POPULATION GENETICS AND ALLELE-SPECIFIC SILENCING OF THE HUNTINGTON DISEASE MUTATION

by

Christopher Kay

B.S., The University of West Florida, 2008

B.A., The University of West Florida, 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)

May 2017

Abstract

Huntington disease (HD) is a devastating neurodegenerative disorder caused by an expanded CAG trinucleotide repeat in the huntingtin gene (*HTT*). A repeat of ≥36 CAG defines the HD mutation and is diagnostic in the presence of motor or psychiatric phenotypes. All patients have an expanded CAG repeat, usually inherited from an affected parent and heterozygous with a normal (≤35) CAG repeat allele. This thesis addresses numerous gaps in our understanding of the HD mutation at the population level, including heterozygote frequency, penetrance, ancestry, the *de novo* mutation rate, and haplotypes useful for therapeutic gene silencing.

Dense SNP genotyping across *HTT* in HD patients and relatives from Canada, Sweden, France, and Italy allowed definition of the three most common HD mutation haplotypes in populations of European ancestry. All common defining alleles of these haplotypes were identified from the 1000 Genomes Project reference data and validated by direct genotyping of HD patients and their families. Haplotypes of the HD mutation provide mutually exclusive sets of target variants for allele-specific *HTT* silencing in the maximum number of European ancestry patients. Haplotypes of the HD mutation were investigated in additional HD patient populations worldwide. The HD mutation in Peru occurs most frequently on the A1 *HTT* haplotype, as in Europe, but on a distinct A1 variant found in Amerindian controls, supporting an indigenous origin of the HD mutation in Latin America. The general population frequency of the HD mutation was estimated from CAG repeat genotyping of 7315 individuals in Canada, the United

States, and Scotland, revealing that approximately 1 in 400 people (0.246%) has a CAG repeat of 36 or greater. This frequency of the HD mutation in the general population is higher than expected from clinical prevalence estimates. The normal CAG repeat distribution was characterized in a range of ethnically distinct populations, allowing comparative estimates of the HD new mutation rate. Estimated as a function of intermediate CAG repeat frequencies below the pathogenic range (27-35 CAG), 1 in 5372 births is a new mutation for HD in European ancestry populations, representing 7.1% of HD mutations (≥36 CAG) in the general population.

Lay Summary

Huntington disease (HD) is a fatal and incurable genetic disease of the brain. The mutation that causes HD is an abnormally long stretch of DNA called the CAG repeat, located in the huntingtin gene (abbreviated *HTT*). In people with HD, the abnormally long CAG repeat is found in one copy of *HTT* and leads to toxic effects in neurons. Everyone has two copies of *HTT* and at least one is needed to maintain brain health. A promising therapy for HD patients is therefore to suppress the mutated copy of *HTT* while sparing the normal second copy. This therapeutic approach, called allele-specific silencing of the HD mutation, requires the identification of small genetic variations near the CAG repeat which can distinguish the mutated *HTT* gene from the normal *HTT* gene. This thesis identifies the most useful genetic variations for allele-specific silencing therapy in HD patient populations around the world.

Preface

Chapter 1

A portions of this chapter were published as a book chapter and review article:

> **Kay C**, Fisher E, Hayden MR. (2014) Epidemiology. In: *Huntington's Disease*. 4th ed. G. Bates, P. Harper, L. Jones (eds), Oxford University Press. pp. 131-164.
>
> **Kay C**, Skotte NH, AL Southwell, Hayden MR. (2014) Personalized Gene Silencing Therapeutics for Huntington Disease. ***Clinical Genetics***. 86: 29-36.

Statement of co-authorship:
I wrote all excerpted portions of the book chapter and review, and made all included figures and tables. Additional sections, figures, and tables authored by Emily Fisher, Niels Skotte, and Amber Southwell can be found in the respective published manuscripts. I authored additional original text found only in this dissertation chapter.

Chapter 2

A version of this chapter was published as:

> **Kay C**, Collins JA, Skotte NH, Southwell AL, Warby SC, Caron NS, Doty CN, Nguyen B, Griguoli A, Ross CJ, Squitieri F, Hayden MR. (2015) Huntingtin Haplotypes Provide Prioritized Target Panels for Allele-Specific Silencing in Huntington Disease Patients of European Ancestry. ***Molecular Therapy***. 23: 1759-71.

Statement of co-authorship:
I prepared all European patient samples in this study for genotyping, reconstructed all haplotypes, performed all *in silico* genotype associations using data from the 1000 Genomes Project, conducted most *in vitro* silencing experiments, wrote the manuscript, and made all figures and tables. Jennifer Collins identified patient samples from the UBC HD Biobank, annotated haplotypes, and prepared haplotype summaries for each population. Niels Skotte guided design of the *in vitro* silencing experiments and performed protein quantitation in treated cell cultures. Amber Southwell guided interpretation of data and generated supplemental *in vivo* data. Simon Warby conducted pilot genotyping of Canadian HD patients from the UBC Biobank. Nicholas Caron conducted follow-up *in vitro* silencing experiments for revision. Crystal Doty guided design of the *in vitro* silencing experiments and assisted Simon Warby in pilot genotyping of Canadian HD patients. Betty Nguyen assisted with *in vitro* silencing experiments. Colin Ross facilitated array SNP genotyping. Anne Griguoli and Ferdinando Squitieri provided Italian HD patient samples.

Ethics statement:
Consent and access procedures were in accordance with institutional ethics approval for human research (UBC certificates H05-70532 and H06-70467).

Chapter 3

A version of this chapter was published as:

> **Kay C,** Tirado-Hurtado I, Cornejo-Olivas M, Collins JA, Wright G, Inca-Martinez M, Veliz-Otani D, Ketelaar ME, Slama RA, Ross CJ, Mazzetti P, Hayden MR. (2016) The Targetable A1 Huntington Disease Haplotype has Distinct Amerindian and European Origins in Latin America. ***European Journal of Human Genetics***. 25: 332-340

Statement of co-authorship:
I prepared most samples in this study for genotyping, reconstructed haplotypes, performed all *in silico* genotype associations using data from the 1000 Genomes Project, wrote the manuscript, and made most figures. Indira Tirado-Hurtado and Miguel Inca-Martinez made components of Figure 1 and assisted in preparation of patient samples in Peru. Mario Cornejo-Olivas, Diego Veliz-Otani, and Pilar Mazzetti collected patient samples in Peru, facilitated sharing of patient material, and assisted interpretation of results. Jennifer Collins annotated all haplotypes and prepared data summaries for different subpopulations and regions. Marlies Ketelaar assisted in CAG repeat genotyping and preparation of samples. Ramy Slama assisted in follow-up genotyping of Amerindian *HTT* variants. Colin Ross facilitated array SNP genotyping.

Ethics statement:
IRB approval for this study was obtained from the Ethical Committee at Instituto Nacional de Ciencias Neurológicas (INCN) in Lima, Peru. Additional HD patients of European ancestry were screened from the UBC HD BioBank under existing research ethics protocols (UBC C&W CREB H06-70467 and H05-70532).

Chapter 4

A version of this chapter was published as:

> **Kay C**, Collins JA, Miedzybrodzka Z, Madore SJ, Gordon ES, Gerry N, Davidson M, Slama RA, Hayden MR. (2016) Huntington disease reduced penetrance alleles occur at high frequency in the general population. ***Neurology*.** 87:282-8.

Statement of co-authorship:
I performed all CAG repeat genotyping in this study, wrote the manuscript, and created all figures and tables. Jennifer Collins created a database for analysis of CAG sizing data and assisted with haplotype analysis. Zosia Miedzybrodzka and Mark Davidson

provided Scottish general population samples. Steven Madore, Erynn Gordon, and Norman Gerry provided samples and demographic data from the Coriell Institute in the United States. Ramy Slama assisted with haplotype analysis.

Ethics statement:
Ethical approval for the secondary use of archived anonymous samples from British Columbia was obtained from the University of British Columbia Children's and Women's Health Centre Clinical Research Ethics Board (UBC C&W CREB) (H06-70356). Ethical approval to examine CAG size distributions was obtained from the UBC C&W CREB (H05-70532).

Chapter 5

A version of this chapter is in preparation for publication as:

> **Kay C**, Collins JA, Baine F, Miedzybrodzka Z, Aminkeng F, Semaka A, McDonald C, Davidson M, Madore SJ, Gordon ES, Gerry N, Cornejo-Olivas M, Tishkoff S, Greenberg LJ, Krause A, Hayden MR. (2017) The molecular epidemiology of Huntington disease is related to intermediate allele frequency and haplotype in the general population.

Statement of co-authorship:
I performed most CAG repeat genotyping, prepared all samples for haplotyping, wrote the manuscript, and made all figures. Jennifer Collins created a database for analysis of CAG sizing data, annotated all haplotypes, and assisted with population-specific CAG and CCG genotype analysis. Fiona Baine obtained and performed CAG repeat genotyping of all samples from South Africa. Zosia Miedzybrodzka and Mark Davidson provided Scottish general population samples. Folefac Aminkeng and Sarah Tishkoff provided samples from other African populations. Alicia Semaka created pilot estimates of the HD new mutation rate in Canada. Cassandra McDonald assisted with CAG repeat and SNP genotyping. Steven Madore, Erynn Gordon, and Norman Gerry provided samples and demographic data from the Coriell Institute in the United States.

Chapter 6

Excerpts of this chapter are in preparation for publication as:

> **Kay C**, Collins JA, Squitieri F, Sumathipala D, Dissanayake V, Hayden MR. (2017) Global Analysis of Huntington Disease Patients Treatable by Allele-Specific Gene Silencing Therapy.

Statement of co-authorship:
I prepared all patient samples in this study for genotyping, reconstructed haplotypes, wrote the manuscript, and made all figures. Jennifer Collins annotated all haplotypes and prepared data summaries for different subpopulations and regions. Ferdinando

Squitieri provided Italian HD patient samples. Dulika Sumathipala and Vajira Dissanayake provided HD patient samples from Sri Lanka.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ALS	amyotrophic lateral sclerosis
ASO	antisense oligonucleotide
CAG	cytosine-adenine-guanine
CCG	cytosine-cytosine-guanine
DM	myotonic dystrophy
DNA	deoxyribonucleic acid
DRPLA	dentatorubral-pallidoluysian atrophy
HD	Huntington disease
HTT	huntingtin (human gene)
HTT	huntingtin (human protein)
Htt	huntingtin (murine gene), formerly *Hdh*
Htt	huntingtin (murine protein), formerly Hdh
IA	intermediate allele
JPH3	junctophilin-3 (gene)
LD	linkage disequilibrium
muHTT	mutant huntingtin
polyQ	polyglutamine
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
RP	reduced penetrance
SCA3	spinocerebellar ataxia type 3
SNP	single nucleotide polymorphism
tSNP	tagging single nucleotide polymorphism
VNTR	variable number tandem repeat
wtHTT	wild-type huntingtin

Acknowledgements

I have many to thank for their support and assistance in completing this dissertation. First I acknowledge my supervisor, Dr. Michael Hayden, who challenged me to be persistent and fearless, both in science and in life. Second, I had no greater colleague and friend in the laboratory than Jennifer Collins, without whom this work would be a fragment of itself. I thank my compatriot, Dr. Amber Southwell, for seeing possibilities in my work but never neglecting to point out the flaws. To Dr. Alicia Semaka, my scholarly sister, I am always indebted to you. I will never forget the opportunities and support that you gave me to open this chapter of my life. I thank Dr. Shaun Sanders and Dr. Dale Martin for their camaraderie and selflessness when it all seemed too much. To the many others in the Hayden Lab who have contributed, reviewed, or merely listened to my work, recognize that part of this belongs to you. I am additionally grateful to the Canadian Institutes of Health Research for funding my research, and to my supervisory committee for providing critical feedback.

The long road to writing a dissertation is often a lonely one. Beyond my circle of colleagues, the challenges of graduate work were largely invisible to those in my personal life. No one shared the burdens of my doctoral studies more than my partner, Glen Phillips, and no one has been so steadfast in the belief that I would succeed and become greater for it. Know that you are always loved for standing by me. Finally, I thank my mother, Donna Kay, who gave me the work ethic, the freedom, and the moral grounding to choose a life of my own.

In memory of Tom Routledge (1937-2011)

Who believed in me and so many other young people

Who insisted we mattered

You were right after all

Chapter 1: Introduction

1.1 Huntington Disease (HD) Before Discovery of the Mutation

Huntington disease (HD) was first widely recognized as a distinct and hereditary form of movement disorder following the publication of “On Chorea” in 1872 by George Huntington, a physician working on Long Island, New York. The disease presents with a spectrum of uncontrolled choreiform movements, rigidity, cognitive impairments, and behavioural disturbances, usually with involuntary movement (chorea) as the first presenting symptom (Ross et al., 2014). HD demonstrates an autosomal dominant pattern of inheritance with 50% risk to offspring and unbroken lines of affected relatives.

Numerous clinical aspects of HD eluded explanation before identification of the causative mutation. HD can manifest at any age, from early childhood to old age, but usually emerges in midlife. Among juvenile HD patients (<20 years old), inheritance is predominantly paternal. Age of onset tends to decrease over successive generations, in a phenomenon known as anticipation. Individuals with the mutation could only be ascertained through development of characteristic signs. Measures of disease onset and early progression were lacking, given the inability to identify and track premanifest individuals who had inherited the mutation. New mutations for HD were considered rare or nonexistent, since sporadic cases often presented late in life and after the death of one or both parents. Non-paternity could not be excluded in such cases, and

undiagnosed HD in one parent was suspected even in the absence of characteristic neurological symptoms in the family history.

1.2 Genetic Testing for the HD Mutation

The mutation causative of Huntington disease was discovered in 1993 by a consortium of investigators following a decade of painstaking positional cloning from the linked G8 marker (Gusella et al., 1983; , 1993). All Huntington disease patients have an expanded CAG repeat of 36 trinucleotides or greater in exon 1 of the huntingtin gene (*HTT*) (Kremer et al., 1994). The expanded CAG repeat is translated into repetitive glutamine residues in the corresponding huntingtin (HTT) protein, and translation of the expanded repeat is necessary to elicit its toxic effects (Trottier et al., 1995; Goldberg et al., 1996). An expanded CAG repeat genotype became the accepted molecular standard for HD diagnosis in those presenting with suggestive signs of the disease.

The development of a rapid, inexpensive, and precise test for the HD mutation, based on PCR amplification of the CAG repeat region and analysis of the resulting fragment length, enabled unambiguous differential diagnosis of HD patients (Warner et al., 1993; Andrew et al., 1994c). In addition, the genetic test allowed for definitive pre-symptomatic testing, revealing the HD genotype years or decades before clinical diagnosis. The effect of the genetic test on population-based studies of HD has been profound. Many sporadic cases, often with onset of symptoms in old age, have been genotyped with a CAG repeat expansion in the pathogenic range (James et al., 1994). Such cases,

previously grouped under the nonspecific diagnosis of senile chorea due to negative family history, instead represent HD patients of late onset. Elderly individuals who carry the CAG repeat expansion without signs of HD have also been uncovered (Rubinsztein et al., 1996).

Approximately 1% of individuals presenting with signs of HD and characteristic pathology on post-mortem examination do not have an expanded CAG repeat in *HTT* (Andrew et al., 1994a; Rosenblatt et al., 1998; Persichetti et al., 1994; Wild et al., 2008). Some of these cases lacking an expanded CAG repeat have been revealed to be autosomal dominant genocopies caused by heritable repeat expansions in other transcripts such as *JHP3* (Wild et al., 2008; Holmes et al., 2001). However, the *HTT* CAG expansion represents the vast majority of patients with suggestive signs of the disease, perhaps as high as 99% of referred cases, showing high sensitivity for the characteristic HD phenotype (Andrew et al., 1994b). In clinical and genetic practice, the expanded CAG repeat in *HTT* is considered definitional of HD, with phenocopies and genocopies instead classified as HD-like pathologies.

1.3 Relationship of CAG Repeat Length to HD Age of Onset

After discovery of the causative CAG repeat expansion in *HTT*, an inverse relationship of disease onset to CAG repeat length was recognized (Snell et al., 1993; Duyao et al., 1993; Andrew et al., 1993b). This inverse correlation was validated in many subsequent clinical cohorts and has been further developed into probabilistic survival models by

expanded CAG repeat length, using data from premanifest individuals (Brinkman et al., 1997; Langbehn et al., 2004). Simple regression models of mean CAG size and age of onset in manifest patients are biased for earlier predicted onset, whereas survival models that incorporate data from pre-symptomatic individuals show superior accuracy in prospective validation studies (Langbehn et al., 2010). There is poor agreement between all models at the lower end of the CAG repeat range (36-40) due to the rarity of these alleles in clinical series and the wide variability in patient age of onset, although onset tends to occur over 60 years of age at these repeat lengths (Lee et al., 2012b). Juvenile onset (<20 years of age) typically occurs with HD alleles in the upper CAG repeat range (>55 CAG). Repeats up to an estimated 265 CAG have been reported in extremely early onset infantile patients, requiring specialized long-range PCR amplification or molecular confirmation by Southern blot analysis for genotyping (Milunsky et al., 2003).

1.4 Frequency of the HD Mutation

The frequency of the HD mutation is usually expressed in terms of individuals heterozygous for the mutation, whether or not these individuals presently manifest HD. The heterozygote frequency of the disease is higher than its prevalence, since it includes premanifest individuals who have the mutation but do not show signs of HD. The frequency of heterozygotes is therefore equal to the prevalence of manifest HD patients plus the prevalence of premanifest individuals in the population. Various models have attempted to indirectly estimate the heterozygote frequency for HD from

disease prevalence, using life tables of average disease onset by age and theoretical autosomal dominant transmission rates (Conneally, 1984). Such indirect methods approximate the heterozygote frequency to be 2.5 to 3.0 times the manifest prevalence rate of disease. The ratio of symptomatic individuals to pre-symptomatic carriers may also fluctuate over time with changing population age and family structure.

Direct estimates of the heterozygote frequency of the HD mutation have been relatively unsuccessful. Several expanded CAG repeats (≥36) were observed in large clinical amyotrophic lateral sclerosis (ALS) and major depressive disorder cohorts, but no HD mutations were detected among several thousand comparator control chromosomes (Ramos et al., 2012; Perlis et al., 2010). One HD mutation was reported among 1772 chromosomes sized for CAG repeat length in the Portuguese general population (Sequeiros et al., 2010). Direct evaluation of HD mutation frequency in the general population has been challenging due to demanding sample sizes and the ethical restrictions associated with such tests. In addition, next-generation sequencing technologies have been largely unable to accurately map reads across typical CAG repeat lengths, necessitating locus-specific strategies. It is possible that some heterozygotes for the HD mutation never clinically manifest, particularly at the lower end of the HD range where age of onset is late and penetrance may be incomplete, but genotyping of the CAG repeat in larger samplings of the general population are necessary to evaluate this possibility.

1.5 Penetrance of the HD Mutation

The expanded CAG repeat demonstrates age- and length-dependent penetrance. The absence of direct estimates of heterozygote frequency in the general population hinders accurate penetrance estimates of the mutation, requiring approximation from clinically ascertained individuals. Survival analysis of manifest and premanifest individuals has enabled calculation of approximate probabilities of asymptomatic survival to specific ages. Prior to molecular testing, Newcombe constructed life tables that estimated a median age (50% likelihood) of disease onset by 46.0 years, with increasing likelihood of onset in successive age brackets (Newcombe, 1981). Following genetic testing, onset likelihood estimates were revealed to vary by expanded CAG repeat length, from an estimated median onset of 66 years at 39 CAG to 27 years at 50 CAG (Brinkman et al., 1997). Data from a consortium of forty centres enabled formulation of a parameterized survival model for calculating the probability of reaching a specific age without neurological symptoms for a given CAG length (Langbehn et al., 2004). The validity of this survival model has been prospectively validated using longitudinal observational data of premanifest individuals over a two-year period (Langbehn et al., 2010). Although penetrance is thought to be complete at inherited disease repeat lengths typically seen in neurological practice, it remains unclear how penetrant the HD mutation is across the entire CAG repeat range.

There remains considerable uncertainty regarding the penetrance of the HD mutation at CAG 36-40, where average onset is latest, since an unknown number of individuals with

repeats in this range may never present to neurological services (Rubinsztein et al., 1996). Langbehn estimated a 21% probability of onset by 80 years for an individual with 36 CAG repeats, rising to 93% at 40 CAG repeats, yet clinical ascertainment of individuals in this range was recognized to inherently overestimate the corresponding onset probabilities (Langbehn et al., 2004). Using a mutational flow model based on germline instability rates of expanded CAG repeats, Falush estimated clinical ascertainment of individuals with 36-39 CAG repeats in the population. Under this model, <50% of individuals with 40 repeats were expected to be clinically ascertained with HD in their lifetime, falling to <5% for individuals with 36-38 repeats (Falush et al., 2001). The penetrance of expanded CAG repeats may also increase in response to improved ascertainment of late-onset HD phenotypes.

1.6 The HD New Mutation Rate

Early attempts to calculate the HD new mutation rate were blind to the molecular basis of the disease. It was assumed that pathogenic mutations arose from random mutational events, subsequently passed down generations. We now know that all pathogenic HD alleles originate from expansions of intermediate alleles (IAs; 27-35 CAG) containing unstable CAG repeats that are longer than most normal repeats in the general population (Goldberg et al., 1993; Myers et al., 1993; Potter et al., 2004; Semaka et al., 2006). Calculating the new mutation rate has become more complicated as a result, since mutability differs between intermediate allele lengths, and some HD heterozygotes representing new mutations do not manifest the HD phenotype with a

low-penetrance mutation (Semaka et al., 2013a; Rubinsztein et al., 1996). Indeed, intermediate allele mutability and variable penetrance of the mutation imply that a single *de novo* mutation rate may not be meaningful, and that dynamic models of mutational flow across the CAG size range may more appropriately represent the emergence of new HD alleles in a population.

Before molecular characterization of the CAG repeat, the HD new mutation rate was estimated by methods from quantitative population genetics, such as direct counting of sporadic cases over the sum of the population, or Haldane's indirect calculation based on a balance of new mutation and reduced fitness (Conneally, 1984). However, sporadic HD cases were likely undercounted prior to molecular testing. The belief that *de novo* HD mutations were rare led to alternative explanations in sporadic cases, such as non-paternity or suspected HD in a deceased parent. New mutation rate estimates before discovery of the CAG repeat expansion (Shaw and Caro, 1982) may therefore be underestimates owing to the lack of genetic testing, misdiagnosis of individuals presenting without family history, and reduced penetrance of the HD mutation at shorter CAG repeat lengths (**Table 1-1**).

Given the quantitative difficulties introduced by the continuous CAG range in calculating the HD new mutation rate, studies after 1993 have instead focused on the more limited estimation of genetically confirmed HD cases presenting without family history. A proportion of such cases have been revealed as *de novo* CAG repeat expansions from parental alleles in the IA range, almost entirely transmitted from the father (Goldberg et

al., 1993; Semaka et al., 2006). An accurate counting of HD cases without family history, supported by molecular diagnostics and detailed medical history of the parents, may therefore approximate the HD new mutation rate in the population as a whole.

In New South Wales, Australia, McCusker showed that 11% of clinically ascertained HD cases reported no recorded family history (McCusker et al., 2000). Based on a comparison of clinical records with the Leiden roster of HD families across the Netherlands, Siesling found that at least 6.3% of patients diagnosed by genetic testing in the Netherlands had no relation to a known HD family (Siesling et al., 2000). These numbers were echoed by Almqvist in British Columbia (BC), Canada, where 8% of individuals testing positive for HD reported negative family history with medical documentation of normal parents both 65 years or older (Almqvist et al., 2001). Ramos-Arroyo reported that molecularly confirmed IA expansions represented a minimum *de novo* mutation rate of 4.7% in a Spanish clinical cohort, increasing to a maximum of 8.1% when cases with confirmatory parent medical records were included (Ramos-Arroyo et al., 2005). Based on these numbers, a plausible *de novo* mutation rate among presenting HD patients is between 5% and 10%. The true *de novo* mutation rate may be higher, given that *de novo* cases are expected to have shorter HD mutation lengths, manifest later in life, and thus be under-represented among all patients at any given point in time. Importantly, these retrospective studies show that new mutations are not rare among HD cases, and that all molecularly confirmed cases derive from expansion of an IA in one parent, nearly always the father.

Quantitative estimates of how IAs contribute to HD prevalence have been attempted. Such calculations necessarily depend on estimates of IA frequency and the germline rate of IA expansion. IA frequency appears to be consistent across populations of European ancestry, averaging 3.2% of CAG repeat alleles in the largest sample sets (Perlis et al., 2010; Sequeiros et al., 2010; Ramos et al., 2012; Semaka et al., 2013b). Using a maximum likelihood estimate of IA frequency in the population, the HD new mutation rate was inferred by dividing the per-year likelihood of a *de novo* HD case by the total number of births originating from a paternal IA transmission (Hendricks et al., 2009). This study estimated that between 1 in 951 to 1 in 6241 births from fathers with an IA genotype will result in transmission of an expanded HD allele. Applying this rate to IA carrier frequency, approximately 3.7×10^{-6} to 2.8×10^{-5} new mutations for HD were estimated to occur among all births. Empirical new mutation rates using size-specific IA CAG repeat instability data, rather than indirect extrapolations from the population, have not been reported.

Table 1-1. New mutation rate estimates for HD in the total population

Direct Estimate	No. of Mutants /10^6 gametes	Indirect Estimate	No. of Mutants /10^6 gametes	Reference
–	—	6.7×10^{-7}	0.67	Kishimoto et al., 1957
5.4×10^{-6}	5.4	9.6×10^{-6}	9.6	Reed & Neel, 1959
—	—	1.5×10^{-6}	1.5	Wendt & Drohm, 1972
5.0×10^{-6}	5	0.8×10^{-6}	0.8	Mattsson, 1974
$0.42–4.0 \times 10^{-6}$	0.42 to 4	7×10^{-7}	0.7	Stevens, 1976
1.3×10^{-7}	0.13	—	—	Hayden et al., 1979
5.0×10^{-7}	0.5	—	—	Shaw & Caro, 1982
$0–1.0 \times 10^{-6}$	0 to 1.0	—	—	Walker et al., 1983
2.0×10^{-5}	20	—	—	Falush et al., 2001
3.7×10^{-6} to 2.8×10^{-5}	3.7 to 28	—	—	Hendricks et al., 2009

Adapted from Hayden, 1981.

1.7 Haplotypes of the HD Mutation

Haplotypes of the *HTT* gene region assisted in mapping of the causative trinucleotide CAG repeat expansion (MacDonald et al., 1992). The HD mutation occurs on distinct haplotypes at the same locus, suggesting multiple mutational origins (Andrew et al., 1993a). Haplotypes in which the HD mutation is found also differ markedly by ethnic group. These differences have been proposed to underlie differences in the prevalence of HD between European ancestry populations and ethnically distinct populations in East Asia and Africa.

Core haplotypes of the *HTT* gene region were originally constructed from the dinucleotide repeat at D4S127 and the multiallele VNTR and *Taq*I RFLP at D4S95

(Wasmuth et al., 1988; MacDonald et al., 1992). A haplotype present on approximately 30% of HD chromosomes but only 5% of controls demonstrated strong linkage disequilibrium within a disease-linked 500 kb region between D4S180 and D4S182 (MacDonald et al., 1992). Within a year of identification of the polyglutamine CAG codon repeat, a polyproline CCG codon repeat immediately downstream of the CAG repeat was also identified, ranging from 7 to 12 repeats in length with 7 CCG and 10 CCG being most common alleles in the general population (Rubinsztein et al., 1993; Andrew et al., 1994c; Barron et al., 1994; Novelletto et al., 1994b). In European ancestry individuals the CCG 7 allele was found to be in linkage disequilibrium with the HD mutation, occurring on 93% of HD mutation chromosomes as opposed to 67% of control chromosomes. The CCG 10 allele, present on 30% of control chromosomes, instead occurred on only 7% of HD mutation chromosomes. Discovery of the glutamic acid codon deletion among four repeated GAG codons in exon 58 of *HTT*, denoted as ∆2642, provided another HD-enriched polymorphism present on 38% of HD mutation chromosomes but only 7% of controls (Ambrose et al., 1994; Almqvist et al., 1995; Squitieri et al., 1994; Rubinsztein et al., 1995).

European ancestry HD haplotypes with the ∆2642 marker were universally found on chromosomes with the CCG 7 allele (Squitieri et al., 1994; Rubinsztein et al., 1995). Compellingly, normal chromosomes on this CCG7-∆2642 haplotype demonstrated significantly higher average CAG repeat length than on other haplotype backgrounds. Exploration of CAG repeat variation and the surrounding genetic haplotype in European, black South African, and East Asian groups yielded the finding that HD haplotypes differ

both within and between these groups (Squitieri et al., 1994; Almqvist et al., 1995; Rubinsztein et al., 1995). Therefore, the genetic origins of HD are diverse not merely within populations, as revealed by positional cloning efforts, but also between groups of distinct ethnic origin. Whereas European ancestry HD mutations are linked to CCG 7 haplotypes, both with and without the ∆2642 marker, Japanese HD haplotypes are in high linkage disequilibrium with CCG 10 (Masuda et al., 1995). Intriguingly, ∆2642 was found to be completely absent among all Japanese HD and control chromosomes. These haplotype differences were immediately postulated to account for the high HD prevalence in European ancestry populations, perhaps by *cis*-acting modification of CAG repeat instability by ∆2642, the CCG 7 background, or other linked polymorphic elements (Vuillaume et al., 1998).

Association of European ancestry HD mutations with the CCG 7 and ∆2642 markers was replicated over the following years in other Western populations, including Greece (Yapijakis et al., 1995), Hungary (Jakab et al., 1999), Turkey (Ataç et al., 1999), Croatia (Hecimovic et al., 2002), Russia (Kutuev et al., 2004), Spain (Garcia-Planells et al., 2005), and Portugal (Costa et al., 2006). Intriguingly, all HD chromosomes in a sample of 48 Venezuelan families were found on the CCG 7 background, but only one pedigree carried the ∆2642 deletion (Paradisi et al., 2008), suggesting a single European origin to the majority of cases in that country. The association of CCG 10 with HD mutations in East Asia was replicated in Taiwan (Wang et al., 2004) mainland China (Ma et al., 2010), and in a second Japanese cohort (Morovvati et al., 2008). The HD-associated ∆2642 marker common to Europeans but absent in East Asian populations appears to

be present in India, albeit at lower frequencies (Saleem et al., 2003). Indian HD chromosomes appear to be predominantly associated with CCG 7, like European populations (Pramanik et al., 2000).

With the advent of public single nucleotide polymorphism (SNP) databases, new resources became available for the generation of detailed haplotypes at the *HTT* locus. The first SNP-based haplotypes of the HD gene region were constructed by Scholefield & Greenberg in South Africa to discern the molecular origins of HD in the mixed subpopulation of that country (Scholefield and Greenberg, 2007). Although only one SNP of a genotyped six was available intragenic to *HTT*, common haplotypes were found in two Afrikaans-speaking Caucasian families and four of six mixed ancestry families, supporting a genetic link of Dutch origin. Dense intragenic SNP haplotyping has confirmed a common European background of HD in these two groups, whereas the majority of HD mutations in black people are on haplotypes indicative of endemic African origin (Baine et al., 2013).

Comprehensive SNP-based haplotype analysis of the *HTT* gene was first performed by Warby (Warby et al., 2009), in which 22 tagging SNPs from International HapMap Project, each representing informative linkage disequilibrium blocks, were genotyped and phased to the CAG repeat in 65 Caucasian HD patients and 58 controls individuals. This approach, in which most SNPs were intragenic, allowed for detailed description of haplotypes with downstream potential for SNP-based silencing of the mutant *HTT* transcript. The two most strongly HD-associated intragenic haplotypes, A1 and A2, were

also enriched on IAs and high-normal alleles. From these results, a hypothesis favoring expansion-prone predisposing haplotypes was developed, but an alternative explanation that A1 and A2 represent allele populations originating from high-normal CAG repeat founders in the distant past could also be proposed (Falush, 2009; Warby et al., 2009). Such high-normal founders would lead over time by stepwise expansion to a pool of intermediate CAG repeat alleles on the same haplotype background, without instability-prone *cis*-elements intrinsic to each haplotype. Similar patterns of SNP-based haplotype association in a different European ancestry HD population were reported by Lee (Lee et al., 2012a).

Warby expanded his study of HD haplotypes to East Asians, showing that Caucasian A1 and A2 haplotypes were entirely absent among Japanese and Chinese chromosomes (Warby et al., 2011). Further, haplotypes of the HD mutation in East Asian patients differs substantially from that of European ancestry Caucasians, suggesting a different mutational origin with associated effects on prevalence. Whether disparities in prevalence result from haplotype-specific differences in mutability, as suggested by Warby, remains unclear. But different haplotypes appear to mark different mutational origins of the HD mutation in ethnically distinct groups.

1.8 The Normal CAG Repeat Distribution

One plausible connection between HD-associated haplotypes and prevalence concerns the pool of intermediate alleles from which new mutations originate. Together with early

haplotyping studies, it was found that populations with elevated HD prevalence rates are likewise enriched for high-normal CAG repeat alleles (Rubinsztein et al., 1994). Intriguingly, this connection between high-normal trinucleotide repeat lengths and disease prevalence has been observed in other trinucleotide repeat disorders, including dentatorubral-pallidoluysian atrophy (DRPLA) (Yanagisawa et al., 1996) and myotonic dystrophy (Goldman et al., 1996). The common patterns observed across these disorders suggest a common mechanism contributing to disease prevalence, resulting from the mutational properties of trinucleotide repeats at progressively longer lengths. Differences in mean CAG length were recognized to associate with prevalence rates early after identification of the CAG repeat (Kremer et al., 1994) and were first tabulated in detail by Squitieri (Squitieri et al., 1994). A significantly lower mean CAG size was observed in black South Africans, Chinese, Japanese, and Finnish control samples compared with consistently high averages among European samples (**Table 1-2**). These differences correspond to rates of HD prevalence in the respective populations. Moreover, this bias in European populations was found to be directly attributable to HD-associated haplotypes containing CCG 7 and ∆2642, themselves present in conjunction with CAG lengths longer than the population average. These findings were reproduced by Rubinsztein, who additionally characterized a positive skew in Caucasian CAG repeat distributions, suggesting a mutational bias in favor of expansion with increasing CAG repeat length (Rubinsztein et al., 1994). Ethnic differences in mean CAG size were confirmed in a Chinese population (Soong and Wang, 1995) and in multiple local haplotyping studies described earlier. The consistent pattern across these reports is that normal CAG lengths are longer, on average, in populations of European ancestry, and

that this high average is attributable to haplotypes associated with high-normal CAG repeat alleles. Haplotypes of high-normal CAG repeat alleles in European populations are likewise enriched among HD alleles, suggesting that high-normal alleles may act as the reservoir from which the intermediate and HD alleles ultimately arise.

Table 1-2. Reported mean normal CAG repeat length by ethnic group

Ethnicity	Mean CAG Length	Reference
Caucasian		
Canadian, Swedish, Italian	18.4 ± 3.7	Squitieri et al., 1994
German	18.69	Deka, 1999
Turkish	19.3 ± 2.9	Ataç, 1999
Croatian	18.5 ± 3.2	Hećimović, 2002
Portuguese	18.4 ± 0.1	Costa, 2006
Finnish	17.1 ± 1.8	Squitieri et al., 1994
East Asian		
Chinese	16.4 ± 1.5	Squitieri et al., 1994
Japanese	16.6 ± 1.3	Squitieri et al., 1994
Chinese	17.49	Deka, 1999
Taiwan Chinese	17.7 ± 2.0	Wang, 2004
Mainland Chinese	17.6 ± 1.4	Ma, 2010
Black African		
Black South African	16.2 ± 2.5	Squitieri et al., 1994
Beninese	17.36	Deka, 1999
Other/Mixed		
New Guinean	15.95	Deka, 1999
Indian	16.8 ± 2.08	Pramanik, 2000
Indian	17.99 ± 2.66	Saleem, 2003
Brazilian	17.7	Raskin, 2000
African Brazilian	17.9	Raskin, 2000
Mexican	19.0	Alonso, 2009

Expanded trinucleotide repeat length also associates with prevalence in other trinucleotide repeat disorders, demonstrated in two comparative studies that examined multiple repeat loci across control populations from different ethnic groups. High-normal alleles at the HD and myotonic dystrophy (DM) loci are enriched among Caucasians, in whom those diseases are more common, as compared to Africans or East Asians in whom they are rare (Watkins et al., 1995). Likewise, high-normal alleles for the DRPLA locus are much more common among Asians, in whom prevalence is dramatically higher than for Caucasians or Africans. Deka extended this analysis to seven disease-causing trinucleotide repeat loci, including HD, DM, and DRPLA, in sample sets of German, Beninese, Chinese, and New Guinean origins, reproducing similar findings (Deka et al., 1999). Of interest, the lowest average *HTT* CAG repeat length of any sample set was reported in New Guinea highlanders, at 15.95 CAG repeats.

1.9 Origin and Ancestry of the HD Mutation

These studies of mean CAG length, in conjunction with haplotype studies linking high-normal alleles, IAs, and expanded HD alleles in Caucasians, all suggest a mechanism of gradual stepwise expansion across the CAG repeat range. In this context, HD prevalence in established populations can be understood as a long-term consequence of very small expansion biases in progressively longer CAG repeats, which accumulate over many generations into IAs and ultimately expand into the HD range. Populations may therefore be expected to have higher HD prevalence rates as a function of high-

normal allele frequencies, with haplotypes linking the two allele pools (Andrew and Hayden, 1995). Historical founder events, such as that in Venezuela, may represent local exceptions to a baseline new mutation rate inherent to each population.

1.10 Therapeutic Gene Silencing of the HD Mutation

1.10.1 Therapeutic Rationale for Gene Silencing in HD

HD is marked by a dominant gain-of-function pathology, where therapeutic silencing of the mutant gene copy may be preferable to silencing both gene copies non-selectively. Except in rare homozygotes, all HD patients have >35 CAG repeats on one *HTT* copy but a nonpathogenic normal (<26) or intermediate (27-35) number of repeats on the other chromosome (Kremer et al., 1994). Transgenic animal studies have established that translation of the pathogenic repeat expansion is necessary for the development of HD (Goldberg et al., 1996) and that levels of transgenic expanded polyQ protein correlate with neuropathology (Graham et al., 2006). Conditional deactivation of the mutant transgene is sufficient to reverse disease phenotypes (Yamamoto et al., 2000). On the basis of these observations, targeted reduction of the mutant HTT (muHTT) transcript by gene silencing has emerged as a major focus of HD therapeutic research.

Multiple lines of evidence suggest that simultaneous reduction of wild-type HTT (wtHTT) may impact safety or therapeutic efficacy. HTT is a large, highly conserved protein with no homologues (Tartari et al., 2008), and it interacts with a vast array of partners

(Shirasaki et al., 2012). HTT has diverse roles in transcriptional regulation, mitochondrial activity, synaptic function, and other cellular processes (Zuccato et al., 2010). Constitutive knockout of murine huntingtin (Htt) is embryonic lethal (Nasir et al., 1995; Duyao et al., 1995; Zeitlin et al., 1995), and low levels of expression lead to abnormal development (Auerbach et al., 2001). Inducible knockdown of Htt in young adult mice results in neurological deficits and premature death (Dragatsis et al., 2000). Mice constitutively hemizygous for Htt develop to adulthood but develop aberrant neurological and neuropathological phenotypes (Nasir et al., 1995; O'Kusky et al., 1999). A major goal of therapeutic HD gene silencing is, therefore, to reduce expression of human muHTT while retaining expression of the wild-type protein.

In addition to a clear role in development and adult neurological function, protective effects of wtHTT in HD have also been demonstrated. In general, these studies show that lower levels of wtHTT relative to muHTT result in more severe HD phenotypes (Leavitt et al., 2001; Van Raamsdonk et al., 2006; Van Raamsdonk et al., 2005; Leavitt et al., 2006). Levels of wtHTT thus appear to modulate aspects of polyQ toxicity. These studies suggest that absolute levels of muHTT compared to wtHTT may play a role in toxicity, with subtle loss-of-function effects in addition to, or perhaps cumulative to, the primary toxic effects of the polyQ protein. Uniform reduction of muHTT and wtHTT in humans may have unpredictable consequences. In light of the above considerations, allele-specific *HTT* silencing is a preferred strategy to minimize adverse effects of therapeutic *HTT* gene silencing approaches.

Allele-specific silencing uses antisense RNA or DNA designed to selectively bind and downregulate complementary RNA carrying a specific genetic variant. Allele-specific silencing depends on the presence of a defined polymorphism on the target transcript, typically encoded in one gene copy but not the other. In HD, genetic targets for allele-specific silencing include the expanded CAG repeat and single nucleotide polymorphisms (SNPs) in linkage disequilibrium with the expansion.

1.10.2 Targeting Expanded CAG

Nearly all HD patients share heterozygosity at the CAG repeat itself, which is longer than 35 repeats on the pathogenic allele but rarely exceeds that number on the other allele. Targeting the expanded CAG repeat is, therefore, an attractive method for selectively lowering muHTT. CAG repeat-targeted reagents have seen recent advances in selectivity, particularly with the use of CAG-complementary ss-siRNA mismatched at one or more positions (Yu et al., 2012; Fiszer et al., 2011; Fiszer et al., 2013). Greater than 30-fold selectivity against muHTT has been demonstrated in a juvenile-onset HD patient cell line with 69/17 CAG repeats (expanded/normal CAG repeat length), although wtHTT expression was reduced by ~50% in a 44/15 cell line with smaller CAG size difference (Yu et al., 2012). Self-duplexing siRNAs mismatched to the CAG repeat show comparable potency and selectivity for muHTT in 69/17 fibroblasts, but also result in ~50% off-target silencing of the wild-type transcript in 44/21 and 47/18 cells (Fiszer et al., 2013). Selective muHTT knockdown has also been accomplished using CAG-complementary antisense oligonucleotides (ASOs) in similar cell lines (Gagnon et al.,

2010). These findings suggest that advances in CAG-targeted selectivity may be possible with novel nucleic acid chemistries in development. At minimum, CAG-directed silencing may represent a key therapy for juvenile HD patients, where pathogenic CAG repeats are found at extreme lengths beyond 60 CAG. But for the majority of HD patients, SNP-directed silencing may present a more selective alternative.

Off-target silencing is a risk with CAG-targeted antisense reagents given that CAG repeats occur normally in no fewer than 66 protein-coding transcripts and an unknown number of non-coding sequences (Butland et al., 2007). Expression of other CAG repeat-containing transcripts must not be harmfully reduced by CAG-targeted silencing reagents intended for *HTT* transcripts. Indeed, the ability of such methods to simultaneously target the CAG repeat of *ATXN3*, pathogenically expanded in spinocerebellar ataxia type 3 (SCA3), has been proposed as an advantage in the development of these compounds (Hu et al., 2009). While expression of five genes containing CAG or mixed CAG/CAA repeats was not reduced in cells treated with ss-siRNA targeting CAG-expanded *HTT* transcript (Yu et al., 2012), the full range of CAG-coding transcripts, including wild-type *ATXN3*, remains to be investigated. Given the wide species divergence of trinucleotide repeat lengths even in comparison to close primate relatives (Molla et al., 2009), deleterious off-target effects of CAG-targeted reagents will remain unclear until human trials.

1.10.3 Targeting SNPs in Linkage Disequilibrium with the CAG Expansion

SNP-targeted allele-specific silencing *HTT* was first demonstrated in an endogenous cellular context using human HD fibroblasts heterozygous for rs363125 in exon 39 of *HTT* (van Bilsen et al., 2008). In cells transfected with siRNA complementary to the expansion-linked cytosine allele, levels of target *HTT* transcript were reduced approximately 80%, whereas transcript containing the off-target adenine allele remained stable. Further studies with mismatched siRNA (Pfister et al., 2009) and chemically modified ASOs (Carroll et al., 2011; Ostergaard et al., 2013) have expanded the repertoire of oligonucleotide designs enabling potent and selective silencing of muHTT transcript carrying a specific SNP target. In the case of ASOs, modifications aimed at blocking secondary RNase H cleavage sites have now enabled near-complete silencing of muHTT with a given SNP allele over a broad dosing window and no significant silencing of wtHTT (Ostergaard et al., 2013).

siRNA reagents can only be designed against SNPs present in mature mRNA, limiting available targets for development. Conversely, ASOs trigger RNase H activity against both cytoplasmic mRNA and nuclear pre-mRNA (Bennett and Swayze, 2010) and can be designed against polymorphisms present anywhere in the transcribed genomic sequence including introns and UTRs. As the majority of polymorphisms across the *HTT* transcript are found within introns (Southwell et al., 2012), ASOs allow greater choice in allele target selection compared to reagents that only target mature mRNA.

