., 1985. Neuropathological Classification of Huntington's Disease. *Journal of Neuropathology and Experimental Neurology*, 44(6), pp 559-577.

Vonsattel, J.P. & DiFiglia, M., 1998. Huntington disease. *Journal of neuropathology and experimental neurology*, 57(5), pp.369–384.

Wan, J. et al., 2015. Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. *BMC Genomics*, 16(1), pg 49.

Wang, H. et al., 2009. Effects of overexpression of huntingtin proteins on mitochondrial integrity., 18(4), pp.737–752.

Wang, R. et al., 2012. Sp1 Regulates Human Huntingtin Gene Expression. *Journal of molecular neuroscience : MN*.

Warby, S.C. et al., 2009. CAG expansion in the Huntington disease gene is associated with a specific and targetable predisposing haplogroup. *American journal of human genetics*, 84(3), pp.351–366.

Warby, S.C. et al., 2005. Huntingtin phosphorylation on serine 421 is significantly reduced in the striatum and by polyglutamine expansion *in vivo*., 14(11), pp.1569–1577.

Wasserman, W.W. & Fahl, W.E., 1997. Functional antioxidant responsive elements. *Proceedings of the National Academy of Sciences of the United States of America*, 94(10), pp.5361–5366.

Wasserman, W.W. & Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4), pp.276–287.

Watari, K. et al., 2014. Multiple functions of G protein-coupled receptor kinases. *Journal of molecular signaling*, 9(1), p.1.

Wellington, C.L. et al., 1998. Caspase cleavage of gene products associated with triplet expansion disorders generates truncated fragments containing the polyglutamine tract. *The Journal of biological chemistry*, 273(15), pp.9158–9167.

Wendt, K.S. & Peters, J.-M., 2009. How cohesin and CTCF cooperate in regulating gene expression. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 17(2), pp.201–214.

Wexler, N.S. et al., 1987. Homozygotes for Huntington's disease. *Nature*, 326(6109), pp.194–197.

White, J.K. et al., 1997. Huntingtin is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion. *Nature Genetics*, 17(4), pp.404–410.

Wood, J.D. et al., 1996. Partial characterisation of murine huntingtin and apparent variations in the subcellular localisation of huntingtin in human, mouse and rat brain. *Human molecular genetics*, 5(4), pp.481–487.

Wozniak, G. G., & Strahl, B. D., 2014. Hitting the “mark”: interpreting lysine methylation in the context of active transcription. *Biochimica Et Biophysica Acta*, 1839(12), pp. 1353–1361.

Xia, J. et al., 2003. Huntingtin contains a highly conserved nuclear export signal., 12(12), pp.1393–1403.

Yan, C. et al., 2004. A novel homologous recombination system to study 92 kDa type IV collagenase transcription demonstrates that the NF-kappaB motif drives the transition from a repressed to an activated state of gene expression. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 18(3), pp.540–541.

Yin, J.-W., & Wang, G., 2014. The Mediator complex: a master coordinator of transcription and cell lineage development. *Development*, 141(5), pp. 977–987.

Young, F.B. et al., 2012. Putting proteins in their place: palmitoylation in Huntington disease and other neuropsychiatric diseases. *Progress in neurobiology*, 97(2), pp.220–238.

Yu, Z., Teng, X. & Bonini, N.M., 2011. Triplet repeat-derived siRNAs enhance RNA-mediated toxicity in a Drosophila model for myotonic dystrophy. C. E. Pearson, ed. *PLoS genetics*, 7(3), p.e1001340.

Zeitlin, S. et al., 1995. Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. *Nature Genetics*, 11(2), pp.155–163.

Zhang, J, Chung, T. & Oldenburg, K., 1999. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *Journal of biomolecular screening : the official journal of the Society for Biomolecular Screening*, 4(2), pp.67–73.

Zhang, Ying et al., 2009. Primary sequence and epigenetic determinants of *in vivo* occupancy of genomic DNA by GATA1. *Nucleic acids research*, 37(21), pp.7024–7038.

Zhang, Yu et al., 2003. Depletion of wild-type huntingtin in mouse models of neurologic diseases. *Journal of neurochemistry*, 87(1), pp.101–106.

Zhu, X.D. & Sadowski, P.D., 1995. *Cleavage-dependent ligation by the FLP recombinase. Characterization of a mutant FLP protein with an alteration in a catalytic amino acid*,

Zuccato, C. et al., 2003. Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nature Genetics*, 35(1), pp.76–83.

Zuccato, C. et al., 2001. Loss of huntingtin-mediated BDNF gene transcription in Huntington's disease. *Science (New York, NY)*, 293(5529), pp.493–498.

Zuccato, C. et al., 2007. Widespread disruption of repressor element-1 silencing transcription factor/ neuron-restrictive silencer factor occupancy at its target genes in Huntington's disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(26), pp. 6972–6983.