The possibility of selectively silencing muHTT without the need to carefully monitor wtHTT presents an advantage in clinical trials of SNP-targeted reagents, where dosing across tissues of the brain is likely to be variable (Kordasiewicz et al., 2012). In the case of cerebrospinal fluid delivery, the brunt of HTT reduction is likely to occur in the outer cortex. Under current constraints of selectivity, excess dose of CAG-targeted antisense reagents may elicit suppression of wtHTT in an effort to reach therapeutic concentrations in deeper structures such as the caudate and putamen. SNP-targeted ASOs appear relatively resistant to this effect given appropriate design (Ostergaard et al., 2015). Like CAG-targeted approaches, SNP-targeted reagents can also lead to deleterious off-target effects by binding to complementary sequences elsewhere in the transcriptome. However, oligonucleotide sequences spanning a target SNP can be subjected to stringent alignment screens against the human transcriptome, reducing potential for adverse silencing before costly preclinical development.

SNP-specific *HTT* silencing crucially depends on the availability of polymorphisms present on the mutant transcript but absent on the wild type. In addition to high linkage disequilibrium with the CAG expansion, a target SNP allele must also be relatively uncommon on chromosomes with the normal CAG repeat allele, and thus offer high specificity for the mutant transcript. SNP heterozygosity captures both measures, and can help determine the portion of HD patients treatable with an antisense drug targeted to a specific allele. Although many SNPs heterozygous in HD patients have been found, no single target SNP will allow treatment of all HD patients. The highest heterozygosity rates for SNPs in HD patient cohorts are roughly 50%, with additional SNPs providing

smaller cumulative increases in overall patient coverage (Lombardi et al., 2009; Pfister et al., 2009; Warby et al., 2009). If appropriately selected so as to avoid target redundancy, SNP-specific approaches will require a minimum of three different SNPs, each targeted by a different antisense compound, to achieve selective silencing treatment in >80% of HD patients (Warby et al., 2009). Prior to commencing human trials, additional genotyping will also be necessary to determine suitable candidate patients for treatment with a given drug.

1.10.4 Population Genetics of Allele-Specific Silencing in HD

1.10.4.1 Expanded CAG Repeat Silencing

The overall treatable portion of the HD patient population will depend on the ability of CAG-repeat directed methods to discriminate between the mutant CAG expansion and the normal CAG allele at specific size differences. We estimate that potent and selective discrimination of CAG repeats differing by no more than 25 CAG will be necessary to treat at least 50% of the patient population (**Figure 1-1**). A minimum safe threshold may be advised in clinical use, and reduction of wtHTT in patients near this threshold may require careful monitoring. The variety of size combinations at the CAG repeat locus may complicate trial design, as each individual may exhibit a different response to treatment specific to their combination of CAG alleles. While CAG-directed methods have rapidly improved in recent years and may continue to do so, it remains unlikely that CAG repeat-directed silencing can sufficiently or safely treat all HD patients.

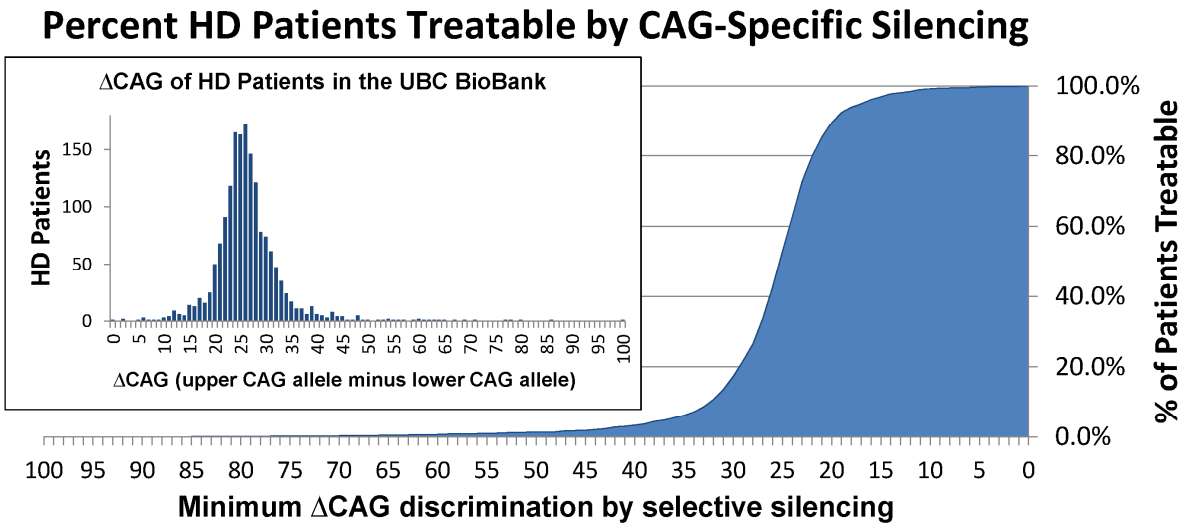


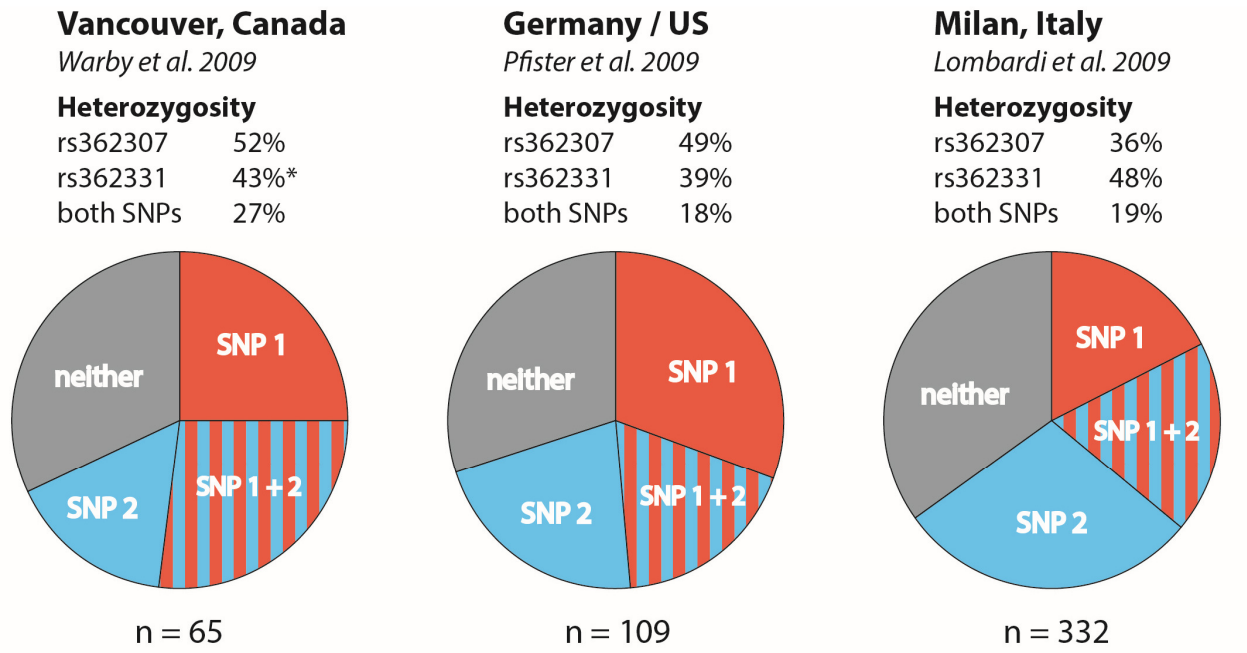
Figure 1-1. Percent of HD patients treatable with CAG-specific silencing as a function of CAG repeat discrimination

The portion of the HD patient population that may be treated by CAG-specific silencing of the HD mutation depends on the minimum difference in CAG lengths that can be selectively discriminated by a CAG-targeted compound. The difference in CAG length between the upper and lower alleles (ΔCAG) is shown for 1638 HD patients from the UBC BioBank, resembling a normal distribution (inset). The treatable percent of this population is calculated as a function of the minimum ΔCAG that can be discriminated by allele-specific CAG-targeting reagents. For example, a patient with 44/17 CAG alleles would require minimum ΔCAG discrimination of 27 CAG units. Treatment of 50% of the HD patient population requires a minimum CAG-specific discrimination of 25 repeats, such as for an individual with 42/17 CAG repeat alleles.

1.10.4.2 SNP-Specific Silencing

Accurate population genetics of transcribed *HTT* polymorphic alleles in linkage disequilibrium with the expanded CAG repeat are necessary to select appropriate therapeutic targets for development of allele-specific antisense compounds. Population genetic studies have lagged behind antisense pharmacology, and the full extent of variants heterozygous across all HD patients remains unclear.

Dozens of polymorphisms have been assessed for heterozygosity in HD cohorts to guide allele-specific target identification. For example, sequencing 26 SNP sites located in the exons or 3'UTR of 332 Italian HD patients revealed that rs362331 was heterozygous in 47.6% of patients (Lombardi et al., 2009). A similar list of 24 exonic and 3'UTR SNPs were sequenced in a US/German cohort, where rs362331 was heterozygous in only 39.4% of patients but rs362307 was heterozygous in 48.6% (Pfister et al., 2009). Interestingly, rs362307 was found to be heterozygous in 36.0% of Italian patients, suggesting fine-scale genetic differences between these two cohorts (**Figure 1-2**). In a series of 65 Canadian HD trios, phased genotypes of rs362307 were found to tag the most common local haplotype of the HD mutation, denoted A1, present on 55% of HD chromosomes but only 9% of control chromosomes (Warby et al., 2009). It is evident from these studies that specific target SNP alleles in linkage disequilibrium with the HD mutation are similar across Caucasian patient populations and that half of any given cohort is treatable with a single target. However, the optimal target allele in a given cohort or subpopulation is not necessarily the same across all populations.



*rs362331 heterozygosity in Warby et al. 2009 inferred from rs3856973 (pairwise $r^2 = 0.948$, D' = 0.982, HapMap3)

Figure 1-2. HD patients heterozygous for specific SNP targets in different populations

rs362307 and rs362331 are highly heterozygous SNPs in three different HD patient cohorts of European ancestry. A portion of HD patients are heterozygous for both SNPs. ASOs targeted to rs362307 and rs362331 could treat at least two-thirds of the HD patient population in three different cohorts of European ancestry: 68% in Canada, 70% in Germany/US , 65% in Italy. rs362307 tags the most common HD haplotype, A1.

Individual SNP alleles often denote haplotypes of variants in linkage disequilibrium, and the CAG expansion is known to preferentially occur on specific haplotypes of the *HTT* gene in patients of European ancestry (Warby et al., 2009; Lee et al., 2012a). Differences in SNP heterozygosity between Caucasian subpopulations may represent local differences in *HTT* haplotype distribution of control chromosomes, CAG-expanded chromosomes, or both. These local haplotype differences can crucially inform efficient

SNP-targeted silencing of the mutant HD transcript, in that more than one allele target may be necessary to treat the majority of HD patients across European ancestry populations on account of differences in haplotype frequency.

Because SNP alleles are organized into constituent haplotypes, multiple variants with high heterozygosity in HD patients may not necessarily treat a higher percentage of patients in combination. For example, although 46.2% of Italian patients were heterozygous for rs362331 and 36.0% for rs362307, the percentage of patients treatable with either target, including those targetable with both, was found to be 65.1% (**Figure 1-2**). This is likely due to common haplotypes on which both target alleles occur. Some patients will be heterozygous for only rs362307 or only rs362331, and hence selectively treatable at only one target, while others may be homozygous for both polymorphisms and be limited to nonselective treatment with reagents targeting those SNPs. Genotyping studies of Caucasian HD patients have estimated the combination of specific SNPs enabling the greatest selective treatment coverage within each study cohort, approximated by heterozygosity rates. Across these studies, two-thirds of HD patients are estimated to be treatable with only two SNP targets, rising to 75-80% of patients with three targets.

A further consideration of SNP-specific silencing approaches is genetic diversity in non-Caucasian populations, where HD is present at low prevalence rates and known to occur on distinct haplotypes. The A1 haplotype, as represented by rs362307, is absent in East Asian and Black African HD cohorts as well as in corresponding general

population controls (Warby et al., 2011; Baine et al., 2013). This suggests that optimal SNP-specific targets in Caucasians may not be useful in HD patients of diverse ethnic backgrounds. However, HD of European origin constitutes the major proportion of cases worldwide, and HD in South Asia may occur on similar haplotypes at comparable prevalence (Saleem et al., 2003). HD in admixed Caucasians, such as Hispanic patients, also appears to share a Caucasian genetic origin, and is likely to share similar haplotypes and allele targets for therapeutic use.

ASOs have demonstrated high potency, selectivity, and *in vivo* distribution among SNP-specific *HTT* silencing approaches (Ostergaard et al., 2013). *In vivo* efficacy trials in fully humanized Hu97/18 mice (Southwell et al., 2013), which have appropriate SNP heterozygosity, will be required to test the therapeutic potential of ASOs directed to specific HD-associated SNPs. Should these trials yield striking changes in disease phenotypes, the population limitations of SNP-specific silencing may be accepted for the sake of advancing promising drug candidates to human use.

1.10.5 Clinical Implementation of Personalized *HTT* Silencing

HD population genetics suggest that a primary SNP allele target would allow allele-specific gene silencing treatment in approximately 50% of the Caucasian HD patient population, with a secondary target covering nearly 20% more. Trials for therapeutic exon-skipping in Duchenne Muscular Dystropy (DMD) have recently been pursued, despite targeting only 13% of the DMD patient population (Goemans et al., 2011),

suggesting that clinical translation of personalized *HTT* gene silencing strategies is feasible in HD. Trial design for additional SNP-directed antisense molecules may also be streamlined by initial studies developed in cooperation with regulatory agencies, increasing the speed of approval for follow-up targets after an initial drug. Should SNP-selective ASOs prove to be therapeutically beneficial, the rapid development and clinical availability of ASOs to selectively treat 85% of the HD population may be possible. As personalized therapeutics advance in other domains of medicine, the ability to perform trials with multiple molecules targeted to SNP allele in high linkage disequilibrium across the same haplotype may also be permitted, in order to advance the best candidate molecule for a given target block of patients.

Beyond the development of clinical trials for allele-specific silencing, routine clinical use of personalized antisense reagents will also require a novel treatment paradigm. Individuals from the HD patient population can be genotyped for available polymorphic targets in parallel to diagnostic or predictive testing. The most appropriate drug for an individual patient genotype may then be selected from an existing panel. CAG-targeted silencing may work in conjunction with SNP-specific drugs by offering a therapeutic alternative in patients who are homozygous for all available SNP targets, or who are found to have especially large CAG repeat expansions. For the small minority of patients who are both homozygous for SNP-specific targets and carry a second CAG repeat allele near the HD range, nonselective *HTT* silencing or small-molecule approaches may be necessary.

1.11 Thesis Objectives

Detailed population genetic studies of the HD mutation allow for overlapping investigation of unresolved questions in HD epidemiology and therapeutic development. Complete haplotypes of the HD mutation enable the classification and prioritization of *HTT* alleles for therapeutic silencing in the greatest number of patients. Haplotype analysis in additional populations enables evaluation of the utility of specific allele targets in those groups as well as exploration of the ancestral origin of the mutation. Estimation of the HD mutation heterozygote frequency allows for inference of the penetrance and overall burden of the mutation in the population that could be addressed by gene silencing strategies. Finally, comparative study of the normal CAG repeat distribution in different populations enables estimation of the HD new mutation rate with insight into the underlying molecular determinants of disease prevalence. The objectives of this thesis were therefore to:

- **(Chapter 2) Define common haplotypes of the HD mutation for identification of allele-specific gene silencing targets in European ancestry populations**
- **(Chapter 3) Investigate targetable haplotypes of the HD mutation in novel patient populations worldwide**
- **(Chapter 4) Estimate the frequency and penetrance of the HD mutation in the general population**

- **(Chapter 5) Evaluate IA frequency and the HD new mutation rate in distinct populations worldwide**

Chapter 2: Huntingtin Haplotypes Provide Prioritized Target Panels for Allele-Specific Silencing in Huntington Disease Patients of European Ancestry

2.1 Synopsis

Huntington disease (HD, [MIM 143100]) is the most common monogenic movement disorder and among the most common diseases resulting from a specific mutation (Fisher and Hayden, 2014; Kay et al., 2014a). HD is caused by an expanded CAG repeat in exon 1 of the huntingtin gene (*HTT*), molecularly defined by more than 35 tandem CAG triplets in one copy of the *HTT* gene (1993; Kremer et al., 1994; Rubinsztein et al., 1996). Expanded CAG triplets encode similarly repetitive glutamine residues in the HTT protein, leading to multiple downstream pathogenic effects and selective neuropathology (Zuccato et al., 2010). The progressive symptoms of HD typically manifest in adulthood, deteriorating to death approximately 15-20 years after onset (Roos et al., 1993; Ross et al., 2014). Despite more than 20 years of research since discovery of the causative mutation, there are no therapies at present to alter the course of HD.

The defined monogenetic cause of HD, and its consequent gain-of-function toxicity, allow suppression of *HTT* as a therapeutic strategy (Southwell et al., 2012). Preclinical studies have shown reversal of HD phenotypes by inducible or exogenous silencing of

transgenic mutant *HTT* (Yamamoto et al., 2000; Kordasiewicz et al., 2012; Harper et al., 2005; Rodriguez-Lebron et al., 2005). However, reagents that suppress both wild-type HTT and mutant HTT may have detrimental long-term consequences. Constitutive loss of murine homolog *Htt* is embryonic lethal, and reduced *Htt* expression is associated with developmental and cognitive deficits (Nasir et al., 1995; Duyao et al., 1995; Zeitlin et al., 1995; Auerbach et al., 2001). Postnatal inactivation of *Htt* leads to neurodegenerative phenotypes, suggesting that adult reduction of HTT may be poorly tolerated over long-term treatment (Dragatsis et al., 2000). Wild-type HTT has also been shown to be protective against toxic effects of mutant HTT in a dose-dependent manner (Leavitt et al., 2001). Preferential silencing of the mutant *HTT* allele, preserving normal *HTT* expression, may therefore minimize loss-of-function effects and yield greater therapeutic benefit than nonspecific suppression of both *HTT* copies. Forthcoming trials will establish whether transient suppression of HTT is tolerated in the human brain, but the long-term safety and efficacy of nonspecific HTT suppression in HD patients remains unclear.

Single nucleotide polymorphism (SNP)-targeted silencing of mutant *HTT*, acting to degrade a mutant transcript bearing a specific target allele, has achieved potent reduction of mutant *HTT* with negligible effect on normal *HTT* expression (Carroll et al., 2011; Southwell et al., 2014; Skotte et al., 2014). Careful structure-activity studies of antisense oligonucleotides (ASOs) suggest that suppression of wild-type *HTT* may be avoided with SNP-targeted reagents given appropriate preclinical screens in model systems (Ostergaard et al., 2013; Southwell et al., 2014; Skotte et al., 2014). However,

specificity of therapy requires a patient to be heterozygous for the target allele and for that allele to be present on the CAG-expanded copy of *HTT*. A crucial question in the development of SNP-targeted reagents is therefore the identification of target alleles with the greatest heterozygosity in the HD patient population, allowing allele-specific therapy in the greatest number of patients. Various transcribed *HTT* polymorphisms have been genotyped in local patient cohorts (Warby et al., 2009; Pfister et al., 2009; Carroll et al., 2011; Lee et al., 2012a), but phased heterozygosity estimates, which are required to prioritize development of allele-specific therapeutics, are unavailable for large numbers of alleles across diverse patient groups. Heterozygosity of a given target allele can vary considerably between patient populations. The ∆2642 codon deletion present in exon 58 of *HTT* has been targeted for selective *HTT* silencing *in vitro* by siRNA (Rubinsztein et al., 1995; Zhang et al., 2009), but the frequency of this polymorphism among HD chromosomes varies from 59% in an American cohort (Lee et al., 2012a) to 18.6% in Italy (Novelletto et al., 1994a). No study has examined the phased heterozygosity and haplotype relationship of all common alleles in the *HTT* transcript and it therefore remains unclear which *HTT* polymorphism, or combination of polymorphisms, would offer treatment for the greatest number of patients across different ancestry groups.

The goal of this study was to identify the most efficient and useful combination of targets for allele-specific *HTT* suppression in the greatest number of HD patients of European ancestry. We present the first comprehensive investigation of haplotypes spanning the entire ~170kb *HTT* gene sequence using all common (>1% MAF) SNPs from the 1000

Genomes Project (Abecasis et al., 2010; Abecasis et al., 2012) and dense SNP genotyping in 1285 HD patients and relatives from Canada, Sweden, France, Italy, and Finland. We reveal a regular genetic architecture across *HTT*, enabling rational, haplotype-specific targeting of mutant *HTT* in the greatest portion of HD patients of European descent. We identify multiple novel targets for silencing mutant *HTT* and show potent, selective silencing of the mutant transcript using ASOs complementary to a biallelic indel (insertion-deletion polymorphism) on the primary HD haplotype. Our work provides a general target selection strategy for dominant genetic disorders occurring on specific founder haplotypes, and provides a prioritized panel of targets for allele-specific therapeutic approaches to HD.

2.2 Materials and Methods

2.2.1 Genotyping and Haplotype Assignment in Canadian Subjects

91 SNPs were genotyped in 858 Canadian Caucasian HD patients and relatives using the Illumina GoldenGate genotyping array and BeadXpress platform. Genotypes were called using Illumina GenomeStudio software, and 91-SNP haplotypes reconstructed using PHASE v2.1 (Stephens and Donnelly, 2003). Haplotypes were manually annotated, then phased to CAG repeat length and confirmed for sequence identity by familial relationship. 28/91 SNPs in our original panel are rare or occur predominantly in non-Caucasian ethnic groups, leaving 63 SNPs of >1% frequency in European populations (1000 Genomes Phase I). 51 of these 63 common SNPs occur within the

HTT gene sequence, and were used for annotation of intragenic haplotypes within the extended 63 SNP haplotype.

2.2.2 Analysis of *HTT* Haplotypes in the 1000 Genomes Project Cohort

Variant Call Format (VCF) files encompassing the *HTT* gene region (GRCh37 Chr4: 3034088-3288007, SHAPEIT haplotypes (Delaneau et al., 2008)) were downloaded from the 1000 Genomes Project Consortium (Phase I) (Abecasis et al., 2012) using the Data Slicer tool and analyzed in the R statistical computing environment. 2166 haplotypes of chromosome 4 were available from 1083 individuals. Chromosomes bearing the intragenic A1 haplotype were identified using rs362307, a previously defined tagging SNP (tSNP) (Warby et al., 2009). SNPs present on at least 90% of the chromosomes bearing rs362307 (i.e. at least 70 of 76 chromosomes with rs362307[T]) but less than 100 of all 2166 chromosomes were identified as candidate A1 markers for further analysis. A2 chromosomes were similarly identified using rs2798235 and rs363080, the defining A2 markers from manual 63-SNP haplotype annotation of the Canadian Caucasian cohort. Discovery of A2 variants in linkage disequilibrium followed a similar strategy as for linked A1 variants. A3 chromosomes were identified among A haplogroup chromosomes by exclusion of all chromosomes bearing any specific A haplotype-defining SNPs in our 63-SNP panel. A1, A2, and A3 subtype markers were defined as any SNP present on a subset of each haplotype, but on no other chromosomes in the complete 1000 Genomes data set. Following identification of all A1, A2, and A3 variants, phased genotypes of intragenic *HTT* SNPs present at ≥5%

EUR frequency in the 1000 Genomes Phase I data set were extracted from all 738 European chromosomes and manually annotated in comparison to our directly genotyped 63-SNP haplotype data.

2.2.3 Genotyping and Haplotype Assignment in European Subjects

200 Swedish, 100 French, 33 Finnish, and 291 Italian HD family members were identified from the UBC HD BioBank and in cooperation with IRCCS Neuromed in Pozzilli, Italy. All Swedish, French, and Finnish samples were collected in their respective countries of origin for HD gene mapping studies in the 1990s (Andrew et al., 1993a). Of these samples, 120 Swedish, 76 French, 22 Finnish, and 209 Italian family members were identified as phaseable for haplotype and CAG repeat length. All 63 haplotype-defining SNPs genotyped in the Canadian Caucasian cohort were genotyped in the selected European samples, with addition of six novel A1 and A1 subtype SNPs, five novel A2 SNPs, and one novel A3 subtype SNP. European samples were additionally genotyped at 15 SNPs not present in the 63 SNP panel but necessary for reconstruction of haplotypes inferred in prior 4p16.3 locus genotyping (Lee et al., 2012a). Haplotypes in European samples were reconstructed with PHASE v2.1 and manually annotated as for the Canadian Caucasian cohort.

2.2.4 Direct Genotyping of *HTT* A1 Variants

A1 markers rs149109767 and rs72239206 are biallelic indels, and were genotyped by fragment analysis in phaseable samples from the UBC HD BioBank with 63-SNP haplotype data. Genotypes of rs149109767 and rs72239206 were phased to SNP haplotype and CAG repeat length by familial relationship. In total, 454 phased, nonredundant HD chromosomes and 652 nonredundant control chromosomes were directly genotyped and phased to CAG repeat length. PCR products containing rs149109767 were amplified using dye-labeled del2642F (6FAM-GCTGGGGAACAGCATCACACCC) and del2642 R (CCTGGAGTTGACTGGAGACTTG). Products containing rs72239206 were amplified with delACTT 3F (GAGGATTGACCACACCACCT) and dye-labeled delACTT 3R (HEX-ATGTGGCCATTTGACACGATA). Primers were multiplexed for ease of genotyping, and PCR products analyzed by ABI 3730xl BioAnalyzer with GeneMapper software.

2.2.5 Design of A1-Targeted ASOs

Locked nucleic acid gapmer ASOs targeting the mutant ΔACTT (rs72239206) and reference sequence were designed in-house and synthesized by Exiqon on a fee-for-service basis. Oligos were resuspended in 1X TE and stored at -20C between transfection experiments.

2.2.6 Passive Transfection of HD Patient Cells with A1-Targeted ASOs

Human HD lymphoblasts previously haplotyped as A1/C1 (GM03620, CAG 59/18) were cultured in 2mL complete RPMI media (500,000 cells in 15% FBS + 1% pen-strep) with 78 nM, 312 nM, or 1250nM of ASO. Cells were incubated 120h, and harvested for western blot analysis as described previously (Skotte et al., 2014). Anti-non-muscle myosin IIA (Abcam ab24762) immunoblotting was used as a loading control.

2.2.7 Allele-Specific *HTT* mRNA Quantification

Human HD lymphoblasts previously haplotyped as A1/C1 (GM02176, CAG 44/18) were cultured in complete RPMI media (15% FBS + 1% pen-strep). $5x10^6$ cells were transfected by electroporation using the Amaxa Nucelofector Kit C for each ASO dose (5000, 1250, 312, and 78 nM) in 100 uL nucleofector solution. Transfected cells were cultured for 24h, and half of each culture pelleted for RNA extraction, cDNA synthesis, and allele-specific qPCR. Remaining cell culture was propagated for protein analysis at 72h. A1 and C1 *HTT* mRNA transcript was quantified in quadruplicate for each dose of each experiment using an allele-specific TaqMan probe designed to rs362331 (ABI, C___2231945_10) and normalized to GAPDH (ABI, 4333764F). Each transfection experiment was performed three times, with two transfection replicates for each dose in each experiment (n=4-6 for each data point).

2.2.8 Allele-Specific HTT Protein Quantification

Transfected lymphoblast culture was harvested at 72h and pelleted for quantitative western blot analysis. Cells were pelleted by centrifugation at 250 g for 5 min at 4C and stored at -80C. Proteins were extracted by lysis with SDP+ buffer and 70 µg of total protein were resolved on 10% low-BIS acrylamide gels and transferred to 0.45 µm nitrocellulose membrane as previously described (Skotte et al., 2014). Membranes were blocked with 5% milk in PBS, and then blotted with anti-HTT antibody 2166 (Millipore) for detection of HTT. Anti-non-muscle myosin IIA (Abcam ab24762) immunoblotting was used as a loading control. Secondary antibodies, IR dye 800CW goat anti-mouse (Rockland 610-131-007) and AlexaFluor 680 goat anti-rabbit (Molecular Probes A21076), were used for detection and membranes were scanned using the LiCor Odyssey Infrared Imaging system. Licor Image Studio Lite was used to quantify the intensity of the individual bands (n=5-6 for each data point). Figure data are presented as mean +/- SEM. Two way ANOVA with Bonferroni post hoc test was performed for each dose series and p-values illustrated with ** and *** for p=0.01 and p=0.001, respectively. Representative images for HTT were chosen.

2.2.9 In Vivo ASO Treatment

YAC128 HD model mice (Slow et al., 2003) were maintained under a 12 h light:12 h dark cycle in a clean facility and given free access to food and water. Experiments were performed with the approval of the animal care committee of the University of British

Columbia. ASOs were delivered by intracerebroventricular injection as in (Southwell et al., 2014) at the indicated doses diluted to a final volume of 10 µl in sterile PBS. Four weeks later, brains were collected and sectioned in a 1 mm coronal rodent brain matrix (ASI Instruments). The most anterior 2 mm section, containing mostly olfactory bulb, was discarded. The next most anterior 2 mm section, containing mostly cortex and striatum, was divided into hemispheres and lysed as previously described (Southwell et al., 2014). 40 µg total protein was used for allele-specific HTT protein quantification as above.

2.2.10 Ethics Statement

Consent and access procedures were in accordance with institutional ethics approval for human research (UBC certificates H05-70532 and H06-70467). Publically available human lymphoblast cell lines were obtained from NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research (http://ccr.coriell.org).

2.3 Results

2.3.1 SNPs Across *HTT* Represent Specific Gene-Spanning Haplotypes

In order to determine the frequency and heterozygosity of all common allele-specific *HTT* targets relative to one another, we sought to establish haplotypes for a large number of common polymorphisms across the gene region. Various partial haplotypes

have been constructed across *HTT,* which overlap ambiguously due to low marker density in each study. We previously genotyped 91 SNPs across the *HTT* gene region (Carroll et al., 2011), of which 63 are present at greater than 1% frequency in European populations (Abecasis et al., 2012). Of these 63 common SNPs, 51 are located between the start of the *HTT* 5'UTR and the end of the 3'UTR (chr4:3076408-3245687, GRCh37). In total, 527 Canadian HD patients and 305 unaffected (control) relatives from the UBC HD Biobank were genotyped and phased at all 63 SNPs for this study. Using patterns of familial segregation, we reconstructed gene-spanning haplotypes at all 63 SNPs for 293 unrelated CAG-expanded chromosomes (CAG > 35) and 829 control chromosomes (CAG ≤ 35) from Canadian individuals of European ancestry.

Annotation of dense 63-SNP *HTT* haplotypes revealed the same haplogroup assignments using 22 tSNPs across the *HTT* gene region (Baine et al., 2013; Warby et al., 2009), and revealed that recombination between common haplotypes rarely occurs within the transcribed *HTT* gene sequence (**Figure 2-1a**). For example, the A3 haplotype is frequently associated with an extragenic 5' crossover with the C1 haplotype, whereas no common haplotype recombines C1 within the *HTT* gene sequence. Only 9/283 (3.2%) HD chromosomes and 25/829 (3.0%) control chromosomes in our Canadian cohort represent intragenic recombinants of gene-spanning *HTT* haplotypes, confirming that recombination within *HTT* is rare.

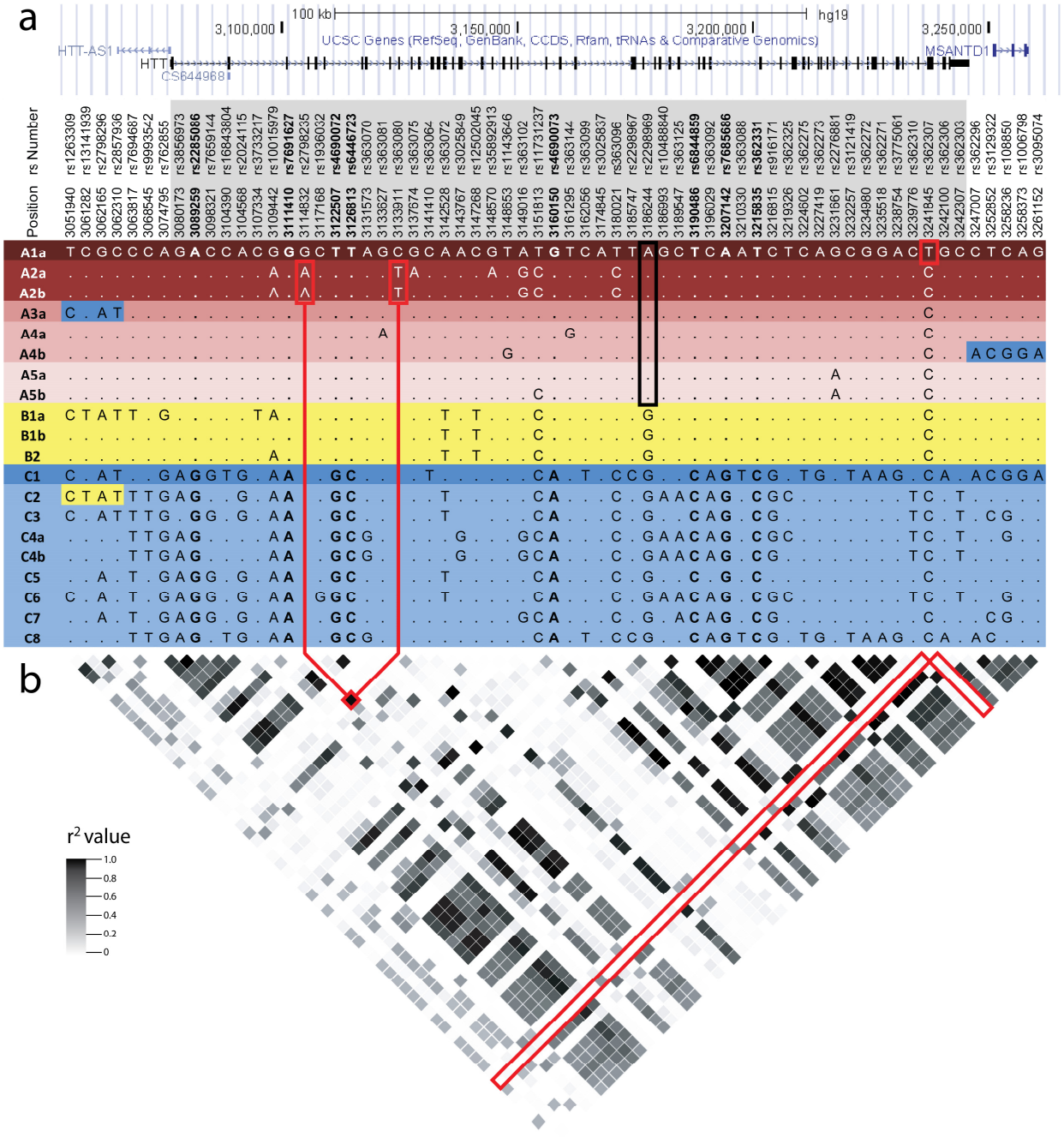


Figure 2-1. SNPs across *HTT* represent specific gene-spanning haplotypes.

(a) Single-nucleotide polymorphisms (SNPs) across *HTT* belong to gene-spanning haplotypes representing three major haplogroups A, B, and C. Transcribed (intragenic) SNPs are shaded gray. The primary Huntington disease (HD) haplotype, A1, is defined by rs362307. The secondary HD haplotype A2

is defined by rs2798235 and rs363080. One SNP (rs2298969) distinguishes the A haplogroup from the B and C haplogroups (black box). In our panel, 8 of 51 intragenic SNPs exclusively distinguish the A and B haplogroups from the C haplogroup (bold). (b) Pairwise linkage disequilibrium (LD) of SNP genotypes (r^2) reveals a complex haplotype structure across the *HTT* gene region (black, $r^2 = 1$; shades of gray, $1 > r^2 > 0$; white, $r^2 = 0$). Alleles present on similar haplotypes are in high pairwise LD across *HTT*, such as A2-defining rs2798235 and rs363080. Haplotype A1-defining SNP rs362307 is not in LD with any other variant in our initial 63-SNP panel. Positions correspond to GRCh37. Representative extragenic crossover sequences are colored according to the most likely originating haplotype.

Analysis of pairwise linkage disequilibrium (LD) between genotypes of all 63 SNPs in 1664 phased haplotypes from Canadian HD patients and controls reveals a ~170kb region of high LD (D' > 0.9) from rs762855 to rs362303 (chr4: 3074795-3242307), indicating a haplotype block of low recombination across the entire transcribed *HTT* sequence (**Appendix A, Figure A-1**). In contrast, stringent pairwise LD by correlation coefficient (r^2) reveals a punctuated pattern of SNP disequilibrium within the *HTT* haplotype block, reflecting a diversity of haplotypes spanning the gene locus despite the absence of recombination (**Figure 2-1b**). Strikingly, SNPs in high pairwise correlation within *HTT* mark specific intragenic haplotypes. For example, rs2798235 and rs363080 represent unique markers of the A2 haplotype and are found in near-perfect pairwise correlation ($r^2 = 0.98$), whereas both SNPs show low pairwise correlation with all other variants in our common 63-SNP panel. The observed pattern of LD across *HTT* indicates that SNPs within *HTT* tag specific gene-spanning haplotypes encompassing the entire transcribed *HTT* sequence, without evidence of common recombination events. Among Canadian subjects, 95.8% (271/283) of HD chromosomes and 95.9%

(795/829) of control chromosomes conform to 20 specific nonrecombinant haplotypes at 51 common intragenic SNPs and at exon 1 CCG repeat length (**Appendix A, Figure A-2**). A and B haplogroups are always marked by 7 CCG in the Canadian cohort and C haplotypes are always 8, 9, or 10 CCG.

2.3.2 A1, A2, and A3 are the Most Common Gene-Spanning HD Haplotypes

We next determined the most frequent gene-spanning haplotypes occurring on HD chromosomes. Among 283 unrelated Canadian HD chromosomes, 48.1% (136/283) are found on the A1 haplotype marked by rs362307, 32.2% (91/283) are found on closely related A2a or A2b, and 12.0% (34/283) are found on A3 (**Figure 2-2a**). In total, 92.2% (261/283) of Canadian HD chromosomes are found on A1, A2, or A3 haplotypes spanning *HTT*. Among control chromosomes, 8.0% are A1, 16.4% are A2a or A2b, and 13.1% are A3. Haplotypes A4 and A5, each present on 6.3% of control chromosomes, are never observed on Canadian HD chromosomes. Notably, A1 and A2a represent the most genetically distant haplotypes within the A haplogroup, despite representing the most frequent HD haplotypes. Haplogroup B is a distinct genetic lineage in 5.3% of *HTT* controls, present on only 3/283 HD chromosomes in the Canadian cohort (1.1%). Haplogroup C is a complex collection of haplotypes constituting nearly half of unrelated control chromosomes (42.6%), but is found on only 3.2% of HD chromosomes. The most common intragenic haplotype among all annotations is C1, present on 29.8% of unrelated control chromosomes in the Canadian cohort.

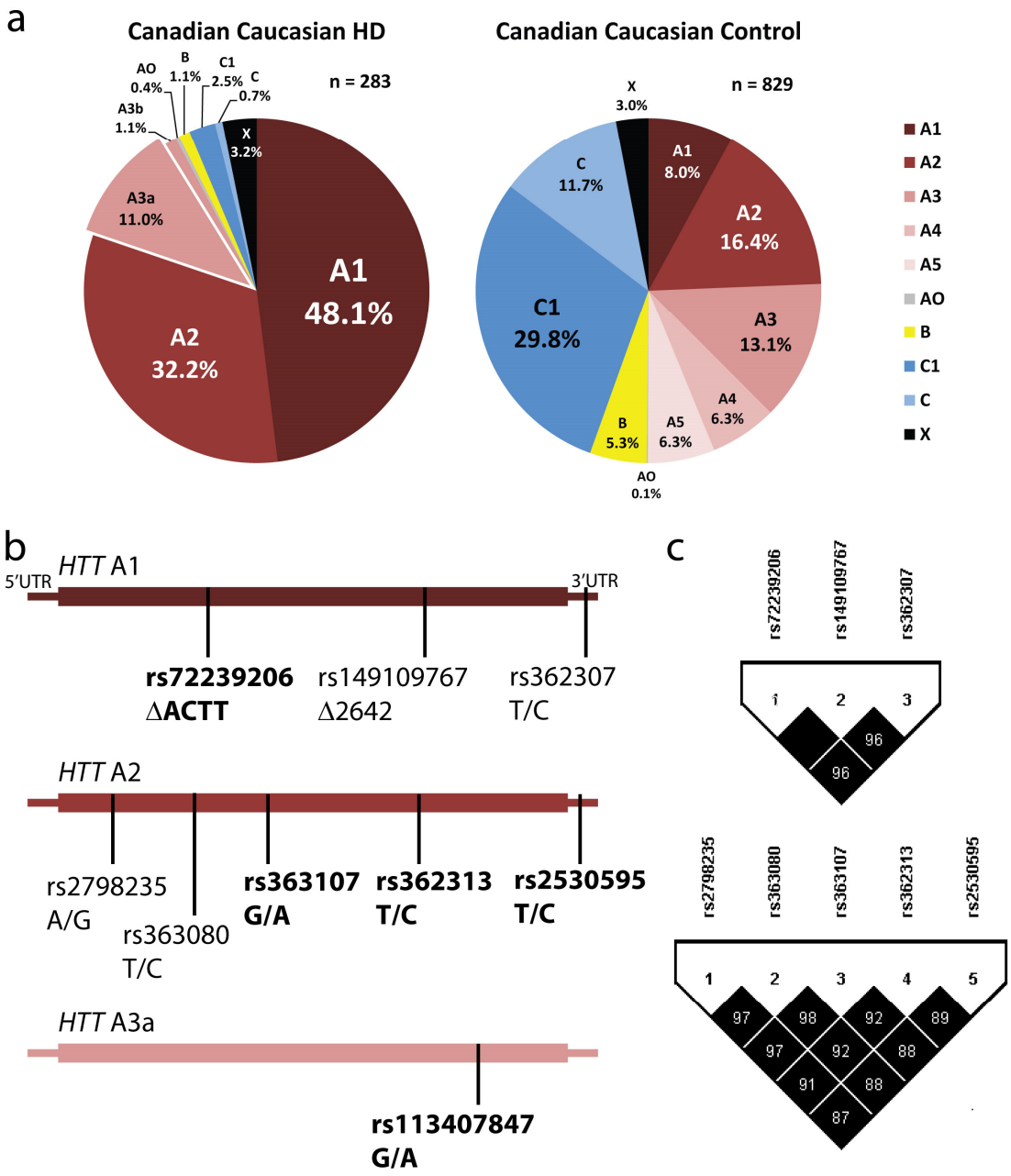


Figure 2-2. A1, A2, and A3a are the most common gene-spanning HD haplotypes.

(a) In the UBC HD BioBank, the expanded CAG repeat (CAG > 35) is found on the A1 haplotype in 48.1% of phased, unrelated Canadian Huntington disease (HD) chromosomes. Intragenic recombinant haplotypes (X) are rare (3.0% of controls) whereas >95% of *HTT* haplotypes show no evidence of

recombination within the transcribed gene sequence. C1 is the most common intragenic haplotype among Canadian control chromosomes, representing 29.8%. (b) The most common HD haplotype, A1, is uniquely defined by three transcribed polymorphisms in high pairwise linkage disequilibrium (LD) across *HTT*. The 4 bp indel rs72239206 represents a novel polymorphism associated with the CAG expansion (bold). The second most common HD haplotype, A2, is defined by five intragenic single-nucleotide polymorphisms (SNPs), three of which are novel (bold). HD-associated A3a, the third most common HD haplotype, is specifically marked by the novel SNP rs113407847. (c) Pairwise LD plot (*r*2) of A1 and A2 haplotype-defining polymorphisms as calculated from 700 phased haplotypes of European Caucasians.

2.3.3 Identification of All Defining Intragenic Alleles on HD-Associated Haplotypes

CAG-expanded *HTT* is enriched for gene-spanning A1 and A2 haplotypes relative to normal control chromosomes. This suggests that alleles found exclusively on these haplotypes are potential targets for allele-specific silencing of mutant *HTT*. To determine all polymorphisms uniquely found on the most frequent HD-associated haplotypes (A1 and A2), we identified all chromosomes in the 1000 Genomes Project with defining SNPs for these haplotypes. In total, 2297 intragenic polymorphisms are annotated across *HTT* (chr4:3076408-3245687, GRCh37) in the 1000 Genomes Phase I data set across all ethnicities (Abecasis et al., 2012).

In our 51-SNP *HTT* panel, A1 is uniquely defined by rs362307, which is a [T/C] SNP present in both exon 67 and the 3'UTR of *HTT*. Among all 1000 Genomes control chromosomes, 3.5% (76/2166) carry rs362307[T]. Among the 76 chromosomes bearing

rs362307[T], 75 (98.7%) carry the glutamic acid deletion known as ∆2642 (rs149109767) and 74 (97.4%) carry a 4bp intronic deletion (rs72239206). Among all 2166 chromosomes, including those 76 bearing rs362307[T], ∆2642 is present on 77 (3.5%) and rs72239206 is present on 83 (3.8%). Therefore 95% (75/77) of chromosomes with ∆2642 and 89% (74/83) of those with rs72239206 also carry rs362307[T]. Both polymorphisms are therefore sensitive proxy markers of rs362307 (**Figure 2-2b**). No other SNPs were strongly associated with rs362307. The next most common SNP also present on at least 90% of the A1 chromosomes was found nonspecifically on 633/2166 chromosomes in the 1000 Genomes set, including the subset of A1 chromosomes. Among SNPs less frequent than rs362307 but specific to A1 chromosomes, rs187059132 is present on 32/76. Variants rs362307, rs149109767, and rs72239206 are therefore highly specific for the A1 haplotype, having high pairwise correlation (r^2>0.9) with each other but with no other SNPs in *HTT*. 1000 Genomes Phase 1 chromosomes bearing all three A1-defining polymorphisms are found exclusively on individuals of European or admixed European ethnicity (**Appendix A, Table A-1**). Each individual variant is rare or absent in other ethnic groups, in agreement with the reported absence of ∆2642 and rs362307 in individuals of East Asian and black South African ancestry (Baine et al., 2013; Warby et al., 2011; Almqvist et al., 1995).

The A2 haplotype, comprised of closely related subtypes A2a and A2b, is uniquely tagged by rs2798235 and rs363080 in our 51-SNP *HTT* panel. 100 chromosomes in 1000 Genomes Phase I carry rs363080, of which 98 carry rs2798235. The latter SNP is

exclusively found on chromosomes tagged by rs363080. High pairwise correlation between these two markers is similarly observed in direct genotyping of our Canadian *HTT* chromosomes (r^2 = 0.98, **Figure 2-1a**). In 1000 Genomes, three additional intragenic SNPs – rs363107, rs362313, and rs2530595 – are found on 100, 99, and 99 of chromosomes bearing rs363080, respectively, and are likewise present only on these chromosomes. All five polymorphisms are present on 98% of chromosomes bearing any of the five variants, and therefore represent specific markers of the A2 haplotype (**Figure 2-2b**).

HD also commonly occurs on A3. In our 51-SNP panel, the A3 haplotype is defined by intragenic markers of the A haplogroup in the absence of SNPs specific for the other A haplotypes (**Figure 2-1a**). 119 A3 haplotypes were identified out of 738 control chromosomes of European ancestry from 1000 Genomes Phase I (16.1%). No identifying SNPs were found that uniquely encompass all 119 A3 chromosomes. However, a specific subtype SNP was observed on 45.4% (54/119) of A3 chromosomes – rs113407847 – designating A3a (**Figure 2-2b**). In the 738 European individuals, rs113407847 is found only in the subset of A3 haplotypes. Despite common association with HD, no SNPs specific to both A1 and A3 were found, except when shared with other, non-HD associated A haplotypes.

2.3.4 Validation of Polymorphisms Specific for the *HTT* A1 Haplotype

To validate our *in silico* association of all three A1-defining polymorphisms from whole genome sequencing data in 1000 Genomes, ∆2642 and rs72239206 were directly genotyped and phased by familial segregation to the *HTT* CAG repeat in chromosomes previously genotyped for rs362307. These comprised all Canadian HD chromosomes genotyped with the original 63-SNP panel as well as other previously haplotyped samples from various ethnic groups. In total, 454 phased, unrelated HD chromosomes and 652 unrelated control chromosomes from HD families were successfully genotyped and phased to the CAG repeat at rs362307, ∆2642, and rs72239206. Pairwise LD analysis of direct genotyping data indicates that all three polymorphisms are present in HD and control chromosomes in near-perfect LD ($r^2 > 0.99$) and that all three minor alleles are highly enriched on HD chromosomes versus controls (**Appendix A, Table A-2**).

2.3.5 Validation of HD Haplotype-Specific Polymorphisms in Distinct HD Patient Populations

Marker studies of the ∆2642 codon deletion suggest that the frequency of the A1 haplotype varies considerably between Caucasian HD patient populations (**Appendix A, Figure A-3**). A key question following our definition of specific gene-spanning HD haplotypes was therefore to determine the distribution of these haplotypes among different patient populations of European ancestry. A revised SNP panel was designed

to include the prior 63-SNP panel as well as novel defining SNPs of the A1, A2, and A3a haplotypes derived from the 1000 Genomes Project. Using this revised panel, we genotyped 120 Swedish, 76 French, and 209 Italian HD family members, derived from respective countries of origin. Haplotypes were reconstructed and phased to CAG repeat size, as for our Canadian Caucasian cohort. All common 63-SNP haplotypes found in the Canadian Caucasian cohort were replicated by genotyping of the European HD cohorts using our revised panel. All three A1 variants and all five A2 variants conformed to high expected pairwise correlation in direct genotyping of the European cohorts with the revised panel (**Figure 2-2c**). Among all European patients, the CAG expansion on A3 was found exclusively in phase with the unique A3a-identifying SNP rs113407847, but not on A3b lacking this SNP, suggesting that A3a is a disease-associated haplotype. Direct genotyping of rs113407847 in Canadian HD A3 chromosomes similarly revealed that the CAG expansion occurs almost exclusively on A3a when present on A3 (31 of 34 A3 Canadian HD chromosomes) (**Figure 2-2a**). Common HD-associated haplotypes A1, A2, and A3a therefore share uniform sets of defining markers in ethnically distinct European HD patient cohorts, implying deep ancestral relationship of these disease-associated haplotypes across different European populations.

2.3.6 Frequency of HD Haplotypes Differs by European Ancestry

While intragenic haplotypes of *HTT* are consistently A1, A2, and A3a across Caucasian HD chromosomes, our direct genotyping reveals striking differences in the frequency of

specific HD-associated haplotypes among CAG-expanded chromosomes in different European populations (**Figure 2-3**).

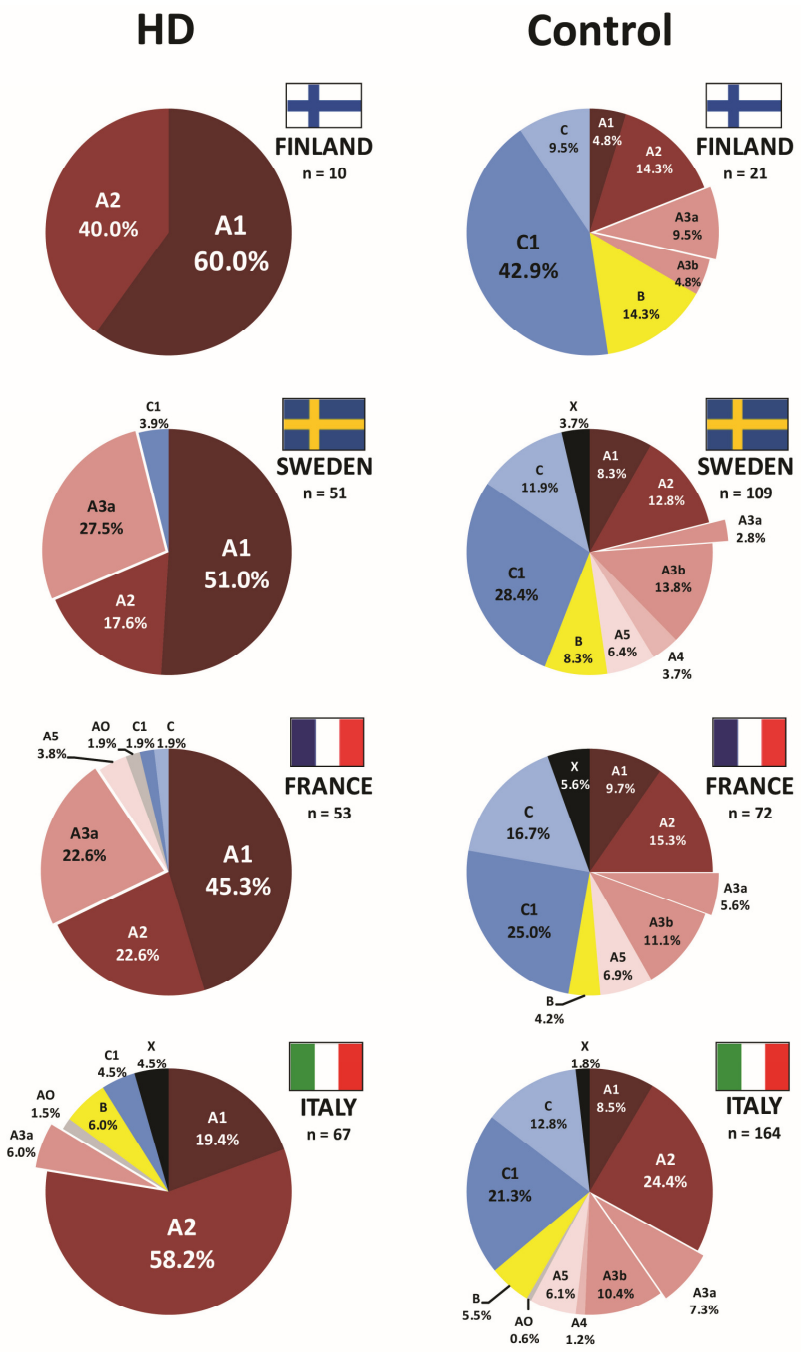


Figure 2-3. European *HTT* haplotype distributions. In four European Huntington disease (HD) cohorts, distinct distributions of HD haplotypes are observed. As in Canada, A1 is the most frequent HD haplotype in Finland (60%) and Sweden (51%), and France (45%). In contrast, A2 is the most frequent HD haplotype in Italy (58%).

Among unrelated Swedish HD chromosomes, 51% (26/51) are found on A1, similar to our previously genotyped Canadian HD cohort (p=0.7616, Fisher's Exact). The frequency of A2 among Swedish HD is comparatively lower than in Canada (18% versus 32%, p=0.0455) and A3a is more frequent (28% versus 11% in Canadians, p=0.0033). French HD chromosomes are also most frequently A1 (45% versus 48%, p=0.7654) with A2 present at similar frequencies and A3a more common than in Canadian HD (A2, p=0.1957; A3a, p=0.0256). In contrast, Italian HD chromosomes are predominantly found on A2 (58%, p<0.0001 versus Canadian), with a much smaller proportion of HD on A1 versus the Canadian cohort (19%, p<0.0001) and a similar proportion of A3a (7%, p=0.2647). In a small set of Finnish HD families, haplotyped with our original 63-SNP panel, all unrelated disease chromosomes are A1 (6/10, 60%) or A2 (4/10, 40%). Despite differences in specific haplotype frequency between our Canadian and European cohorts, >90% of HD chromosomes are found on A1, A2, and A3a haplotypes in all four populations of Northern European ancestry and in 84% of Italian HD chromosomes.

HTT haplotypes on control chromosomes from HD families also differ between European populations, though less dramatically than CAG-expanded chromosomes. The A haplogroup trends toward higher frequency in Italian controls versus Canadian (p=0.0597), but is found at similar frequency among Swedish (p=0.6838) and French control chromosomes (p=0.8073). A1 occurs at statistically similar frequencies in all four control cohorts, whereas A2 occurs at higher frequency among Italian controls than in

Canadian (24% versus 16%, p=0.0185) or Swedish controls (13%, p=0.0203), mirroring its elevated frequency among Italian HD chromosomes.

2.3.7 A1, A2, and A3a Haplotypes are Target Panels to Optimize Population Coverage for Allele-Specific *HTT* Silencing

High pairwise correlation of specific haplotype-defining polymorphisms allows for targeting of the A1 and A2 haplotypes as a selective *HTT* silencing strategy. As all three A1 markers are present in near-perfect LD, targeting any single A1 polymorphism will allow allele-specific *HTT* silencing in a nearly equal number of HD patients heterozygous for this haplotype. Heterozygosity of A1 in HD patients, when phased to the CAG expansion, is highest in Sweden (47%), Canada (44%), and France (43%), but much lower in Italy (15%), suggesting greater utility in patients of Northern European ancestry (**Figure 2-4**).

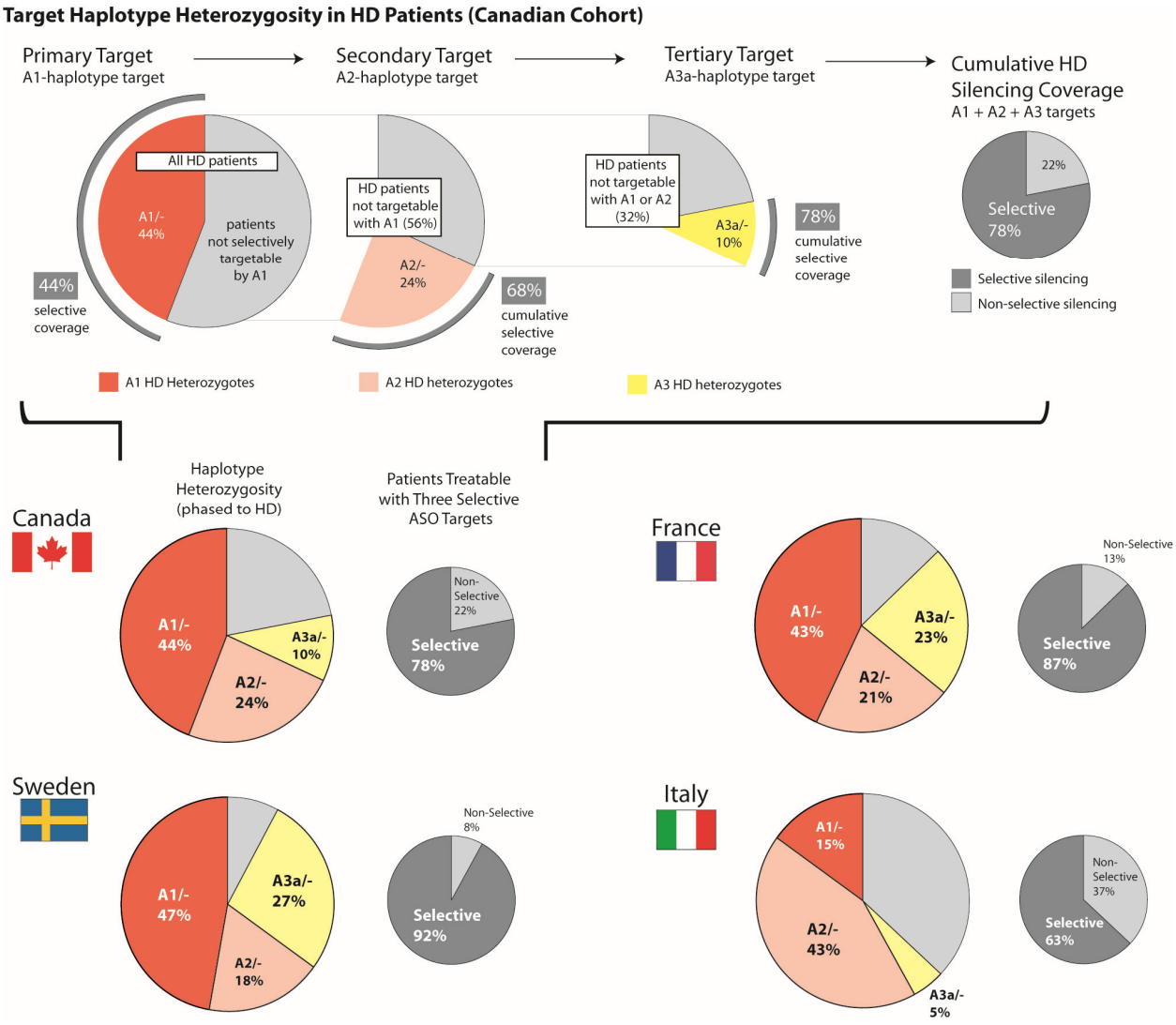


Figure 2-4. Cumulative treatment coverage of Huntington disease (HD) patients by A1, A2, and A3a allele-specific *HTT* silencing targets in four patient populations of European ancestry.

Targeting one A1 allele (among three) and one A2 allele (among five) would permit selective *HTT* silencing treatment in 68% of HD patients in the Canadian cohort. In combination, A1 + A2 + A3a targets would permit selective silencing of an average of 80% of HD patients across all four populations.

An estimated 98% of patients with HD on A2 (phased to rs363080) will have all five A2 targets present. The percent of patients heterozygous for A2, phased to the CAG expansion, range from to 18% in Sweden to 43% in Italy, suggesting a greater utility in Southern European populations. Tertiary targeting of rs113407847 would allow treatment of patients bearing the CAG expansion on A3a, ranging from a maximum of 27% of patients in Sweden to only 5% in Italy.

On average across all four patient populations, the A1 haplotype may allow selective silencing treatment of 37% of HD patients (**Table 2-1**), with targeting of the A2 haplotype permitting treatment of an additional 27%. Since no patient will share both A1 and A2 targets on the same disease chromosome, percentages of patients treatable with each haplotype target are additive. Across all four cohorts, a maximum of 64% of patients will be treatable with two allele targets representing A1 and A2, rising to approximately 70% if Finnish HD chromosomes are included. In total, targeting three specific polymorphisms representing A1, A2, and A3a may allow selective silencing treatment of ~80% of HD patients overall from the Canadian, Swedish, French, and Italian patient populations. Defining SNPs of these HD-associated haplotypes therefore represent panels of targets that could achieve ~80% patient treatment by allele-specific *HTT* silencing strategies.

Table 2-1. Percent of Huntington disease patients treatable by different haplotype targets in four populations of European ancestry

		One Target					Two Targets			Three Targets			Homozygous for all SNPs
	n	A1	A2	A3a	A	AB	A1+A2	A1+A	A2+A	A1+A2+A3a	A1+A2+A	A1+A2+AB	(untreatable)
Canadians	283	44%	24%	10%	**48%**	44%	**68%**	66%	55%	**78%**	73%	73%	10%
Swedish	51	47%	18%	27%	**51%**	47%	65%	**69%**	59%	**92%**	76%	76%	4%
French	53	43%	21%	23%	**55%**	51%	64%	**75%**	58%	**87%**	79%	79%	2%
Italian	67	15%	**43%**	5%	27%	24%	**58%**	37%	48%	**63%**	58%	60%	15%
Average		37%	27%	16%	**45%**	42%	**64%**	62%	55%	**80%**	72%	72%	8%

An alternative strategy is targeting SNPs specific for the A haplogroup that includes haplotypes A1, A2 and A3a. In populations of European ancestry, greater than 90% of HD patients carry an expanded CAG repeat on the A haplogroup and would be treatable if heterozygous with a B or C haplotype on the normal chromosome. In the Canadian, Swedish, and French cohorts, we find that a higher percentage of patients are heterozygous in phase with expanded CAG for the A haplogroup than for either A1 or A2 haplotypes (**Table 2-1**). Only in the Italian cohort does coverage with A2 exceed that of the A haplogroup given only one allele target. However, when two targets are considered cumulatively, only A1 and A2 alleles in combination allow treatment of the majority of patients in all four populations. Further, if one excludes patients eligible for A1 and A2 treatment, A3a is heterozygous in the largest proportion of European patients remaining. While defining polymorphisms of the A haplogroup provide attractive heterozygosity rates due to deep ancestry of constituent A1, A2, and A3a haplotypes, targeting these specific, mutually exclusive disease haplotypes individually would enable treatment of more patients overall.

Both the 63-SNP Canadian data and the 1000 Genomes reference haplotypes indicate that the A haplogroup is defined by a single SNP, rs2298969. In contrast, 24 SNPs distinguish the A and B haplogroups from the C haplogroup (**Table 2-2**), offering many more potential targets for development. Averaged across all four cohorts, rs2298969 would offer treatment for 45% of patients, whereas each of 24 SNPs present on both A and B haplogroups are expected to be heterozygous in 42%. As with the A haplogroup,

the overall percent of patients treatable with A1 and A2 in combination exceeds any corresponding combination of AB haplogroup alleles with either A1 or A2. The percent of patients treatable with A1, A2, and any AB target is similar to that achievable with A1, A2, and the A haplogroup (72% in both cases), but still fewer than the maximum percent treatable with A1, A2, and A3a in combination (80%).

Table 2-2. Prioritized intronic and exonic target alleles specific for the most common Huntington disease haplotypes and haplogroups

A1	A2	A3a	A haplogroup	A+B haplogroups
rs72239206	rs2798235	rs113407847	rs2298969	rs28867436
rs149109767	rs363080			rs2285086
rs362307	rs363107			rs6446722
	rs362313			rs7691627
	rs2530595			rs17780797
				rs2071655
				rs4690072
				rs6446723
				rs11409456
				rs4690073
				rs4690011
				rs28403215
				rs6844859
				rs7685686
				rs362331
				rs2071662
				rs362319
				rs82333
				rs110501
				rs2276882
				rs2237006
				rs362314
				rs2237007
				rs2071704

Alleles in **bold** are exonic, all others are intronic. Variants are ordered by distance from the expanded CAG repeat.

2.3.8 ASOs Targeting ∆ACTT (rs72239206) Potently and Selectively Silence A1 HTT mRNA and Protein in Human Cells

Among all genotyped HD patients in this study, A1 is the most frequently heterozygous haplotype in *cis* with the expanded CAG repeat. Targeting the A1 haplotype is also crucial to achieving maximum cumulative patient coverage given two or three allele targets. The defining A1 markers, rs362307, ∆2642, and rs72239206, therefore represent key sites for achieving allele-specific treatment in the greatest number of HD patients of European ancestry. Both rs362307 and ∆2642 are found in mature mRNA, have known association with the CAG repeat expansion, and have been investigated as targets of siRNA-mediated selective *HTT* silencing (Zhang et al., 2009; Pfister et al., 2009). Small interfering RNA (siRNA) designed against the ∆2642 allele induced 51% reduction in mutant *HTT* mRNA in transfected juvenile HD fibroblasts, with ~20% off-target reduction of wild-type *HTT* mRNA (Zhang et al., 2009). Potency and selectivity of siRNA complementary to the rs362307[T] variant have been shown using reporter constructs, but not in the context of full-length *HTT* (Pfister et al., 2009).

In addition to offering a novel A1 target not previously associated with HD, we hypothesized that targeting of the 4bp rs72239206 indel sequence may offer greater selectivity than discrimination by a single nucleotide polymorphism, and sought to evaluate the potential of rs72239206 as a selective mutant *HTT* silencing target. Unlike rs362307 and ∆2642, rs72239206 is located in an intron (intron 22 of *HTT*) and is

therefore only targetable by agents complementary to unspliced pre-mRNA. ASOs can induce RNAse H-mediated degradation of complementary pre-mRNA as well as mRNA (Bennett and Swayze, 2010) and we therefore designed ASO sequences incorporating a gapmer design with locked nucleic acid (LNA) wings and phosphorothioate linkages complementary to the rs72239206 deletion sequence. (**Figure 2-5a**).

ASOs are passively taken up by neurons in primary culture (Carroll et al., 2011). In the absence of transgenic *HTT* neurons bearing rs72239206, we sought to test the silencing potential of our LNA gapmers by passive uptake in human HD lymphoblasts bearing the A1 haplotype (GM03620, CAG 59/18). Remarkably, A1 HTT is selectively silenced in human lymphoblasts grown with rs72239206-targeted LNA gapmers in media, suggesting that lymphoblasts also passively take up ASO in culture. (**Appendix A, Figure A-4**). On the basis of these preliminary experiments, we sought to examine dose-dependent knockdown of A1 *HTT* mRNA and protein in HD patient lymphoblasts of typical CAG length using active transfection to maximize effective dose.

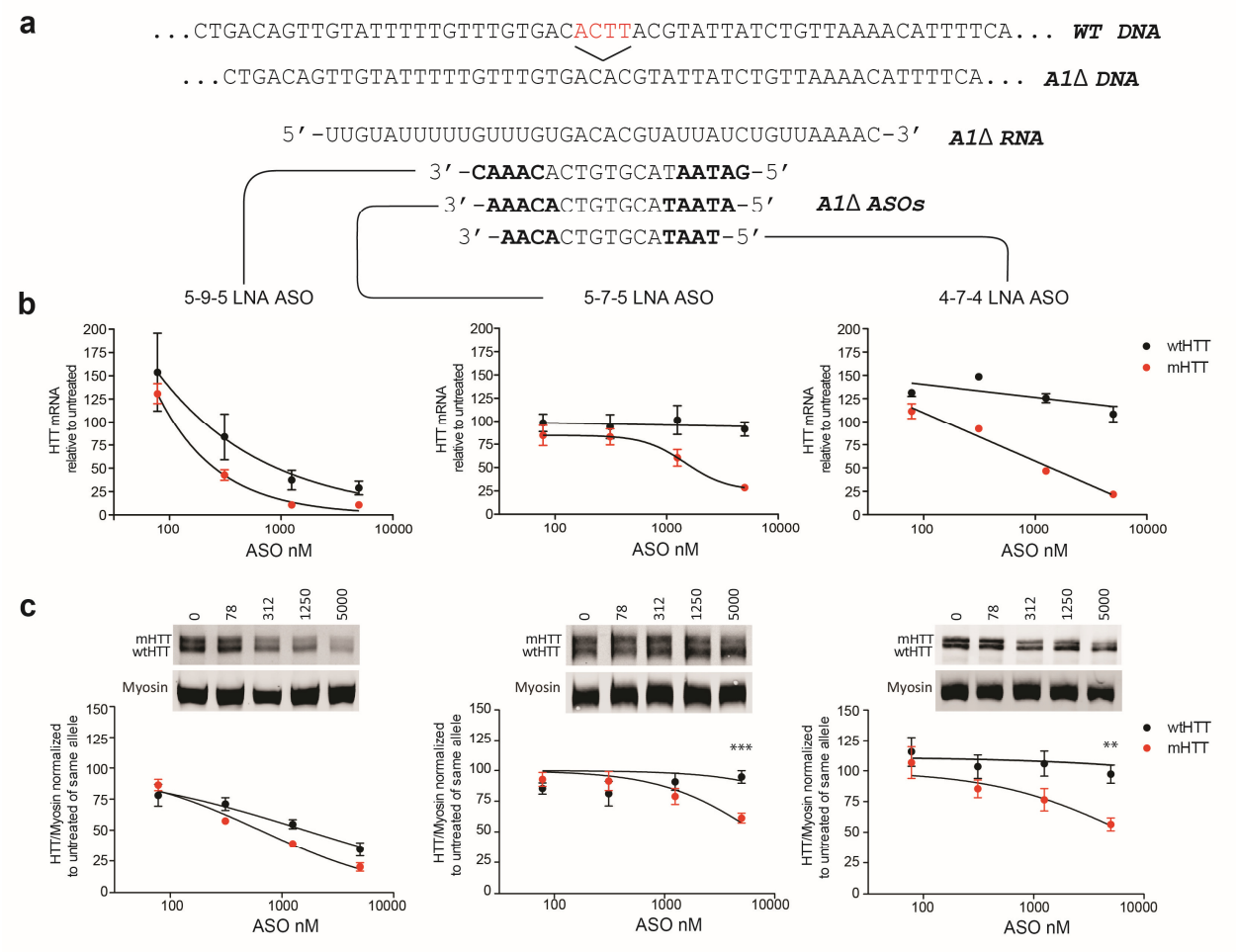


Figure 2-5. ASOs targeting ΔACTT (rs72239206) potently and selectively silence A1 *HTT* mRNA and protein in human cells

(a) Design of antisense oligonucleotide (ASO) gapmers selectively targeting mutant *HTT* A1 mRNA at the ΔACTT sequence. (b) Transfection of patient-derived lymphoblasts (44/18 CAG) with ΔACTT-complementary ASOs selectively reduces mutant *HTT* mRNA. Patient lymphoblasts transfected with 5-9-5, 5-7-5, and 4-7-4 LNA gapmers show dose-dependent reduction of mutant *HTT* mRNA expressed as a percent of *HTT* mRNA levels in untreated controls, falling to 21.5% mutant *HTT* mRNA at the highest 4-7-4 dose. Wild-type *HTT* mRNA levels do not fall below untreated levels at any tested dose of 5-7-5 or 4-7-4 LNA gapmer. (c) Dose-dependent reduction of mutant HTT protein relative to untreated controls, sparing wtHTT at all tested 5-7-5 and 4-7-4 LNA gapmer doses. ** and *** represent $P = 0.01$ and $P = 0.001$ by analysis of variance with Bonferroni *post hoc*. LNA, locked nucleic acid.

Transfection of adult human A1/C1 lymphoblasts (GM02176, CAG 44/18) with a 5-9-5 LNA gapmer complementary to rs72239206 resulted in potent *HTT* mRNA silencing (**Figure 2-5b**), but only minimal discrimination between A1 and C1 transcripts (11% A1 and 29% C1 *HTT* mRNA remaining at the highest dose versus untreated cells). As expected, reduction of the DNA gap by two nucleotides to a 5-7-5 configuration improved selectivity but reduced potency (Ostergaard et al., 2013), with the *HTT* A1 transcript reduced to 29% of untreated mRNA levels at the highest dose versus 92% *HTT* C1 control. Shortening this molecule to a 4-7-4 LNA gapmer design further improved selectivity for the *HTT* A1 transcript, reducing A1 *HTT* mRNA to 22% of untreated levels at the highest dose while sparing *HTT* C1 mRNA at untreated levels (**Figure 2-5b**). Western blot analysis using allelic separation of CAG 44/18 bands revealed reduction at the protein level for all three LNA gapmer designs, inducing dose-dependent reduction of mutant HTT with 5-7-5 and 4-7-4 gapmers while sparing normal HTT (**Figure 2-5c**). Targeting the A1-specific rs72239206 deletion sequence with complementary ASOs can therefore potently and selectively silence mutant *HTT* mRNA and protein in patient-derived cells that are genetically representative of HD patients bearing the A1 haplotype.

2.3.9 Targeting the rs72239206 Deletion Site is Efficacious and Tolerated In Vivo

In the absence of transgenic mice bearing the rs72239206 deletion in *cis* with expanded CAG, *in vivo* silencing of A1 *HTT* mRNA and protein could not be directly evaluated. However, the wild-type analog of the 5-9-5 LNA gapmer, designed against reference

sequence that includes the four bases deleted in rs72239206, also elicited potent reduction of human HTT in brains of YAC128 mice bearing transgenic full-length mutant *HTT* (**Appendix A, Figure A-5**). The rs72239206 deletion site is therefore accessible to ASO-mediated *HTT* mRNA silencing *in vivo*.

2.4 Discussion

The translation of allele-specific *HTT* silencing to therapeutic application requires clarity as to which transcribed SNPs are the most useful targets in the HD patient population. The frequency of specific polymorphic targets is known to vary between clinical cohorts, while secondary and tertiary targets that maximize the total number of patients treatable have been incompletely described. Our study provides the first comprehensive heterozygosity estimates across the *HTT* transcript in multiple patient populations, identifying specific allele targets of highest priority for development of selective antisense therapies. We have fully described the most common gene-spanning haplotypes relevant for selective suppression of mutant *HTT* in patients of European ancestry – A1, A2, and A3a – and identify all common polymorphisms specific for these haplotypes. In four different patient populations, these gene-spanning haplotypes represent panels of allele-specific targets that would achieve treatment of the greatest proportion of HD patients. We show that as few as three gene silencing reagents targeting the A1, A2, and A3a haplotypes may offer allele-specific *HTT* silencing therapy for 80% of all patients of European descent. A1 may be silenced using one of three defining polymorphisms, and A2 using one of five defining polymorphisms. If only one

allele target can be chosen for development, silencing the A haplogroup by rs2298969 may offer treatment in the greatest proportion of patients. But when two targets are considered additively, A1 and A2 targets in combination allow for treatment of the majority of patients in all four major populations evaluated in this study. When three targets are considered, no combination of intragenic polymorphisms allows for selective silencing in a greater proportion of cases than defining polymorphisms of the A1, A2, and A3a haplotypes. A1 and A2 haplotypes therefore represent sets of priority targets for preclinical evaluation of allele-specific *HTT* silencing reagents, with rs113407847 a priority tertiary candidate. However, due to RNA structure or kinetics of pre-mRNA splicing, not all SNP targets are available for ASO-mediated silencing (Skotte et al., 2014). Should A3a-defining rs113407847 fail as a target, our results suggest that rs2298969 or any of 24 SNPs present on the A and B haplogroups would then represent tertiary targets for treating the next greatest proportion of patients.

Only four of our 34 priority alleles are located within exons: two specific to the A1 haplotype (rs149109767 and rs362307), one specific to the A2 haplotype (rs2530595), and one on the A and B haplogroups (rs362331) (**Table 2-2**). For allele-specific antisense strategies restricted to mature mRNA, such as siRNA, our data suggest that targeting an exonic A1 allele (rs149109767 or rs362307) in combination with the A2-specific allele rs2530595 will enable treatment of the most patients. It is notable, however, that rs149109767 is a codon deletion within a tandem GAG glutamic acid repeat, limiting effective design of short oligonucleotide sequences that can span and distinguish the mutant and wild-type alleles at this site. Therefore rs362307 remains a

high-priority SNP for the development of A1-specific antisense reagents acting against mature mRNA, with rs2530595 representing a secondary exonic target specific for the A2 haplotype.

A polyadenylated exon 1 transcript of *HTT* containing expanded CAG repeats is translated into an N-terminal polyglutamine fragment in HD patients (Sathasivam et al., 2013). Pathogenicity of mutant HTT exon 1 in N-terminal fragment models of HD raises the concern as to whether suppression of the exon 1 transcript, in addition to suppression of full-length mutant HTT, will be necessary to achieve therapeutic outcomes. No HD haplotype-defining SNP among our 34 target alleles occurs within exon 1 of *HTT*, although we note association of the CCG 7 repeat allele in exon 1 with the disease-associated A haplogroup (**Appendix A, Figure A2**). The mutant exon 1 transcript is found in multiple transgenic models of HD, including full-length YAC128 and BACHD, yet therapeutic rescue of HD phenotypes has been observed in these models by ASO-mediated suppression of mutant *HTT* outside exon 1 (Kordasiewicz et al., 2012). This suggests that suppression of full-length mutant *HTT* is sufficient for phenotypic reversal in full-length murine models, despite continued production of exon 1 transcript containing expanded CAG repeats. Indeed, we detect no change in exon 1 transcript levels in YAC128 brains when full-length *HTT* mRNA expression is abolished by ASO treatment (data not shown). Exogenous suppression of full-length mutant *HTT* by ASO is also sufficient for rescue of transcriptional abnormalities in YAC128 mice (Stanek et al., 2013). While the pathogenic impact of the exon 1 transcript remains unclear in HD patients, the efficacious reversal of HD phenotypes by suppression of full-

length *HTT* in transgenic models suggests that reduction of the exon 1 transcript may not be required to elicit therapeutic benefits of ASO treatment.

Expansion of the CAG repeat has been shown to occur on multiple haplotypes in different Caucasian populations (Warby et al., 2009; Lee et al., 2012a). Here we demonstrate that three intragenic *HTT* haplotypes, identical across four different populations of European ancestry, account for approximately 90% of HD chromosomes across these groups. This suggests that haplotypes on which repeated CAG expansion events occur are ancestral to all populations of European descent, and may perhaps be shared by other related populations. The ∆2642 codon deletion (rs149109767), identified here as an exclusive marker of the A1 haplotype, has been observed in HD patients and controls from India (Saleem et al., 2003), whereas A1 is entirely absent among both HD and control chromosomes of black South Africans and East Asians where prevalence of HD is dramatically lower (Warby et al., 2011; Baine et al., 2013). This suggests that association of the A1 haplotype with HD may occur in all populations of Indo-European ancestry, spanning South Asia, Europe, and American populations of European descent. The frequency of HD on A1, A2, and A3a haplotypes requires detailed haplotype analysis in patient populations from the Middle East, Central Asia, South Asia, and Africa to evaluate the global therapeutic impact of these targets. The high prevalence of expanded CAG on A1, A2, and A3a among all patient populations of European descent, as well as the presence of these haplotypes in other ancestrally related populations, suggests that these haplotypes may allow allele-specific silencing in the maximum proportion of HD patients worldwide.

In all populations, there remains a portion of HD patients not eligible for SNP-selective silencing due to *HTT* haplotype homozygosity. 15% of Italian HD patients in our study are homozygous at all SNPs genotyped, whereas only 2% and 4% are homozygous in the French and Swedish cohorts, respectively (**Table 2-1**). On average, 8% of patients are ineligible for allele-specific therapy due to the absence of heterozygosity at any common SNP in *HTT*. Only a small proportion of HD patients of European ancestry (<10%) are therefore likely to benefit from more than three allele-specific targets, given that ~80% of patients of European descent are heterozygous for A1, A2, and A3a and ~8% are homozygous at all SNPs. It is nonetheless theoretically possible to achieve treatment of all patients heterozygous for at least one SNP, constituting 92% on average between the four patient populations examined.

Among all patients lacking heterozygous *HTT* SNPs in this study, all but one are homozygous for specific target alleles of A1, A2, or A3a. ASOs specifically designed for nonselective *HTT* silencing have shown short-term tolerability and efficacy in transgenic models (Kordasiewicz et al., 2012; Boudreau et al., 2009), and nonselective treatment of patients homozygous for disease-associated haplotypes may be possible using antisense reagents developed for selective silencing in heterozygous patients. Alternatively, targeting both SNP alleles at one polymorphic site, as suggested for rs7685686 (Skotte et al., 2014), would enable selective treatment of approximately half (45.9%) of European HD patients in our study while also allowing the possibility of nonselective treatment of all remaining patients. In vivo efficacy testing of allele-specific

silencing reagents, as currently underway in humanized 97/18 mice (Southwell et al., 2013; Southwell et al., 2014), will also crucially inform the future prospects of allele-specific therapy for HD. Should these model studies yield robust therapeutic outcomes, the potential to safely treat over three quarters of the global HD patient population with only three allele-specific drugs, or half of patients selectively and half nonselectively with only two drugs, may present attractive strategies for development.

Antisense reagents designed only for nonselective suppression of *HTT*, complementary to the *HTT* transcript but not selective for disease-associated alleles, are also being developed as a therapeutic approach to HD. However, it remains unclear whether adult loss of HTT protein can be adequately compensated in humans over the long-term duration of *HTT* suppressive therapy. The protective effects of wild-type HTT in various models of HD suggest that specific reduction of mutant HTT may offer a safer and more efficacious alternative to indiscriminate suppression of both wild-type and mutant alleles in patients. Although transient suppression of *Htt* appears to be tolerated in mice, adverse effects of human *HTT* suppression may emerge during clinical trials or long-term treatment. Options for selective HTT suppression should be considered given these potential risks. Any adverse outcomes of long-term total HTT suppression will also need weighed against the cost of developing allele-specific drugs that may be safer but useful only in a subset of genetically eligible patients. The expense of developing secondary and tertiary targets for selective treatment must further be considered versus the number of additional patients that each new target may treat.

The disease-associated alleles identified in this study provide a prioritized list for validation and preclinical screening of allele-specific *HTT* silencing compounds. One ASO targeting a single allele may treat a maximum of 45% of patients across all patients of European ancestry, whereas three haplotype-specific ASOs targeting A1, A2, and A3a may enable selective treatment of a maximum of 80%. If only one allele-specific drug can be developed, our data supports targeting the A haplogroup to selectively treat the most patients. However, if total *HTT* suppression results in poor patient outcomes, necessitating selective suppression of mutant *HTT* in order to achieve clinical benefit, our data supports targeting the A1 haplotype as the necessary first step to achieving maximum allele-specific treatment of the patient population until multiple antisense drugs can be developed for each disease-associated haplotype.

In summary, we show that *HTT* is defined by a gene-spanning haplotype block in populations of European descent, and that specific sets of SNPs define gene-spanning haplotypes in both HD patients and controls. To our knowledge, this is the first annotation of dense haplotypes encompassing the *HTT* gene using whole-genome sequencing data. We identify and validate all polymorphisms specific for the three most common HD haplotypes, comprising >90% of HD chromosomes in four distinct populations of European ancestry. The defining polymorphisms of these haplotypes constitute optimal targets for development of allele-specific silencing compounds, with all other transcribed SNPs providing treatment for fewer patients of European descent. Fine-scale *HTT* haplotype analysis in other patient populations will be necessary to guide a global allele-specific targeting strategy, particularly where the number of

patients may be high, to further clarify the clinical impact worldwide. Targetable *HTT* haplotypes revealed by this study represent a crucial step toward that objective, and toward safe gene silencing treatment of the greatest number of HD patients.

Chapter 3: The Targetable A1 Huntington Disease Haplotype has Distinct Amerindian and European Origins in Latin America

3.1 Synopsis

HD occurs worldwide, but shows marked geographic differences in prevalence. Studies using defined clinical and genetic diagnostic standards report 10.6-13.7 cases per 100,000 inhabitants in Western populations, with higher rates among subpopulations of European ancestry (Fisher and Hayden, 2014). In contrast, Black African and East Asian populations are reported to have the lowest rates worldwide at 0.25-0.7 cases per 100,000 inhabitants (Kay et al., 2014a; Baine et al., 2016). In populations of European ancestry where prevalence is highest, the expanded CAG repeat occurs preferentially on haplotypes A1, A2, and A3a, whereas in African and East Asian patients the HD mutation occurs on a heterogeneous mix of local haplotypes (Warby et al., 2009; Kay et al., 2015; Warby et al., 2011; Baine et al., 2013). European HD haplotypes A1 and A2 do not occur in African and East Asian populations, suggesting that these haplotypes may account in part for higher HD prevalence rates in populations of European ancestry (Kay et al., 2014a; Kay et al., 2015).

HD has been reported in many countries of Latin America, but detailed genetic investigations are lacking (Castilhos et al., 2015). Clinical descriptions were first reported in Cuba and Brazil, and subsequently in Argentina, Mexico, Peru, Chile, and Colombia (Torres Ramirez et al., 2008). No population-based prevalence estimates of HD are

available in Latin America, but a major focus of HD in Latin America and the world is in the state of Zulia, surrounding Lake Maracaibo in Venezuela, where hundreds of affected individuals are known to belong to a single kindred (Wexler et al., 2004). Across Latin America, the HD mutation is believed to originate from European founders with subsequent admixture in indigenous populations (Castilhos et al., 2015; Wexler et al., 2004).

Limited haplotyping studies of HD in Latin America suggest that the HD mutation occurs on specific haplotypes, although not necessarily from European founders (Castilhos et al., 2015). *HTT* haplotypes in Venezuelan patients, including the CCG repeat in exon 1 (GRCh37 chr4:g.3076673_3076675[7]) and ∆2642 codon deletion in exon 58 (GRCh37 chr4:g.3230411_3230413delGAG) (Ambrose et al., 1994), suggest a common origin of the HD mutation among affected families of Venezuela but without definitive evidence of European ancestry (Paradisi et al., 2008). (*HTT* exon numbering follows Ensembl transcript HTT-001 / ENST00000355072.9, NCBI reference mRNA NM_002111.8, and exon numbering originally described in Ambrose et al. 1994.)

The Cañete Valley of southern Lima has been reported as a focus of HD in Peru (Cuba, 1986), and the HD mutation in Peru has been hypothesized to originate in Cañete, possibly from European sources (Cornejo-Olivas et al., 2015). Today HD is known to be widespread in Lima and across northern and southern regions of Peru, with additional reports of HD within native communities of the Peruvian Amazon (Cornejo-Olivas M et al., 2012). Peru is mainly composed of a mestizo population with predominant Amerindian

genetic ancestry and heterogeneous levels of non-indigenous ancestry, principally from Europe (Sandoval et al., 2013). The term “mestizo” is used to refer to these Peruvian individuals of mixed ethnic ancestry (Wang et al., 2008). Genetic data from mestizo Peruvians are consistent with historically documented post-Columbian immigration and admixture from European sources, with minor contributions of African and East Asian immigrants (García, 1991; Herrera, 1986). Relative European and indigenous ancestry of HD is unknown in Peru and elsewhere in Latin America. In the absence of detailed haplotype data for the HD mutation in Latin America, the potential to therapeutically target *HTT* SNPs for allele-specific gene silencing in Latin American patients also remains unclear (Kay et al., 2014b).

To illuminate the origins of HD in Peru and other Latin American populations, here we comprehensively haplotype the *HTT* gene region in a mestizo Peruvian HD patient cohort, in a collection of HD families originating from across Latin America, and in an unaffected comparison population of defined Amerindian ancestry. We uncover a surprising indigenous variant of the A1 *HTT* haplotype in unaffected Amerindian controls, which is strongly associated with the HD mutation in Peruvian patients and in HD families of diverse Latin American origins. In contrast, this Amerindian A1 haplotype is rare in Caucasian European patients and controls. We propose that the parent A1 *HTT* haplotype may be the most frequent haplotype of the HD mutation in Latin America, and show that distinct A1 variant haplotypes of European and Amerindian origin are present in contemporary mestizo populations. Alleles shared by European A1 and Amerindian A1

haplotypes may offer optimal targets for the development of allele-specific therapies in ethnically diverse patient populations across North America, South America, and Europe.

3.2 Material and Methods

3.2.1 HD Patient and Control Cohorts

Mestizo Peruvian HD patients have been genetically diagnosed at the Neurogenetics Research Center (NRC) at the Instituto Nacional de Ciencias Neurológicas (INCN) in Lima, Peru since 2000. DNA samples from Peruvian patients and controls were collected and stored for diagnostic and research purposes at INCN. IRB approval for this study was obtained from the Ethical Committee at INCN. HD and control donors gave informed consent that allowed for further studies on HD. Unaffected Amerindian Peruvian samples were originally collected as part of a genetic study of Parkinson disease, and were granted proper authorization for use in further studies related to neurodegenerative diseases (Cornejo-Olivas et al., 2014). All Peruvian samples were de-identified at the NRC before shipment to the Centre for Molecular Medicine and Therapeutics (CMMT) in Vancouver. Additional Latin American HD families were identified in the UBC HD BioBank from samples collected through the Centre for HD at the University of British Columbia. Latin American HD families were defined by confirmed origin of the proband or both parents from Mexico, Colombia, Ecuador, El Salvador, Argentina or Venezuela. Additional HD patients of European ancestry were screened from the UBC HD BioBank under existing research ethics protocols (UBC C&W CREB H06-70467 and H05-70532).

3.2.2 CAG and CCG Repeat Sizing

CAG repeat genotyping of 265 Peruvian HD and 43 Amerindian unaffected samples was performed at both the NRC and CMMT. CAG and CCG repeat lengths were confirmed in Peruvian, Amerindian, and Latin American individuals against controls of known CAG and CCG length at the CMMT. CAG and CCG repeat sizes were determined as previously described (Baine et al., 2013) using fluorescently labelled primers flanking the CAG repeat (HD344F, 5'-HEX-CCTTCGAGTCCCTCAAGTCCTTC-3' and HD450R, 5'-GGCGGCGGTGGCGGCTGTTG-3'), the CCG repeat (HD419F, 5'-AGCAGCAGCAGCAACAGCC-3' and HD482R, 5'-6FAM-GGCTGAGGAAGCTGAGGAG-3'), and both CAG and CCG repeats (HD344F, 5'-HEX-CCTTCGAGTCCCTCAAGTCCTTC-3' and HD482R, 5'-GGCTGAGGAAGCTGAGGAG-3') following established HD diagnostic testing guidelines (Bean and Bayrak-Toydemir, 2014).

3.2.3 *HTT* SNP Genotyping and Haplotype Reconstruction

Overlapping custom Illumina GoldenGate arrays of 92 SNPs and 77 SNPs spanning the *HTT* gene region (Illumina, San Diego, CA) were directly genotyped in 142 mestizo Peruvian, 43 Amerindian Peruvian, and 63 Latin American individuals (Kay et al., 2015). Curated genotype clusters for each SNP were exported as genotype calls using GenomeStudio (Illumina, San Diego, CA) and formatted for haplotype inference with

PHASE v2.1 for each assay design. Accuracy of reconstructed haplotypes was verified where possible by pedigree analysis of genotyped family members. Haplotypes of all Peruvian HD alleles were confirmed by familial segregation. Haplotypes of Latin American HD alleles from the CMMT were confirmed by familial segregation and by CCG repeat association as previously described (Kay et al., 2015). Following phasing and comparison of haplotype sequences across all genotyped individuals, common haplotypes consisting of 125 SNPs could be inferred from individual 92 SNP or 77 SNP data, of which 79 SNPs had informative heterozygosity in our sample populations (**Appendix B, Table B-1**). Genotype and haplotype data from our 62 mestizo Peruvian HD probands have been submitted to the Leiden Open Variation Database (www.LOVD.nl/HTT; individuals 81358-81419).

3.2.4 Identification and Genotyping of Amerindian A1 SNPs

Chromosomal Native American ancestry annotations from 1000 Genomes were examined from previously published data (Gravel et al., 2014). Among 25 reference Latin American controls from 1000 Genomes Project Phase 3 with homozygous Native American ancestry across *HTT*, one was homozygous for the A1 *HTT* haplotype previously defined by rs72239206 (GRCh37 chr4:g.3142661_3142664delACTT), rs149109767 (GRCh37 chr4:g.3230411_3230413delGAG), and rs362307 (GRCh37 chr4:g.3241845C>T) (Kay et al., 2015). Three additional *HTT* SNPs in this individual had homozygous variant alleles exclusive to the A1 haplotype among all 1000 Genomes Project samples: rs12508079 (GRCh37 chr4:g.3080238T>C), rs188072823 (GRCh37

chr4:g.3188616C>T), and rs186719032 (GRCh37 chr4:g.3255302G>A). SNP rs12508079:T>C was genotyped in all Peruvian HD chromosomes, all Amerindian control chromosomes, and all Latin American HD chromosomes by TaqMan assay (C___2480924_10, Applied Biosystems, Foster City, CA). SNPs rs188072823:C>T and rs186719032:G>A were genotyped by direct amplicon sequencing (rs188072823: forward primer TCGTCACTCCAAACACAATGG and reverse primer ACATAAGTCACAGCTGAAGAAAAA, rs186719032: forward primer GGCCGCTTCGAGGATGAT and reverse primer AACTTTCCGTGCAGCTCAA).

3.2.5 Ancestry Analysis of Mestizo Peruvian and Amerindian Individuals

Twelve Amerindian control samples (6 individuals with Amerindian A1 or derivative crossover haplotype and 6 individuals without A1) and 3 mestizo Peruvian HD probands with C1 HD haplotypes were selected for ancestry analysis using 1000 Genomes Project Phase 3 samples as a reference. Selected samples were genotyped for 7907 markers using a custom Infinium ADME Core Panel (Illumina, San Diego, CA), an approach that has been shown to accurately infer genetic ancestry (Visscher et al., 2009; Jackson et al., 2016). Genotypes were generated in GenomeStudio (Illumina) following standard quality control protocols. PLINK (v1.07) was then employed to remove markers with <95% call rate and deviations from Hardy-Weinberg equilibrium (P<0.001) as well as ensure that all samples had a >95% call rate. The intersection of markers between the genotyped samples and the 1000 Genomes Project samples was subsequently examined. Markers with minor allele frequencies of >1% were subjected to linkage disequilibrium pruning in

PLINK (pairwise genotypic correlation: window 50, step 5, r^2 0.2). This pruned dataset was then assessed using principal component analysis (PCA) with EIGENSOFT (v5.0) and ADMIXTURE (v1.2) employing cross validation to determine the optimal *K* value.

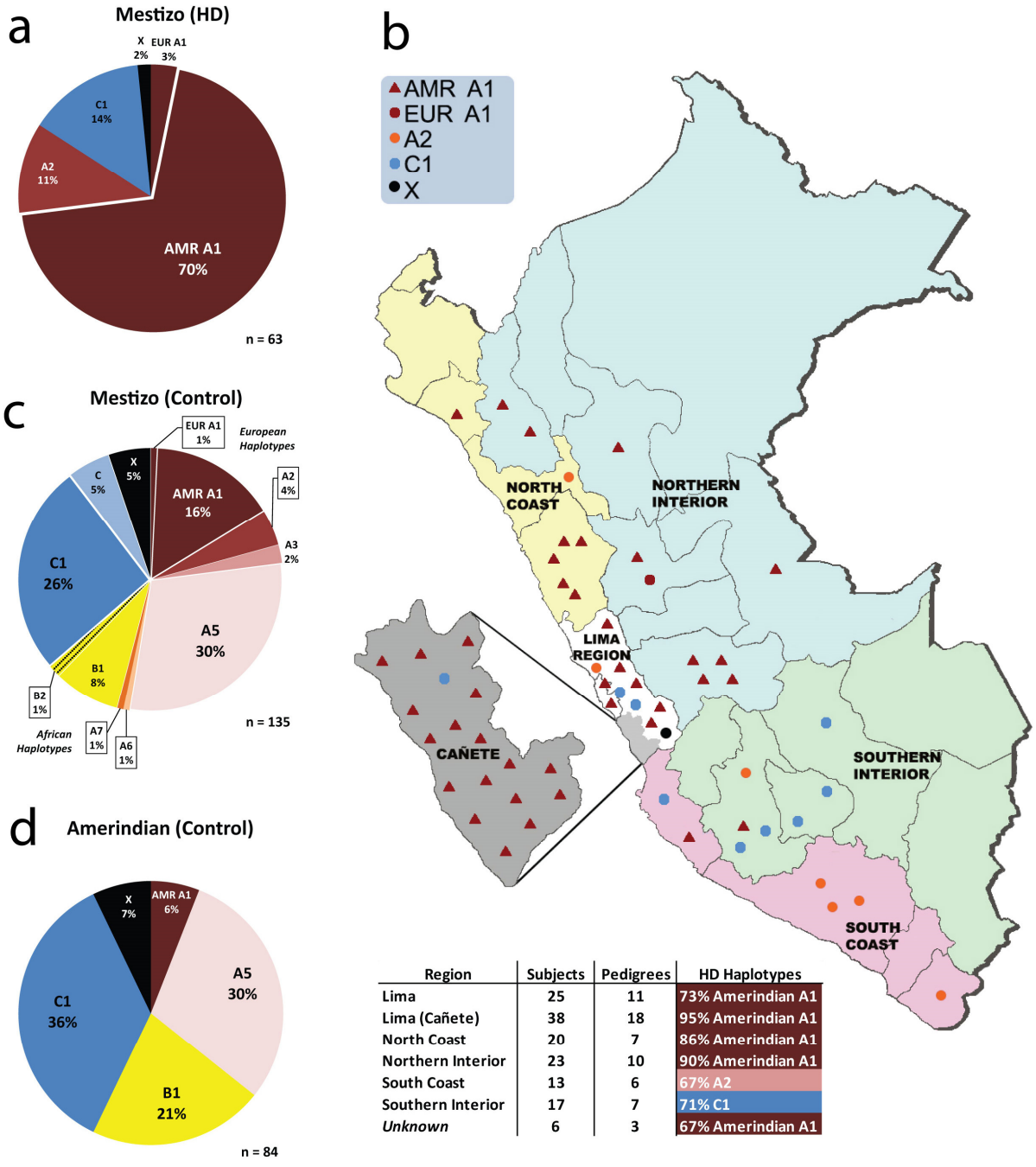
3.3 Results

3.3.1 HD in Peru Occurs Predominantly on the A1 *HTT* Haplotype

To determine the genetic ancestry of the expanded CAG repeat in Peruvian HD patients, we genotyped 142 individuals from 62 unrelated mestizo Peruvian HD families at either 92 SNPs or 77 SNPs spanning the *HTT* gene, as previously genotyped in HD patients of European ancestry (Kay et al., 2015). These mestizo Peruvian HD families spanned five broad geographic regions of the country: Lima, Northern Interior, North Coast, Southern Interior, and South Coast. One HD homozygote was found in one family, and was counted as two independently inherited HD chromosomes. In total, gene-spanning *HTT* haplotypes of 79 informative SNPs (**Appendix B, Table B-1**) were fully phased to CAG repeat length in 63 unrelated HD chromosomes and 135 unrelated control chromosomes from mestizo Peruvian patient families.

Forty-six out of 63 (73%) Peruvian HD mutations were found on the A1 *HTT* haplotype (**Figure 3-1a**) previously shown to be the most common gene-spanning *HTT* haplotype in HD chromosomes of European ancestry (Kay et al., 2015). A1 HD haplotypes occur at an even higher frequency in Peruvian patients than in patients of European ancestry

(p=0.0008, chi-square, 73% versus 48% in Peru and the UBC BioBank, respectively). Seven of 63 (11%) Peruvian HD mutations were found on A2, previously shown to be the second most common HD haplotype in patients of European ancestry. However, Peruvian HD mutations also occurred on the C1 haplotype at similar frequency to A2 (9/63, 14%), whereas in European patients the HD mutation on C1 is rare (3.2-4.5%). In East Asian and Southeast Asian patient populations, the HD mutation has also been shown to occur frequently on C1 (Warby et al., 2011; Pulkes et al., 2014). Among Black South African patients, the HD mutation occurs on a heterogenous collection of A, B, and C haplotypes, usually informative of African origin (Baine et al., 2013). In our mestizo Peruvian HD families, no HD haplotypes indicative of African origin were observed.



Region	Subjects	Pedigrees	HD Haplotypes
Lima	25	11	73% Amerindian A1
Lima (Cañete)	38	18	95% Amerindian A1
North Coast	20	7	86% Amerindian A1
Northern Interior	23	10	90% Amerindian A1
South Coast	13	6	67% A2
Southern Interior	17	7	71% C1
Unknown	6	3	67% Amerindian A1

Figure 3-1. *HTT* haplotype distributions in the Peruvian population.

(a) The HD mutation in Peru occurs predominantly on the A1 haplotype (73%), of which nearly all are the Amerindian A1 variant haplotype. (b) Haplotypes of the HD mutation differ by geographic region of Peru. (c) Control *HTT* haplotypes in Mestizo Peruvians reflect indigenous ancestry with major European

admixture and minor African contributions. (d) Unaffected Amerindian *HTT* haplotypes reflect low genetic diversity in this subpopulation relative to Mestizo controls, and reveal the Amerindian A1 haplotype in the indigenous population of Peru.

There are significant regional differences in the frequency of HD haplotypes across Peru (**Figure 3-1b**). A1 predominates in Lima, the North Coast, and the Northern Interior, whereas A2 and C1 are more common in the South Coast and Southern Interior, respectively. In Cañete, where a historical focus of HD is reported, HD occurs almost exclusively on A1.

Control chromosomes from Peruvian HD pedigrees represent diverse ethnic origins, including typically European, East Asian, and African *HTT* haplotypes (**Figure 3-1c**). Interestingly, A1 occurs at a higher frequency in mestizo Peruvian controls (16.3%) than in European controls (8.0-9.7%), whereas this haplotype is absent in East Asian populations (Kay et al., 2015). A5 also occurs at a higher frequency in mestizo Peruvian controls (29.6%) than in European controls (6.1-6.9%), but at a rate comparable to East Asian controls (27%) (Warby et al., 2011).

3.3.2 The A1 Haplotype Occurs in the Indigenous Amerindian Population of Peru

In order to investigate indigenous Amerindian contributions to the ancestry of HD in mestizo Peruvian patients, we additionally haplotyped 43 unaffected Peruvian

Amerindian individuals with our 77 SNP panel (**Figure 3-1d**). Haplotype diversity among the Amerindian individuals was low, with only 5 distinct gene-spanning haplotypes revealed by our SNP panel versus 21 in mestizo Peruvians (**Appendix B, Table B-2**), reflecting prior findings of low genetic diversity and serial founder effects in Native American populations (Sandoval et al., 2013; Wang et al., 2007). Unexpectedly, the A1 haplotype defined by rs72239206 (GRCh37 chr4:g.3142661_3142664delACTT), rs149109767 (GRCh37 chr4:g.3230411_3230413delGAG), and rs362307 (GRCh37 chr4:g.3241845C>T), known to be associated with HD in European populations, was found in multiple Amerindian chromosomes (5/84, 6.0%), whereas it is entirely absent in East Asian reference populations to which indigenous Americans are more closely related. Other common *HTT* haplotypes that would indicate European ancestry, such as A2, are absent among our Amerindian individuals, arguing against European admixture as the source of A1 in this population.

To examine the possibility of cryptic admixture in our Amerindian individuals, which could support a European origin of A1 in these individuals, we genotyped 7907 SNPs in a subset of 12 Amerindian individuals and three mestizo HD patients using a custom genetic diversity panel. All five Amerindian A1 haplotype carriers were included, in addition to one putative A1 crossover haplotype. It has been shown by our group and others that genotypes from the ADME panel enable identification of individuals with misattributed ancestry by principal component analysis (PCA) (Visscher et al., 2009). We extracted genotypes representing the ADME panel from Phase 3 of the 1000 Genomes Project and performed PCA and ADMIXTURE analysis with our Amerindian

and mestizo Peruvian HD samples. PCA revealed that the Amerindian samples formed an extreme genetic cluster closest to Peruvian reference samples (PEL) from the 1000 Genomes Project (**Figure 3-2a**). There was no difference in the clustering of Amerindian A1 carriers versus Amerindian individuals without A1. Our three selected mestizo Peruvian HD patients, each with HD on C1, were intermediate between the Amerindian cluster and individuals of European ancestry, and overlapped with reference Peruvian samples. ADMIXTURE analysis confirmed the predominant indigenous ancestry of our Amerindian individuals, in contrast to varying degrees of European admixture in reference Peruvian individuals and the mestizo Peruvian HD probands (**Figure 3-2b**; **Appendix B, Figure B-1**). These data strongly suggest that our unaffected Amerindian individuals are of uniform indigenous descent despite the presence of the A1 *HTT* haplotype typical of Europeans. From these data, we hypothesized that the A1 haplotype in Amerindians and in some mestizo Peruvians may be of ancient Amerindian origin rather than recent European origin via admixture.

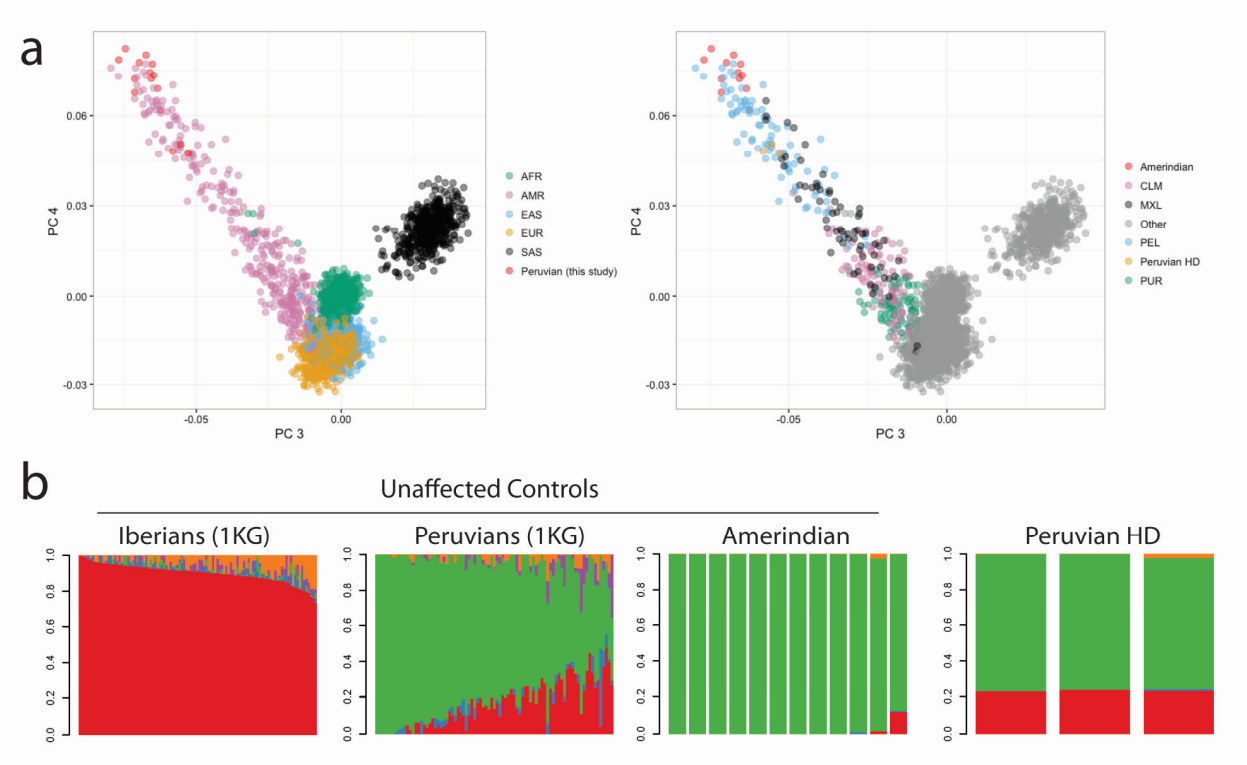


Figure 3-2. Genome-wide admixture analysis of selected Peruvian Amerindian controls and mestizo HD patients relative to the 1000 Genomes Project reference samples.

(a) Plotting of principal components 3 and 4 from Amerindian Peruvian individuals reveals extreme clustering by ethnicity and partial overlap with reference Peruvians from the 1000 Genomes Project. Left panel depicts selected Peruvian samples compared to 1000 Genomes superpopulation groups: AFR, African, AMR, admixed American, EAS, East Asian, EUR, European, and SAS, South Asian. Right panel depicts selected Peruvian Amerindian samples and mestizo Peruvian HD samples in comparison to 1000 Genomes populations from the AMR superpopulation: CLM, Colombian, MXL, Mexican, PEL, Peruvian, and PUR, Puerto Rican. Amerindian Peruvian controls are most similar to reference Peruvians, reflecting known heterogeneity of indigenous genetic ancestry across Latin America. (b) ADMIXTURE analysis suggests a nearly exclusive indigenous ancestry among Amerindian Peruvian controls, including all Amerindian individuals with the A1 haplotype. Mestizo Peruvian HD probands show similar levels of admixture as reference Peruvians relative to Europeans from Spain and Portugal (*K*=5, see **Appendix B, Figure B-1**).

3.3.3 The Amerindian A1 Haplotype is Marked by Specific SNPs and Represents the Majority of HD Mutations in Peru

To determine if additional genetic variants could distinguish A1 haplotypes of Amerindian origin from A1 haplotypes of European origin, we examined *HTT* variants in A1 haplotypes of admixed Latin American individuals from the 1000 Genome Project. Specific chromosomal segments from these individuals have been previously identified as Native American, European, or African by local genomic ancestry inference (Gravel et al., 2014). Among 25 reference Latin American individuals from 1000 Genomes Project Phase 3 with homozygous Native American ancestry across *HTT*, one was homozygous for the A1 *HTT* haplotype defined by rs72239206 (GRCh37 chr4:g.3142661_3142664delACTT), rs149109767 (GRCh37 chr4:g.3230411_3230413delGAG), and rs362307 (GRCh37 chr4:g.3241845C>T). This individual was found to be homozygous for three additional SNP variants specific to the A1 haplotype: rs12508079:T>C in *HTT* intron 1, rs188072823:C>T, and rs186719032:G>A.

In the 1000 Genomes Project Phase 3 data set, the rs12508079:T>C variant allele is highly enriched in A1 from Latin American individuals (**Figure 3-3**; **Appendix B, Figure B-2 and Table B-3**). A1 with rs12508079:T>C represents 9/10 (90%), 8/10 (80%), and 22/22 (100%) of Colombian, Mexican, and Peruvian A1 control haplotypes, respectively. In contrast, the rs12508079:T>C variant allele is rare among A1 *HTT* chromosomes from European reference populations, as well as in South Asians where the A1

haplotype is infrequently found. The rs12508079:T>C variant allele is not found in individuals of exclusive African or East Asian ancestry from 1000 Genomes Project Phase 3, in agreement with the previously established absence of A1 in these populations (Warby et al., 2011; Baine et al., 2013; Kay et al., 2015).

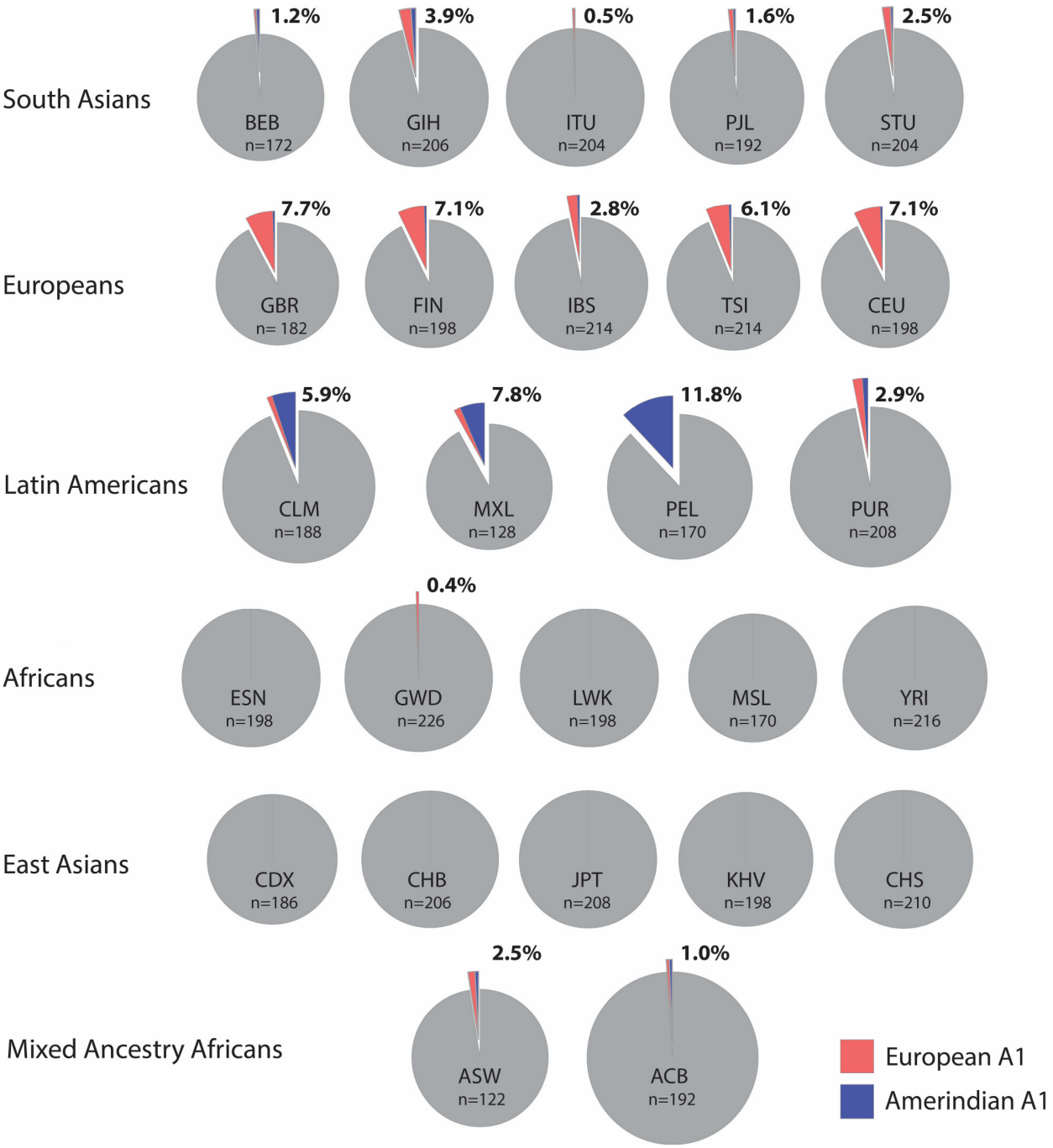


Figure 3-3. European and Amerindian A1 *HTT* haplotype frequencies across all reference populations of the 1000 Genomes Project

Latin American reference populations have the highest frequency of the Amerindian A1 variant haplotype, defined by the rs12508079:T>C variant allele in conjunction with the parent A1 haplotype marker

rs149109767 (GRCh37 chr4:g.3230411_3230413delGAG), constituting the majority of all A1 haplotypes among Colombians (CLM), Mexicans (MXL), and Peruvians (PEL). Amerindian A1 also constitutes a third of all A1 haplotypes among Puerto Ricans (PUR). In contrast, Amerindian A1 is rare in European populations where the parent A1 haplotype is common: British (GBR), Finnish (FIN), Iberian (IBS), Toscani Italian (TSI), Utah Caucasian (CEU). Amerindian A1 is also rare in South Asian populations where the parent A1 haplotype is also infrequently found: Bengali (BEB), Gujarati (GIH), Telugu (ITU), Punjabi (PJL), and Sri Lankan Tamil (STU). The parent A1 haplotype is entirely absent in Black African and East Asian populations, except for one instance in a Gambian individual. European and Amerindian A1 are both found among admixed Africans from the United States (ASW) and Barbados (ACB).

Variant alleles of rs186719032:G>A and rs188072823:C>T were found on 20/22 (91%) A1 control haplotypes in the Peruvian reference population, but only half of Mexican and Colombian A1 control haplotypes, suggesting a diversity of Amerindian subtypes among A1 haplotypes bearing rs12508079:T>C. We thus define the *Amerindian A1* haplotype as any *HTT* haplotype carrying the rs12508079:T>C variant allele in addition to the canonical A1 variant alleles at rs72239206, rs149109767, and rs362307. Some, but not all Amerindian A1 haplotypes also bear rs186719032:G>A and rs188072823:C>T variant alleles, representing Amerindian A1 subtypes.

We directly genotyped rs12508079:T>C, rs186719032:G>A, and rs188072823:C>T in all unaffected Amerindian samples and all Peruvian HD patients, then phased to 79-SNP *HTT* haplotypes and CAG repeat length by familial segregation (**Figures 3-1 and 3-4; Appendix B, Table B-4**). All five normal Amerindian chromosomes with A1 variant alleles at rs72239206, rs149109767, and rs362307 also carried variant alleles at

rs12508079, rs186719032, and rs188072823, suggesting a common Amerindian A1 haplotype by descent. Strikingly, 41/45 (91.1%) mestizo Peruvian A1 HD chromosomes occur on the Amerindian A1 haplotype bearing all three Amerindian A1-specific alleles, and not on the European A1 haplotype lacking these three Amerindian-specific variants, supporting a pre-Colombian origin of HD in this population rather than a proximate European origin as previously supposed. Specific Amerindian A1 variant alleles at rs12508079:T>C, rs186719032:G>A and rs188072823:C>T did not occur on any other haplotypes than A1 in direct genotyping of our Peruvian HD patients and unaffected Amerindian individuals.

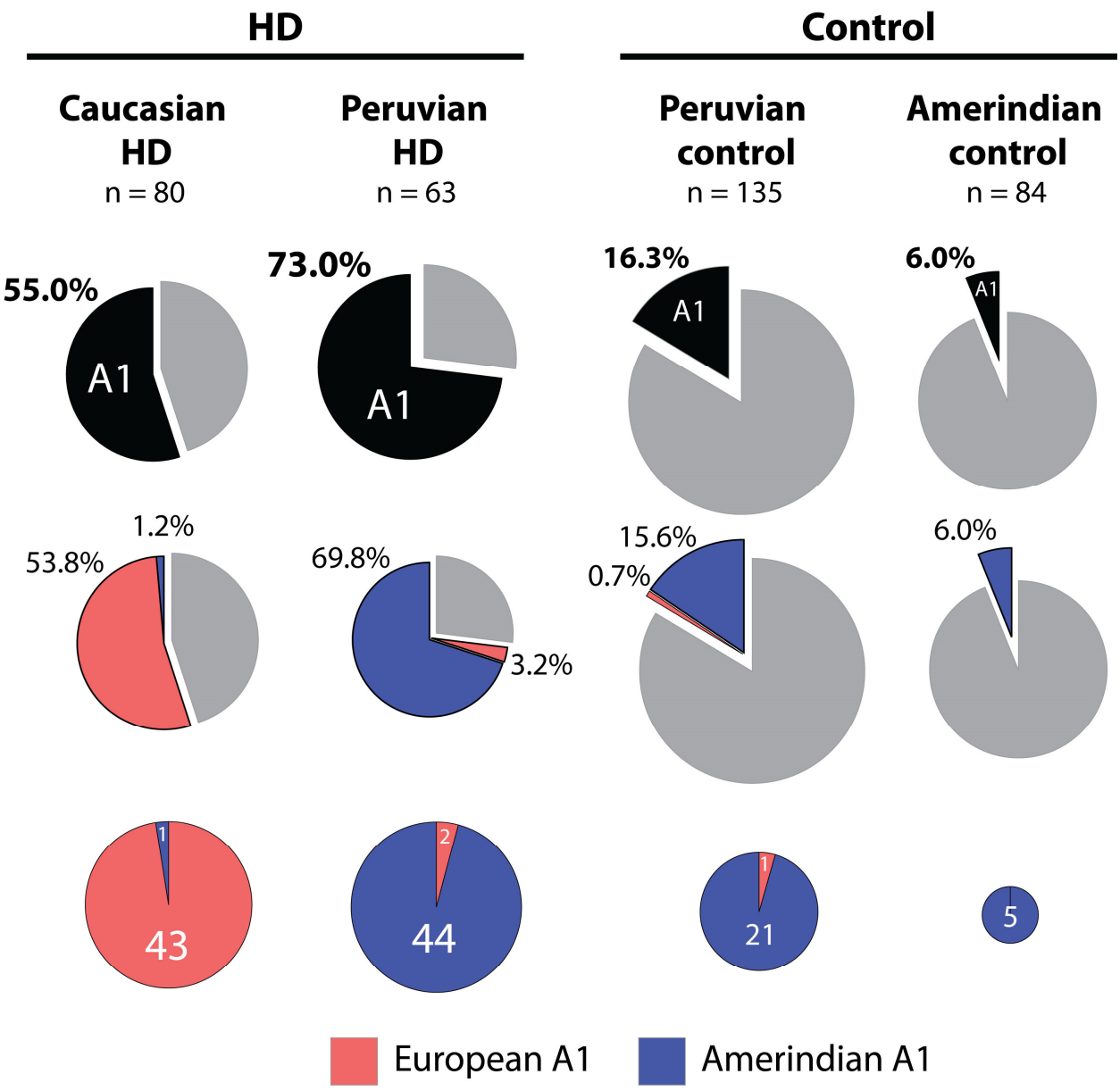


Figure 3-4. The frequency of European and Amerindian A1 *HTT* haplotypes differs dramatically between HD chromosomes from Caucasian Canadian patients and mestizo Peruvian patients. The A1 *HTT* haplotype in mestizo Peruvian HD and control chromosomes is almost exclusively the Amerindian A1 variant. In unaffected Amerindian Peruvian individuals, the A1 *HTT* haplotype is exclusively the Amerindian A1 variant. In contrast, the Amerindian A1 haplotype variant occurs very rarely in Caucasian Canadian HD chromosomes. The A1 *HTT* haplotype is defined by variant alleles at

rs72239206 (GRCh37 chr4:g.3142661_3142664delACTT), rs149109767 (GRCh37 chr4:g.3230411_3230413delGAG), and rs362307 (GRCh37 chr4:g.3241845C>T). The Amerindian A1 variant haplotype is additionally defined by the C allele at rs12508079 (GRCh37 chr4:g.3080238T>C).

Additionally, we genotyped patients representing 80 unrelated HD chromosomes of European ancestry for rs12508079:T>C, 44 (55%) of which were previously found to have the A1 *HTT* haplotype defined by variant alleles at rs72239206, rs149109767, and rs362307 (**Figure 3-4**). In stark contrast to Peruvian HD chromosomes with the A1 haplotype, only 1/44 European HD chromosome with the A1 haplotype carried the Amerindian A1-defining allele of rs12508079:T>C. Fully 43/44 (97.7%) European A1 HD chromosomes lack the Amerindian A1-defining variant allele, in agreement with the rarity of rs12508079:T>C in A1 chromosomes from European reference samples in the 1000 Genomes Project.

3.3.4 Amerindian A1 is the Most Common HD haplotype in HD families of Latin American Origin

To extend our findings of an Amerindian-specific A1 variant to HD chromosomes from other Latin American populations, we haplotyped 63 subjects from 17 HD families of diverse Latin American origins in the UBC HD BioBank (**Figure 3-5**). Ten out of 17 (59%) Latin American HD chromosomes were found on the A1 *HTT* haplotype marked by variant alleles at rs72239206, rs149109767, and rs362307. Eight of these 10 Latin American A1 HD chromosomes (80%) also carried the rs12508079:T>C variant allele,

similar to the enrichment for Amerindian A1 observed in Peruvian HD chromosomes. One HD mutation was observed on the A7 haplotype, suggesting African ancestry, and one on the C4 haplotype found in both European and African patients. Among seven unrelated A1 control chromosomes from our Latin American HD families, four carried the Amerindian-specific variant allele of rs12508079:T>C, similar to the enrichment for Amerindian A1 in Latin American reference samples from the 1000 Genomes Project. Control chromosomes from Latin American HD families were also enriched for A5 (20%) relative to European controls, as in mestizo Peruvian (30%) and unaffected Amerindian individuals (30%). A1 and A2 of European ancestry were observed in our admixed Latin American controls but few haplotypes of definitive African ancestry, suggesting predominant Amerindian and European ancestry in these families.

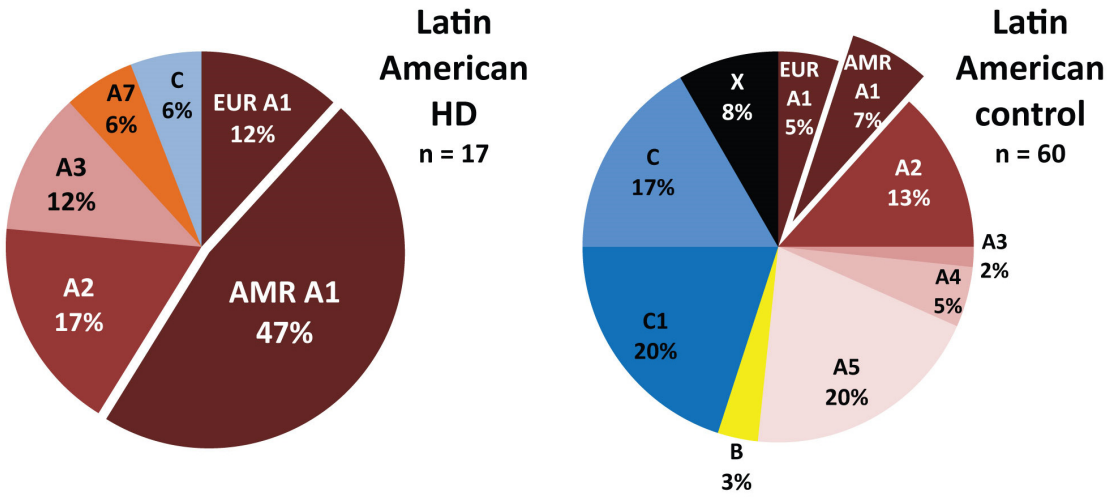


Figure 3-5. *HTT* haplotype distribution of HD and control chromosomes in Latin American HD families from the UBC HD Biobank.

Amerindian A1 is the most frequent HD haplotype in a collection of patients from across Latin America (47%). The parent A1 haplotype, which could be therapeutically targeted for allele-specific *HTT* silencing

in European patient populations, constitutes the majority (59%) of HD chromosomes in Latin American patients. Latin American control haplotypes reflect indigenous Amerindian and European admixture.

The occurrence of the HD mutation on the parent A1 haplotype in both European and Latin American patients suggests that common defining A1 variant alleles could represent shared targets for allele-specific silencing in both populations. In European patients, A1 and A2 allele targets in combination allow allele-specific treatment of the most HD patients (Kay et al., 2015). We find that the same defining A1 alleles as identified in European HD patients (Chapter 2), specifically rs72239206 (GRCh37 chr4:g.3142661_3142664delACTT), rs149109767 (GRCh37 chr4:g.3230411_3230413delGAG), and rs362307 (GRCh37 chr4:g.3241845C>T), allow allele-specific treatment of the most Peruvian HD patients given one allele target (35/62, 56.5%), and allow treatment of the most Peruvian HD patients overall in combination with A2 (42/62, 67.7%). A1 and A2 alleles in combination also allow treatment of 71% (12/17) of Latin American HD probands among our UBC BioBank families. These data suggest that A1 and A2 allele targets may represent optimal targets for allele-specific therapy in Latin American HD patient populations, as in patient populations of predominant European ancestry.

3.4 Discussion

To date, the origin of HD in Latin America has remained unclear in the absence of genetic ancestry markers within *HTT*. To our knowledge, this is the first analysis of dense

haplotypes of the HD mutation in Latin America, allowing description of the inferred ancestry of the HD mutation in Peru and other patients of Latin American origin. We show that HD in Latin America occurs most frequently on the A1 haplotype, as in European HD patients. Remarkably, a subtype of the A1 haplotype occurs in individuals of confirmed Amerindian ancestry, and represents the most frequent HD haplotype in patients from Peru and elsewhere in Latin America. HD in Latin American patients is also found on haplotypes of European ancestry and in rare instances on haplotypes suggestive of African origin, reflecting extensive admixture in Latin America (Ruiz-Linares et al., 2014). HD mutations on C1 may represent either an indigenous origin or a recent founder effect resulting from East Asian immigrants that arrived in Peru over the past two centuries (Herrera, 1986). However, ADMIXTURE analysis shows minimal East Asian ancestry among Peruvian patients with the HD mutation on C1, supporting an indigenous origin of the C1 disease haplotype in these families. Peruvian HD chromosomes on A2, frequent in the South Coast, may represent recent admixture from European sources.

Ancestry admixture in Latin America is known to occur in distinct proportions by country and region. Interestingly, our *HTT* haplotype data and admixture analysis of individuals from the Peruvian population reflect the known demographic and genetic history of Peru. For example, the Peruvian population has one of the highest Amerindian ancestry components (83%) among all evaluated groups, with admixed Spanish European ancestry and marginal African and Asian contributions (Sandoval et al., 2013). Indeed, our *HTT* haplotyping data of Peruvian HD patients shows a predominant Amerindian ancestry at this locus, with European admixture and isolated African contributions among

HD and control chromosomes. Interestingly, a recent epidemiological study of the incidence of HD among American Indians across the United States revealed an absence of the disease among the Navajo (Gordon et al., 2016). In addition to admixture differences, there may be considerable genetic heterogeneity of HD among different Native American subpopulations.

The presence of the A1 haplotype in Amerindian individuals with no discernable European admixture, and the identification of ancestry-specific A1 alleles that are frequent in Latin American individuals but rare in Europeans, suggests that the parent A1 haplotype has deep common ancestry between European and Native American populations. This is further supported by the presence of the A1 haplotype in all reference populations of South Asian ancestry. Shared ancestry of the parent A1 haplotype between Europeans and Native Americans is unexpected given the absence of A1 among contemporary East Asian individuals to which Native Americans are believed to be related. Recent reports have suggested that Native American populations may have closer ancestry to indigenous Siberian and Central Asian populations than to contemporary East Asians, and that gene flow between Eurasian populations and Native Americans may have occurred. Nuclear DNA extracted from 24,000-year old remnants of a boy in south-central Siberia indicated a shared ancestry between contemporary Native Americans and Europeans but divergent from modern East Asians (Raghavan et al., 2014). We hypothesize that the A1 haplotype was present in a similar ancient population of Central Asia, distinct from the ancestors of contemporary East Asians. Alternatively, A1 may have been lost in contemporary East Asians by bottleneck effects and genetic drift.

The defining Amerindian A1 alleles at rs12508079:T>C, rs186719032:G>A, and rs188072823:C>T are observed at low frequency outside Latin American populations, and these alleles may represent reverse admixture from New World sources. Alternatively, these alleles may have been present in ancestral Indo-European populations and become enriched in Amerindian populations by serial founder events. Detailed examination of Amerindian A1 alleles in HD and control cohorts from other defined indigenous populations would be necessary to clarify and test different migration models of this haplotype.

Regardless of the ancestral origin of A1 worldwide, the common occurrence of HD on the parent A1 haplotype in both Europe and Latin America suggests that allele-specific therapies targeting A1 could be designed for patients in both populations. We previously demonstrated that approximately 40% of HD patients of European ancestry could be treated with antisense reagents selective for the A1 *HTT* haplotype (Kay et al., 2015). In our Peruvian HD cohort the percentage of treatable patients is higher than in European populations, with 57% of Peruvian HD probands heterozygous for A1 and phased to the HD mutation. Among our Latin American HD pedigrees, 53% would be treatable with A1 targets, also exceeding the heterozygosity observed in European patients. This suggests that antisense reagents developed for allele-specific suppression of the A1 haplotype (i.e. against the deleted allele of rs72239206, the deleted allele of rs149109767, or the T allele of rs362307) may have similar or greater utility in treatment of Latin American HD patients as in patients of major European descent. Further, as in European patients, alleles

specific to the A1 and A2 haplotype allow treatment of the greatest number of Peruvian patients in combination (68%) among all possible combinations of two allele targets, suggesting that A1 and A2 may also represent prioritized panels of primary and secondary allele targets in Latin American patient populations. In conclusion, our study shows that the targetable A1 HD haplotype is common in Latin America, but that this haplotype has genetically distinct Amerindian and European origins in contemporary mestizo American populations.

Chapter 4: Huntington Disease Reduced Penetrance Alleles Occur at High Frequency in the General Population

4.1 Synopsis

A CAG repeat longer than 35 trinucleotides in exon 1 of the huntingtin gene (*HTT*) defines an allele capable of causing the HD phenotype, with reduced penetrance at 36-39 CAG repeats (Kremer et al., 1994; , 1998; Bean and Bayrak-Toydemir, 2014). The frequency of repeat alleles ≥36 CAG in the general population is unknown. Although many control cohorts have been assessed by exome or whole-genome sequencing (Auton et al., 2015), short-read technologies are unable to accurately assess *HTT* CAG repeat lengths (McIver et al., 2011). In this study, we estimate the frequency of individuals in the general population with ≥36 CAG repeats in several large samples from three Western countries, using current genetic diagnostic standards for HD (Bean and Bayrak-Toydemir, 2014).

The smallest CAG repeat length associated with the HD phenotype is 36, but repeat lengths of 36-39 CAG have been found in asymptomatic elderly individuals, indicating that the lower end of the HD CAG repeat range is incompletely penetrant over the average human lifespan (Rubinsztein et al., 1996; McNeil et al., 1997). Using our general population frequency of reduced penetrance CAG repeat alleles, we directly estimate the penetrance of 36-38 CAG repeats among individuals ≥65 years of age. We

compare our direct penetrance rates to estimates from clinical cohorts in order to evaluate penetrance bias that may result from clinical ascertainment. To our knowledge, this study offers the first large-scale assessment of ≥36 CAG repeats in the general population and the first attempt to directly assess the penetrance of alleles in the reduced penetrance range.

4.2 Methods

4.2.1 Study Populations

3952 saliva-extracted DNA samples were obtained from the Coriell Personalized Medicine Collaborative (CPMC) Community Cohort (Keller et al., 2010). The CPMC advertised to healthcare providers at local hospitals as well as to the general public through coverage in local media and by word of mouth. Participants were 18 years of age or older. Participants were not recruited based on the presence or absence of any disease phenotype. Samples used for this study were limited to those who explicitly consented to have their samples used in research beyond the scope of the CPMC. Samples were anonymized before use in this study.

2016 anonymous general population samples were obtained from a population-based research DNA sample set held by NHS Grampian, Scotland. All DNA was originally extracted from buccal swabs taken from participants of a historic antenatal cystic fibrosis screening study conducted in Aberdeen, Scotland.

General population controls from British Columbia, Canada were obtained from a large archived cohort of samples previously collected for the purpose of lipid screening as part of routine clinical blood work. Donors were not recruited or collected on the basis of any specific lipid profile or disease and were reported to be representative of British Columbia's general population in that they were largely Caucasian individuals of Northern European ancestry (Semaka et al., 2013b). A subset of 1600 permanently anonymized samples was randomly selected from the archived collection of approximately 6000 DNA samples.

4.2.2 Standard Protocol Approvals, Registrations, and Patient Consents

For the CPMC, saliva samples were collected in person following a group informed consent session. Only individuals who consented to share their DNA were included in this analysis. The CPMC protocol was reviewed and approved by the Institutional Review Board (IRB) of the Coriell Institute for Medical Research. In addition, the use of samples in this analysis was also reviewed and approved by the IRB of the Coriell Institute for Medical Research. Ethical approval to share fully anonymized material from the archived University of Aberdeen cohort was obtained from North of Scotland Research Ethics Committee (NOSRES: 13/ NS/ 0025). Ethical approval for the secondary use of archived anonymous samples from British Columbia was obtained from the UBC C&W Clinical Research Ethics Board (H06-70356). Ethical approval to

examine CAG size distributions in all three populations at the UBC C&W site was obtained from UBC C&W CREB (H05-70532).

4.2.3 CAG and CCG Repeat Sizing

All samples in this study were sized by an identical assay at the Centre for Molecular Medicine and Therapeutics at the University of British Columbia in Vancouver, Canada. CAG and CCG repeat sizes were determined as previously described (Semaka et al., 2013b) using fluorescently labelled primers flanking the CAG repeat (HD344F, 5'-HEX-CCTTCGAGTCCCTCAAGTCCTTC-3' and HD450R, 5'-GGCGGCGGTGGCGGCTGTTG-3') and CCG repeat (HD419F, 5'-AGCAGCAGCAGCAACAGCC-3' and HD482R, 5'-6FAM-GGCTGAGGAAGCTGAGGAG-3'). A third PCR encompassing both CAG and CCG sequences (HD344F, 5'-HEX-CCTTCGAGTCCCTCAAGTCCTTC-3' and HD482R, 5'-GGCTGAGGAAGCTGAGGAG-3') was used to exclude allelic dropout and to phase CAG and CCG sizes from the first two sizing assays (1998; Potter et al., 2004; Bean and Bayrak-Toydemir, 2014). Sizing was performed relative to a control panel of sequenced CAG and CCG repeat lengths. CAG repeat alleles were classified as normal (<27 CAG), intermediate (27-35 CAG), or HD-associated (≥36 CAG). HD-associated CAG repeat alleles were classified as reduced penetrance (36-39 CAG) or full penetrance (≥40 CAG). From the original DNA sample sets, 46 CPMC samples, 201 University of Aberdeen samples, and 6 University of British Columbia samples failed CAG repeat sizing at both alleles.

4.2.4 Haplotype Analysis

General population individuals with at least one CAG repeat ≥36 were genotyped at 90 SNPs across the Huntington gene as previously described (Kay et al., 2015). Reconstructed risk haplotypes (A1, A2, or A3a) were phased to CAG repeat length by haplotype-specific CCG and CAG repeat length associations previously described (Semaka et al., 2013b; Kay et al., 2015) or by haplotype homozygosity.

4.2.5 Frequency and Age-Dependent Penetrance Estimates

Bayesian binomial 95% confidence intervals for all CAG allelic and genotypic frequencies were calculated using *binom* and *ggplot2* in R. For allelic frequency differences between general population samples, chi-square tests were performed using contingency tables. The number of individuals ≥65 years old in British Columbia was derived from Statistics Canada for the 2012 prevalence year and multiplied by the genotypic frequency of CAG 36, CAG 37, and CAG 38 in our general population sample to estimate the number of individuals ≥65 years old in British Columbia with each genotype. Patients ascertained in our 2012 British Columbia prevalence cohort with diagnostic CAG repeat lengths confirmed by the Molecular Genetics Laboratory at the British Columbia Children and Women's Hospital were included for estimates of age-dependent penetrance of CAG 36-39 alleles (Fisher and Hayden, 2014). To derive age-dependent penetrance, the number of clinically ascertained, genetically confirmed

patients ≥65 years old with each reduced penetrance CAG repeat length was divided by the number of individuals ≥65 years old expected to have that repeat length in British Columbia in 2012. Extrapolated clinical penetrance rates for CAG 36, 37, and 38 by age 65 were taken from multicentre data published by Langbehn and colleagues (Langbehn et al., 2004).

4.3 Results

We assessed *HTT* CAG repeat length in 3906 individuals from the adult general population of the United States, 1815 individuals from the North of Scotland, and 1594 previously reported individuals from British Columbia, Canada (Semaka et al., 2013b). Out of a total of 7315 individuals (14360 alleles), 18 individuals had HD-associated CAG repeat lengths ≥36 (0.246%, 95%CI: 0.151-0.380%) (**Table 4-1**). Among individuals with HD-associated alleles, CAG 36-37 genotypes were the most common (36, 0.096%; 37, 0.082%; 38, 0.027%; 39, 0.000%; ≥40, 0.041%) (**Figure 4-1a**). Approximately 1 in 16 individuals (6.20%) have an IA genotype (453/7315, 95%CI: 5.66-6.76%), a small fraction of whom carry two IAs (5/7315, 0.068%). Across all three populations, IAs occurred at an allelic frequency of 3.13% (458/14630, 95%CI: 2.86-3.42%). HD-associated allele frequency (p=0.485, chi-square) and intermediate allele frequency (p = 0.080, chi-square) did not differ between samples from the United States, Scotland, and Canada.

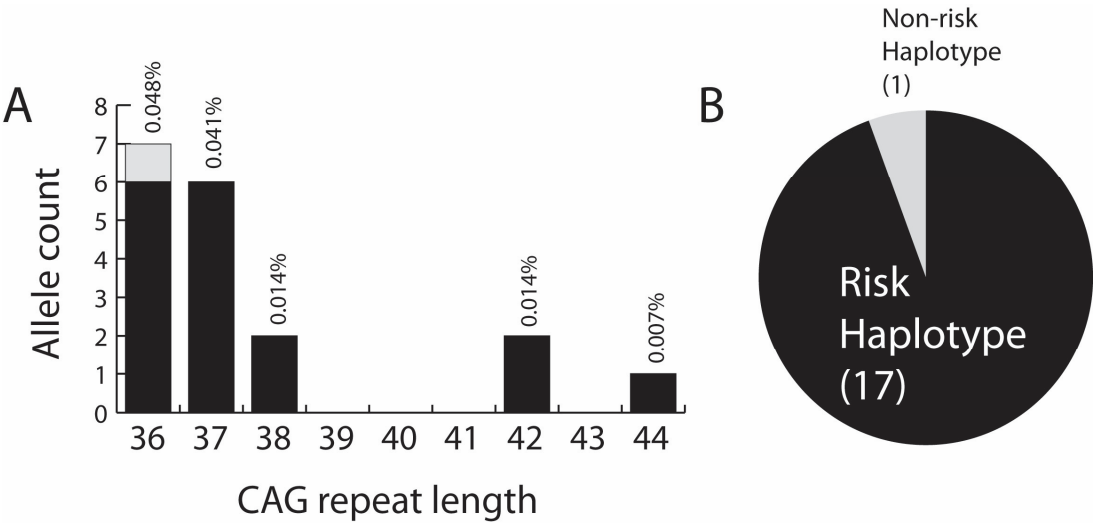


Figure 4-1. Distribution and haplotype of general population CAG repeat alleles in the HD range (≥36 CAG)

(A) HD CAG repeat distribution and frequency out of 14630 total general population alleles. **(B)** Haplotypes of HD alleles in the general population conform to risk haplotypes (A1, A2, or A3a) of clinically ascertained, disease-causing HD mutations in patients of European ancestry.

We additionally sought to determine whether the ≥36 CAG repeat alleles found in our general population sample occur on the same genetic background as those found in clinically manifest HD patients. We performed *HTT* haplotype analysis in all 18 general population individuals with alleles ≥36 CAG repeats. All but one general population ≥36 CAG chromosome conformed to our previously established *HTT* haplotypes of disease-causing mutations in manifest HD patients of European ancestry, A1, A2, or A3a (**Figure 4-1b**) (Kay et al., 2015; Warby et al., 2009), suggesting that HD-associated CAG repeat alleles in the general population occur on a similar *cis*-genetic background to those in patients from the clinic.

Fifteen of 18 (83.3%) incidentally ascertained ≥36 CAG repeats are considered HD alleles of reduced penetrance (CAG 36-39) by current ACMG diagnostic criteria (**Table 4-2**) (Bean and Bayrak-Toydemir, 2014). However, in our recent multisource prevalence study of HD in British Columbia, Canada, a much smaller proportion (15/353, 4.25%) of genetically diagnosed patients with validated CAG repeat lengths have CAG repeats in this range (**Figure 4-2a**). Onset of HD typically occurs in adulthood, with length of the expanded CAG repeat inversely related to mean age at disease onset (Langbehn et al., 2004; Lee et al., 2012b). Average age of onset is latest in patients with CAG repeats between 36-41, suggesting that individuals with reduced penetrance alleles are more likely to manifest HD in old age. Indeed, among symptomatic patients with a confirmed reduced penetrance genotype in British Columbia, 80% (12/15) were ≥65 years of age as of the April 1st, 2012 prevalence date. We therefore sought to estimate the minimum penetrance of this allele class among individuals ≥65 years of age by dividing the number of clinically ascertained, symptomatic patients with each reduced penetrance CAG genotype by the expected number of individuals with each genotype in British Columbia (**Figure 4-2b**).

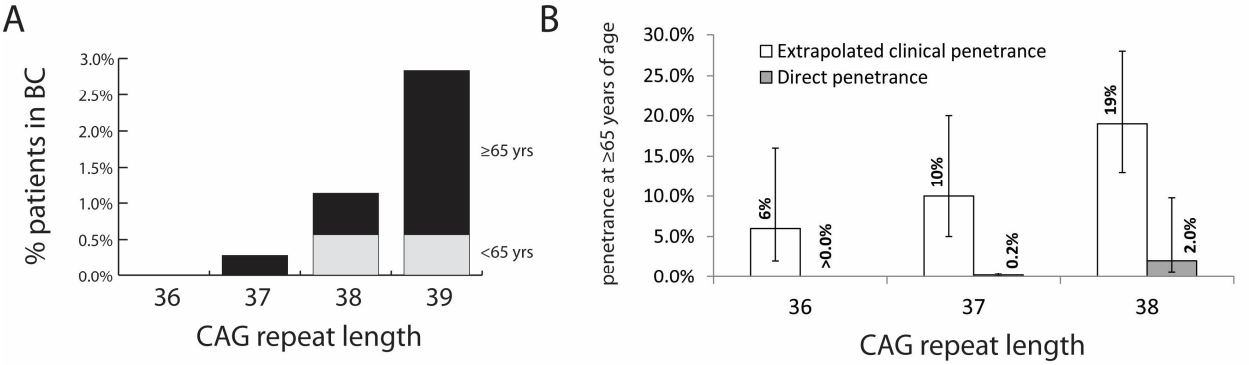


Figure 4-2. Estimated minimum penetrance of CAG repeats in the CAG 36-39 range at ≥65 years of age

(**a**) Frequency of alleles with reduced penetrance for HD in a multisource ascertainment of patients from British Columbia, Canada. Proportions of patients ≥65 years of age are shown in black. (**b**) Estimated age-dependent penetrance rates of CAG 36-38 alleles in the general population as compared to clinical penetrance rates. Extrapolated clinical penetrance rates by 65 years of age derive from previously reported survival analysis of clinical onset data (Langbehn et al., 2004). Direct penetrance rate estimates are calculated as the number of symptomatic patients ≥65 years of age in British Columbia in 2012 with each CAG repeat genotype divided by the number of individuals ≥65 years of age in British Columbia expected to have each expanded CAG repeat genotype.

CAG 37 is penetrant for HD in a minimum of 0.2% (95% CI: 0.1-0.4%) of individuals ≥65 years of age. CAG 38 is penetrant for HD in a minimum of 2.0% (95% CI: 0.6-9.8%) of individuals ≥65 years of age. No patient with a validated CAG 36 genotype was identified in our clinical prevalence cohort, precluding a penetrance estimate at this CAG length. However, one deceased patient with confirmed CAG 36 has been observed in the province since genetic testing began in 1995, and seven CAG 36 alleles were observed in our general population sample, suggesting that the penetrance of this allele for the HD phenotype is less than CAG 37 but greater than zero. No CAG 39

alleles were observed in our general population control sample, and thus no penetrance estimate of CAG 39 genotypes among individuals ≥65 years of age could be calculated. However, two-thirds (10/15) of genetically diagnosed BC patients with validated 36-39 CAG repeat lengths were CAG 39, suggesting that the penetrance of this allele for HD is greater than for CAG 38. Extrapolated CAG 36-38 penetrance rates for HD by 65 years of age, derived from survival analysis of a large, multicentre clinical cohort, exceed our direct penetrance estimates across the entire reduced penetrance allele class, at 6%, 10%, and 19% by age 65 for CAG 36, 37, and 38, respectively (**Figure 4-2b**) (Langbehn et al., 2004).

4.4 Discussion

Here we report an unexpectedly high frequency of HD-associated alleles in the reduced penetrance range of 36-39 CAG repeats. These data suggest that approximately 1 in 400 individuals have an expanded CAG repeat associated with HD in Western populations and that approximately 1 in 2500 have an HD allele in the full penetrance range.

Previous attempts to calculate the frequency of the HD mutation have relied upon indirect methods of extrapolating heterozygote frequency from the prevalence of diagnosed HD patients (Walker et al., 1981). Our direct estimates show that the heterozygote frequency of ≥36 CAG in the general population is up to ten-fold higher than indirect estimates, but that the majority of these are HD alleles of reduced

penetrance (CAG 36-39). The frequency of reduced penetrance alleles in the general population thus appears to be underestimated by clinical ascertainment, and individuals at risk of developing HD late in life may be more numerous than previously believed. Low clinical ascertainment of CAG 36-38 alleles is supported by an inferred ascertainment rate of ≤5% in mutational flow models (Falush et al., 2001). It is notable that the frequency of fully penetrant HD alleles in our sample (0.041%, 95% CI: 0.012-0.109%) is compatible with prior indirect estimates (Walker et al., 1981). For example, we estimate that approximately 1 in 2500 individuals has a full penetrance HD allele in the general population, or approximately 3.0 times the multisource prevalence rate for British Columbia (Fisher and Hayden, 2014).

Our data also suggest the penetrance of CAG 36-39 for the typical HD phenotype may be lower than estimated from clinical data sets. CAG-specific penetrance rates of HD alleles have been attempted using clinical cohorts of mutation-positive individuals both with and without clinical signs of HD (Langbehn et al., 2004; Quarrell et al., 2007), but these estimates are thought to be biased for symptomatic individuals, resulting in a corresponding bias toward higher penetrance rates (Langbehn et al., 2010). Analysis of clinical records of manifest and premanifest individuals with 36-39 CAG estimated a 60% chance of clinical diagnosis by 65 years of age (Quarrell et al., 2007). We show a much higher frequency of 36-39 CAG repeat alleles in our general population sample than are clinically ascertained at this rate. For example, 15.8% of the population of British Columbia was ≥65 years of age in 2012. Given a 0.205% prevalence rate of 36-39 CAG genotypes in the general population, approximately 1400 British Columbians

≥65 years of age would have had a 36-39 CAG allele in 2012, and at least 840 would have shown signs of HD given a 60% penetrance rate by the age of 65. In contrast, our comprehensive 2012 prevalence study of HD in British Columbia revealed only 15 affected patients with CAG 36-39, comprising only 4.25% of all verified genetic diagnoses of HD in the province, 13 of whom were ≥65 years of age (Fisher and Hayden, 2014). This modest proportion of 36-39 CAG alleles among genetically confirmed patients is comparable to clinical cohorts in Portugal, the Netherlands, and COHORT sites across the United States, Australia, and Canada (Maat-Kievit et al., 2001; Sequeiros et al.; Dorsey, 2012).

HD patients with expanded CAG repeats between 36 and 41 have the oldest average age of onset in clinical cohorts, with many patients manifesting motor symptoms over 60 years of age (Langbehn et al., 2004; Lee et al., 2012b). Late-onset HD may also be characterized by milder signs and slower progression which elude diagnosis as HD and be confounded by comorbid diseases of aging (James et al., 1994; Lipe and Bird, 2009). The absence of family history or overt motor features may further preclude clinicians from considering a diagnosis of HD in late-onset cases. It is therefore possible that many individuals with reduced penetrance alleles escape diagnosis or do not seek medical services for early signs of HD.

Given the high frequency of reduced penetrance alleles in our study, genetic testing for HD should be considered in those who present with suggestive but mild clinical features, particularly among the elderly. In premanifest individuals with ≥36 CAG

repeats, clinical diagnosis of HD is known to be preceded by a prodromal stage of subtle motor and behavioral signs associated with neuronal loss in the caudate nucleus (Ross et al., 2014). At present, only clinically ascertained individuals with ≥36 CAG repeats have been evaluated for progressive prodromal phenotypes preceding the onset of HD (Paulsen et al., 2014; Tabrizi et al., 2013). It remains unclear how many elderly asymptomatic individuals with CAG 36-39 genotypes show prodromal signs of HD, and whether such individuals would inevitably progress to HD later in life.

As sequencing technology improves and clinical whole genome sequencing becomes more widespread, premanifest individuals without a family history of HD could incidentally discover the presence of an HD-associated allele in their genome. In the absence of population-based penetrance estimates for incidentally ascertained HD alleles, particularly in the 36-39 CAG repeat range, it will be challenging to counsel these individuals about their chance of developing HD (Maat-Kievit et al., 2001). In this study, we estimate that 0.2% of individuals with CAG 37 and 2.0% of individuals with CAG 38 are symptomatic for HD among individuals ≥65 years of age. It is crucial to note that the number of individuals who manifest HD with a reduced penetrance allele will change with age structure, both within the ≥65 years of age category and in the population as a whole. Our penetrance estimate should therefore be considered in the context of the demographic structure of a specific population in a specific year. Additionally, our penetrance estimates only consider patients with a CAG 36-38 repeat allele confirmed within British Columbia. Some patients in our prevalence cohort with strictly clinical diagnoses or out-of-province genetic tests may have CAG 36-38 but are

not included in our penetrance estimates. Our data nonetheless show a bias toward higher penetrance estimates derived from clinical onset data, and indicate that the age-specific penetrance in people with ≥36 CAG repeat alleles for HD, particularly alleles CAG 36-38, may be lower than previously reported.

The frequency of intermediate alleles reported in this study is comparable to smaller samples (Sequeiros et al., 2010; Semaka et al., 2013b). It is established that rates of paternal germline CAG repeat instability increase across the intermediate allele range, leading to a higher risk of expansion to ≥36 CAG repeat alleles in offspring (Semaka et al., 2013a; Semaka et al., 2006). The unexpectedly high frequency of individuals with CAG 36-37 in this study suggests that fathers with reduced penetrance alleles, while at relatively low risk of developing HD themselves, may play a larger role in transmitting full penetrance HD alleles to the next generation than previously appreciated. A large proportion of HD patients with full penetrance alleles but no family history may have inherited the CAG expansion from a father with CAG 36-37 rather than from a father with an IA.

It is unclear why some individuals bearing HD alleles with reduced penetrance (36-39 CAG) manifest the classic HD phenotype as early as midlife, while others remain entirely asymptomatic through advanced ages (Quarrell et al., 2007). Age of onset can vary widely between patients with the same expanded CAG repeat allele (Wexler et al., 2004; Lee et al., 2012b). The comparatively high frequency of reduced penetrance alleles in the general population suggests that additional genetic and environmental

factors may modify the likelihood of manifesting HD from these alleles within a normal human lifespan (Lee et al., 2015a). Given the small proportion of individuals with CAG 36-38 repeat alleles who manifest with clinical HD, genetic background and environmental factors may play major roles in the likelihood of developing the disease in patients with reduced penetrance genotypes.

We have shown an unexpectedly high frequency of HD alleles with reduced penetrance (CAG 36-39) in the general population. The manifestation of HD in individuals with CAG repeat alleles in the reduced penetrance range is likely low, but prodromal disease phenotypes may be underascertained among the elderly. Individuals with reduced penetrance alleles may also have an increased role in the expansion of CAG repeats into full penetrance alleles in their offspring.

Table 4-1. HD-associated and intermediate allele frequency in the general population

			Intermediate Alleles (CAG 27-35)			HD-Associated Alleles (≥36)		
	Individuals	Mean		*alleles*	*individuals*		*alleles*	*individuals*
Population	**(Alleles)**	**CAG**	**n**	**% (95% CI)**	**% (95% CI)**	**n**	**% (95% CI)**	**% (95% CI)**
United States	3906 (7812)	18.44	233	2.98% (2.62-3.37%)	5.91% (5.21-6.69%)	8	0.102% (0.048-0.193%)	0.205% (0.097-0.386%)
Scotland, United Kingdom	1815 (3630)	18.71	132	3.63% (3.06-4.28%)	7.16% (6.04-8.42%)	4	0.110% (0.037-0.262%)	0.220% (0.074-0.523%)
British Columbia, Canada	1594 (3188)	18.45	93	2.92% (2.38-3.55%)	5.77% (4.71-7.00%)	6	0.188% (0.079-0.388%)	0.376% (0.157-0.774%)
Total	**7315 (14630)**	**18.51**	**458**	**3.13% (2.86-3.42%)**	**6.20% (5.66-6.76%)**	**18**	**0.123% (0.076-0.190%)**	**0.246% (0.151-0.380%)**

Table 4-2. Reduced penetrance and full penetrance HD allele frequency in the general population

		Reduced Penetrance (CAG 36-39)			Full Penetrance (≥40 CAG)		
	Individuals		*alleles*	*individuals*		*alleles*	*individuals*
Population	**(Alleles)**	**n**	**% (95% CI)**	**% (95% CI)**	**n**	**% (95% CI)**	**% (95% CI)**
United States	3906 (7812)	6	0.077% (0.032-0.158%)	0.154% (0.064-0.316%)	2	0.026% (0.005-0.082%)	0.051% (0.011-0.164%)
Scotland, United Kingdom	1815 (3630)	3	0.083% (0.023-0.220%)	0.165% (0.047-0.441%)	1	0.028% (0.003-0.129%)	0.055% (0.006-0.257%)
British Columbia, Canada	1594 (3188)	6	0.188% (0.060-0.165%)	0.376% (0.157-0.775%)	0	0.000% (0.000-0.060%)	0.000% (0.000-0.120%)
Total	**7315 (14630)**	**15**	**0.103% (0.060-0.165%)**	**0.205% (0.120-0.330%)**	**3**	**0.021% (0.006-0.055%)**	**0.041% (0.012-0.109%)**

Chapter 5: The Variable Molecular Epidemiology of Huntington Disease is Related to Intermediate Allele Frequency and Haplotype in the General Population

5.1 Synopsis

Huntington disease (HD) is an autosomal dominant disorder caused by an expanded CAG repeat in the huntingtin gene (*HTT*). Prevalence studies incorporating both genetic and clinical diagnostic standards show that up to one in 7300 people are affected with HD in Western populations (13.7 per 100,000) (Fisher and Hayden, 2014). Strikingly, HD manifests at less than one-twentieth this rate in East Asian populations from Japan, Taiwan, and Hong Kong (0.1-0.7 per 100,000) (Adachi and Nakashima, 1999; Chen and Lai, 2010; Chang et al., 1994; Kay et al., 2014a). In South Africa, black people also present with HD at approximately one-tenth the rate in the white subpopulation, and approximately one-third the rate in the mixed ancestry subpopulation (Baine et al., 2016). In British Columbia, Canada, HD is much more prevalent among individuals of European descent (17.2 per 100,000) than in the ethnically diverse remainder of the population (2.1 per 100,000) (Fisher and Hayden, 2014). Taken together, these epidemiological observations suggest that major population differences in the prevalence of HD are attributable to ethnic ancestry and are principally genetic in nature.

Differences in the prevalence of HD by ancestry may be related to the new mutation rate for HD. New mutations occur as a result of CAG repeat expansion, usually in the paternal germline, resulting in a disease-associated genotype of 36 CAG repeats or more in offspring (Goldberg et al., 1993; Duyao et al., 1993). Intermediate alleles (IAs), defined as 27-35 CAG repeats, have the potential for germline expansion into the HD range, with risk of expansion increasing across the IA repeat range (Semaka et al., 2006; Semaka et al., 2013a). IA frequency has been estimated from 2.9 to 3.4% of alleles in representative population samples of European ancestry, with an IA genotype in as many as 1 in 15 individuals (Semaka et al., 2013b; Sequeiros et al., 2010; Ramos et al., 2012; Kay et al., 2016a). The dramatically lower prevalence of HD in non-European populations may result from a lower IA frequency and lower new mutation rate in these ancestry groups, with fewer CAG expansion events and a consequent lower incidence of the disease (Andrew and Hayden, 1995). However, the frequency of IAs has not been assessed in non-European ancestry groups with lower prevalence rates of HD, and thus no comparative estimates of IA frequency or the HD new mutation rate are currently available.

In European ancestry populations where the prevalence of HD is highest, expanded CAG repeats causative of HD are found preferentially on specific gene-spanning *HTT* haplotypes (A1, A2, and A3a) (Kay et al., 2015). IAs and high normal CAG repeat alleles (20-26 CAG) are similarly enriched for these haplotypes among all normal CAG repeat alleles, suggesting that high-normal CAG repeats expand into the IA and HD repeat ranges (Warby et al., 2009; Rubinsztein et al., 1995; Almqvist et al., 1995). In

East Asian and black South African populations where the prevalence of HD is low, these European ancestry HD-associated haplotypes are entirely absent, high normal CAG repeats are less frequent, and the HD mutation instead occurs on a diverse assortment of ancestry-specific haplotypes (Warby et al., 2011; Baine et al., 2013). It remains unclear whether *HTT* haplotypes of IAs are reflective of ancestry-specific HD haplotypes in non-European populations. In admixed ancestry patient groups with suspected European origins of the mutation, such as the mixed ancestry subpopulation of South Africa, the HD mutation may result from a comparatively recent HD mutation founder or from numerous European ancestry IAs that expanded after admixture (Scholefield and Greenberg, 2007). In the latter case, IA haplotypes in the mixed ancestry South African population would be expected to occur on haplotypes of European ancestry. Evaluation of IA frequency and haplotype in distinct ethnic groups may therefore clarify the genetic origin of HD in various non-European ancestry populations.

In this comparative genetic study, we genotype the CAG repeat in general population chromosomes from a wide range of ancestry groups in North America, Europe, South America, Africa, and East Asia. We characterize the CAG repeat distribution, IA frequency, and HD new mutation rate in a series of populations where recent molecular prevalence estimates are available for comparison. For the first time, we estimate IA frequency among individuals of African, East Asian, and admixed European ancestry. We show that population differences in disease prevalence are associated with both IA frequency and the inferred HD new mutation rate. We further show that the prevalence

of HD correlates with the proportion of IAs on the A1 haplotype in different European and admixed European ancestry populations. Our study thus provides an empirical framework for understanding the molecular epidemiology of HD worldwide, showing that both IA frequency and the HD new mutation rate are closely linked to epidemiological differences in disease prevalence both within and between major ancestry groups.

5.2 Material and Methods

5.2.1 Study Populations

Saliva-extracted DNA samples from North Americans with self-declared ethnicity were obtained from the Coriell Personalized Medicine Collaborative (CPMC) Community Cohort (Keller et al., 2010). Participants were not recruited on the presence or absence of any disease phenotype. 3907 individuals were genotyped from the CPMC cohort as previously described (Kay et al., 2016a), of whom 3221 provided data on self-declared ethnicity. Individuals with self-declared ethnicity as Hispanic American, African American, or East Asian were used for subpopulation analysis of CAG repeat length by ancestry. For analysis of the North American general population, 1592 previously genotyped samples from the general population of British Columbia, Canada were added to genotyped samples from the CPMC Community Cohort, yielding a total of 10,998 alleles from 5499 individuals. General population samples from the North of Scotland were obtained from a population-based DNA sample set held by NHS Grampian and the University of Aberdeen. 1815 individuals were genotyped from this

DNA sampling as previously described (Kay et al., 2016a). DNA was collected in the 1990s and represents individuals of predominantly European ancestry from the North of Scotland. Population-specific European chromosomes from Norway, Sweden, Denmark, and Finland were identified from the UBC HD Biobank from cohorts originally recruited for mapping studies of the HD mutation. South African DNA samples were obtained from the Department of Pathology, University of Cape Town and the School of Pathology, University of the Witwatersrand. These samples were sourced and stored from routine referrals through various research collaborating clinics, and were used with appropriate authorization. All general population individuals consented for their sample to be used for research purposes. In this study, the term ‘mixed ancestry’ is used to describe a dynamic, historically distinct group of South Africans with significant genetic contributions from Khoisan- and Bantu-speaking peoples in addition to smaller contributions from Europe and Asia. The term ‘black South African’ refers to Bantu speaking peoples of South Africa with no reported European or Asian ancestry, including individuals from Tswana, Xhosa, and Venda ethnic groups. General population samples from East and West African individuals were obtained from a diverse DNA bank of African individuals held at the University of Pennsylvania. Ethnic grouping of African samples into West and East African populations was confirmed by principal component analysis (PCA) using representative genome-wide markers in the ADME Core Panel (Illumina, San Diego, CA) as previously described (Kay et al., 2016b). Samples from individuals of Peruvian Amerindian ethnicity were originally collected as part of a genetic study of Parkinson disease, and granted proper authorization for use in further studies related to neurodegenerative diseases.

5.2.2 CAG and CCG Repeat Genotyping

All samples in this study were sized by an identical assay at the Centre for Molecular Medicine and Therapeutics at the University of British Columbia in Vancouver, Canada. CAG and CCG repeat sizes were determined as previously described using fluorescently labelled primers flanking the CAG repeat (HD344F, 5'-HEX-CCTTCGAGTCCCTCAAGTCCTTC-3' and HD450R, 5'-GGCGGCGGTGGCGGCTGTTG-3') or the adjacent CCG repeat (HD419F, 5'-AGCAGCAGCAGCAACAGCC-3' and HD482R, 5'-6FAM-GGCTGAGGAAGCTGAGGAG-3') (Semaka et al., 2013b). A third PCR encompassing both CAG and CCG sequences (HD344F, 5'-HEX-CCTTCGAGTCCCTCAAGTCCTTC-3' and HD482R, 5'-GGCTGAGGAAGCTGAGGAG-3') was used to exclude allelic dropout and to phase CAG and CCG sizes from the first two sizing assays. Sizing was performed relative to a control panel of sequenced CAG and CCG repeat lengths.

5.2.3 Statistical Analysis of CAG Repeat Allele Distributions

Comparative analysis of non-Gaussian distributions, such as observed for CAG repeat alleles in a population, requires the use of non-parametric statistics. Non-parametric 95% confidence intervals for mean CAG repeat length were calculated by bootstrap resampling using the *boot* function of package *boot* in the R statistical computing environment. Bayesian binomial 95% confidence intervals for IA frequency were

calculated using the *binom.bayes* function of package *binom*. Non-parametric comparisons of mean CAG repeat length between populations were approximated by one-way Kruskal-Wallis Tests with posthoc Nemenyi Tests for pairwise multiple comparisons of mean rank sums, using the *posthoc.kruskal.nemenyi.test* function in package *PMCMR*. Comparisons of IA frequency between populations were performed by two-sided Fisher's Exact Test or Chi-Square Test where applicable.

5.2.4 *HTT* Haplotyping

General population individuals with at least one repeat allele of 27-35 CAG were genotyped at 90 SNPs across *HTT* as previously described (Kay et al., 2015). Reconstructed haplotypes were phased to CAG repeat length by haplotype-specific CCG and CAG repeat length associations (Semaka et al., 2013b; Kay et al., 2015) or by haplotype homozygosity.

5.2.5 Estimation of the HD New Mutation Rate

For each sample group, the general population frequency of each intermediate allele (CAG 27, 28, 29, etc.) was multiplied by the corresponding male germline expansion rate of that CAG repeat length into the HD range (Semaka et al., 2013a; Kay et al., 2015), yielding a population rate of paternal IA expansions to the HD range for each specific IA CAG repeat length. The population rate of paternal IA expansion to the HD range for each specific IA CAG repeat length was then summed across the IA range

(CAG 27-35), yielding an estimated total population rate of paternal expansions from IAs to the HD range (≥36 CAG) in each population. Paternal germline transmissions are expected to occur randomly among different CAG repeat alleles in the population and maternal IA expansions into the HD range are extremely rare (Semaka et al., 2014). The total rate of paternal IA expansions to the HD range therefore approximates the absolute rate of HD new mutations in the population. Empirical estimates of the HD new mutation rate assume no IA instability differences by haplotype or ethnic background.

5.3 Results

5.3.1 Mean CAG Repeat Length and IA Frequency Correlate with the Prevalence of HD

To evaluate population differences in CAG repeat distribution and IA frequency, we genotyped the CAG repeat in general population controls from 14 ethnically or geographically distinct groups worldwide (**Table 5-1**). Mean CAG repeat length was highest in self-identified Hispanic individuals from the United States (19.03, 95% CI 18.42-19.65) and lowest in black South Africans (16.90, 95% CI 16.76-17.03), and the ranked means of these two populations were the most significantly different among all pairwise population comparisons (p=9.98E-14, **Appendix C, Table C-1**). Mean CAG repeat length was higher in the North Scottish and North Americans than in mixed ancestry South African, black South African, and East Asian ancestry groups. Mean

CAG repeat length consistently exceeded 18.0 in all European ancestry groups except for the Finnish population. Average CAG length was consistently below 17.5 in all black African and East Asian ancestry groups.

Table 5-1. Average CAG repeat length in general population alleles from 14 ethnically distinct populations

	Alleles	IA Frequency			Mean CAG	
Population	**n**	**n**	**%**	**(95% CI)**	**mean**	**(95% CI)**
Hispanic American	160	9	5.6%	(2.8-10.0%)	19.03	(18.42-19.65)
North Scottish	3634	135	3.7%	(3.1-4.4%)	18.71	(18.60-18.83)
White South African	436	16	3.7%	(2.2-5.8%)	18.48	(18.18-18.77)
North American	10998	324	2.9%	(2.6-3.3%)	18.44	(18.38-18.50)
Swedish	108	2	1.9%	(0.4-5.8%)	18.43	(17.84-19.01)
Danish	142	4	2.8%	(1.0-6.6%)	18.25	(17.72-18.80)
Finnish	127	1	0.8%	(0.1-3.6%)	17.96	(17.51-18.42)
African American	224	4	1.8%	(0.6-4.2%)	17.72	(17.29-18.16)
Mixed South African	1147	21	1.8%	(1.2-2.7%)	17.60	(17.43-17.78)
East Asian	261	0	0.0%	(0.0-0.7%)	17.48	(17.29-17.68)
Black West African	180	2	1.1%	(0.2-3.5%)	17.46	(17.02-17.90)
Black East African	362	8	2.2%	(1.1-4.1%)	17.36	(17.05-17.69)
Amerindian	86	1	1.2%	(0.1-5.3%)	16.97	(16.37-17.53)
Black South African	1575	22	1.4%	(0.9-2.1%)	16.90	(16.76-17.03)

IAs were observed in all population groups except for East Asians. IAs are significantly more frequent in North Scottish and North Americans than in mixed ancestry South Africans (p=0.0011 and p=0.0272, Fisher's Exact), black South Africans (p=0.0001 and p=0.0001, Fisher's Exact), and East Asians (p=0.0001 and p=0.0009, Fisher's Exact), mirroring the higher prevalence rate of HD in populations of predominant European ancestry. Interestingly, there are more IAs among North Scottish than among North Americans (p=0.0473, Chi-square). Both mean CAG repeat length and IA frequency are lower in Finnish controls than in all other European ancestry groups, in agreement with the reported lower prevalence of HD in Finnish people relative to other European ancestry populations (Sipila et al., 2014), although these comparisons do not achieve statistical significance.

HD prevalence estimates within the past five years are available in the North Scottish, North American, Finnish, white South African, mixed ancestry South African, black South African, and East Asian populations (**Table 5-2**). The prevalence of HD was plotted in relation to mean CAG repeat length and IA frequency in each of these seven population samples (**Figure 5-1**). Prevalence and mean CAG repeat length data were additionally available from the Taiwanese population and included for comparison (Wang et al., 2004; Chen and Lai, 2010). No matched data sets of HD prevalence and mean CAG repeat length or IA frequency were available from other populations reported in the literature. Across populations, mean CAG repeat length and IA frequency correlate with the reported prevalence rate of HD in each group.

Table 5-2. Intermediate allele frequency and reported prevalence of HD by population

Population	Total Alleles n	Intermediate Alleles n	Allele Frequency	(95% CI)	Carriers (1 in x)	Population Prevalence (per 100,000)	Reference
North Scottish	3634	135	3.7%	(3.1-4.4%)	13	16.10	(Evans et al., 2013)
White South African	436	16	3.7%	(2.2-5.8%)	14	8.27	(Baine et al., 2016)
North American	10998	324	2.9%	(2.6-3.3%)	17	13.70	(Fisher and Hayden, 2014)
Mixed Ancestry African	1147	21	1.8%	(1.2-2.7%)	27	2.42	(Baine et al., 2016)
Black South African	1575	22	1.4%	(0.9-2.1%)	36	0.75	(Baine et al., 2016)
Finnish	127	1	0.8%	(0.1-3.6%)	64	2.12	(Sipila et al., 2014)
East Asian	261	0	0.0%	(0.0-0.7%)	0	0.80	(Kay et al., 2014a)

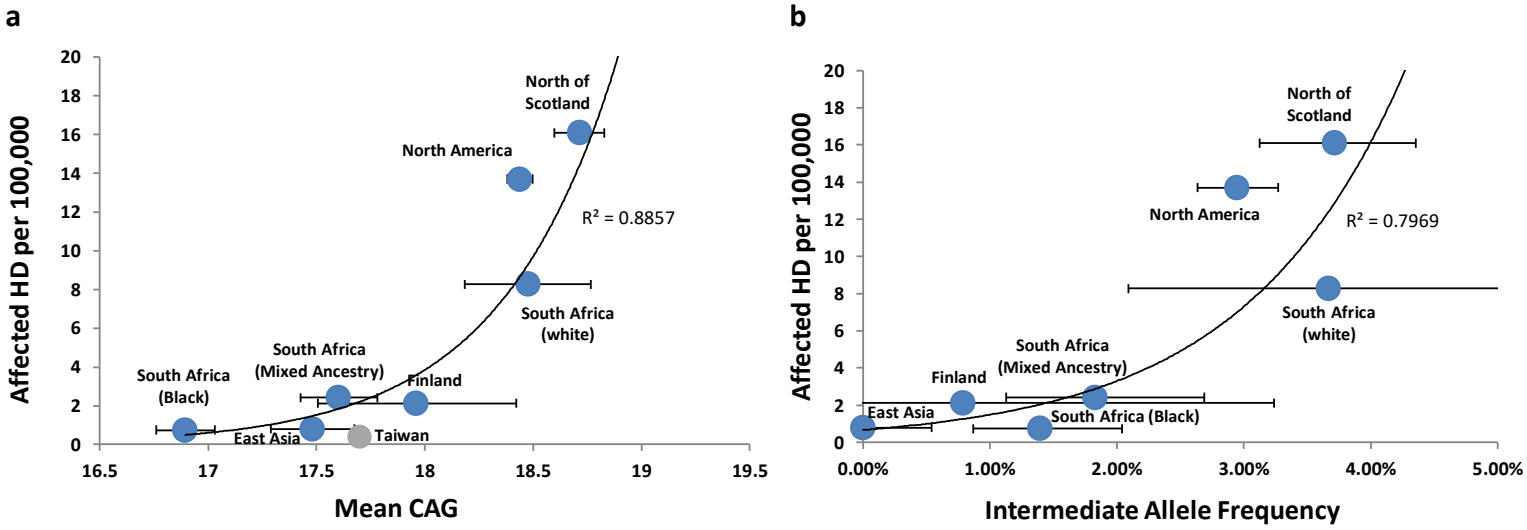


Figure 5-1. Mean CAG repeat length and IA frequency correlate with prevalence of HD by population

5.3.2 High Mean CAG Repeat Length is Attributable to HD-Associated Haplotypes in European Ancestry Populations

Dense *HTT* haplotyping of patients and relatives in the UBC HD Biobank allowed for detailed examination of the distribution of CAG repeat lengths by haplotype in normal chromosomes. In our prior investigation of four European ancestry patient cohorts, we identified 20 common gene-spanning *HTT* haplotypes constituting >95% of all normal chromosomes (Kay et al., 2015). Four of these haplotypes – A1, A2a, A2b, and A3a – were found to represent >90% of HD mutation haplotypes in decreasing order of frequency among disease chromosomes. Among 804 unrelated normal chromosomes haplotyped from individuals in the UBC HD Biobank and analyzed for this study, the A1, A2a, A2b, and A3a haplotypes represent the highest mean CAG repeat lengths among all 20 common *HTT* haplotypes (**Figure 5-2; Appendix C, Figure C-1**). Strikingly, mean CAG repeat length of each HD-associated haplotype decreases in the same order of association with the HD mutation. For example, A1 has a mean CAG repeat length of 23.6 CAG (**Figure 5-2b**), A2a has mean CAG repeat length of 22.2 (**Figure 5-2c**), and A2b has mean CAG repeat length of 21.2 (**Figure 5-2d**). Related haplotypes A4, A5a, and A5b have mean CAG repeat lengths of 16.95, 17.07, and 14.92 among normal chromosomes (**Figure 5-2e; Appendix C, Figure C-1**), respectively, and are never found on chromosomes with the HD mutation despite sharing common variant markers of haplogroup A with HD-associated haplotypes A1, A3, and A3a. The most frequent normal haplotype in the population, C1, has an average CAG length of 17.60 and represents 29.8% of unrelated normal chromosomes in the UBC HD Biobank, but is

found on only 3.2% of HD chromosomes (Kay et al., 2015). Thus mean CAG repeat length of each *HTT* haplotype among normal chromosomes is closely linked to the frequency of that haplotype among HD chromosomes. The HD mutation never or rarely occurs on haplotypes with low average CAG repeats in the normal population, but commonly occurs on haplotypes with the highest average CAG repeat lengths.

The mean CAG repeat length of all 804 unrelated normal *HTT* haplotypes in the UBC HD Biobank is 18.70, comparable to other European ancestry general population samples reported here and in previous studies (**Table 5-1; Table 1-2**). When HD-associated haplotypes (A1, A2, and A3a) are excluded from this pool of normal chromosomes, mean CAG repeat length declines to 17.37, comparable to mean CAG repeat length in East Asian and black African populations where the prevalence of HD is much lower (**Table 5-2**). These data suggest that longer mean CAG repeat length and higher IA frequency in European ancestry populations relative to black Africans and East Asians are attributable to gene-spanning haplotypes with high normal CAG repeat length in the European ancestry general population (A1, A2, and A3a).

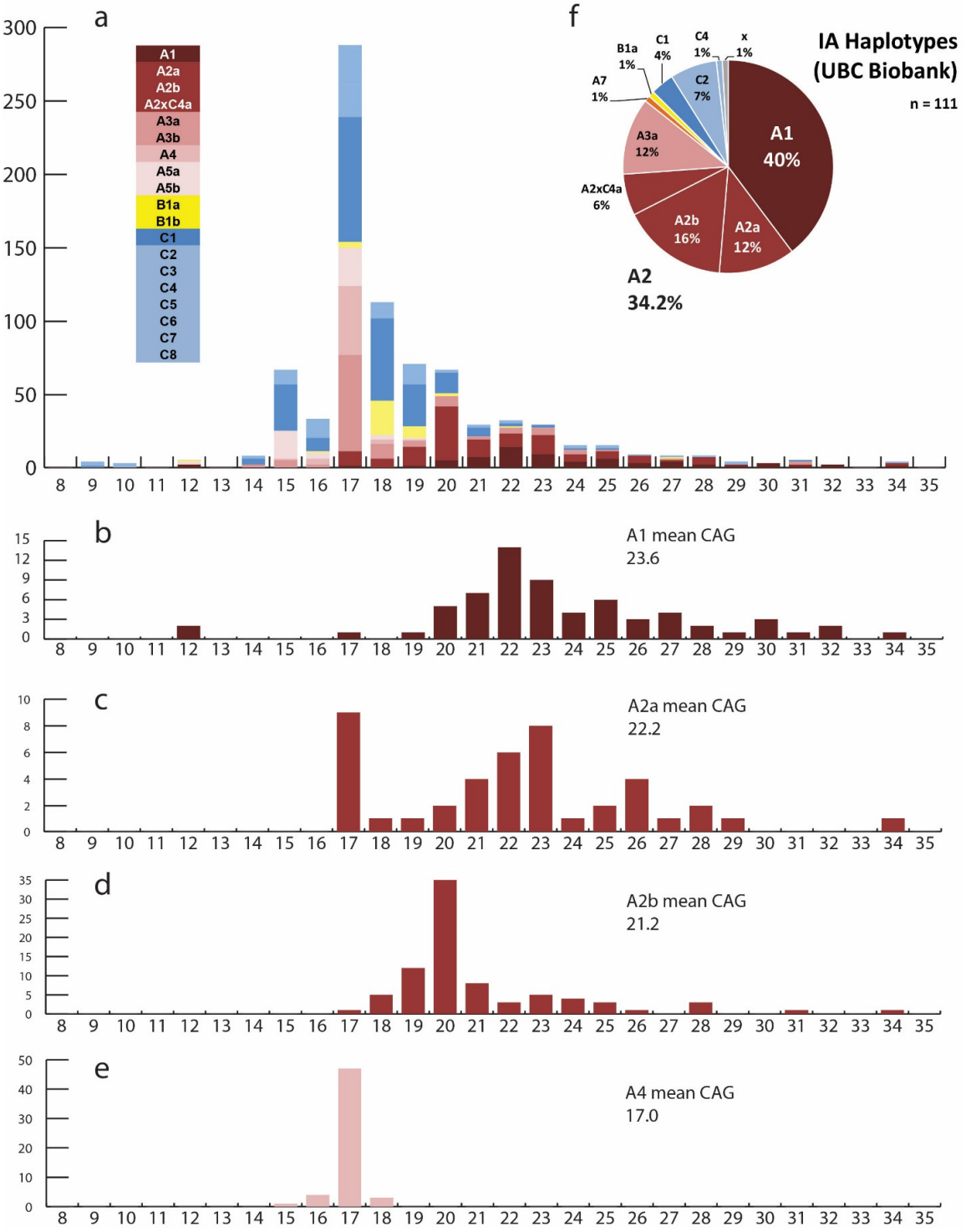


Figure 5-2. Distributions of CAG repeat alleles by *HTT* haplotype

(a) Distribution of CAG repeat alleles by *HTT* haplotype among control chromosomes in the UBC HD Biobank. (b) Distribution of CAG repeat lengths among A1 haplotype controls. (c) Distribution of CAG

repeat lengths among A2a haplotype controls. (d) Distribution of CAG repeat lengths among A2b haplotype controls. (e) Distribution of CAG repeat lengths among A4 haplotype controls. (f) *HTT* haplotype distribution of IAs from the UBC HD Biobank.

5.3.3 IAs Occur on Ancestry-Specific *HTT* Haplotypes in Ethnically Distinct Populations

Haplotypes of IAs in the UBC HD Biobank reflect our previously reported haplotypes of the HD mutation in the Canadian population (**Figure 5-2f**) (Kay et al., 2015). A1 is the most common IA haplotype (40%), followed by A2 variant haplotypes (34.2%) and the A3a haplotype (12%). To compare IA haplotypes from the general population of different ethnic groups, we genotyped IAs drawn from our general population samples of North Scottish, North Americans in the United States and Canada, Hispanic Americans in the United States, mixed ancestry South Africans, and black South Africans (**Figure 5-3**). IAs from the North of Scotland occurred more frequently on A1 than in North Americans (p= 0.0072, Fisher's Exact), reflecting the higher mean CAG repeat length and IA frequency observed in North Scottish individuals compared to other population groups (**Table 5-1**). A1 was also the most frequent IA haplotype among Hispanic Americans. In contrast, European HD-associated haplotypes A1 and A2 are uncommon among mixed ancestry South African IAs and entirely absent among black South African IAs. IA haplotypes in the black South African subpopulation resemble our previously reported distribution of HD haplotypes in this subpopulation, with all haplotypes conforming to African ancestry (Baine et al., 2013). Haplotypes in mixed

ancestry South African IAs are predominantly of African origin, whereas African ancestry haplotypes are uncommon among our previously reported mixed ancestry South African HD chromosomes (Baine et al., 2013). More specifically, 79% of mixed ancestry South African HD haplotypes were previously found to occur on European ancestry A1, A2, or A3a haplotypes, in comparison to only 29% of mixed ancestry South African IAs in this study. The HD-associated A3a haplotype observed in IAs from European ancestry populations is completely absent among IAs from mixed ancestry and black South Africans. No IAs were observed among all 261 East Asian control chromosomes examined in this study, precluding haplotype analysis of IAs in this population.

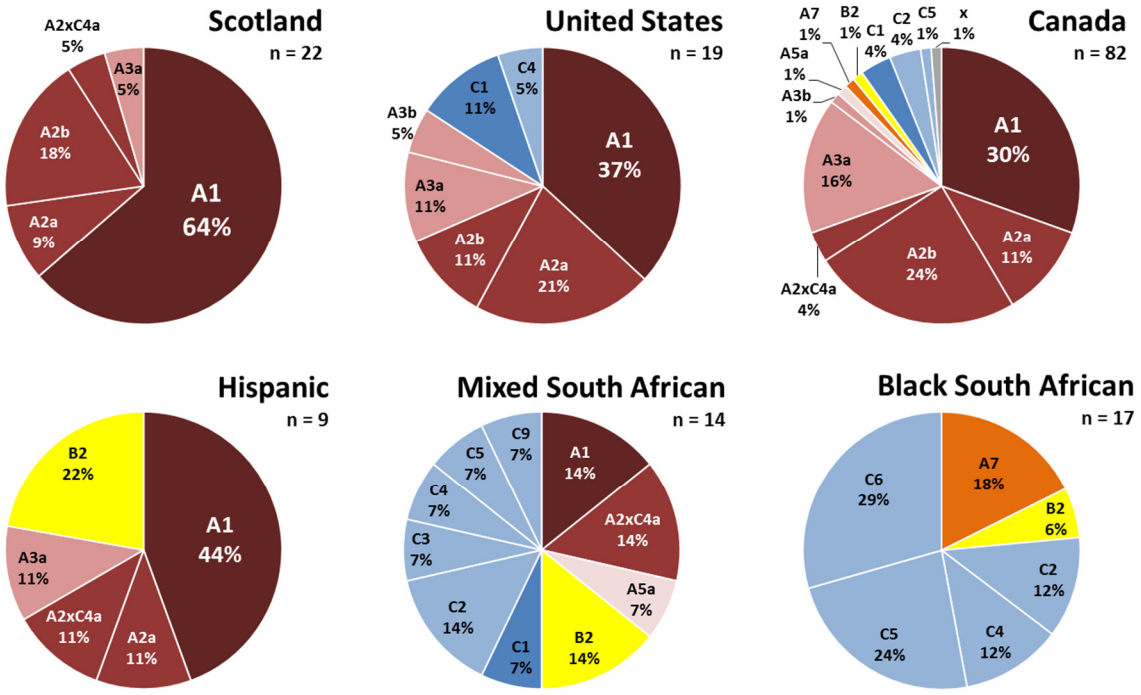


Figure 5-3. General population IA haplotypes from ethnically distinct populations

IA haplotypes in the North Scottish, American, Canadian, Hispanic American, mixed ancestry South African, and black South African general population are ancestry-specific. A2xC4a represents an intragenic recombinant haplotype of A2 and C4a, with the CAG repeat located within the A2 segment.

5.3.4 The HD New Mutation Rate Correlates with the Manifest Prevalence Rate of HD

We recently reported empirical CAG repeat length-specific IA expansion rates in the male germline, enabling evidence-based standards of IA expansion risk in genetic counselling practice (Semaka et al., 2013a; Semaka and Hayden, 2013). These male germline instability rates are comparable to paternal transmission instability rates observed in the UBC HD Biobank at similar CAG repeat lengths (**Appendix C, Table C-**

2). Using CAG-specific IA frequencies from the largest population samples in this study in combination with our reported CAG-specific IA expansion rates, we can estimate the molecular HD new mutation rate in each population. We calculated a molecular HD new mutation rate among North Scottish, Americans, Canadians, mixed ancestry South Africans, and black South Africans using the product of CAG-specific IA frequencies and germline HD expansion risk at each specific IA length (**Figure 5-4**). The comparatively high IA frequency among North Scottish (3.71%) results in a correspondingly high inferred new mutation rate of 0.0307% HD expansion events among all paternal transmissions, or 1 new HD mutation per 3257 births. Black South Africans and East Asians have the lowest IA frequency and therefore the lowest inferred new mutation rates, estimated to be less than one new HD expansion per 25,000 births. As maternal transmissions of *de novo* HD mutations are rare (Semaka et al., 2014), these figures approximate the overall HD new mutation rate in each population.
It is notable that the proportion of IAs on the HD-associated A1 haplotype, which has the highest average CAG repeat length of all haplotypes in the general population, is associated with the new mutation rate in each population (**Figure 5-4**). This suggests that populations with a higher frequency of the A1 haplotype in the general population and among IA chromosomes may likewise have a higher new mutation rate for HD.

Our recently reported general population frequency of the HD mutation, when compared to the molecular HD new mutation rate, allows for approximation of the proportion of individuals with a disease genotype (≥36 CAG) resulting from a transmitted new mutation (Kay et al., 2016a). In the Western general population, 0.246% of individuals,

or approximately 1 in 400, have an HD mutation genotype of ≥36 CAG repeats. Using CAG-specific IA frequencies from our combined general population CAG repeat distributions in the North of Scotland, the United States, and Canada, we estimate that the overall HD new mutation rate in the Western general population is 0.01745%, implying that 1 in 5372 births is expected to result in a paternally transmitted new mutation for HD. Given that 1 in 400 individuals in the general population have an HD mutation genotype, and 1 in 5372 individuals have a paternally transmitted new mutation for HD, it is therefore expected that approximately 7.1% of individuals with a disease genotype inherited the HD mutation *de novo* from an unaffected father with an IA genotype.

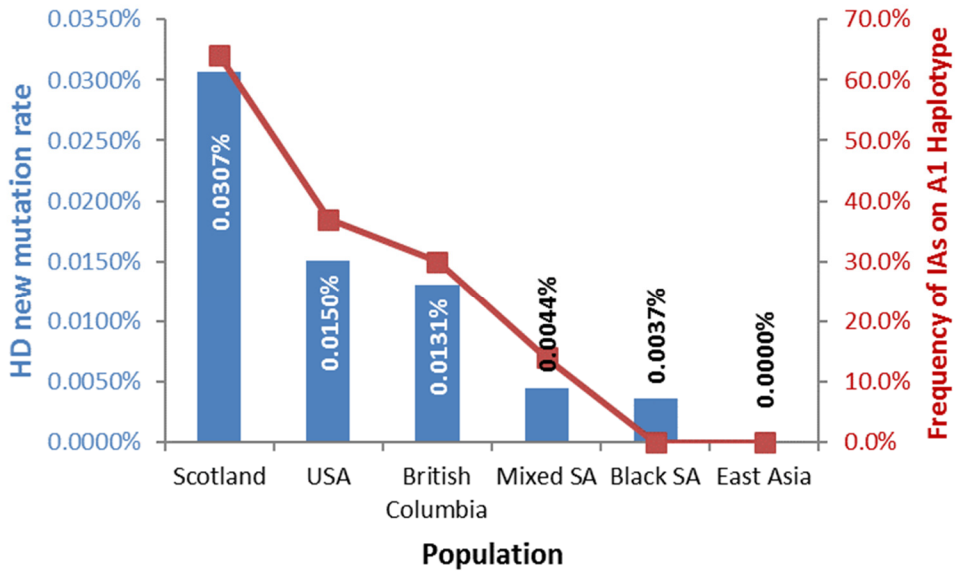


Figure 5-4. The inferred new mutation rate for HD in different populations

Population-specific new mutation rates for HD, as a percentage of expansions into the HD range among all paternal transmissions. The new mutation rate for HD is associated with the proportion of IAs on the A1 haplotype in each population.

5.4 Discussion

This is the first study to quantitatively compare CAG repeat allele data from distinct populations with corresponding prevalence rates of HD. Haplotype studies in populations of European ancestry have previously linked high normal and intermediate CAG repeat alleles with expanded HD alleles, suggesting that HD mutations arise from stepwise expansion of CAG repeat alleles in the population (Andrew and Hayden, 1995; Warby et al., 2009). In this context, differing prevalence rates of HD may be a long-term consequence of small differences in CAG repeat allele frequencies between populations, wherein high normal CAG repeat alleles expand over many generations into IAs and disease-causing HD mutations. Populations with higher prevalence rates of HD may therefore be expected to have higher mean CAG repeat lengths and higher IA frequencies. Our data demonstrate this relationship across a range of global populations. We show that populations with a higher prevalence of HD have (1) higher mean CAG repeat length among normal chromosomes, (2) higher IA frequency, and (3) a higher inferred HD new mutation rate. High mean CAG repeat length in European ancestry populations is directly attributable to HD-associated A1, A2, and A3a haplotypes with normal repeats exceeding the population average. These haplotypes are absent in East Asian and black African populations where average CAG repeat length and the prevalence of HD are correspondingly lower.

Among HD-associated haplotypes in European ancestry populations, the A1 haplotype has the highest mean CAG repeat length and the strongest association with the HD

mutation (Kay et al., 2015). We show that the proportion of IAs on the A1 haplotype is associated with IA frequency and the HD new mutation rate, suggesting that a higher frequency of the A1 haplotype in the general population may result in a higher prevalence of the disease. The A1 haplotype is found in South Asian individuals (Kay et al., 2015) and a distinct A1 variant haplotype is found at high frequency in admixed Latin American populations (Kay et al., 2016b), suggesting that *de novo* HD mutations may occur in these populations at a similar rate to populations of European ancestry. It is notable that the highest mean CAG repeat length across all 15 populations examined in this study was in Hispanic Americans (19.03), nearly identical to that observed in normal chromosomes genotyped in a clinical cohort from Mexico (19.04) (Alonso et al., 2009). Hispanic Americans in our study were also found to have the highest IA frequency of all evaluated groups (5.6%). These data suggest that the genetic prevalence rate of HD in admixed populations of Latin America may in fact exceed the prevalence rate in populations of predominant European ancestry. To date, population-based prevalence estimates of HD in Latin America are unavailable (Castilhos et al., 2015). Our comparative genetic investigation suggests that the prevalence of HD in Latin America may be high relative to other populations, and that accurate estimation of the prevalence of HD in admixed populations of Latin America is needed.

Among European ancestry populations, the lowest reported molecular prevalence rate is in the Finnish population (Sipila et al., 2014). This is thought to result from the genetic isolation of the Finnish people relative to other European populations, leading to divergent disease allele frequencies (Kere, 2001). As previously reported, our data

shows lower mean CAG repeat length among Finnish general population controls relative to other European ancestry populations, suggesting that a lower IA frequency and lower HD new mutation rate may account for the lower prevalence of HD in Finnish people.

Our study presents the first assessment of IA frequency and haplotype in ethnically distinct populations of African and admixed African ancestry. Comparison of IA haplotypes from mixed ancestry and black South Africans with our previously reported HD haplotypes from these populations offers novel insights to the origin of HD in different South African ethnic groups (Baine et al., 2013). The HD mutation in mixed ancestry South Africans occurs mostly on European ancestry A1 and A2 haplotypes, thought to originate from Dutch ancestry founders, with a small proportion of mutations occurring on African-ancestry haplotypes. In contrast, we show that mixed ancestry South African IAs are found predominantly on haplotypes of African origin. Taken together, these data support a European founder effect for the HD mutation in the mixed ancestry South African population. The occurrence of IAs on A5 and C1 haplotypes in mixed ancestry South Africans, resembling HD haplotypes found in Southeast Asian patients (Pulkes et al., 2014), may reflect minor Asian genetic and demographic contributions as previously shown at the genome level (de Wit et al., 2010).

IAs from black South Africans occur on an assortment of haplotypes of African origin, similar to the diverse background of African haplotypes observed in black South African HD chromosomes (Baine et al., 2013). Interestingly, our previously reported enrichment

for the B2 haplotype in black South African HD chromosomes is not reflected among IAs from the black South African general population, suggesting a major role for stochastic and founder effects in the demographic history of HD in this ethnic group. European ancestry populations, by contrast, show similar haplotype distributions of IAs and HD alleles, suggesting that regular CAG expansion events on specific haplotypes play a major role in the frequency of HD as opposed to chance expansions occurring among all haplotypes in the population. Expressed differently, the higher new mutation rate of HD in European ancestry populations results from a higher IA frequency on ancestry-specific A1, A2, and A3a haplotypes, whereas these haplotypes are absent in black African populations where both IA frequency and the new mutation rate are correspondingly lower. In the latter situation, chance expansion events into the IA and HD range are likely to play a greater role in the origin of new HD alleles.

Known clusters of HD, such as in the Zulia state of Venezuela, are thought to represent founder events where rapid demographic expansion from affected ancestors has resulted in a locally high frequency of the mutation (Wexler et al., 2004). In such populations, haplotypes of the HD mutation are expected to be shared between affected descendants of a founder. We observe the HD mutation on the European ancestry A3a haplotype in DNA from four distantly related Venezuelan patients, known to be part of the extended founder pedigree in this population. We hypothesize that the founder mutation in this population occurred on the A3a haplotype, and that all affected descendants of this founder likewise share the HD mutation on A3a. As such, *de novo* HD mutations may have played a minor role in the high prevalence of HD in the

Maracaibo region of Venezuela relative to high birth rates from a single pedigree within a limited geographic area.

Our analysis suggests that differences in IA frequency alone can explain variable HD prevalence rates in major ethnic groups. However, differences in *cis* or *trans* genetic background may nonetheless modify the instability of high normal CAG repeat alleles or IAs. An instability-modifying haplotype has been reported in SCA3 (Martins et al., 2008) and undiscovered *cis* factors in *HTT* CAG repeat instability cannot be excluded by our data. However, such *cis* modifiers do not appear to be necessary to explain the molecular epidemiology of HD as a function of IA frequency by population. High normal CAG repeat lengths on specific haplotypes in European ancestry populations (A1, A2a, A2b, and A3a) may have resulted from drift of CAG repeat length early in the evolution of these haplotype lineages, which over many generations resulted in higher population frequencies of IAs occurring on these haplotypes (Falush, 2009). It is notable that the A1 and A2a haplotypes are the most genetically diverged European ancestry haplotypes within the A haplogroup, indicating that these haplotypes acquired high normal CAG repeat lengths independently rather than by common inheritance of instability-promoting sequences in the surrounding genomic region.

Ancestry-specific differences in prevalence are a hallmark of many trinucleotide repeat disorders (TNRs), including myotonic dystrophy (DM) (Theadom et al., 2014), Friedreich's ataxia, dentatorubral-pallidoluysian atrophy (DRPLA), and various spinocerebellar ataxias (SCAs) (Ruano et al., 2014). Where studied, a higher

prevalence of a given TNR disorder typically coincides with enrichment for longer normal alleles in the same population, often attributable to one or more haplotypes on which long normal alleles are preferentially found to occur (Imbert et al., 1993) (Liquori et al., 2003) (Cossee et al., 1997). Our study, the first to comparatively analyze IA frequency across populations with differing prevalence rates of HD, suggests that populations enriched for long normal alleles in other trinucleotide repeat disorders may also have higher premutation allele frequencies and new mutation rates at their respective loci.

Our population-based estimate of the HD new mutation rate among individuals of predominant European ancestry is substantially higher than prior extrapolations in this population. Conditional probability estimates of *de novo* HD transmission risk constructed from IA frequency and the prevalence of North American patients without family history suggested that 1/6,241 to 1/951 transmissions from a father with an IA genotype will result in a new mutation for HD (Hendricks et al., 2009). Our population-based estimates show that the risk of a new HD mutation for any given paternal transmission in the population already overlaps this range (1 in 5372), and is much higher among transmissions from fathers with a confirmed IA, particularly at repeat lengths approaching the HD range. Our population-based estimate of the HD new mutation rate is roughly two orders of magnitude higher than indirect estimates predating mapping of the CAG repeat expansion (Shaw and Caro, 1982). Our data data suggests that new mutations for HD may be common in the population relative to *de novo* mutations causative of other genetic diseases.

Future molecular and clinical studies will be necessary to validate our inferred HD new mutation rates. In particular, our rates include expansion events into the reduced penetrance range (36-39 CAG) where manifestation of HD occurs later in life or not at all. It is therefore important to consider not only the rate of expansion into the HD range (≥36 CAG) but also the penetrance of the resulting allele in individuals inheriting a *de novo* HD genotype. We recently revealed a high frequency of reduced penetrance alleles in the European ancestry general population, suggesting that the population prevalence of individuals with a *de novo* HD mutation could exceed the number of individuals who go on to clinically manifest HD (Kay et al., 2016a). Since individuals with reduced penetrance HD alleles often do not manifest HD but are at high risk of further repeat expansion in offspring, the number of individuals who present HD symptoms without a family history may not correspond to the inferred genetic new mutation rate presented in our study.

In conclusion, we show that the mean CAG repeat length and IA frequency in a population correlate with the manifest prevalence of HD in that population. In populations of European and admixed American ancestry, this relationship is mediated by the presence of high normal CAG repeat alleles which occur preferentially on HD-associated haplotypes A1, A2, and A3a. The inferred new mutation rate for HD is higher in populations with a higher IA frequency. We estimate that up to 1 in 3260 births may result in a *de novo* CAG repeat expansion into the HD range (≥36 CAG) in populations of predominant European ancestry. New mutations for HD are frequent and vary

substantially between populations as a function of IA frequency and haplotype composition, providing an empirical framework for the observed differences in prevalence of HD by ancestry worldwide.

Chapter 6: Discussion

6.1 Overview

This dissertation expands the scope of knowledge about the HD mutation in human populations worldwide. My investigations have advanced understanding of the therapeutic targetability of the mutation by haplotype, the ancestral origin of the mutation in different populations, and the underlying genetic epidemiology of HD.

In Chapter 2, I describe all common alleles marking distinct haplotype lineages of the mutation: A1, A2a, A2b, and A3a. These HD-associated haplotypes subsequently define target panels for allele-specific *HTT* silencing in a maximum proportion of European ancestry patients. The novelty and utility of this work is underscored by my recently issued patent on antisense compositions that lower mutant *HTT* expression by targeting haplotype-defining A1, A2, and A3a alleles in appropriately heterozygous patients (WO2016041058A1).

In Chapter 3, I show by dense *HTT* haplotyping in Latin American patients that the targetable A1 haplotype is the most common HD haplotype in this population, and also represents the most useful target panel for allele-specific silencing of the HD mutation in the most Latin American patients. Follow-up analysis of *HTT* haplotypes in unaffected Amerindian individuals and mestizo Peruvian control chromosomes revealed that A1 haplotypes in Latin America are predominantly marked by Amerindian ancestry-specific

alleles, representing a distinct Amerindian A1 haplotype variant in New World populations. This Amerindian A1 haplotype is found on 70% of HD mutation chromosomes in Peruvian patients, and 47% of HD mutation chromosomes in families of mixed Latin American descent, arguing for a major indigenous origin of HD in admixed Latin American populations.

In Chapter 4, I show the first direct prevalence estimate of individuals heterozygous for the HD mutation in the general population. Approximately 1 in 400 individuals have a CAG repeat genotype associated with HD, far more than the 1 in 7300 who manifest the disease (Fisher and Hayden, 2014). Most of the HD mutation alleles found in the general population show reduced penetrance in clinical samples, and therefore many individuals with an HD-associated CAG repeat genotype may not develop the disease in a normal lifespan. It remains unclear how many asymptomatic individuals with HD-associated alleles show subclinical phenotypes, such as prodromal striatal volume loss and psychiatric disturbances. More individuals may seek treatment options for HD in the future than are currently diagnosed, particularly among the elderly where age-specific penetrance is likely to be higher.

In Chapter 5, I establish that the frequency of intermediate *HTT* CAG repeat alleles (IAs; CAG 27-35), which lead to *de novo* HD mutations in offspring, is correlated with the observed prevalence of HD in different populations. I show that CAG repeats of A1 and A2 haplotypes in the general population are markedly longer than CAG repeats on other

haplotypes, offering validation of the hypothesis that specific haplotypes act as reservoirs for CAG repeat expansion in populations where HD is more prevalent.

6.2 HD-Associated A1 and A2 Haplotypes in Different Global Populations

HD-associated A1 and A2 haplotypes occur in 1000 Genomes Project reference subjects with ancestry from Europe, South Asia, and the Americas, but not in subjects with exclusive ancestry from Africa or East Asia (**Figure 3-3**). As in Europe, the Amerindian A1 haplotype occurs rarely in South Asian reference subjects suggesting that most A1 haplotypes in the South Asian population share common descent with European A1, possibly predating divergence of the Amerindian A1 haplotype. In a clinical investigation of South Asian HD patients, 4/25 patients of Northern Indian ancestry and 4/10 patients of Southern Indian ancestry had a Δ2642 codon deletion genotype (Saleem et al., 2003), suggesting that the HD mutation occurs on the A1 *HTT* haplotype in these populations. Alleles representing the A2 haplotype are also found in South Asian subjects from 1000 Genomes Project Phase 3. We therefore performed *HTT* haplotype analysis in HD patients and relatives of South Asian ancestry from the UBC HD Biobank, with the hypothesis that the HD mutation in this ancestry group occurs on A1 and A2 haplotypes as in European patients (**Figure 6-1**).

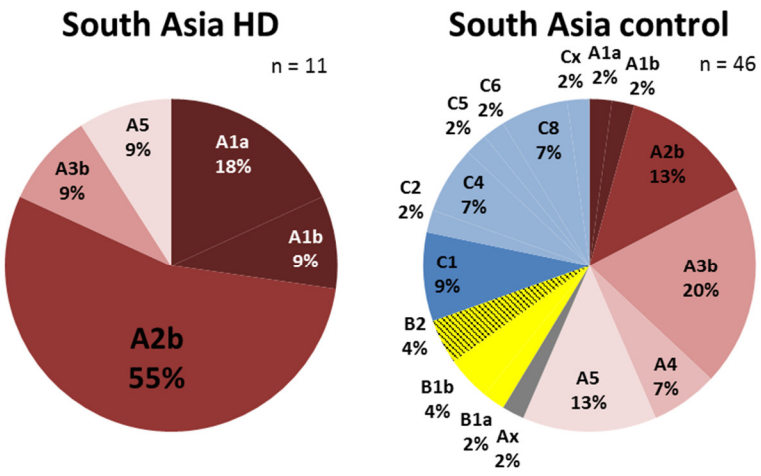


Figure 6-1. *HTT* haplotypes from HD and control chromosomes in South Asian HD patients

A comparatively small number of HD families of South Asian ancestry (n=11) were available from the UBC HD Biobank on the basis of local clinical referrals. Patients had confirmed ancestry of the HD mutation from parents or grandparents of South Asian origin. Haplotype analysis of *HTT* in these families revealed that the HD mutation in South Asian patients occurs on the A1 and A2 haplotypes as in patients of European ancestry, most commonly on the A2b haplotype (55%) with the A1 haplotype less frequently represented (27%). A3a, marked by the rs113407847 variant allele, is absent in South Asian reference controls from the 1000 Genomes Project as well as in South Asian haplotypes from our families. Amerindian A1, marked by rs12508079:T>C, was entirely absent in South Asian HD and control chromosomes from our families.

In collaboration with the University of Colombo, we additionally performed *HTT* haplotype analysis in a small clinical cohort of HD patients from Sri Lanka (**Figure 6-2**). The HD mutation in Sri Lankan patients also occurs on the A1 and A2 haplotypes, with

enrichment for the A2 haplotype in HD chromosomes as in South Asian HD families from the UBC HD Biobank. Interestingly, although the parent A2 haplotype is enriched on HD chromosomes in both the South Asian and Sri Lankan patient cohorts, HD in the South Asian cohort presents on the A2b subtype and HD in the Sri Lankan cohort presents on the A2a subtype. A1 and A2 were not observed among Sri Lankan control chromosomes, although sample size was small and the absence of A1 and A2 in this sample could be attributable to chance. Amerindian A1 was likewise absent among our Sri Lankan *HTT* haplotypes.

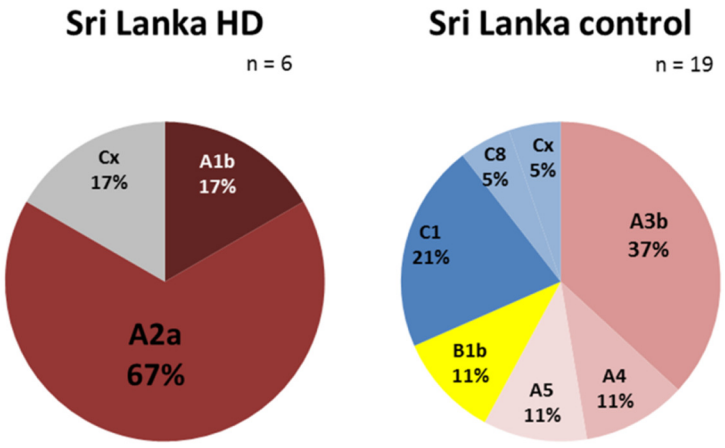


Figure 6-2. *HTT* haplotypes from HD and control chromosomes in Sri Lankan HD patients

These data suggest that A1 and A2 are the dominant haplotypes of the HD mutation in all populations where those haplotypes are found, including populations having common descent with European and South Asian ancestry groups. This may include populations of Eastern European, Middle Eastern, and possibly Central Asian ancestry. A1 and A2 *HTT* haplotypes are absent in Sub-Saharan African and East Asian populations. The emergence of A1 and A2 therefore occurred after migration of humans out of Africa. It is

unclear whether the ancestors of East Asians diverged prior to the emergence of the A1 and A2 haplotypes in common ancestors of contemporary Caucasian populations, or whether A1 and A2 haplotypes were instead lost in East Asians by drift or serial bottleneck effects of migration. Considering the absence of A1 and A2 among Sub-Saharan Africans and East Asians, these haplotypes are also likely to be absent in populations of Australoid ancestry, such as the Aboriginals of Australia, who may predate the later human migrations out of Africa where A1 and A2 emerged (Rasmussen et al., 2011).

The presence of an Amerindian variant of the A1 haplotype is surprising, given the longstanding presumption that Indigenous Americans have closer relation to East Asian populations than with European or South Asian ancestry groups. Our discovery of a distinct A1 haplotype variant in Amerindian Peruvians supports recent evidence that the ancestral origins of indigenous Americans may substantially derive from Central Asian populations sharing deep ancestry with modern Europeans (Raghavan et al., 2014). These findings suggest an important Eurasian contribution to the ancestry of indigenous Americans, possibly admixed with a more recent East Asian divergence. The A1 haplotype may have been present in ancestral Eurasian populations an estimated 24,000 years ago, but absent in earlier common ancestors of Caucasians and East Asians who diverged an estimated 41,000 years ago (Tateno et al., 2014).

6.3 Allele-Specific *HTT* Silencing in Different Global Populations

The detailed description of A1 and A2 *HTT* haplotype frequencies by population, as well as the apparent association between A1 and A2 haplotypes in the general population and high prevalence rates of HD, critically informs allele-specific *HTT* silencing therapies in development. In populations where HD is most frequent, the HD mutation predominantly occurs on A1 and A2 *HTT* haplotypes. As shown in Chapter 2, alleles that are the most heterozygous in European ancestry patients and phased with the HD mutation are in fact the defining alleles of the A1 and A2 haplotypes. This is true despite the relatively common frequency of A1 and A2 haplotypes among European control chromosomes, due to their very high frequencies among HD chromosomes in patients. Defining alleles of the parent A1 haplotype also enable allele-specific *HTT* silencing treatment of the most patients in Peru, despite the ancestrally distant relationship of Amerindian A1 versus European A1 in this population.

The global association of the HD mutation with A1 and A2 in populations with the highest prevalence of the disease suggests that targeting of these haplotypes may enable allele-specific gene silencing treatment of the greatest number of patients worldwide. HD is found preferentially on A1 and A2 *HTT* haplotypes in patients from Europe, North America, South America, and South Asia. Among the populations studied, the HD mutation occurs on distinct haplotypes only in East Asians and black South Africans, where the disease is comparatively rare. To estimate the proportion of patients treatable by targeting both the A1 and A2 haplotypes, we calculated

heterozygosity of these haplotypes in phase with the HD mutation for each patient population for which we had available haplotype data. A small number of HD families with confirmed disease ancestry in the Middle East were also available for haplotype analysis (n=8). In these families of confirmed Middle Eastern origin, the HD mutation was always found to occur on the A2b haplotype.

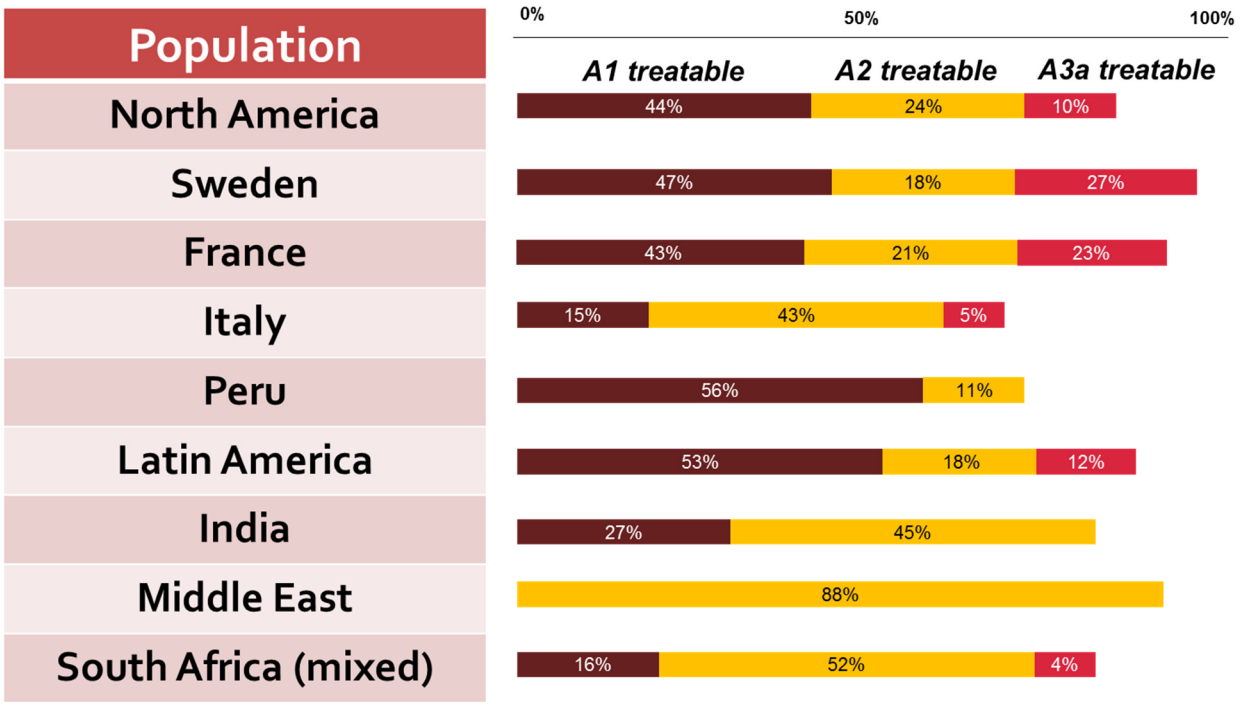


Figure 6-3. Combinatorial coverage of the patient population for allele-specific silencing by targeting A1, A2, and A3a *HTT* haplotypes

Remarkably, combinatorial targeting of the A1 and A2 *HTT* haplotypes for allele-specific gene silencing enables treatment of >60% of patients in all populations where these haplotypes are found (**Figure 6-3**). Addition of A3a as a tertiary target increases overall

coverage to approximately 70% of patients on average across the same populations, notably increasing the number of treatable patients in Western Europe.

Additional haplotyping in patient populations from Eastern Europe, Central Asia, and different regions of Africa will be necessary to guide a comprehensive allele-specific silencing strategy. However, a global pattern of optimal A1 and A2 targeting is evident among populations where the prevalence of HD is highest. This suggests that defining alleles of A1 and A2 may represent optimal allele-specific *HTT* silencing targets for treating the most HD patients worldwide.

6.4 Frequency of the HD Mutation and Pre-Manifest Treatment

The unexpectedly high frequency of the HD mutation described in Chapter 4 raises questions regarding the true health burden of the mutation in the general population. It is possible that many individuals with a reduced penetrance allele, while not diagnosed with HD, manifest subtle signs or symptoms resulting from the mutation. Such undiagnosed individuals may experience morbidity not currently recognized as a consequence of the HD mutation. Large-scale genotyping of well-phenotyped patient cohorts, recruited from specialists and general practice, will be necessary to evaluate this possibility. At present, massive sequencing studies in national health care systems or private regional providers do not genotype the *HTT* CAG repeat due to limitations of short-read platforms and associated analysis. Targeted genotyping of the *HTT* CAG

repeat in such cohorts may be necessary, with evident ethical concerns of uncovering the HD mutation in people without family history of HD.

Manifest HD is known to occur after an extended prodromal period, with measurable caudate volume loss up to 10 years before clinical diagnosis (Tabrizi et al., 2013). It is possible that improved ascertainment of pre-manifest individuals with the HD mutation may lead to prophylactic therapies aimed at delaying onset, such as gene silencing candidates in early clinical stages. Haplotype analysis of HD-associated alleles from the general population (Chapter 4) suggests that a similar proportion of individuals from the general population could be eligible for allele-specific therapy as in clinical cohorts. Such pre-manifest intervention would crucially depend on the success and safety of gene-silencing therapies now in clinical trials for altering the progressive course of manifest HD, as well as the development of clinical standards for pre-manifest intervention. Given appropriate endpoints in pre-manifest patients, trials in the pre-manifest population may eventually be warranted, and could dramatically expand the number of people who benefit from HD-specific treatments.

6.5 Mutational Flux of the CAG Repeat Expansion in HD

We have previously shown that germline CAG repeat instability and the risk of *de novo* HD mutation increases exponentially with increasing parental CAG repeat length across the intermediate range (CAG 27-35) (Semaka et al., 2013a). It is therefore unsurprising that populations with a higher frequency of intermediate alleles have a higher

prevalence of HD, as shown in Chapter 5. Extensive haplotype analysis of *HTT* in the general population illuminates the underlying reasons for higher IA prevalence in specific ethnic groups. Where frequency of IAs on the A1 and A2 haplotypes is highest, the overall average CAG length in the population is also high. Based on data from the UBC HD Biobank, this elevated average CAG repeat length appears to result from enrichment for A1 and A2 haplotypes bearing high normal repeat lengths (CAG 20-26). Given a steadily increasing rate of germline CAG repeat instability across the high normal range, enrichment of A1 and A2 haplotypes in the general population therefore increases the likelihood of expansion to progressively higher CAG repeat lengths across generations. A consequent increase in IA frequency will lead to a subsequently higher HD new mutation rate. It is unclear whether the overall frequency of A1 and A2 haplotypes in a population is directly proportional to the frequency of IAs and HD mutation alleles.

The rate of *de novo* HD mutation may be balanced by the loss of juvenile HD mutations not transmitted to successive generations. The point prevalence of HD can therefore be thought of as a snapshot of ongoing mutational flux in a population, with *de novo* CAG repeat expansion adding new HD mutations to the allele pool and anticipation gradually eliminating expanded derivative alleles several generations later. It is unclear how many generations, on average, a given *de novo* HD mutation persists in the population before limiting juvenile HD repeat lengths are reached. A higher HD new mutation rate, resulting from enrichment of high-normal CAG repeat alleles on the A1 and A2 haplotypes, may be counterbalanced by an elevated rate of juvenile HD cases who do

not transmit the mutation to the next generation, leading to a higher point prevalence of the mutation but little net change in mutation frequency over time in a population. Alternatively, imbalance in the HD new mutation rate and the elimination of extreme early-onset alleles may result in increases or decreases in the prevalence of HD in a population over time.

It is notable that the highest average CAG repeat lengths are found in populations of Latin American ancestry, such as in Mexico (19.04 CAG) (Alonso et al., 2009) and in Latino individuals from our United States sample (19.03 CAG). We also show that Latino North Americans may have the highest IA frequency of all studied populations (9/160, 5.6%). No population-based prevalence estimates of HD are available in Latin America (Castilhos et al., 2015), however these genetic data argue that HD may be more frequent in Latin American populations than in people of predominant Western European ancestry.

6.6 Origin of the HD Mutation in Different Populations

Data in this dissertation allows for description of the different origins of the HD mutation in greater detail than previously possible. My investigations show multiple origins of the HD mutation, both within populations and between ethnically distinct populations around the world.

In populations with a high prevalence of HD, the mutation results from ongoing and continuous expansion of high normal and intermediate CAG repeat alleles found on the A1, A2, and A3a haplotypes. The exact mutation rate by population, and consequent prevalence, likely depends on the frequency of intermediate alleles. The mutation rate is expected to be invariant of birth rate and aging in a population, whereas manifest prevalence of HD is clearly dependent on such demographic factors.

In populations with a relatively low prevalence of the disease, by contrast, chance CAG repeat expansions appear to play a greater role in determining the prevalence rate of HD. In black South Africans, for example, our haplotype analysis of IAs from the general population revealed an assortment of haplotypes specific to black Africans, without clear enrichment for any specific haplotype variant. This reflects our previously published findings of HD mutation haplotypes in black South Africans, with the exception of significant enrichment for the B2 haplotype (Baine et al., 2013). We observed no analogous enrichment of the B2 haplotype among IAs from black South African individuals, suggesting that HD mutation alleles on the B2 haplotype may represent a local founder effect. Given the rarity of HD among black South Africans, the impact of such chance repeat expansions on the overall haplotype distribution is likely to be higher in comparison to populations where *de novo* mutation events are more frequent.

Known clusters of HD, such as in the Zulia state of Venezuela, are thought to represent founder events where rapid demographic expansion from affected ancestors has resulted in a locally high frequency of the mutation. Haplotypes of the HD mutation are

expected to be shared between affected descendants of a founder. We observe the HD mutation on the European ancestry A3a haplotype in DNA from four distantly related Venezuelan patients, known to be part of the extended founder pedigree in this population. We hypothesize that the founder mutation in this population occurred on the A3a haplotype, and that all affected descendants of this founder will likewise share the HD mutation on A3a. As such, *de novo* HD mutations appear to have played a negligible role in the high prevalence of HD observed in the Maracaibo region of Venezuela.

It remains unclear why the A1 and A2 *HTT* haplotypes occur predominantly with high-normal CAG repeat alleles. It has been proposed that unidentified *cis*-elements specific to these haplotypes may predispose CAG repeat alleles to instability or expansion over many generations (Warby et al., 2009). The complete identification of all defining alleles on these haplotypes across the *HTT* gene region, described in Chapter 2, argues against this hypothesis. A1 and A2 are defined by divergent sets of alleles in different regions of the gene, with common haplogroup A alleles also occurring on haplotypes of extremely low average CAG repeat length, such as A5. Instability-modifying *cis*-elements are also unlikely to occur upstream or downstream of these intragenic haplotypes, since recombination events immediately upstream and downstream of *HTT* are common, including among A1 and A2 haplotypes. While detailed study of CAG repeat instability by haplotype among general population alleles is notably lacking, available data described in this dissertation suggest that haplotype-specific alleles acting to modify instability are unlikely to be found. It remains theoretically possible,

though unparsimonious, that defining A1 and A2 alleles may modify CAG repeat instability by distinct mechanisms.

An alternative hypothesis described by Falush, wherein founder A1 and A2 chromosomes on comparatively high CAG repeats led to stepwise expansion in descendent haplotype lineages, is more compatible with the observed haplotype and CAG repeat distribution data presented in this dissertation (Falush, 2009). Based on length-dependent CAG repeat instability alone, Falush modeled that a founder haplotype of 23 CAG, representing one third of the founder alleles, would lead to 72% frequency among HD mutation chromosomes after 2000 generations, and 17% frequency among control chromosomes, without bottleneck or *cis*-element variables. Addition of instability *cis*-elements to this model resulted in bimodal distributions for haplotypes on which such elements emerged, not reflected in the European ancestry haplotype data. It is important to stress that the underlying CAG repeat instability data for this model was extrapolated from HD mutation repeat lengths and does not directly correspond to empirical data recently generated for *HTT* CAG repeat instability in the male germline. The application of empirical instability rates across the control and intermediate allele ranges to such evolutionary models will further elucidate the demographic and mutational history of A1 and A2 in different human populations. Given the relatively small effective human population sizes during the estimated origin of the A1 and A2 haplotypes, it is possible that serial demographic bottleneck effects resulted in high frequencies of A1 and A2 founders with comparatively high CAG repeat lengths, compounded by genetic drift in the resulting populations.

6.7 Future Directions

Whole-genome sequencing, coupled with genotyping of the *HTT* CAG repeat and access to clinical data, may address further questions in the population genetics of the HD mutation pertaining to the penetrance and frequency of the mutation worldwide. In Chapter 4, I show that 0.246% (0.151-0.380%) of individuals in three predominantly European ancestry populations have the HD mutation genotype. It remains possible that this figure is higher or lower in other populations of similar ancestry. The precise frequency of the HD mutation in ethnically distinct populations, such as Sub-Saharan Africans or East Asians where the prevalence of HD is documented to be lower, remains unknown. However, the lower frequency of IAs observed in these populations (Chapter 5) suggests that the HD mutation frequency is similarly low.

Accurately phased variant calls from whole-genome sequencing of HD patients and unaffected controls will improve our understanding of the ancestry of HD in different populations. Direct sequencing of the major HD haplotype, corresponding to the A1 haplotype, in four affected European ancestry families revealed private *HTT* variants (Lee et al., 2015b). Such rare variants may be missed in general population reference cohorts, such as in the 1000 Genomes Project, and may offer evidence of common HD mutation ancestry in groups of families, particularly where a founder effect is suspected. In particular, the application of private variant calling to *HTT* haplotypes in HD families from Sub-Saharan Africa and East Asia will illuminate the origin and relatedness of

cases where the mutation presents on similar common haplotypes, such as the B2 haplotype in black South African HD patients.

Target development for allele-specific silencing will crucially depend on the success of *HTT* silencing modalities in development. Ionis Pharmaceuticals now leads a Phase 1/2a trial for ASO-mediated silencing of *HTT*, targeting both copies of the gene non-selectively to achieve partial reduction of mutant HTT in the brain. This trial was designed with ascending dose by intrathecal delivery to the CSF, the preferred route of administration of ASOs in HD, leading to passive diffusion of the ASO to affected brain regions. Ionis recently announced that the highest dose in this safety trial had been achieved in HD patients without any reportable adverse reactions to date. Should HTT-lowering ASOs prove safe throughout the Ionis trial, future safety trials for allele-specific ASOs targeting mutant *HTT* will likely follow. Choice of the optimal allele target for development and commercialization will then take precedence. The data presented in this dissertation suggests that a lead ASO targeting any of the three defining alleles of the A1 haplotype, in combination with a secondary ASO targeting any of the five defining alleles of the A2 haplotype, will enable allele-specific *HTT* silencing in at least 60% of patients in populations where HD is most prevalent. It is possible that alternative allele targets, such as those defining the A and B haplogroups, will have greater utility in other populations where HD is comparatively less common. The commercial potential of ASOs targeting alleles other than those that define the A1 and A2 haplotypes will depend on our improved understanding of the HD mutation haplotype, frequency, and penetrance around the world. Additional genetic and epidemiological studies of the HD

mutation in diverse populations, particularly across Africa and Asia, will be necessary to evaluate the complete scope of allele targets best suited to therapeutic gene-silencing applications in the maximum number of HD patients worldwide.

References

(1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72: 971-983.

Abecasis GR, Altshuler D, Auton A, et al. (2010). A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

Abecasis GR, Auton A, Brooks LD, et al. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

ACMG/ASHG (1998). ACMG/ASHG statement. Laboratory guidelines for Huntington disease genetic testing. The American College of Medical Genetics/American Society of Human Genetics Huntington Disease Genetic Testing Working Group. Am J Hum Genet 62: 1243-1247.

Adachi Y & Nakashima K (1999). [Population genetic study of Huntington's disease--prevalence and founder's effect in the San-in area, western Japan]. Nihon Rinsho 57: 900-904.

Almqvist E, Spence N, Nichol K, et al. (1995). Ancestral differences in the distribution of the delta 2642 glutamic acid polymorphism is associated with varying CAG repeat lengths on normal chromosomes: insights into the genetic evolution of Huntington disease. Hum Mol Genet 4: 207-214.

Almqvist EW, Elterman DS, MacLeod PM, et al. (2001). High incidence rate and absent family histories in one quarter of patients newly diagnosed with Huntington disease in British Columbia. Clin Genet 60: 198-205.

Alonso ME, Ochoa A, Boll MC, et al. (2009). Clinical and genetic characteristics of Mexican Huntington's disease patients. Mov Disord 24: 2012-2015.

Ambrose C, Duyao M, Barnes G, et al. (1994). Structure and expression of the Huntington's disease gene: Evidence against simple inactivation due to an expanded CAG repeat. Somatic Cell and Molecular Genetics 20: 27-38.

Andrew S, Theilmann J, Almqvist E, et al. (1993a). DNA analysis of distinct populations suggests multiple origins for the mutation causing Huntington disease. Clin Genet 43: 286-294.

Andrew SE, Goldberg YP, Kremer B, et al. (1994a). Huntington disease without CAG expansion: phenocopies or errors in assignment? Am J Hum Genet 54: 852-863.

Andrew SE, Goldberg YP, Kremer B, et al. (1994b). Huntington Disease without CAG Expansion: Phenocopies or Errors in Assignment? American Journal of Human Genetics 54: 852-863.

Andrew SE, Goldberg YP, Kremer B, et al. (1993b). The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat Genet 4: 398-403.

Andrew SE, Goldberg YP, Theilmann J, et al. (1994c). A CCG repeat polymorphism adjacent to the CAG repeat in the Huntington disease gene: implications for diagnostic accuracy and predictive testing. Hum Mol Genet 3: 65-67.

Andrew SE & Hayden MR (1995). Origins and evolution of Huntington disease chromosomes. Neurodegeneration 4: 239-244.

Auerbach W, Hurlbert MS, Hilditch-Maguire P, et al. (2001). The HD mutation causes progressive lethal neurological disease in mice expressing reduced levels of huntingtin. Hum Mol Genet 10: 2515-2523.

Auton A, Brooks LD, Durbin RM, et al. (2015). A global reference for human genetic variation. Nature 526: 68-74.

Baine FK, Kay C, Ketelaar ME, et al. (2013). Huntington disease in the South African population occurs on diverse and ethnically distinct genetic haplotypes. Eur J Hum Genet 21: 1120-1127.

Baine FK, Krause A & Greenberg LJ (2016). The Frequency of Huntington Disease and Huntington Disease-Like 2 in the South African Population. Neuroepidemiology 46: 198-202.

Barron LH, Rae A, Holloway S, et al. (1994). A single allele from the polymorphic CCG rich sequence immediately 3' to the unstable CAG trinucleotide in the IT15 cDNA shows almost complete disequilibrium with Huntington's disease chromosomes in the Scottish population. Hum Mol Genet 3: 173-175.

Bean L & Bayrak-Toydemir P (2014). American College of Medical Genetics and Genomics Standards and Guidelines for Clinical Genetics Laboratories, 2014 edition: technical standards and guidelines for Huntington disease. Genet Med 16: e2.

Bennett CF & Swayze EE (2010). RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. Annual Review of Pharmacology and Toxicology 50: 259-293.

Boudreau RL, McBride JL, Martins I, et al. (2009). Nonallele-specific Silencing of Mutant and Wild-type Huntingtin Demonstrates Therapeutic Efficacy in Huntington's Disease Mice. Molecular therapy 17: 1053-1063.

Brinkman RR, Mezei MM, Theilmann J, et al. (1997). The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. Am J Hum Genet 60: 1202-1210.

Butland S, Devon R, Huang Y, et al. (2007). CAG-encoded polyglutamine length polymorphism in the human genome. BMC Genomics 8: 126.

Carroll JB, Warby SC, Southwell AL, et al. (2011). Potent and Selective Antisense Oligonucleotides Targeting Single-Nucleotide Polymorphisms in the Huntington Disease Gene / Allele-Specific Silencing of Mutant Huntingtin. Mol Ther 19: 2178-2185.

Castilhos RM, Augustin MC, Santos JA, et al. (2015). Genetic aspects of Huntington's disease in Latin America. A systematic review. Clin Genet.

Chang CM, Yu YL, Fong KY, et al. (1994). Huntington's disease in Hong Kong Chinese: epidemiology and clinical picture. Clin Exp Neurol 31: 43-51.

Chen YY & Lai CH (2010). Nationwide population-based epidemiologic study of Huntington's Disease in Taiwan. Neuroepidemiology 35: 250-254.

Conneally PM (1984). Huntington disease: genetics and epidemiology. Am J Hum Genet 36: 506-526.

Cornejo-Olivas M, Mori N, Alva I, et al. (2012). Huntington's Disease in an Indigenous Village in the Peruvian Amazon Jungle. 16th International Congress of Parkinson´s Disease and Movement Disorders. Dublin, Ireland.

Cornejo-Olivas M, Mata I, Dorschner M, et al. (2014). Target Sequencing Analysis of Parkinson's disease genes in a healthy Amerindian population from Puno-Peru. 64th Annual Meeting of the American Society of Human Genetics. San Diego, CA.

Cornejo-Olivas MR, Inca-Martinez MA, Espinoza-Huertas K, et al. (2015). Clinical and Molecular Features of Late Onset Huntington Disease in a Peruvian Cohort. J Huntingtons Dis 4: 99-105.

Cossee M, Schmitt M, Campuzano V, et al. (1997). Evolution of the Friedreich's ataxia trinucleotide repeat expansion: founder effect and premutations. Proc Natl Acad Sci U S A 94: 7452-7457.

Costa MC, Magalhaes P, Guimaraes L, et al. (2006). The CAG repeat at the Huntington disease gene in the Portuguese population: insights into its dynamics and to the origin of the mutation. J Hum Genet 51: 189-195.

Cuba JM (1986). [A focus of Huntington's chorea in Peru]. Rev Neurol (Paris) 142: 151-153.

de Wit E, Delport W, Rugamika CE, et al. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. Hum Genet 128: 145-153.

Deka R, Guangyun S, Smelser D, et al. (1999). Rate and directionality of mutations and effects of allele size constraints at anonymous, gene-associated, and disease-causing trinucleotide loci. Mol Biol Evol 16: 1166-1177.

Delaneau O, Coulonges C & Zagury JF (2008). Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics 9: 540.

Dorsey E (2012). Characterization of a large group of individuals with huntington disease and their relatives enrolled in the COHORT study. PLoS One 7: e29522.

Dragatsis I, Levine MS & Zeitlin S (2000). Inactivation of Hdh in the brain and testis results in progressive neurodegeneration and sterility in mice. Nat Genet 26: 300-306.

Duyao M, Ambrose C, Myers R, et al. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat Genet 4: 387-392.

Duyao MP, Auerbach AB, Ryan A, et al. (1995). Inactivation of the Mouse Huntington's Disease Gene Homolog Hdh. Science 269: 407-410.

Evans SJW, Douglas I, Rawlins MD, et al. (2013). Prevalence of adult Huntington's disease in the UK based on diagnoses recorded in general practice records. Journal of Neurology, Neurosurgery & Psychiatry.

Falush D (2009). Haplotype background, repeat length evolution, and Huntington's disease. Am J Hum Genet 85: 939-942.

Falush D, Almqvist EW, Brinkmann RR, et al. (2001). Measurement of mutational flow implies both a high new-mutation rate for Huntington disease and substantial underascertainment of late-onset cases. Am J Hum Genet 68: 373-385.

Fisher ER & Hayden MR (2014). Multisource ascertainment of Huntington disease in Canada: prevalence and population at risk. Mov Disord 29: 105-114.

Fiszer A, Mykowska A & Krzyzosiak WJ (2011). Inhibition of mutant huntingtin expression by RNA duplex targeting expanded CAG repeats. Nucleic Acids Res 39: 5578-5585.

Fiszer A, Olejniczak M, Galka-Marciniak P, et al. (2013). Self-duplexing CUG repeats selectively inhibit mutant huntingtin expression. Nucleic Acids Res 41: 10426-10437.

Gagnon KT, Pendergraff HM, Deleavey GF, et al. (2010). Allele-Selective Inhibition of Mutant Huntingtin Expression with Antisense Oligonucleotides Targeting the Expanded CAG Repeat. Biochemistry 49: 10166-10178.

Garcia-Planells J, Burguera JA, Solis P, et al. (2005). Ancient origin of the CAG expansion causing Huntington disease in a Spanish population. Hum Mutat 25: 453-459.

García LN (1991). Historia de las Américas Madrid, Editorial Alhambra (Universidad de Sevilla).

Goemans NM, Tulinius M, van den Akker JT, et al. (2011). Systemic Administration of PRO051 in Duchenne's Muscular Dystrophy. New England Journal of Medicine 364: 1513-1522.

Goldberg YP, Kalchman MA, Metzler M, et al. (1996). Absence of Disease Phenotype and Intergenerational Stability of the Cag Repeat in Transgenic Mice Expressing the Human Huntington Disease Transcript. Human Molecular Genetics 5: 177-185.

Goldberg YP, Kremer B, Andrew SE, et al. (1993). Molecular analysis of new mutations for Huntington's disease: intermediate alleles and sex of origin effects. Nat Genet 5: 174-179.

Goldman A, Ramsay M & Jenkins T (1996). Ethnicity and myotonic dystrophy: a possible explanation for its absence in sub-Saharan Africa. Ann Hum Genet 60: 57-65.

Gordon PH, Mehal JM, Rowland AS, et al. (2016). Huntington disease among the Navajo: A population-based study in the Navajo Nation. Neurology 86: 1552-1553.

Graham RK, Slow EJ, Deng Y, et al. (2006). Levels of mutant huntingtin influence the phenotypic severity of Huntington disease in YAC128 mouse models. Neurobiology of Disease 21: 444-455.

Gravel S, Zakharia F, Moreno-Estrada A, et al. (2014). Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet 9: e1004023.

Gusella JF, Wexler NS, Conneally PM, et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306: 234-238.

Harper SQ, Staber PD, He X, et al. (2005). RNA interference improves motor and neuropathological abnormalities in a Huntington's disease mouse model. PNAS 102: 5820-5825.

Hecimovic S, Klepac N, Vlasic J, et al. (2002). Genetic background of Huntington disease in Croatia: Molecular analysis of CAG, CCG, and Delta2642 (E2642del) polymorphisms. Hum Mutat 20: 233.

Hendricks AE, Latourelle JC, Lunetta KL, et al. (2009). Estimating the probability of de novo HD cases from transmissions of expanded penetrant CAG alleles in the Huntington disease gene from male carriers of high normal alleles (27-35 CAG). Am J Med Genet A 149A: 1375-1381.

Herrera I (1986). Los Inmigrantes Chinos en la Amazonia Peruana. Bulletin de l'Institut Français d'Etudes Andines XV: 49-60.

Holmes SE, O'Hearn E, Rosenblatt A, et al. (2001). A repeat expansion in the gene encoding junctophilin-3 is associated with Huntington disease-like 2. Nat Genet 29: 377-378.

Hu J, Matsui M, Gagnon KT, et al. (2009). Allele-specific silencing of mutant huntingtin and ataxin-3 genes by targeting expanded CAG repeats in mRNAs. Nature Biotechnology 27: 478-484.

Imbert G, Kretz C, Johnson K, et al. (1993). Origin of the expansion mutation in myotonic dystrophy. Nat Genet 4: 72-76.

Jackson JN, Long KM, He Y, et al. (2016). A comparison of DMET Plus microarray and genome-wide technologies by assessing population substructure. Pharmacogenet Genomics.

Jakab K, Gardian G, Endreffy E, et al. (1999). Analysis of CAG repeat expansion in Huntington's disease gene (IT 15) in a Hungarian population. Eur Neurol 41: 107-110.

James CM, Houlihan GD, Snell RG, et al. (1994). Late-onset Huntington's disease: a clinical and molecular study. Age Ageing 23: 445-448.

Kay C, Collins JA, Miedzybrodzka Z, et al. (2016a). Huntington disease reduced penetrance alleles occur at high frequency in the general population. Neurology 87: 282-288.

Kay C, Collins JA, Skotte NH, et al. (2015). Huntingtin Haplotypes Provide Prioritized Target Panels for Allele-specific Silencing in Huntington Disease Patients of European Ancestry. Mol Ther 23: 1759-1771.

Kay C, Fisher ER & Hayden M (2014a). Epidemiology (in Huntington's Disease, 4th ed), Oxford University Press.

Kay C, Skotte NH, Southwell AL, et al. (2014b). Personalized gene silencing therapeutics for Huntington disease. Clin Genet 86: 29-36.

Kay C, Tirado-Hurtado I, Cornejo-Olivas M, et al. (2016b). The targetable A1 Huntington disease haplotype has distinct Amerindian and European origins in Latin America. Eur J Hum Genet.

Keller MA, Gordon ES, Stack CB, et al. (2010). Coriell Personalized Medicine Collaborative: a prospective study of the utility of personalized medicine. Personalized Medicine 7: 301-317.

Kere J (2001). Human population genetics: lessons from Finland. Annu Rev Genomics Hum Genet 2: 103-128.

Kordasiewicz HB, Stanek LM, Wancewicz EV, et al. (2012). Sustained Therapeutic Reversal of Huntington's Disease by Transient Repression of Huntingtin Synthesis. Neuron 74: 1031-1044.

Kremer B, Goldberg P, Andrew SE, et al. (1994). A Worldwide Study of the Huntington's Disease Mutation: The Sensitivity and Specificity of Measuring CAG Repeats. New England Journal of Medicine 330: 1401-1406.

Kutuev IA, Khusainova RI, Khidiyatova IM, et al. (2004). Analysis of the IT15 Gene in Huntington's Disease Families. Russian Journal of Genetics 40: 919-925.

Langbehn DR, Brinkman RR, Falush D, et al. (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clin Genet 65: 267-277.

Langbehn DR, Hayden MR & Paulsen JS (2010). CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. Am J Med Genet B Neuropsychiatr Genet 153B: 397-408.

Leavitt BR, Guttman JA, Hodgson JG, et al. (2001). Wild-Type Huntingtin Reduces the Cellular Toxicity of Mutant Huntingtin In Vivo. American journal of human genetics 68: 313.

Leavitt BR, van Raamsdonk JM, Shehadeh J, et al. (2006). Wild-type huntingtin protects neurons from excitotoxicity. J Neurochem 96: 1121-1129.

Lee J-M, Gillis T, Mysore Jayalakshmi S, et al. (2012a). Common SNP-Based Haplotype Analysis of the 4p16.3 Huntington Disease Gene Region. American Journal of Human Genetics 90: 434-444.

Lee J-M, Wheeler Vanessa C, Chao Michael J, et al. (2015a). Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. Cell 162: 516-526.

Lee JM, Kim KH, Shin A, et al. (2015b). Sequence-Level Analysis of the Major European Huntington Disease Haplotype. Am J Hum Genet 97: 435-444.

Lee JM, Ramos EM, Lee JH, et al. (2012b). CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. Neurology 78: 690-695.

Lipe H & Bird T (2009). Late onset Huntington Disease: clinical and genetic characteristics of 34 cases. J Neurol Sci 276: 159-162.

Liquori CL, Ikeda Y, Weatherspoon M, et al. (2003). Myotonic dystrophy type 2: human founder haplotype and evolutionary conservation of the repeat tract. Am J Hum Genet 73: 849-862.

Lombardi MS, Jaspers L, Spronkmans C, et al. (2009). A majority of Huntington's disease patients may be treatable by individualized allele-specific RNA interference. Experimental Neurology 217: 312-319.

Ma M, Yang Y, Shang H, et al. (2010). Evidence for a predisposing background for CAG expansion leading to HTT mutation in a Chinese population. Journal of the Neurological Sciences 298: 57-60.

Maat-Kievit A, Losekoot M, Van Den Boer-Van Den Berg H, et al. (2001). New problems in testing for Huntington's disease: the issue of intermediate and reduced penetrance alleles. J Med Genet 38: E12.

MacDonald ME, Novelletto A, Lin C, et al. (1992). The Huntington's disease candidate region exhibits many different haplotypes. Nat Genet 1: 99-103.

Martins S, Coutinho P, Silveira I, et al. (2008). Cis-acting factors promoting the CAG intergenerational instability in Machado-Joseph disease. Am J Med Genet B Neuropsychiatr Genet 147B: 439-446.

Masuda N, Goto J, Murayama N, et al. (1995). Analysis of triplet repeats in the huntingtin gene in Japanese families affected with Huntington's disease. J Med Genet 32: 701-705.

McCusker EA, Casse RF, Graham SJ, et al. (2000). Prevalence of Huntington disease in New South Wales in 1996. Med J Aust 173: 187-190.

McIver LJ, Fondon JW, 3rd, Skinner MA, et al. (2011). Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. Genomics 97: 193-199.

McNeil SM, Novelletto A, Srinidhi J, et al. (1997). Reduced penetrance of the Huntington's disease mutation. Hum Mol Genet 6: 775-779.

Milunsky JM, Maher TA, Loose BA, et al. (2003). XL PCR for the detection of large trinucleotide expansions in juvenile Huntington's disease. Clin Genet 64: 70-73.

Molla M, Delcher A, Sunyaev S, et al. (2009). Triplet repeat length bias and variation in the human transcriptome. Proceedings of the National Academy of Sciences of the United States of America 106: 17095-17100.

Morovvati S, Nakagawa M, Osame M, et al. (2008). Analysis of CCG repeats in Huntingtin gene among HD patients and normal populations in Japan. Arch Med Res 39: 131-133.

Myers RH, MacDonald ME, Koroshetz WJ, et al. (1993). De novo expansion of a (CAG)n repeat in sporadic Huntington's disease. Nat Genet 5: 168-173.

Nasir J, Floresco SB, O'Kusky JR, et al. (1995). Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. Cell 81: 811-823.

Newcombe RG (1981). A life table for onset of Huntington's chorea. Ann Hum Genet 45: 375-385.

Novelletto A, Persichetti F, Sabbadini G, et al. (1994a). Analysis of the trinucleotide repeat expansion in Italian families affected with Huntington disease. Hum Mol Genet 3: 93-98.

Novelletto A, Persichetti F, Sabbadini G, et al. (1994b). Polymorphism analysis of the huntingtin gene in Italian families affected with Huntington disease. Hum Mol Genet 3: 1129-1132.

O'Kusky JR, Nasir J, Cicchetti F, et al. (1999). Neuronal degeneration in the basal ganglia and loss of pallido-subthalamic synapses in mice with targeted disruption of the Huntington's disease gene. Brain Research 818: 468-479.

Ostergaard ME, Southwell AL, Kordasiewicz H, et al. (2013). Rational design of antisense oligonucleotides targeting single nucleotide polymorphisms for potent and allele selective suppression of mutant Huntingtin in the CNS. Nucleic Acids Res 41: 9634-9650.

Ostergaard ME, Thomas G, Koller E, et al. (2015). Biophysical and biological characterization of hairpin and molecular beacon RNase H active antisense oligonucleotides. ACS Chem Biol 10: 1227-1233.

Paradisi I, Hernandez A & Arias S (2008). Huntington disease mutation in Venezuela: age of onset, haplotype analyses and geographic aggregation. J Hum Genet 53: 127-135.

Paulsen JS, Long JD, Ross CA, et al. (2014). Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study. Lancet Neurol 13: 1193-1201.

Perlis RH, Smoller JW, Mysore J, et al. (2010). Prevalence of incompletely penetrant Huntington's disease alleles among individuals with major depressive disorder. Am J Psychiatry 167: 574-579.

Persichetti F, Srinidhi J, Kanaley L, et al. (1994). Huntington's disease CAG trinucleotide repeats in pathologically confirmed post-mortem brains. Neurobiol Dis 1: 159-166.

Pfister EL, Kennington L, Straubhaar J, et al. (2009). Five siRNAs Targeting Three SNPs May Provide Therapy for Three-Quarters of Huntington's Disease Patients. Current Biology 19: 774-778.

Potter NT, Spector EB & Prior TW (2004). Technical standards and guidelines for Huntington disease testing. Genet Med 6: 61-65.

Pramanik S, Basu P, Gangopadhaya PK, et al. (2000). Analysis of CAG and CCG repeats in Huntingtin gene among HD patients and normal populations of India. Eur J Hum Genet 8: 678-682.

Pulkes T, Papsing C, Wattanapokayakit S, et al. (2014). CAG-Expansion Haplotype Analysis in a Population with a Low Prevalence of Huntington's Disease. J Clin Neurol 10: 32-36.

Quarrell OW, Rigby AS, Barron L, et al. (2007). Reduced penetrance alleles for Huntington's disease: a multi-centre direct observational study. J Med Genet 44: e68.

Raghavan M, Skoglund P, Graf KE, et al. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505: 87-91.

Ramos-Arroyo MA, Moreno S & Valiente A (2005). Incidence and mutation rates of Huntington's disease in Spain: experience of 9 years of direct genetic testing. J Neurol Neurosurg Psychiatry 76: 337-342.

Ramos EM, Keagle P, Gillis T, et al. (2012). Prevalence of Huntington's disease gene CAG repeat alleles in sporadic amyotrophic lateral sclerosis patients. Amyotroph Lateral Scler 13: 265-269.

Rasmussen M, Guo X, Wang Y, et al. (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. Science 334: 94-98.

Rodriguez-Lebron E, Denovan-Wright EM, Nash K, et al. (2005). Intrastriatal rAAV-mediated delivery of anti-huntingtin shRNAs induces partial reversal of disease progression in R6/1 Huntington's disease transgenic mice. Mol Ther 12: 618-633.

Roos RA, Hermans J, Vegter-van der Vlis M, et al. (1993). Duration of illness in Huntington's disease is not related to age at onset. J Neurol Neurosurg Psychiatry 56: 98-100.

Rosenblatt A, Ranen NG, Rubinsztein DC, et al. (1998). Patients with features similar to Huntington's disease, without CAG expansion in huntingtin. Neurology 51: 215-220.

Ross CA, Aylward EH, Wild EJ, et al. (2014). Huntington disease: natural history, biomarkers and prospects for therapeutics. Nat Rev Neurol 10: 204-216.

Ruano L, Melo C, Silva MC, et al. (2014). The global epidemiology of hereditary ataxia and spastic paraplegia: a systematic review of prevalence studies. Neuroepidemiology 42: 174-183.

Rubinsztein DC, Amos W, Leggo J, et al. (1994). Mutational bias provides a model for the evolution of Huntington's disease and predicts a general increase in disease prevalence. Nat Genet 7: 525-530.

Rubinsztein DC, Barton DE, Davison BC, et al. (1993). Analysis of the huntingtin gene reveals a trinucleotide-length polymorphism in the region of the gene that contains two CCG-rich stretches and a correlation between decreased age of onset of Huntington's disease and CAG repeat number. Hum Mol Genet 2: 1713-1715.

Rubinsztein DC, Leggo J, Coles R, et al. (1996). Phenotypic Characterization of Individuals with 30–40 CAG Repeats in the Huntington Disease (HD) Gene Reveals HD Cases with 36 Repeats and Apparently Normal Elderly Individuals with 36–39 Repeats. American journal of human genetics 59: 16-22.

Rubinsztein DC, Leggo J, Goodburn S, et al. (1995). Haplotype analysis of the delta 2642 and (CAG)n polymorphisms in the Huntington's disease (HD) gene provides an explanation for an apparent 'founder' HD haplotype. Hum Mol Genet 4: 203-206.

Ruiz-Linares A, Adhikari K, Acuna-Alonzo V, et al. (2014). Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. PLoS Genet 10: e1004572.

Saleem Q, Roy S, Murgood U, et al. (2003). Molecular analysis of Huntington's disease and linked polymorphisms in the Indian population. Acta Neurol Scand 108: 281-286.

Sandoval JR, Salazar-Granara A, Acosta O, et al. (2013). Tracing the genomic ancestry of Peruvians reveals a major legacy of pre-Columbian ancestors. J Hum Genet 58: 627-634.

Sathasivam K, Neueder A, Gipson TA, et al. (2013). Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. Proc Natl Acad Sci U S A 110: 2366-2370.

Scholefield J & Greenberg J (2007). A common SNP haplotype provides molecular proof of a founder effect of Huntington disease linking two South African populations. Eur J Hum Genet 15: 590-595.

Semaka A, Creighton S, Warby S, et al. (2006). Predictive testing for Huntington disease: interpretation and significance of intermediate alleles. Clin Genet 70: 283-294.

Semaka A & Hayden MR (2013). Evidence-based genetic counselling implications for Huntington disease intermediate allele predictive test results. Clin Genet.

Semaka A, Kay C, Belfroid RD, et al. (2014). A new mutation for Huntington disease following maternal transmission of an intermediate allele. Eur J Med Genet 58: 28-30.

Semaka A, Kay C, Doty C, et al. (2013a). CAG size-specific risk estimates for intermediate allele repeat instability in Huntington disease. J Med Genet 50: 696-703.

Semaka A, Kay C, Doty CN, et al. (2013b). High frequency of intermediate alleles on huntington disease-associated haplotypes in British Columbia's general population. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 162: 864-871.

Sequeiros J, Ramos EM, Cerqueira J, et al. (2010). Large normal and reduced penetrance alleles in Huntington disease: instability in families and frequency at the laboratory, at the clinic and in the population. Clin Genet 78: 381-387.

Shaw M & Caro A (1982). The mutation rate to Huntington's chorea. J Med Genet 19: 161-167.

Shirasaki DI, Greiner ER, Al-Ramahi I, et al. (2012). Network organization of the huntingtin proteomic interactome in mammalian brain. Neuron 75: 41-57.

Siesling S, Vegter-van de Vlis M, Losekoot M, et al. (2000). Family history and DNA analysis in patients with suspected Huntington's disease. J Neurol Neurosurg Psychiatry 69: 54-59.

Sipila JO, Hietala M, Siitonen A, et al. (2014). Epidemiology of Huntington's disease in Finland. Parkinsonism Relat Disord 21: 46-49.

Skotte NH, Southwell AL, Ostergaard ME, et al. (2014). Allele-specific suppression of mutant huntingtin using antisense oligonucleotides: providing a therapeutic option for all Huntington disease patients. PLoS One 9: e107434.

Slow EJ, van Raamsdonk J, Rogers D, et al. (2003). Selective striatal neuronal loss in a YAC128 mouse model of Huntington disease. Hum Mol Genet 12: 1555-1567.

Snell RG, MacMillan JC, Cheadle JP, et al. (1993). Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. Nat Genet 4: 393-397.

Soong BW & Wang JT (1995). A comparison of the Huntington's disease associated trinucleotide repeat between Chinese and white populations. J Med Genet 32: 404-405.

Southwell AL, Skotte NH, Bennett CF, et al. (2012). Antisense oligonucleotide therapeutics for inherited neurodegenerative diseases. Trends in Molecular Medicine 18: 634-643.

Southwell AL, Skotte NH, Kordasiewicz HB, et al. (2014). In vivo evaluation of candidate allele-specific mutant huntingtin gene silencing antisense oligonucleotides. Mol Ther 22: 2093-2106.

Southwell AL, Warby SC, Carroll JB, et al. (2013). A fully humanized transgenic mouse model of Huntington disease. Human Molecular Genetics 22: 18-34.

Squitieri F, Andrew SE, Goldberg YP, et al. (1994). DNA haplotype analysis of Huntington disease reveals clues to the origins and mechanisms of CAG expansion and reasons for geographic variations of prevalence. Hum Mol Genet 3: 2103-2114.

Stanek LM, Yang W, Stuart A, et al. (2013). Antisense Oligonucleotide-Mediated Correction of Transcriptional Dysregulation is Correlated with Behavioral Benefits in the YAC128 Mouse Model of Huntington's Disease. Journal of Huntington's disease 2: 217-228.

Stephens M & Donnelly P (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73: 1162-1169.

Tabrizi SJ, Scahill RI, Owen G, et al. (2013). Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. The Lancet Neurology 12: 637-649.

Tartari M, Gissi C, Lo Sardo V, et al. (2008). Phylogenetic Comparison of Huntingtin Homologues Reveals the Appearance of a Primitive polyQ in Sea Urchin. Molecular Biology and Evolution 25: 330-338.

Tateno Y, Komiyama T, Katoh T, et al. (2014). Divergence of East Asians and Europeans estimated using male- and female-specific genetic markers. Genome Biol Evol 6: 466-473.

Theadom A, Rodrigues M, Roxburgh R, et al. (2014). Prevalence of muscular dystrophies: a systematic literature review. Neuroepidemiology 43: 259-268.

Torres Ramirez L, Cosentino Esquerre C & Mori Quispe N (2008). Actualización sobre la enfermedad de Huntington y experiencia de 30 años en el Instituto Nacional de Ciencias Neurológicas. Diagnostico 47: 65-69.

Trottier Y, Devys D, Imbert G, et al. (1995). Cellular localization of the Huntington's disease protein and discrimination of the normal and mutated form. Nat Genet 10: 104-110.

van Bilsen PH, Jaspers L, Lombardi MS, et al. (2008). Identification and allele-specific silencing of the mutant huntingtin allele in Huntington's disease patient-derived fibroblasts. Hum Gene Ther 19: 710-719.

Van Raamsdonk J, Pearson J, Murphy Z, et al. (2006). Wild-type huntingtin ameliorates striatal neuronal atrophy but does not prevent other abnormalities in the YAC128 mouse model of Huntington disease. BMC Neuroscience 7: 80.

Van Raamsdonk JM, Pearson J, Rogers DA, et al. (2005). Loss of wild-type huntingtin influences motor dysfunction and survival in the YAC128 mouse model of Huntington disease. Human Molecular Genetics 14: 1379-1392.

Visscher H, Ross CJ, Dube MP, et al. (2009). Application of principal component analysis to pharmacogenomic studies in Canada. Pharmacogenomics J 9: 362-372.

Vuillaume I, Vermersch P, Destee A, et al. (1998). Genetic polymorphisms adjacent to the CAG repeat influence clinical features at onset in Huntington's disease. J Neurol Neurosurg Psychiatry 64: 758-762.

Walker DA, Harper PS, Wells CE, et al. (1981). Huntington's Chorea in South Wales. A genetic and epidemiological study. Clin Genet 19: 213-221.

Wang CK, Wu YR, Hwu WL, et al. (2004). DNA haplotype analysis of CAG repeat in Taiwanese Huntington's disease patients. Eur Neurol 52: 96-100.

Wang S, Lewis CM, Jakobsson M, et al. (2007). Genetic variation and population structure in native Americans. PLoS Genet 3: e185.

Wang S, Ray N, Rojas W, et al. (2008). Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet 4: e1000037.

Warby SC, Montpetit A, Hayden AR, 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: 351-366.

Warby SC, Visscher H, Collins JA, et al. (2011). HTT haplotypes contribute to differences in Huntington disease prevalence between Europe and East Asia. Eur J Hum Genet 19: 561-566.

Warner JP, Barron LH & Brock DJ (1993). A new polymerase chain reaction (PCR) assay for the trinucleotide repeat that is unstable and expanded on Huntington's disease chromosomes. Mol Cell Probes 7: 235-239.

Wasmuth JJ, Hewitt J, Smith B, et al. (1988). A highly polymorphic locus very tightly linked to the Huntington's disease gene. Nature 332: 734-736.

Watkins WS, Bamshad M & Jorde LB (1995). Population genetics of trinucleotide repeat polymorphisms. Hum Mol Genet 4: 1485-1491.

Wexler NS, Lorimer J, Porter J, et al. (2004). Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. Proc Natl Acad Sci U S A 101: 3498-3503.

Wild EJ, Mudanohwo EE, Sweeney MG, et al. (2008). Huntington's disease phenocopies are clinically and genetically heterogeneous. Mov Disord 23: 716-720.

Yamamoto A, Lucas JJ & Hen R (2000). Reversal of Neuropathology and Motor Dysfunction in a Conditional Model of Huntington's Disease. Cell 101: 57-66.

Yanagisawa H, Fujii K, Nagafuchi S, et al. (1996). A unique origin and multistep process for the generation of expanded DRPLA triplet repeats. Hum Mol Genet 5: 373-379.

Yapijakis C, Vassilopoulos D, Tzagournisakis M, et al. (1995). Linkage disequilibrium between the expanded (CAG)n repeat and an allele of the adjacent (CCG)n repeat in Huntington's disease patients of Greek origin. Eur J Hum Genet 3: 228-234.

Yu D, Pendergraff H, Liu J, et al. (2012). Single-Stranded RNAs Use RNAi to Potently and Allele-Selectively Inhibit Mutant Huntingtin Expression. Cell 150: 895-908.

Zeitlin S, Liu JP, Chapman DL, et al. (1995). Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. Nature genetics 11: 155-163.

Zhang Y, Engelman J & Friedlander RM (2009). Allele-specific silencing of mutant Huntington’s disease gene. Journal of Neurochemistry 108: 82-90.

Zuccato C, Valenza M & Cattaneo E (2010). Molecular Mechanisms and Potential Therapeutical Targets in Huntington's Disease. Physiol. Rev. 90: 905-981.

Appendices

Appendix A

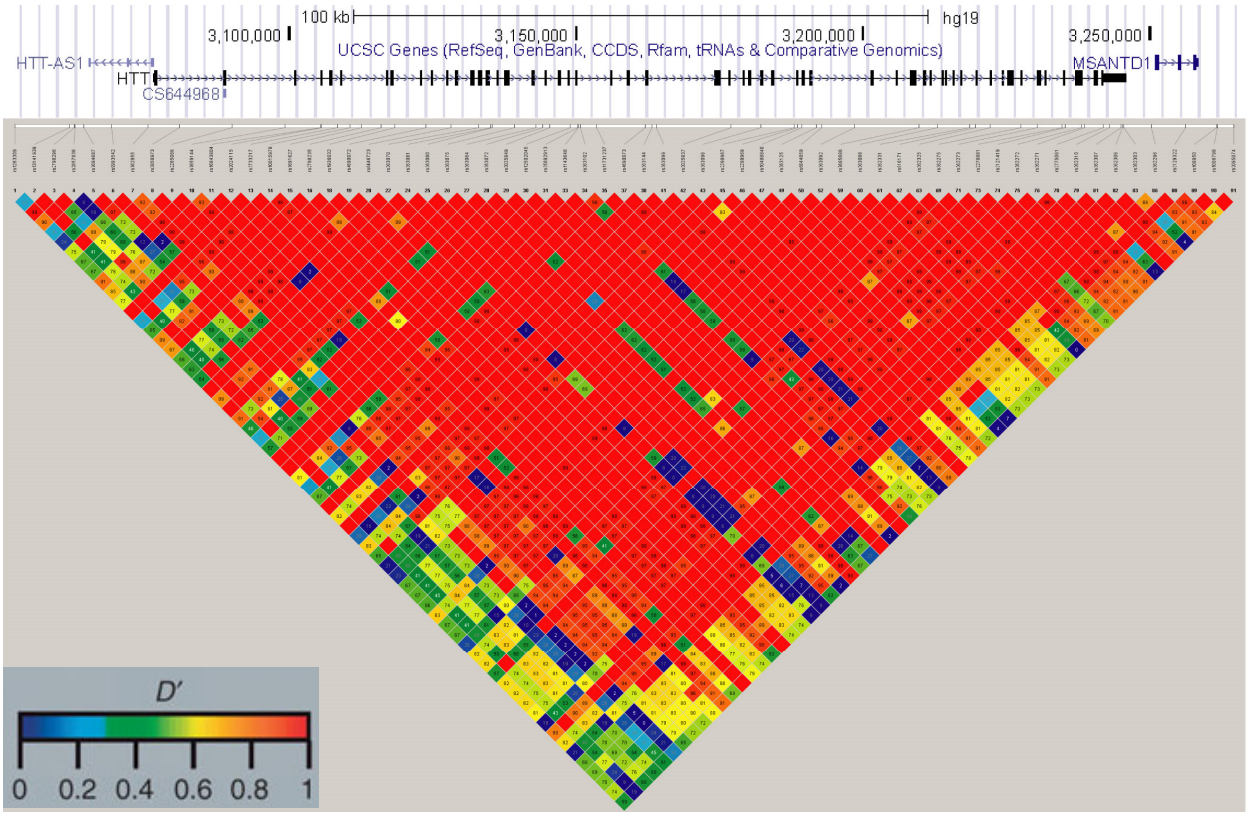


Figure A-1. Linkage disequilibrium across the *HTT* locus in the Canadian patient population, expressed as *D'* using the 63-SNP panel.

The *HTT* gene is found within a haplotype block of very low recombination, shown in red.

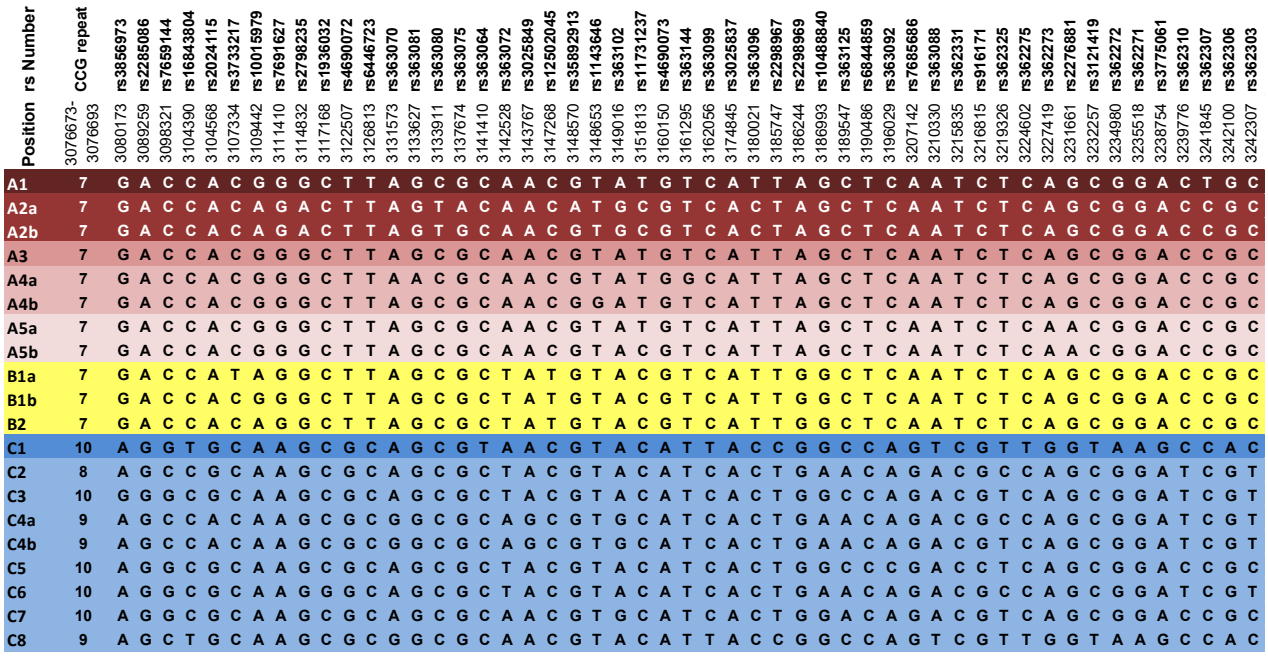


Figure A-2. All 20 common intragenic haplotypes from the Canadian Caucasian cohort including CCG repeat length (51 SNPs)

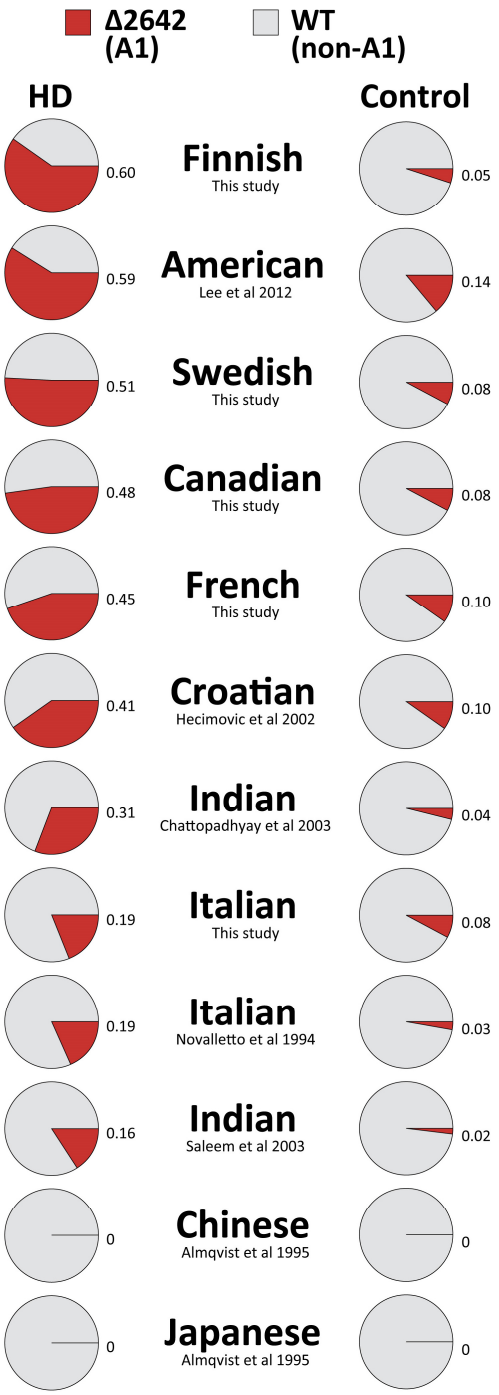


Figure A-3. A1 haplotype frequencies vary among HD and control chromosomes in various populations, as inferred from Δ2642 genotypes.

A1 is present across European, North American, and South Asian populations, but entirely absent among East Asian patients and controls.

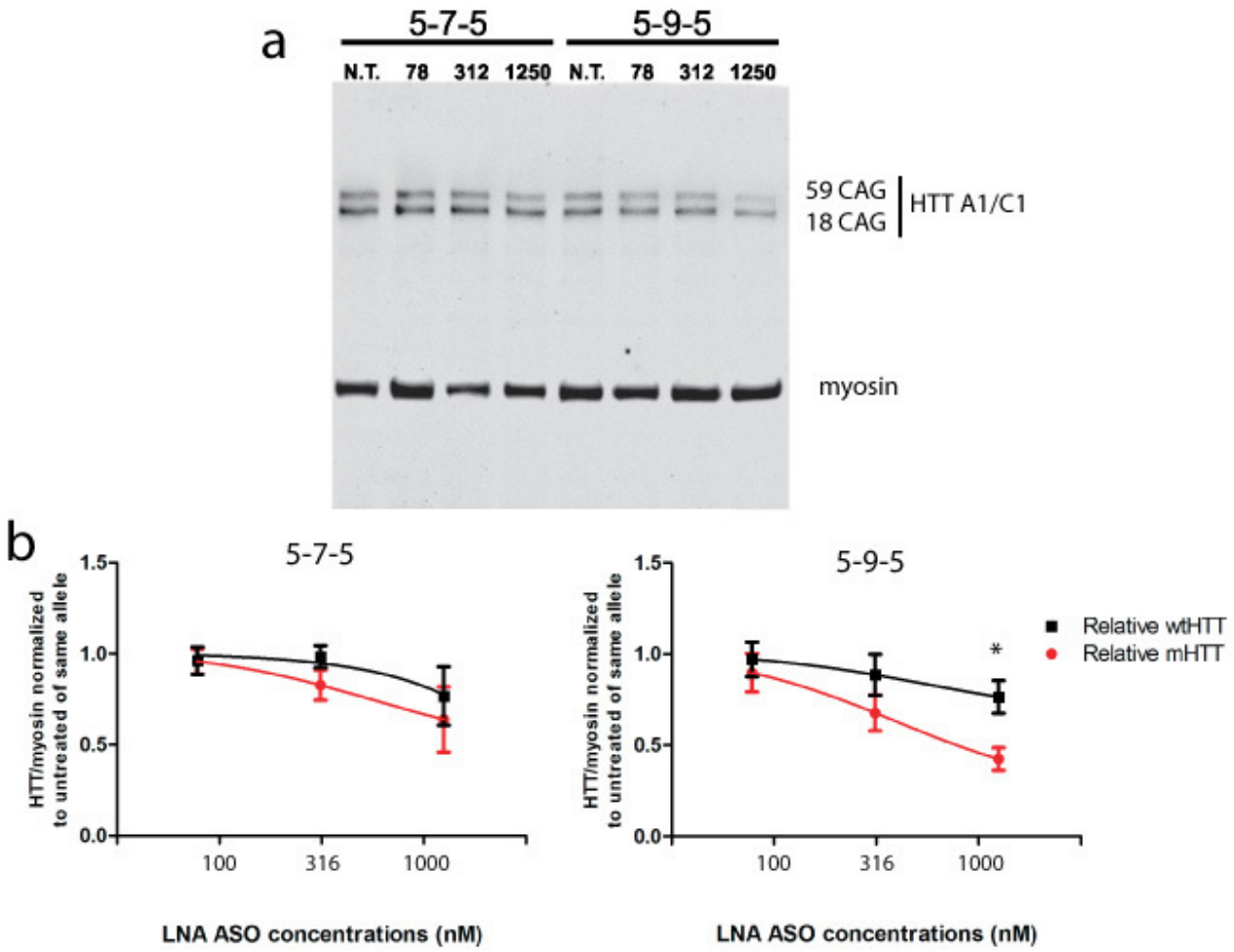


Figure A-4. Passive transfection of patient-derived lymphoblasts with ΔACTT-complementary ASOs.

(a) Passive transfection of patient-derived lymphoblasts with ΔACTT-complementary ASO selectively reduces mutant HTT protein. Patient lymphoblasts transfected with 5-7-5 and 5-9-5 LNA gapmers show dose-dependent reduction of mutant HTT protein relative to untreated controls. Non-muscle myosin was used as a loading control. (b) Quantification of relative mHTT and wtHTT levels following 120hr treatment of juvenile A1 lymphoblasts at 78, 312, and 1250 nM ASO in media (n=4, * $p < 0.05$).

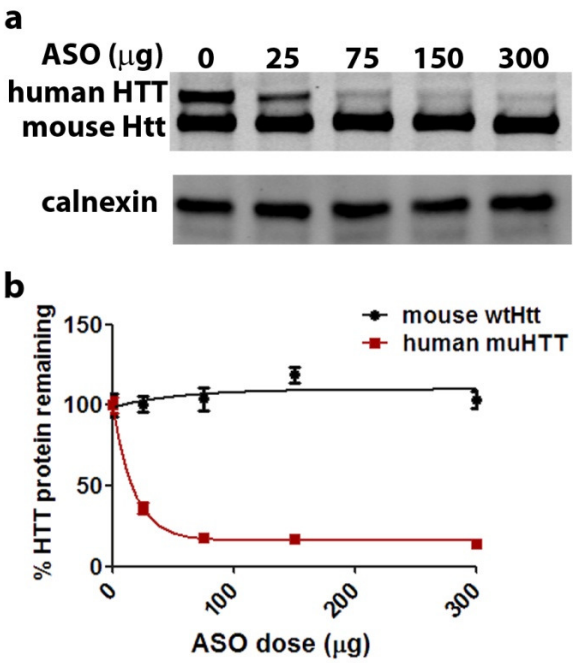


Figure A-5. ASOs complementary to the rs72239206 wild-type sequence potential reduce mutant HTT *in vivo*.

(a) ICV injection of YAC128 mice with WT 5-9-5 LNA gapmer complementary to the rs72239206 major allele results in potent reduction of mutant HTT protein *in vivo* relative to untreated controls. (b) Quantification of relative human HTT and mouse Htt levels following treatment with the indicated doses of WT 5-9-5 LNA gapmer.

Table A-1. A1 allele counts and relative frequencies among HD and control chromosomes in the UBC HD Biobank and 1000 Genomes Phase I

		chromosomes	rs362307		Δ2642		ΔACTT	
		n	n	%	n	%	n	%
CMMT HD chromosomes								
Caucasian		454	214	47.1%	217	47.8%	216	47.6%
East Asian		53	0	0.0%	0	0.0%	0	0.0%
Black African		19	0	0.0%	0	0.0%	0	0.0%
CMMT normal chromosomes								
Caucasian		652	42	6.4%	43	6.6%	40	6.1%
East Asian		94	0	0.0%	0	0.0%	0	0.0%
Black African		211	0	0.0%	0	0.0%	0	0.0%
1000 Genomes normal chromsomes								
MXL	Mexican	132	11	8.33%	11	8.33%	12	9.09%
CEU	CEPH (Caucasian, Utah)	170	14	8.24%	13	7.65%	13	7.65%
GBR	British	178	13	7.30%	13	7.30%	14	7.87%
FIN	Finnish	186	13	6.99%	13	6.99%	13	6.99%
CLM	Colombian	120	7	5.83%	7	5.83%	12	10.00%
TSI	Toscan	178	10	5.62%	11	6.18%	11	6.18%
PUR	Puerto Rican	110	4	3.64%	4	3.64%	5	4.55%
ASW	African American	122	3	2.46%	3	2.46%	3	2.46%
CHB	Han Chinese in Beijing	194	1	0.52%	0	0.00%	0	0.00%
CHS	Han Chinese South	200	0	0.00%	0	0.00%	0	0.00%
IBS	Iberian	28	0	0.00%	1	3.57%	0	0.00%
JPT	Japanese	178	0	0.00%	1	0.56%	0	0.00%
LWK	Luhya	194	0	0.00%	0	0.00%	0	0.00%
YRI	Yoruban	176	0	0.00%	0	0.00%	0	0.00%
1000 Genomes Total		2166	76	3.51%	77	3.55%	83	3.83%

Table A-2. Direct genotyping of A1 haplotype-defining alleles in HD and control chromosomes from the UBC HD BioBank

		HD		Control		Chi-Square
		n	%	n	%	*p*=
rs72239206	Δ	214	47.1%	42	6.4%	$<10^{-10}$
(ΔACTT)	ACTT	240	52.9%	610	93.6%	
	Total	**454**		**652**		
rs149109767	Δ	217	47.8%	43	6.6%	$<10^{-10}$
(Δ2642)	GAG	237	52.2%	609	93.4%	
	Total	**454**		**652**		
rs362307	T	216	47.6%	40	6.1%	$<10^{-10}$
	C	238	52.4%	612	93.9%	
	Total	**454**		**652**		

Appendix B

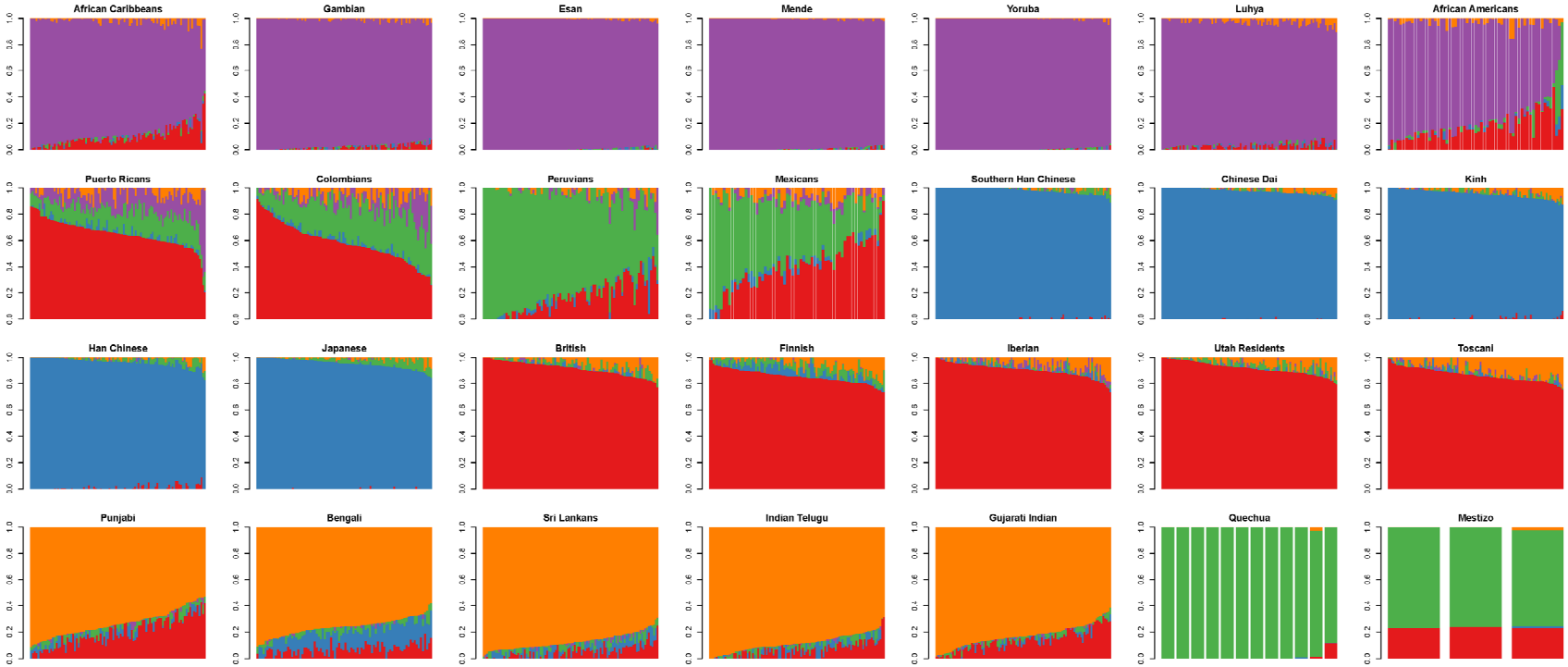


Figure B-1. ADMIXTURE analysis of selected mestizo Peruvian HD patients and Quechua Amerindian controls with reference populations from the 1000 Genomes Project Phase 3

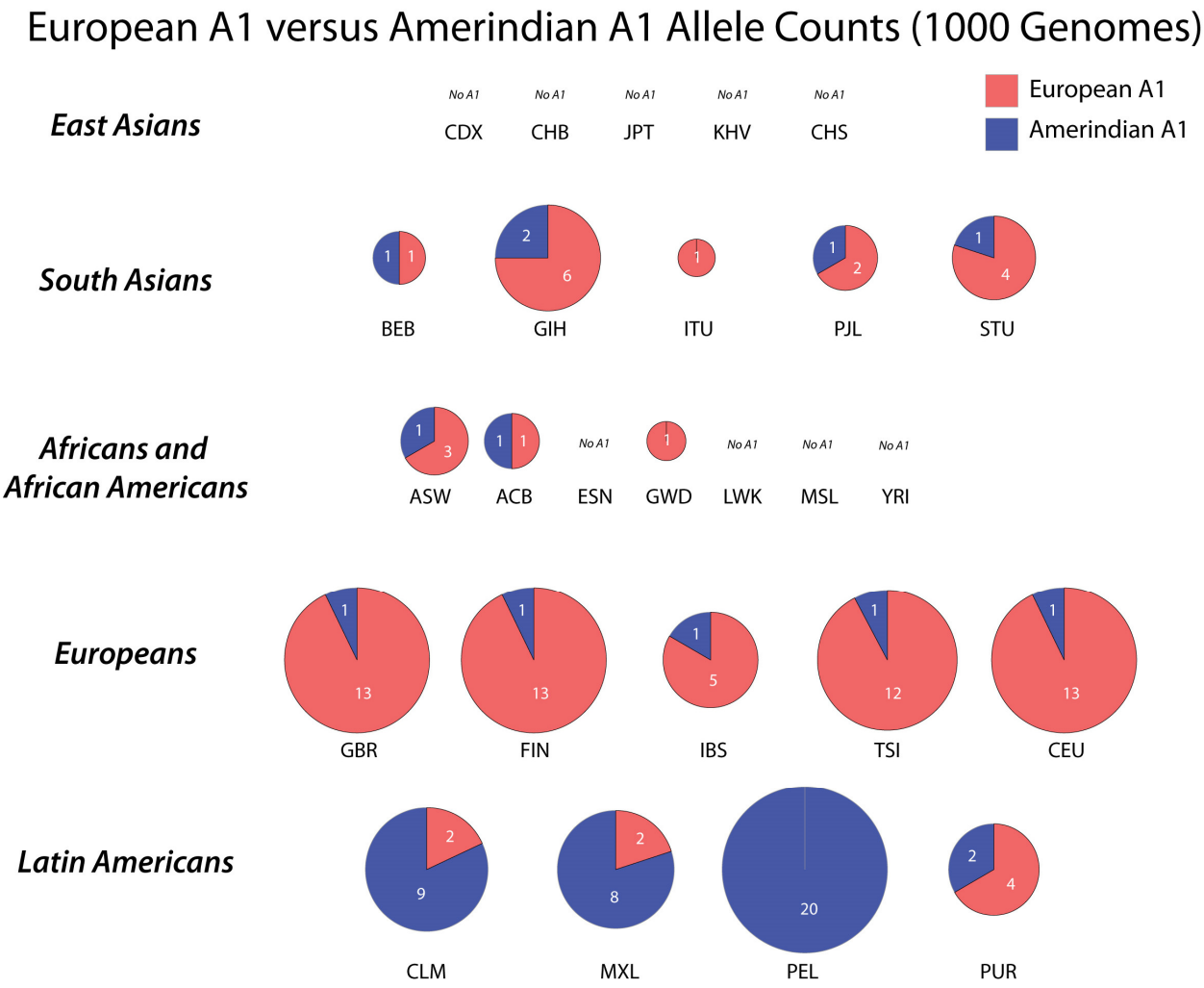


Figure B-2. European and Amerindian A1 *HTT* haplotype counts across reference populations from the 1000 Genomes Project Phase 3

Table B-1. Detailed *HTT* variant descriptions

dbSNP ID	GRCh37 position	HGVS variant name	major allele (+)	minor allele (+)	
rs11248108	2481490	GRCh37 chr4:g.2481490G>A	G	A	
rs2857845	3028113	GRCh37 chr4:g.3028113A>T	A	T	
rs2471347	3044435	GRCh37 chr4:g.3044435G>A	A	G	
rs1263309	3051940	GRCh37 chr4:g.3051940C>T	T	C	
rs13141939	3061282	GRCh37 chr4:g.3061282C>T	C	T	
rs2798296	3062165	GRCh37 chr4:g.3062165A>G	G	A	
rs2857936	3062310	GRCh37 chr4:g.3062310T>C	C	T	
rs7694687	3063817	GRCh37 chr4:g.3063817C>T	C	T	
rs762855	3074795	GRCh37 chr4:g.3074795A>G	A	G	
rs13102260	3076405	GRCh37 chr4:g.3076405G>A	G	A	
rs12508079	3080238	GRCh37 chr4:g.3080238T>C	T	C	intragenic
rs2285086	3089259	GRCh37 chr4:g.3089259A>G	A	G	intragenic
rs7659144	3098321	GRCh37 chr4:g.3098321C>G	C	G	intragenic
rs16843804	3104390	GRCh37 chr4:g.3104390C>T	C	T	intragenic
rs2024115	3104568	GRCh37 chr4:g.3104568A>G	A	G	intragenic
rs3733217	3107334	GRCh37 chr4:g.3107334C>T	C	T	intragenic
rs10015979	3109442	GRCh37 chr4:g.3109442A>G	A	G	intragenic
rs7691627	3111410	GRCh37 chr4:g.3111410G>A	G	A	intragenic
rs2798235	3114832	GRCh37 chr4:g.3114832G>A	G	A	intragenic
rs1936032	3117168	GRCh37 chr4:g.3117168C>G	C	G	intragenic
rs4690072	3122507	GRCh37 chr4:g.3122507T>G	T	G	intragenic
rs6446723	3126813	GRCh37 chr4:g.3126813T>C	T	C	intragenic
rs363080	3133911	GRCh37 chr4:g.3133911C>T	C	T	intragenic
rs363075	3137674	GRCh37 chr4:g.3137674G>A	G	A	intragenic
rs363064	3141410	GRCh37 chr4:g.3141410C>T	C	T	intragenic
rs363072	3142528	GRCh37 chr4:g.3142528A>T	A	T	intragenic
rs72239206	3142661	GRCh37 chr4:g.3142661_3142664delACTT	ACTT	-	intragenic
rs363107	3144441	GRCh37 chr4:g.3144441A>G	A	G	intragenic
rs12502045	3147268	GRCh37 chr4:g.3147268C>T	C	T	intragenic
rs35892913	3148570	GRCh37 chr4:g.3148570G>A	G	A	intragenic
rs363102	3149016	GRCh37 chr4:g.3149016A>G	A	G	intragenic
rs11731237	3151813	GRCh37 chr4:g.3151813C>T	C	T	intragenic
rs4690073	3160150	GRCh37 chr4:g.3160150G>A	G	A	intragenic
rs3025838	3161446	GRCh37 chr4:g.3161446C>T	C	T	intragenic
rs363099	3162056	GRCh37 chr4:g.3162056C>T	C	T	intragenic
rs3025837	3174845	GRCh37 chr4:g.3174845A>C	A	C	intragenic
rs363096	3180021	GRCh37 chr4:g.3180021T>C	C	T	intragenic
rs2298967	3185747	GRCh37 chr4:g.3185747T>C	T	C	intragenic
rs2298969	3186244	GRCh37 chr4:g.3186244A>G	A	G	intragenic
rs10488840	3186993	GRCh37 chr4:g.3186993G>A	G	A	intragenic
rs188072823	3188616	GRCh37 chr4:g.3188616C>T	C	T	intragenic
rs363125	3189547	GRCh37 chr4:g.3189547C>A	C	A	intragenic
rs6844859	3190486	GRCh37 chr4:g.3190486T>C	T	C	intragenic
rs363092	3196029	GRCh37 chr4:g.3196029A>C	C	A	intragenic
rs7685686	3207142	GRCh37 chr4:g.3207142A>G	A	G	intragenic
rs363088	3210330	GRCh37 chr4:g.3210330A>T	A	T	intragenic
rs362331	3215835	GRCh37 chr4:g.3215835T>C	T	C	intragenic
rs916171	3216815	GRCh37 chr4:g.3216815C>G	C	G	intragenic
rs362325	3219326	GRCh37 chr4:g.3219326T>C	T	C	intragenic
rs362322	3221365	GRCh37 chr4:g.3221365A>G	A	G	intragenic
rs362275	3224602	GRCh37 chr4:g.3224602C>T	C	T	intragenic
rs82333	3225389	GRCh37 chr4:g.3225389A>G	A	G	intragenic
rs110501	3225478	GRCh37 chr4:g.3225478T>C	T	C	intragenic
rs362273	3227419	GRCh37 chr4:g.3227419A>G	A	G	intragenic
rs149109767	3230411	GRCh37 chr4:g.3230411_3230413delGAG	GAG	-	intragenic
rs2276881	3231661	GRCh37 chr4:g.3231661G>A	G	A	intragenic
rs3121419	3232257	GRCh37 chr4:g.3232257C>T	C	T	intragenic
rs3216209	3233395	GRCh37 chr4:g.3233394delT	T	-	intragenic
rs362272	3234980	GRCh37 chr4:g.3234980G>A	G	A	intragenic
rs3216820	3235185	GRCh37 chr4:g.3235185delC	C	-	intragenic
rs362271	3235518	GRCh37 chr4:g.3235518G>A	G	A	intragenic
rs362313	3235589	GRCh37 chr4:g.3235589C>T	C	T	intragenic
rs3775061	3238754	GRCh37 chr4:g.3238754A>G	A	G	intragenic
rs2269499	3239702	GRCh37 chr4:g.3239702C>T	C	T	intragenic
rs362307	3241845	GRCh37 chr4:g.3241845C>T	C	T	intragenic
rs362306	3242100	GRCh37 chr4:g.3242100G>A	G	A	intragenic
rs362303	3242307	GRCh37 chr4:g.3242307C>T	C	T	intragenic
rs115335747	3243804	GRCh37 chr4:g.3243804G>A	G	A	intragenic
rs2530595	3245057	GRCh37 chr4:g.3245057C>T	C	T	intragenic
rs362296	3247007	GRCh37 chr4:g.3247007C>A	C	A	
rs3129322	3252852	GRCh37 chr4:g.3252852T>C	T	C	
rs186719032	3255302	GRCh37 chr4:g.3255302G>A	G	A	
rs108850	3258236	GRCh37 chr4:g.3258236C>G	C	G	
rs1006798	3258373	GRCh37 chr4:g.3258373A>G	A	G	
rs3095074	3261152	GRCh37 chr4:g.3261152G>A	G	A	
rs3095073	3263138	GRCh37 chr4:g.3263138G>A	G	A	
rs76034781	3272782	GRCh37 chr4:g.3272782G>A	G	A	
rs3135066	3278517	GRCh37 chr4:g.3278517C>T	C	T	
rs7658462	3283422	GRCh37 chr4:g.3283422C>T	C	T	
rs1138690	3291401	GRCh37 chr4:g.3291401G>T	G	T	
rs1730768	3409359	GRCh37 chr4:g.3409359A>G	G	A	
rs12641989	3419840	GRCh37 chr4:g.3419840G>A	G	A	

Table B-2. Dense *HTT* haplotypes in Latin America

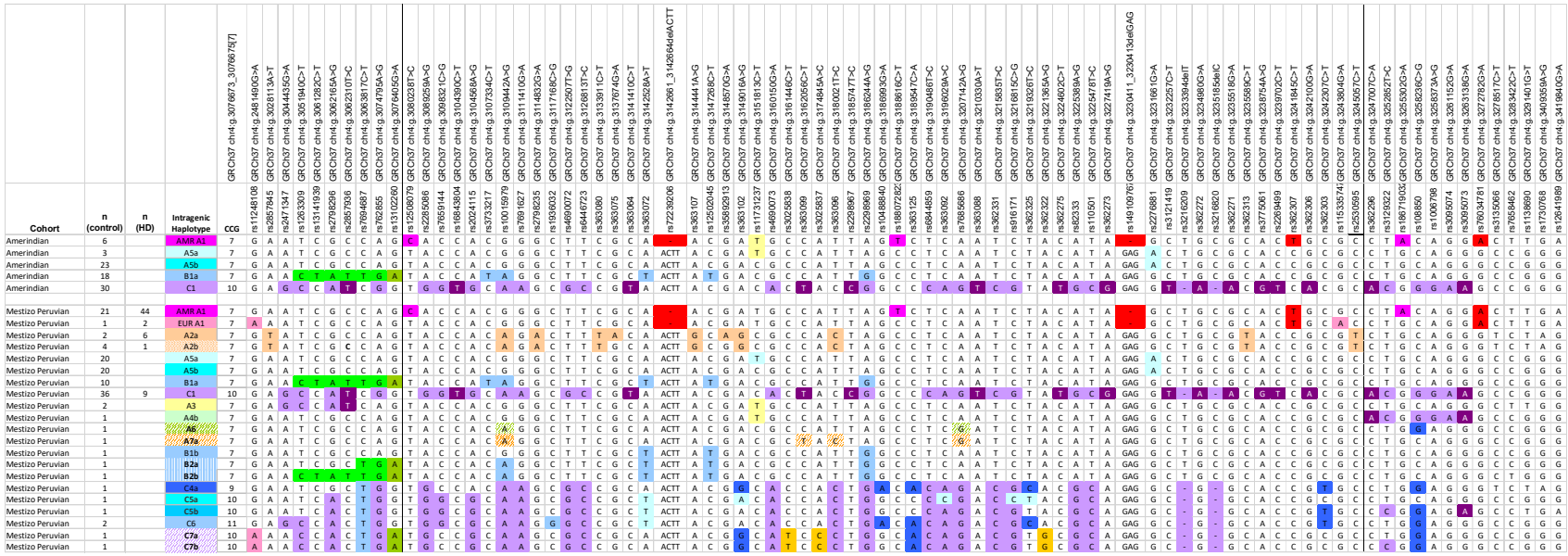
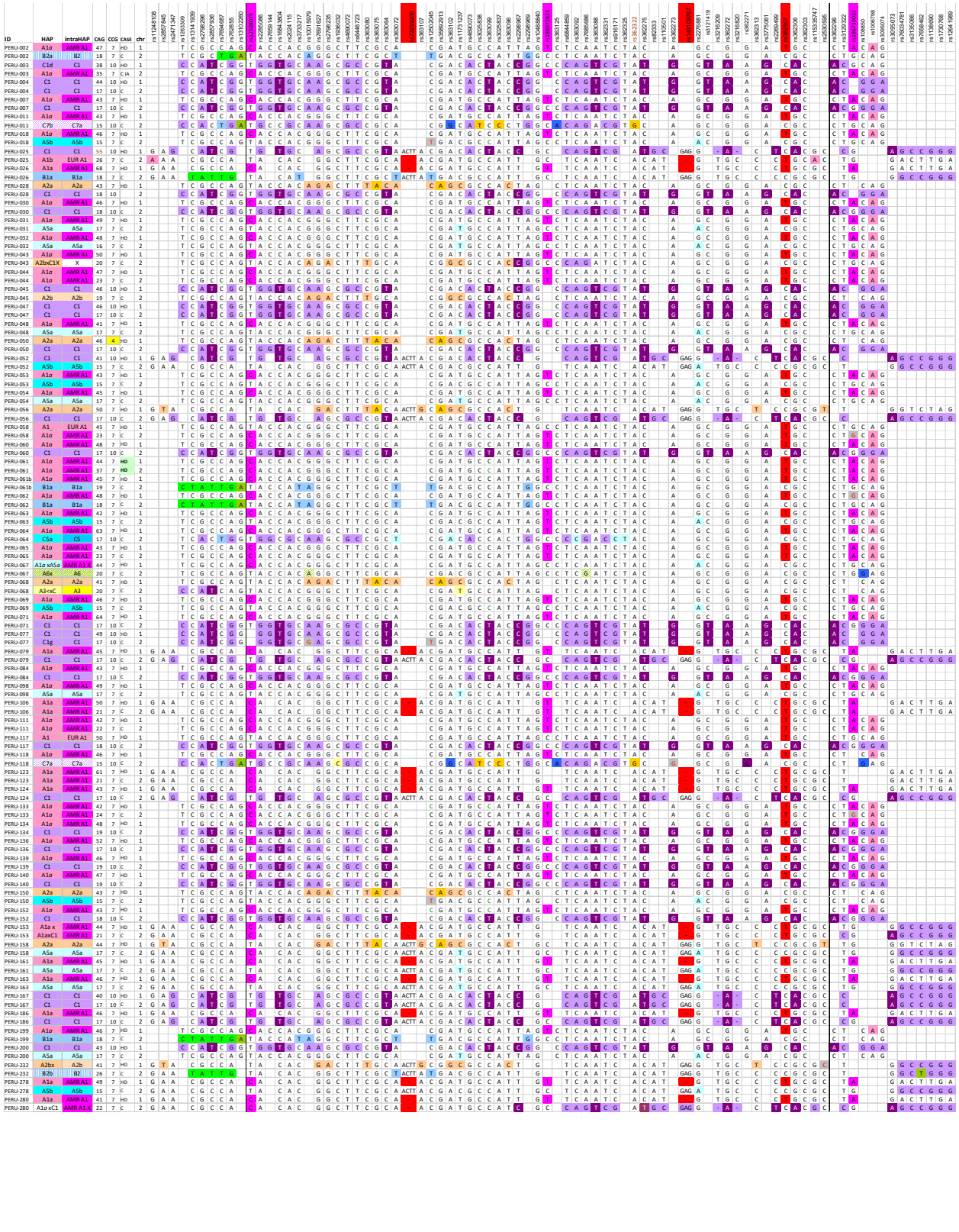


Table B-3. A1 *HTT* haplotype SNPs in 1000 Genomes by ethnic group

			A1			Amerindian A1				A1			Amerindian A1		
Population		**Total chrs**	rs72239206(-)	rs149109767(-)	rs362307(T)	rs12508079(C)	rs188072823(A)	rs186719032(T)		rs72239206(-)	rs149109767(-)	rs362307(T)	rs12508079(C)	rs188072823(A)	rs186719032(T)
Chinese Dai in Xishuangbanna, China(**CDX**)	**CDX**	**186**	0	0	0	0	0	0		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Han Chinese in Bejing, China (**CHB**)	**CHB**	**206**	0	0	1	0	0	0		0.0%	0.0%	0.5%	0.0%	0.0%	0.0%
Japanese in Tokyo, Japan (**JPT**)	**JPT**	**208**	0	0	0	0	0	0		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Kinh in Ho Chi Minh City, Vietnam (**KHV**)	**KHV**	**198**	0	0	0	0	0	0		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Southern Han Chinese, China (**CHS**)	**CHS**	**210**	0	0	0	0	0	0		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Total East Asian Ancestry (EAS)			**0**	**0**	**1**	**0**	**0**	**0**							
Bengali in Bangladesh (**BEB**)	**BEB**	**172**	7	2	3	1	1	1		4.1%	1.2%	1.7%	0.6%	0.6%	0.6%
Gujarati Indian in Houston,TX (**GIH**)	**GIH**	**206**	19	8	8	2	2	2		9.2%	3.9%	3.9%	1.0%	1.0%	1.0%
Indian Telugu in the UK (**ITU**)	**ITU**	**204**	14	1	1	0	0	0		6.9%	0.5%	0.5%	0.0%	0.0%	0.0%
Punjabi in Lahore,Pakistan (**PJL**)	**PJL**	**192**	7	3	3	1	1	1		3.6%	1.6%	1.6%	0.5%	0.5%	0.5%
Sri Lankan Tamil in the UK (**STU**)	**STU**	**204**	13	5	5	1	2	3		6.4%	2.5%	2.5%	0.5%	1.0%	1.5%
Total South Asian Ancestry (SAS)			**60**	**19**	**20**	**5**	**6**	**7**							
African Ancestry in Southwest US (**ASW**)	**ASW**	**122**	3	3	3	1	1	3		2.5%	2.5%	2.5%	0.8%	0.8%	2.5%
African Caribbean in Barbados (**ACB**)	**ACB**	**192**	2	2	2	1	1	1		1.0%	1.0%	1.0%	0.5%	0.5%	0.5%
Esan in Nigeria (**ESN**)	**ESN**	**198**	0	0	0	0	0	0		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Gambian in Western Division, The Gambia (**GWD**)	**GWD**	**226**	1	1	1	0	0	0		0.4%	0.4%	0.4%	0.0%	0.0%	0.0%
Luhya in Webuye, Kenya (**LWK**)	**LWK**	**198**	0	0	0	0	0	0		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Mende in Sierra Leone (**MSL**)	**MSL**	**170**	0	0	0	0	0	0		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Yoruba in Ibadan, Nigeria (**YRI**)	**YRI**	**216**	0	0	0	0	0	0		0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Total African Ancestry (AFR)			**1**	**1**	**1**	**0**	**0**	**0**							
British in England and Scotland (**GBR**)	**GBR**	**182**	15	14	15	1	1	1		8.2%	7.7%	8.2%	0.5%	0.5%	0.5%
Finnish in Finland (**FIN**)	**FIN**	**198**	14	14	14	1	1	2		7.1%	7.1%	7.1%	0.5%	0.5%	1.0%
Iberian populations in Spain (**IBS**)	**IBS**	**214**	6	6	7	1	1	1		2.8%	2.8%	3.3%	0.5%	0.5%	0.5%
Toscani in Italia (**TSI**)	**TSI**	**214**	13	13	12	1	1	1		6.1%	6.1%	5.6%	0.5%	0.5%	0.5%
Utah residents with Northern and Western European ancestry (**CEU**)	**CEU**	**198**	14	14	15	1	1	1		7.1%	7.1%	7.6%	0.5%	0.5%	0.5%
Total European Ancestry (EUR)			**62**	**61**	**63**	**5**	**5**	**6**							
Colombian in Medellin, Colombia (**CLM**)	**CLM**	**188**	16	11	10	9	5	4		8.5%	5.9%	5.3%	4.8%	2.7%	2.1%
Mexican Ancestry in Los Angeles, California (**MXL**)	**MXL**	**128**	11	10	10	8	7	10		8.6%	7.8%	7.8%	6.3%	5.5%	7.8%
Peruvian in Lima, Peru (**PEL**)	**PEL**	**170**	21	20	22	23	20	20		12.4%	11.8%	12.9%	13.5%	11.8%	11.8%
Puerto Rican in Puerto Rico (**PUR**)	**PUR**	**208**	6	6	6	2	3	2		2.9%	2.9%	2.9%	1.0%	1.4%	1.0%
Total Americas Ancestry (AMR)			**54**	**47**	**48**	**42**	**35**	**36**							
Total		**5008**													

Table B-4. Dense *HTT* haplotypes in mestizo Peruvian probands



Appendix C

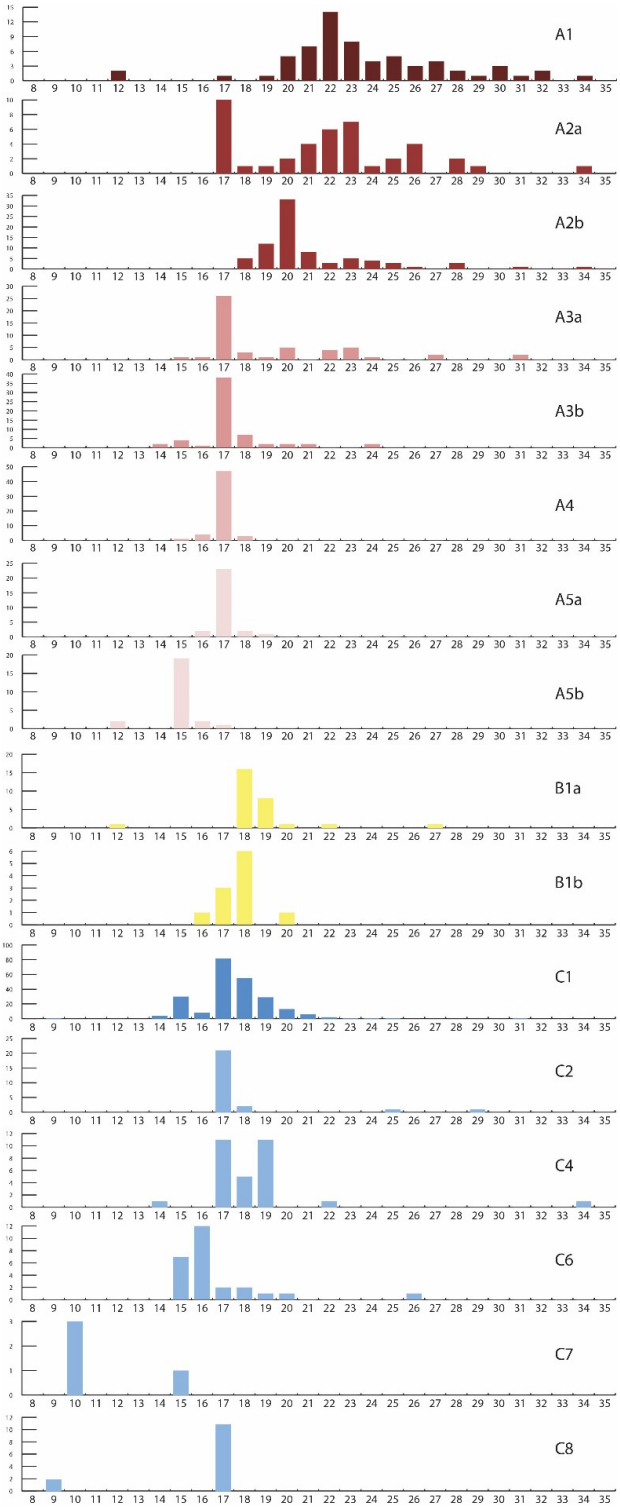


Figure C-1. CAG repeat distribution by intragenic *HTT* haplotype

Table C-1. Pairwise comparison significance values for general population CAG repeat distributions (Kruskal-Wallis test with posthoc Nemenyi tests)

	Scotland	White.SA	N.Amer	E.Asia	Finn	Swed	Denm	Norw	Af.Amer	Mixed.SA	Amer	Black.SA	Latino	E.Afr
White.SA	0.999998	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
N.Amer	0.112769	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E.Asia	**0.001771**	0.125377	**0.044791**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Finn	0.960889	0.999019	0.999823	0.997178	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Swed	1	1	0.999874	0.337263	0.993029	NA	NA	NA	NA	NA	NA	NA	NA	NA
Denm	0.999997	1	1	0.755535	0.999996	0.999997	NA	NA	NA	NA	NA	NA	NA	NA
Norw	0.999179	0.994094	0.909207	**0.039388**	0.841495	1	0.995147	NA	NA	NA	NA	NA	NA	NA
Af.Amer	**1.35E-05**	**0.005338**	**0.000661**	0.999722	0.812021	0.057306	0.223953	**0.002616**	NA	NA	NA	NA	NA	NA
Mixed.SA	**5.33E-14**	**1.06E-07**	**1.26E-13**	0.976624	0.444048	**0.004958**	**0.026702**	**2.63E-05**	1	NA	NA	NA	NA	NA
Amer	**1.42E-09**	**2.47E-07**	**3.89E-08**	**0.015813**	**0.001959**	**1.16E-05**	**6.68E-05**	**1.83E-07**	0.172531	0.08213	NA	NA	NA	NA
Black.SA	**0**	**0**	**0**	**4.66E-07**	**1.53E-06**	**2.94E-10**	**4.20E-10**	**1.53E-13**	**0.002075**	**9.82E-12**	0.999998	NA	NA	NA
Latino	0.999675	0.996598	0.915652	**0.031934**	0.86204	1	0.997296	1	**0.001728**	**6.28E-06**	**9.92E-08**	**9.98E-14**	NA	NA
E.Afr	**3.18E-13**	**9.84E-07**	**9.13E-11**	0.760887	0.201513	**0.001682**	**0.009374**	**1.25E-05**	0.999865	0.999223	0.462103	**0.009935**	**4.04E-06**	NA
W.Afr	**0.000144**	**0.013244**	**0.003684**	0.999752	0.834812	0.074516	0.269291	**0.004608**	1	1	0.235225	**0.012695**	**0.003389**	0.999973

Table C-2. Paternal CAG instability rates by methodology

	Sperm Typing (Semaka et al. 2013)				Familial Transmissions (UBC Biobank)				
CAG	**contract**		**expand**		**contract**		**expand**		**stable**
27-32	472/10467	4.5%	292/10467	2.8%	0/25	0.0%	4/25	16.0%	84.0%
33-35	985/7731	12.7%	1120/7731	14.5%	2/21	9.5%	2/21	9.5%	81.0%
36-38	no data		no data		2/20	10.0%	10/20	50.0%	40.0%
39	96/789	12.3%	481/789	61.0%	0/14	0.0%	8/14	57.1%	42.9%
40	no data		no data		3/31	9.7%	17/31	54.8%	35.5%
41	102/597	17.1%	296/597	49.6%	7/49	14.3%	28/49	57.1%	28.6%
42	72/427	16.9%	297/427	69.6%	1/30	3.3%	13/30	43.3%	53.3%